Least Squares Estimation using Sketched Data with Heteroskedastic Errors

Sokbae Lee[*1] **Serena Ng**[*1]

Abstract

Researchers may perform regressions using a sketch of data of size m instead of the full sample of size n for a variety of reasons. This paper considers the case when the regression errors do not have constant variance and heteroskedasticity robust standard errors would normally be needed for test statistics to provide accurate inference. We show that estimates using data sketched by random projections will behave 'as if' the errors were homoskedastic. Estimation by random sampling would not have this property. The result arises because the sketched estimates in the case of random projections can be expressed as degenerate U-statistics, and under certain conditions, these statistics are asymptotically normal with homoskedastic variance. We verify that the conditions hold not only in the case of least squares regression when the covariates are exogenous, but also in instrumental variables estimation when the covariates are endogenous. The result implies that inference can be simpler than the full sample case if the sketching scheme is appropriately chosen.

1. Introduction

Big data sets can be costly to store and analyze, and one approach around the data bottlenecks is to work with a randomly chosen subset, or a *sketch*, of the data. Data privacy may also dictate that a sketch of the data be made available for public use. The early works of Sarlos (2006), Drineas et al. (2006) and Drineas et al. (2011) consider the algorithmic properties of the least squares estimator using sketched data. Subsequent work extends the analysis to ridge regression (e.g., Wang et al., 2018; Liu & Dobriban, 2020), and logistic regression (e.g., Wang, 2019). See, e.g., Woodruff (2014), Drineas & Mahoney (2018) and Martinsson & Tropp (2020) for a review. However, Ma et al. (2015), Raskutti & Mahoney (2016), Dobriban & Liu (2019) and Ma et al. (2020) have found that an optimal worse-case (algorithmic) error may not yield an optimal mean-squared (statistical) error. This has led to interest in better understanding the sketched least squares estimates in a Bayesian setting as in Geppert et al. (2017), or its asymptotic distribution as in Ahfock et al. (2020) and Ma et al. (2020). Lee & Ng (2020) highlights the tension between a large m required for accurate inference, and a small m for computation efficiency. To date, these results have been derived under the assumptions that the errors are homoskedastic and that the regressors are exogenous. But these assumptions are not innocuous. The estimates will be biased when the regressors are not exogenous, as would normally be the case in causal inference. And if the errors are heteroskedastic, test statistics must use standard errors robust to heteroskedasticity, or else inference will not be accurate even if the regressors are exogenous.

In this paper, we obtain the surprising result that when sketching is based on random projections, robust standard errors will not be needed, meaning that inference using the sketched estimates can proceed as though the errors were homoskedastic. The proof is obtained by analyzing the difference between the full sample and the sketched estimates in terms of degenerate U-statistics. However, the result does not hold when sketching is based on random sampling. Our analysis of the least squares estimator and two-stage least squares estimator shows that these findings hold both when the regressors are exogenous and endogenous.

The following notation will be used. Let $\|a\|$ denote the Euclidean norm of any vector a. Let A_{ij} or $[A]_{ij}$ denote the (i,j) element of a matrix A. For $k = 1, \ldots, d$, let $\sigma_k(A)$ be a singular value of A. Let $\|A\|_2 = \sigma_{\max}(A)$ denote its spectral norm, where $\sigma_{\max}(A)$, and $\sigma_{\min}(A)$ are the largest and smallest singular values of A. The superscript T denotes the transpose of a matrix. For an integer $n \geq 1$, $[n]$ is the set of positive integers from 1 to n. Let $\to_p$ and $\to_d$, respectively, denote convergence in probability and in distribution. For a sequence of random variables A_n and a sequence of positive real numbers a_n, $A_n = o_p(a_n)$ iff $a_n^{-1} A_n \to_p 0$; $A_n = O_p(a_n)$ iff $a_n^{-1} A_n$ is bounded in probability.

An accompanying R package is available on the Comprehensive R Archive Network (CRAN) at `https://`

[*]Equal contribution [1]Department of Economics, Columbia University, New York, USA. Correspondence to: Sokbae Lee <sl3841@columbia.edu>, Serena Ng <sn2294@columbia.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

CRAN.R-project.org/package=sketching and all replication files are available at https://github.com/sokbae/replication-LeeNg-2022-ICML.

2. Sketched Least Squares Estimation with Heteroskedastic Errors

Given n observations $\{(y_i, X_i, Z_i) : i = 1, \dots, n\}$, we consider a linear regression model:

$$y_i = X_i^T \beta_0 + e_i, \tag{1}$$

where y_i is the scalar dependent variable, X_i is a $p \times 1$ vector of regressors, β_0 is a $p \times 1$ vector of unknown parameters. The innovation e_i is said to be (conditionally) homoskedastic if $E[e_i^2|X_i] = E[e_i^2]$. Otherwise, e_i is said to be heteroskedastic. The regressors are said to be exogenous if $E[e_i X_i] = 0$. Otherwise it is endogenous. In that case, we assume a $q \times 1$ vector of instrumental variables, Z_i, satisfying $E[e_i Z_i] = 0$ are available. In matrix form, the model given in (1) can be written as

$$y = X\beta_0 + e,$$

where y and e are $n \times 1$ vectors whose i-th rows are y_i and e_i, respectively, and X is the $n \times p$ matrix of regressors whose i-th row is X_i^T.

We first study the exogenous regressor case when $\mathbb{E}(e_i X_i) = 0$. The least squares estimator $\widehat{\beta}_{OLS} := (X^T X)^{-1} X^T y$ is $\sqrt{n}$ consistent and asymptotically normal, i.e., $\sqrt{n}(\widehat{\beta}_{OLS} - \beta_0) \to_d N(0, V_1)$ as $n \to \infty$, where

$$V_1 := [\mathbb{E}(X_i X_i^T)]^{-1} \mathbb{E}(e_i^2 X_i X_i^T) [\mathbb{E}(X_i X_i^T)]^{-1}$$

is the heteroskedasticity-robust asymptotic variance. Under homoskedasticity, V_1 becomes

$$V_0 := \mathbb{E}(e_i^2) [\mathbb{E}(X_i X_i^T)]^{-1}.$$

The point estimates $\widehat{\beta}$ can be used to test hypothesis, say, $H_0 : \beta_2 = \bar{\beta}_2$ using the t test $\frac{\sqrt{n}(\widehat{\beta}_2 - \bar{\beta}_2)}{\sqrt{[\widehat{V}]_{22}}}$, where $\widehat{V}$ is an estimate of either V_1 or V_0, β_2 is a specific element of β_2, $\bar{\beta}_2$ is the null value, and $[\widehat{V}]_{22}$ the (2,2) diagonal element of $\widehat{V}$. The distribution of this test under the null hypothesis crucially depends on the correct standard error $\sqrt{[\widehat{V}]_{22}}$ being used. Using $\widehat{V}_0$ when the robust estimator $\widehat{V}_1$ should have been used would lead to inaccurate inference, in the sense of rejecting the null hypothesis too often or not enough.

A sketch of the data (y, X) is $(\widetilde{y}, \widetilde{X})$, where $\widetilde{y} = \Pi y$, $\widetilde{X} = \Pi X$, and Π is usually an $m \times n$ random matrix. The sketched least squares estimator is $\widetilde{\beta}_{OLS} := (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \widetilde{y}$. Even though the sketched regression is based on a sample of size m, $\widetilde{X}^T \widetilde{X} = X^T \Pi^T \Pi X$ and $\widetilde{X}^T \widetilde{y} = X^T \Pi^T \Pi y$ can be seen as weighted moments in a sample of size n. Thus let $\widetilde{g}_n := \widetilde{X}^T \widetilde{e}/n$, $\widehat{g}_n := X^T e/n$, $\widetilde{A}_n := (\widetilde{X}^T \widetilde{X}/n)^{-1}$, and $\widehat{A}_n := (X^T X/n)^{-1}$. Then

$$\begin{aligned} \widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS} &= (\widetilde{A}_n - \widehat{A}_n)\widehat{g}_n + \widehat{A}_n(\widetilde{g}_n - \widehat{g}_n) \\ &\quad + (\widetilde{A}_n - \widehat{A}_n)(\widetilde{g}_n - \widehat{g}_n), \end{aligned}$$

By the law of large numbers, $\widehat{A}_n - A = o_p(1)$, where $A := [\mathbb{E}(X_i X_i^T)]^{-1}$, and by the central limit theorem, $\widehat{g}_n = O_p(n^{-1/2})$. We show in Section 3 that for Π with subspace embedding property

$$\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS} = A(\widetilde{g}_n - \widehat{g}_n) + o_p(m^{-1/2}).$$

We study $\widetilde{\beta}_{OLS}$ under the following regularity conditions.

Assumption 2.1. (i) The data $\mathcal{D}_n := \{(y_i, X_i) \in \mathbb{R}^{1+p} : i = 1, \dots, n\}$ are independent and identically distributed (i.i.d.), where p is fixed. Furthermore, X has singular value decomposition $X = U_X \Sigma_X V_X^T$.

(ii) $\mathbb{E}(y_i^4) < \infty$, $\mathbb{E}(\|X_i\|^4) < \infty$, and $\mathbb{E}(X_i X_i^T)$ has full rank p.

(iii) The random matrix Π is independent of $\mathcal{D}_n$.

(iv) $m = m_n \to \infty$ but $m/n \to 0$ as $n \to \infty$.

Assumptions (i) and (ii) are standard. For (iii), note that for a general random Π whose (k, i) element is Π_{ki}, the difference between the full and the sketched moments such as $\widetilde{g}_n - \widehat{g}_n$ and $\widetilde{A}_n - \widehat{A}_n$ are of the form

$$\begin{aligned} &n^{-1}\left(U^T \Pi^T \Pi V - U^T V\right) \\ &= n^{-1} \sum_{i=1}^n \psi_i U_i V_i + n^{-1} \sum_{i=1}^n \sum_{j=1, j \neq i}^n U_i \varphi_{ij} V_j \\ &=: T_{n1} + T_{n2}, \end{aligned}$$

where $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^n$ are vectors of certain i.i.d. random variables $(U_i, V_i) \in \mathbb{R}^2$ that are independent of Π,

$$\psi_i := \sum_{k=1}^{\text{r.dim}(\Pi)} \Pi_{ki}^2 - 1, \ \varphi_{ij} := \sum_{k=1}^{\text{r.dim}(\Pi)} \Pi_{ki} \Pi_{kj},$$

and r.dim$(\Pi) \in \{m, n\}$ denotes the row dimension of Π.

There are two classes of sketching schemes to consider. Random sampling schemes have $\varphi_{ij} = 0$ for all $i \neq j$ because there is only one non-zero entry in each row of Π. In such cases, T_{2n} is negligible and T_{1n} is the leading term. The second class is random projection schemes with which T_{1n} is asymptotically negligible and T_{2n} is the leading term.

To gain intuition, we first provide results for Bernoulli sampling (BS) from the first type and countsketch (CS) from the second type.

Theorem 2.2. *Let Assumption 2.1 hold and $\mathbb{E}(e_i X_i) = 0$.*

(i) Under BS, $m^{1/2}(\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS}) \to_d N(0, V_1)$.

(ii) Under CS, $m^{1/2}(\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS}) \to_d N(0, V_0)$.

Though Theorem 2.2 indicates that both sampling schemes yield asymptotically normal estimates, their variances are different, and normality holds for different reasons. The proof is given in the Appendix. Here, we sketch the main arguments.

First, under BS, the sampling probability is determined by i.i.d. Bernoulli random variables with success probability m/n. Thus, $\Pi = \sqrt{\frac{n}{m}} B$ is an $n \times n$ matrix (not $m \times n$), where B is a diagonal sampling matrix. We have

$$\frac{1}{n}\left(X^T \Pi^T \Pi e - X^T e\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{n}{m}B_{ii} - 1\right) X_i e_i = T_{n1}.$$

Since the summands are i.i.d. with mean zero with variance $(\frac{n}{m} - 1)\mathbb{E}(e_i^2 X_i X_i^T)$, applying central limit theorem for i.i.d. observations yields the sandwich variance V_1.

Consider now CS. Each column of its Π has one non-zero entry taking on value $\{+1, -1\}$ randomly drawn with equal probability and located uniformly at random. For such Π and every nonzero $c \in \mathbb{R}^p$,

$$\begin{aligned} & c^T (X^T \Pi^T \Pi e - X^T e)/n \\ & = n^{-1} \sum_{i=1}^{n} \bar{X}_j(c) \left(\sum_{k=1}^{m} \Pi_{ki}^2 - 1 \right) e_i \\ & \quad + n^{-1} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \sum_{k=1}^{m} \bar{X}_j(c) \Pi_{kj} \Pi_{ki} e_i \\ & = T_{n1} + T_{n2}, \end{aligned}$$

where $\bar{X}_i(c) := \sum_u^q c_u X_{iu}$ is a weighted sum of elements of the i-th row of X. The term T_{n1} is identically zero because there is only one non-zero entry per column of Π.

To analyze T_{n2}, let $W_i = (Y_i, X_i^T, \Pi_{1i}, \ldots, \Pi_{mi})^T$. Since the columns of Π are i.i.d., $\{W_i : i = 1, \ldots, n\}$ are i.i.d. Now let $w = (y, x^T, \pi_1, \ldots, \pi_m)^T$ be a non-random index. Define $\widetilde{H}(w_1, w_2) := \sum_{k=1}^m \bar{x}_1(c) \pi_{k1} S \pi_{k2} e_i$ and $H(w_1, w_2) := \widetilde{H}(w_1, w_2) + \widetilde{H}(w_2, w_1)$. We can write

$$T_{n2} = n^{-1} \sum\sum_{1 \leq i < j \leq n} H(W_i, W_j),$$

noting that $H(w_1, w_2) = H(w_2, w_1)$, and $\mathbb{E}(H(W_1, W_2)|W_1) = \mathbb{E}(H(W_1, W_2)|W_2) = 0$. Importantly, T_{n2} has now been represented as a degenerate U-statistic. In general, the asymptotic distribution of such statistics is either a weighted average of independent, centered chi-square random variables with complex weights, or a centered normal distribution. But if the conditions given in Hall (1984) are satisfied, the latter holds. Precisely,

$$\left\{\frac{1}{2}\mathbb{E}[H^2(W_1, W_2)]\right\}^{-1/2} T_{n2} \to_d N(0, 1).$$

A sufficient condition for this result which we verify in the Appendix is

$$\frac{\mathbb{E}[G^2(W_1, W_2)] + n^{-1}\mathbb{E}[H^4(W_1, W_2)]}{\{\mathbb{E}[H^2(W_1, W_2)]\}^2} \to 0 \quad \text{as } n \to \infty,$$

where $G(w_1, w_2) := \mathbb{E}[H(W_1, w_1) H(W_1, w_2)]$. Furthermore, we also verify that for $W_i \neq W_j$,

$$\frac{1}{2}\mathbb{E}[H^2(W_i, W_j)] = \frac{1}{m}\mathbb{E}(\bar{X}_i^2(c))\mathbb{E}(e_i^2).$$

Note that $\mathbb{E}(e_i^2)$ appears separately from $\mathbb{E}(\bar{X}_i^2(c))$. This is key to the claim in Theorem 2.2 that when $\widetilde{\beta}$ is based on CS, $m^{1/2}(\widetilde{\beta} - \widehat{\beta}) \to_d N(0, E[e_i^2]A) = N(0, V_0)$. Analogous arguments show that each entry of $(\widetilde{X}^T \widetilde{X} - X^T X)$ can also be written as a degenerate U-statistic and $\widetilde{A}_n - \widehat{A} = o_p(1)$, which is needed for consistent estimation of V_0 and V_1.

As discussed in Charikar et al. (2004) and Clarkson & Woodruff (2013; 2017), the sparsity of Π significantly reduces the run time required of the countsketch to compute ΠA to O(nnz(A)), where nnz(A) is the number of non-zero entries of A. Another appeal of countsketch is that the sketches can be obtained by streaming without constructing Π. Here, we show that countsketch removes heteroskedasticity which is appealing because it simplifies inference. In the next section, we study the mean-squared sketching error and show that part (i) of Theorem 2.2 also holds for other Πs in the first class, while part (ii) holds for other Πs in the second class. Section 4 then shows that these results also hold when the regressors are not exogenous.

3. The Mean-Squared Sketching Error

For a random variable G, let $\text{MSE}(G) = [\mathbb{E}(G)]^2 + \text{Var}(G)$ denote the mean squared error. We now analyze the asymptotic behavior of mean squared sketching errors of $(U^T \Pi^T \Pi V - U^T V)/n$, where $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^n$ denote vectors of i.i.d. random variables $(U_i, V_i) \in \mathbb{R}^2$ that are independent of Π, with $\mathbb{E}(U_i^4) < \infty$ and $\mathbb{E}(V_i^4) < \infty$. Recall that $m = m_n \to \infty$ but $m/n \to 0$ as $n \to \infty$.

3.1. Random Sampling with Replacement (RS)

For sketching by random sampling with replacement (RS), we suppose that for each $t = 1, \ldots, m$, we sample k_t from $[n]$ with probability $p_i := \Pr(k_t = i)$ independently and with replacement. The random matrix $\Pi \in \mathbb{R}^{m\times n}$ is then

$$\Pi = \sqrt{\frac{n}{m}} \begin{pmatrix} \iota_{k_1} & \dots & \iota_{k_m} \end{pmatrix}^T,$$

where ι_k is the k-th column vector of the $n \times n$ dimensional identity matrix. Sketching schemes of the RS class have properties characterized by Lemma A.2 in the Appendix. Importantly, each Π in this class has $T_{2n} = 0$ and as a result, T_{n1} is the only term we need to consider. An important example in the RS class is uniform sampling with replacement with $p_i = n^{-1}$.

Theorem 3.1. *(i) If Π is a random matrix satisfying RS and $\sum_{i=1}^n p_i^2 = o(m^{-1})$, then, as $n \to \infty$,*

$$\text{MSE}\left[\sqrt{m}\left(U^T\Pi^T\Pi V - U^T V\right)/n\right] \to \text{Var}(U_i V_i).$$

(ii) If Π is Bernoulli sampling matrix (BS), then, as $n \to \infty$,

$$\text{MSE}\left[\sqrt{m}\left(U^T\Pi^T\Pi V - U^T V\right)/n\right] \to \mathbb{E}(U_i^2 V_i^2).$$

The mean-squared errors for RS and BS are the same if $U_i V_i$ is mean zero.

Theorem 3.1 is useful in two ways. First, let $U = c^T X^T$ and $V = Xc$. Under Assumption 2.1 and for all nonzero $c \in \mathbb{R}^p$, Theorem 3.1 yields

$$\text{MSE}\left[\sqrt{m}(c^T\widetilde{X}^T\widetilde{X}c - c^T X^T X c)/n\right] = O(1).$$

By Chebyshev's Inequality, $(\widetilde{X}^T\widetilde{X} - X^T X)/n = O_p(m^{-1/2})$. Similarly, for $U = Xc$ and $V = e$, the theorem implies $\widetilde{g}_n - \widehat{g}_n = O_p(m^{-1/2})$. Applying continuous mapping theorem gives the result stated earlier that $\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS} = A(\widetilde{g}_n - \widehat{g}_n) + o_p(m^{-1/2})$.

It is known that the efficient estimator under heteroskedasticity is the generalized least squares (GLS), defined as

$$\widehat{\beta}_{GLS} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y,$$

where Ω is $n \times n$ a diagonal matrix with $\Omega_{ii} = \sigma_i^2$. GLS weights each observation with σ_i^{-1} so that the errors in the weighted regression are homoskedastic. Now the OLS estimator applied to sketched data can be written as

$$\widetilde{\beta}_{OLS} = (X^T\Pi^T\Pi X)^{-1}X^T\Pi^T\Pi y.$$

A question of interest is whether $\Pi^T\Pi$ can play the role of Ω^{-1}. The theorem sheds light on this problem as its second use is to obtain the asymptotic variance of the sketched estimator. To this end, let again $U_i = c^T X_i$ and $V_i = e_i$. Assuming $\mathbb{E}(g_i) = 0$ where $g_i = e_i X_i$, RS and BS imply:

$$\text{MSE}\left[\sqrt{m}\, c^T(\widetilde{g}_n - \widehat{g}_n)\right] = \mathbb{E}(e_i^2 c^T X_i X_i^T c).$$

The asymptotic standard error is generally the expectation of a product of e_i^2 and $(c^T X_i)^2$ and becomes the product of two expectations only under homoskedasticity when $\mathbb{E}[e_i^2 c^T X_i X_i^T c] = \mathbb{E}[e_i^2]\mathbb{E}[c^T X_i X_i^T c]$. Thus, under RS and BS, the asymptotic variance of $\widetilde{\beta}_{OLS}$ is V_0 only if homoskedasticity is explicitly imposed, implying that $\Pi^T\Pi$ corresponding to random sampling will not homogenize error variance in the same way that Ω^{-1} can.

It is noteworthy that even under homoskedsaticity, we cannot always use a central limit theorem for i.i.d. data even if the full sample of data are i.i.d. because the sampling scheme may induce dependence in the sketched data. Thus the asymptotic normality result can only be analyzed on a case by case basis. Ma et al. (2020) confronts a similar problem when studying the asymptotic distribution of estimators in linear regressions under random sampling with replacement and homoskedastic errors. Let K_i and p_i, respectively, denote the number of times and the probability that i^{th} observation is sampled. Their estimator has $W = \text{diag}\left\{K_i/(mp_i)\right\}_{i=1}^n$ playing the role of $\Pi^T\Pi$. Our Theorem 3.1 applies to their setup with uniform sampling where $p_i = 1/n, n \geq m$, but it would not apply when p_i is data dependent. In this case, Ma et al. (2020) also cannot use a central limit theorem for i.i.d. data. Instead, they apply Hayék-Sidak central limit theorem and use Poissonization to account for dependence in the sketched data that arises from sampling.

3.2. Random Projection (RP)

Sketching schemes in the RP class have properties characterized by Lemma A.3 in the Appendix if $\Pi \in \mathbb{R}^{m\times n}$ is a random matrix with the following properties:

Assumption 3.2. (i) $\mathbb{E}[\Pi_{ki}] = 0$, $\mathbb{E}[\Pi_{ki}^2] = m^{-1}$ for all $k \in [m]$ and all $i \in [n]$, and $\max_{(k,i)\in[m]\times[n]}\mathbb{E}[\Pi_{ki}^4] = O(m^{-1})$;

(ii) $\mathbb{E}[\Pi_{ki}\Pi_{kj}] = 0$ and $\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] = m^{-2}$ for all $k \in [m]$ and all $i \neq j \in [n]$;

(iii) $\mathbb{E}[\Pi_{ki}\Pi_{kj}\Pi_{\ell p}\Pi_{\ell q}] = 0$ for all $k \neq \ell \in [m]$ and all $i \neq j, p \neq q \in [n]$.

Under Assumption 3.2, $T_{n1} = O_p(n^{-1/2})$ is asymptotically negligible, and T_{n2} becomes the leading term for RP. As discussed above, the Π for the CS only has one non-zero entry in each column. Since $\Pi_{ki}\Pi_{kj} = 0$ for all $k, i \neq j$, it is straightforward to check that the above conditions are satisfied. For Gaussian random projections,

$$\text{GP}: \Pi_{ki} \sim N(0, m^{-1}).$$

Since all elements of Π are i.i.d. with mean zero, variance m^{-1} and the fourth moment $O(m^{-1})$, the conditions are also satisfied. The SRHT has

$$\text{SRHT} : \Pi = \sqrt{\frac{n}{m}} SHD,$$

where $S \in \mathbb{R}^{m\times n}$ is a uniform sampling matrix with replacement, $H \in \mathbb{R}^{n\times n}$ is a normalized Walsh-Hadamard transform matrix, and $D \in \mathbb{R}^{n\times n}$ is a diagonal Rademacher matrix with i.i.d. entries of ± 1. The Appendix shows that the conditions for RP hold for SRHT.

The following theorem gives the asymptotic mean squared sketching errors of RP schemes.

Theorem 3.3. *If Π is a random matrix satisfying RP, then, as $n \to \infty$,*

(i) $\text{MSE}\left[\sqrt{m}\left(U^T\Pi^T\Pi V - U^T V\right)/n\right] \to \{\mathbb{E}(U_i^2)\mathbb{E}(V_i^2) + [\mathbb{E}(U_iV_i)]^2\}$.

(ii) If, in addition, and $\mathbb{E}[e_iX_i] = 0$ and the columns of Π are i.i.d., then $m^{1/2}(\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS}) \to_d N(0, V_0)$.

The limiting MSE of RP is simply the product between two marginal expectations when $\mathbb{E}(U_iV_i) = 0$ (and not the expectation of the product). It implies

$$\text{MSE}\left[\sqrt{m}\, c^T(\widetilde{g}_n - \widehat{g}_n)\right] = \mathbb{E}(e_i^2)\mathbb{E}(c^T X_i X_i^T c)$$

and is the reason why the asymptotic variance for $\widetilde{\beta}_{OLS}$ for RP schemes is V_0.

If e_i^2 and $(c^T X_i)^2$ are positively (respectively, negatively) correlated, the limiting MSE of RP is smaller (respectively, larger) than that of RS and BS. The limiting MSE is the same if e_i^2 and $(c^T X_i)^2$ are uncorrelated.

Asymptotic normality of $\widetilde{\beta}$ can be established by applying a central limit theorem for degenerate U-statistic if the columns of Π are i.i.d., as reported in part (ii) of Theorem 3.3. The SRHT and SRFT are not covered by this result because the columns of their Π matrix are not i.i.d. and requires a limit theorem for a particular type of mixing data. In general, establishing asymptotic normality of $\widetilde{\beta}$ based on SRHT or SRFT require different proof techniques. The approach taken in Ahfock et al. (2020) is to condition on the data $\mathcal{D}_n$ and apply a central limit theorem for a triangular array of random variables. We do not condition on the data and appeal to the theory of degenerate U-statistics. Though deriving distribution theory for the SRHT and SRFT estimates is not straightforward, we will show in simulations that their finite sample properties are similar to those of CS.

4. Two-Stage Least Squares

The 2SLS estimator is appropriate when $\mathbb{E}(X_ie_i) \neq 0$ but exongeous instruments Z_i satisfying $\mathbb{E}(Z_ie_i) = 0$ are available. The 2SLS estimator is

$$\widehat{\beta}_{2SLS} = (X^T P_Z X)^{-1} X^T P_Z y,$$

where $P_Z := Z(Z^TZ)^{-1}Z^T$ is the projection matrix. The estimator first projects on Z to purge the variations in X correlated with e, and in the second step replaces X with $\widehat{X} = P_Z X$. Let $\widehat{g}_n := Z^T e/n$ and $\widehat{A}_n := [(X^TZ/n)(Z^TZ/n)^{-1}(Z^TX/n)]^{-1}(X^TZ/n)(Z^TZ/n)^{-1}$. Analyzing $\widehat{\beta}_{2SLS} - \beta_0 = \widehat{A}_n\widehat{g}_n$ under Assumption 4.3 given below, as $n \to \infty$, we have

$$\sqrt{n}(\widehat{\beta}_{2SLS} - \beta_0) \to_d N(0, W_1),$$

where $W_1 := A\,\mathbb{E}(e_i^2 Z_i Z_i^T)\,A^T$ with

$$A := [\mathbb{E}(X_iZ_i^T)[\mathbb{E}(Z_iZ_i^T)]^{-1}\mathbb{E}(Z_iX_i^T)]^{-1} \times \mathbb{E}(X_iZ_i^T)[\mathbb{E}(Z_iZ_i^T)]^{-1}.$$

Under homoskedasticity, $\mathbb{E}(e_i^2|Z_i) = \sigma^2$ and W_1 reduces to

$$W_0 := \mathbb{E}(e_i^2)[\mathbb{E}(X_iZ_i^T)[\mathbb{E}(Z_iZ_i^T)]^{-1}\mathbb{E}(Z_iX_i^T)]^{-1}.$$

A sketched version of the 2SLS estimator is

$$\widetilde{\beta}_{2SLS} := (\widetilde{X}^T P_{\widetilde{Z}} \widetilde{X})^{-1} \widetilde{X}^T P_{\widetilde{Z}} \widetilde{y}.$$

We now provide some algorithmic results not previously documented in the literature.

Assumption 4.1. Let data $\mathcal{D}_n = \{(y_i, X_i, Z_i) \in \mathbb{R}^{1+p+q} : i = 1, \ldots, n\}$ be fixed, Z^TZ and X^TP_ZX are non-singular, and Z has singular value decomposition $Z = U_Z\Sigma_Z V_Z^T$. For given constants $\varepsilon_1, \varepsilon_2, \varepsilon_3, \delta \in (0, 1/2)$, the following holds jointly with probability at least $1 - \delta$:

$$(i)\ \left\|U_Z^T\Pi^T\Pi U_Z - I_q\right\|_2 \le \varepsilon_1,$$
$$(ii)\ \left\|U_Z^T\Pi^T\Pi U_X - U_Z^TU_X\right\|_2 \le \varepsilon_2,$$
$$(iii)\ \left\|U_Z^T\Pi^T\Pi\widehat{e} - U_Z^T\widehat{e}\right\| \le \varepsilon_3\left\|\widehat{e}\right\|,$$
$$(iv)\ \sigma^2_{\text{min}}(U_Z^TU_X) \ge 2f_1(\varepsilon_1, \varepsilon_2),$$

where $f_1(\varepsilon_1, \varepsilon_2) := [\varepsilon_1 + \varepsilon_2(\varepsilon_2 + 2)]/(1 - \varepsilon_1)$.

Low level conditions for Assumption 4.1(i)-(iii) are given in Cohen et al. (2016), among others. Assumption 4.1(i) is equivalent to the statement that the all eigenvalues of $U_Z^T\Pi^T\Pi U_Z$ are bounded between $[1 - \varepsilon_1, 1 + \varepsilon_1]$. This ensures that $\widetilde{Z}^T\widetilde{Z}$ is non-singular with probability at least $1 - \delta$. Part (iv) strengthens non-singularity of X^TP_ZX to require that $\sigma^2_{\text{min}}(U_Z^TU_X)$ is strictly positive and bounded below by the constant $2f_1(\varepsilon_1, \varepsilon_2)$.

Theorem 4.2. *Under Assumptions 4.1, the following holds with probability at least $1 - \delta$:*

$$\left\|\widetilde{\beta}_{2SLS} - \widehat{\beta}_{2SLS}\right\| \le \frac{f_2(\varepsilon_1, \varepsilon_2) + \varepsilon_3\left\|\widehat{e}\right\|[1 + f_2(\varepsilon_1, \varepsilon_2)]}{\sigma_{min}(X)\sigma^2_{min}(U_Z^TU_X)} \times \left[1 + \frac{2f_1(\varepsilon_1, \varepsilon_2)}{\sigma^2_{min}(U_Z^TU_X)}\right],$$

where $f_2(\varepsilon_1, \varepsilon_2) := \varepsilon_2 + \varepsilon_1/(1 - \varepsilon_1) + \varepsilon_2\varepsilon_1/(1 - \varepsilon_1)$.

The sketched estimator $\widetilde{\beta}_{2SLS}$ involves, firstly, a regression of $\widetilde{X}$ on $\widetilde{Z}$, and then a regression of $\widetilde{y}$ on the fitted values in the first step. The estimator thus depends on adequacy of subspace approximation in both steps. Theorem 4.2 provides a worst-case bound for $\widetilde{\beta}_{2SLS} - \widehat{\beta}_{2SLS}$ with the data $\mathcal{D}_n$ being fixed. It depends on (i) $\varepsilon_j, j = 1, 2, 3$, (ii) variability of $\|\widehat{e}\|$, (iii) the signal from X as given by $\sigma_{\min}(X)$, and (iv) instrument strength as given by $\sigma_{\min}(U_Z^T U_X)$. The sketched estimator can be arbitrarily close to the full sample estimate with high probability, provided that the subsample size m is sufficiently large, X is linearly independent, and the instrument Z is sufficiently relevant for X.

Though 2SLS is a two step estimator, we can still write

$$\begin{aligned}\widetilde{\beta}_{2SLS} - \widehat{\beta}_{2SLS} &= (\widetilde{A}_n - \widehat{A}_n)\widehat{g}_n + \widehat{A}_n(\widetilde{g}_n - \widehat{g}_n)\\ &\quad + (\widetilde{A}_n - \widehat{A}_n)(\widetilde{g}_n - \widehat{g}_n)\end{aligned}$$

as in the OLS case, but now $\widetilde{g}_n := \widetilde{Z}^T\widetilde{e}/n$, $\widetilde{A}_n := [(\widetilde{X}^T\widetilde{Z}/n)(\widetilde{Z}^T\widetilde{Z}/n)^{-1}(\widetilde{Z}^T\widetilde{X}/n)]^{-1}(\widetilde{X}^T\widetilde{Z}/n)(\widetilde{Z}^T\widetilde{Z}/n)^{-1}$. A statistical analysis of $\widetilde{\beta}_{2SLS}$ requires additional assumptions.

Assumption 4.3. (i) The data $\mathcal{D}_n := \{(y_i, X_i, Z_i) \in \mathbb{R}^{1+p+q} : i = 1, \ldots, n\}$ are i.i.d. with $p \leq q$. Furthermore, X and Z have singular value decomposition $X = U_X\Sigma_X V_X^T$ and $Z = U_Z\Sigma_Z V_Z^T$.

(ii) $\mathbb{E}(y_i^4) < \infty$, $\mathbb{E}(\|X_i\|^4) < \infty$, $\mathbb{E}(\|Z_i\|^4) < \infty$, and $\mathbb{E}(X_iX_i^T)$ and $\mathbb{E}(Z_iX_i^T)$ have full rank p.

(iii) The random matrix Π is independent of $\mathcal{D}_n$.

(iv) $m = m_n \to \infty$ but $m/n \to 0$ as $n \to \infty$, while p and q are fixed.

Arguments similar to those used to prove Theorem 2.2 lead to the following.

Theorem 4.4. *Let Assumption 4.3 hold and* $\mathbb{E}(Z_ie_i) = 0$. *If* Π *is RP satisfying RP(i)-(iii) with columns that are i.i.d.*

(i) Under BS, $m^{1/2}(\widetilde{\beta}_{2SLS} - \widehat{\beta}_{2SLS}) \to_d N(0, W_1)$.

(ii) Under RP, $m^{1/2}(\widetilde{\beta}_{2SLS} - \widehat{\beta}_{2SLS}) \to_d N(0, W_0)$.

Theorem 4.4 provides statistical properties of the sketched 2SLS estimator in Theorem 4.2 to complement the algorithmic results.

Theorem 4.4 states that when the data are sketched by RP, $\widetilde{\beta}$ is asymptotically normally distributed with mean $\widehat{\beta}$ and variance W_0/m. Under our assumptions, W_0 can be consistently estimated by

$$\widehat{W}_0 := \widehat{e}^T\widehat{e}\left(X^TZ(Z^TZ)^{-1}Z^TX\right)^{-1},$$

where $\widehat{e} := y - X\widehat{\beta}$ (not the residuals from the second step).

Interestingly, the asymptotic variance W_0 is the same as if the errors in the full sample regression were homoskedastic. But the result follows from estimation using sketched data rather than by assumption. This is not the case when inference is based on the full sample estimates, or the estimates computed from sketched data of the RS type. In such cases, a homoskedastic covariance weighting matrix would be inefficient since $\mathbb{E}(e_i^2|Z_i) \neq \mathbb{E}(e_i^2)$.

In the econometrics literature, the instruments are said to be relevant if $\mathbb{E}[Z_iX_i^T] \neq 0$. The latter is formalized by the rank condition in Assumption 4.3(ii). Tests for instrument relevance usually require robust standard errors corresponding to the parameter estimates in a regression of X on Z unless heteroskedasticity can be ruled out. An implication of our preceding analysis is that this is not necessary when the regression is estimated on data sketched by RP, as will be illustrated below.

5. Practical Inference

In applications, researchers would like to test a hypothesis about β_0 using a sketched estimate, and our results provide all the quantities required for inference. In the exogenous regressor case, we generically have

$$\widetilde{V}_m^{-1/2}(\widetilde{\beta}_{OLS} - \beta_0) \approx N(0, I_p)$$

where the form of $\widetilde{V}_m$ depends on Π. For any Π in BS or RP class, we can use White (1980)'s heteroskedasticity-consistent estimator:

$$\widetilde{V}_m = \widetilde{V}_{1,m} = (\widetilde{X}^T\widetilde{X})^{-1}(\sum_{i=1}^{m}\widetilde{X}_i\widetilde{X}_i^T\widetilde{e}_i^2)(\widetilde{X}^T\widetilde{X})^{-1}.$$

For Π in the RP class, we can let $\widetilde{s}_{OLS}^2 := \frac{1}{m}\sum_{i=1}^{m}(\widetilde{y}_i - \widetilde{X}_i^T\widetilde{\beta}_{OLS})^2$. Then without assuming homoskedasticity,

$$\widetilde{V}_m = \widetilde{V}_{0,m} = \widetilde{s}_{OLS}^2(\widetilde{X}^T\widetilde{X})^{-1},$$

In the endogenous regressor case, $\widetilde{W}_m^{-1/2}(\widetilde{\beta}_{2SLS} - \beta_0) \approx N(0, I_p)$. For RP, we let $\widetilde{s}_{2SLS}^2 := \frac{1}{m}\sum_{i=1}^{m}(\widetilde{y}_i - \widetilde{X}_i^T\widetilde{\beta}_{2SLS})^2$ and $\widetilde{W}_{0,m} := \widetilde{s}_{2SLS}^2\left(\widetilde{X}^T\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T\widetilde{X}\right)^{-1}$. For BS, we define $\widetilde{W}_{1m}$ from $\widetilde{W}_1$.

Sketching estimators require a choice of m. From the algorithmic perspective, m needs to be chosen as small as possible to achieve computational efficiency. If Π is constructed from SRHT, the size of m is roughly (ignoring the log factors) of order q in the best case. The requirement for countsketch is more stringent and is proved in the appendix (see Theorem B.7). In view of this, we may set

$$m_1 = C_m q \log q \quad \text{or} \quad m_1 = C_m q^2,$$

where C_m is a constant that needs to be chosen by a researcher. However, statistical analysis often cares about the variability of the estimates in repeated sampling and a larger m may be desirable from the perspective of statistical efficiency. An *inference-conscious* guide m_2 can be obtained in Lee & Ng (2020) by targeting the power at $\bar{\gamma}$ of a one-sided t-test for given nominal size $\bar{\alpha}$. In particular, let $\widetilde{\beta}$ be $\widetilde{\beta}_{OLS}$ if $\mathbb{E}[X_i e_i] = 0$ and let $\widetilde{\beta}$ be $\widetilde{\beta}_{2SLS}$ when $\mathbb{E}[X_i e_i] \neq 0$ but $\mathbb{E}[Z_i e_i] = 0$. For pre-specified effect size $c^T(\beta^0 - \beta_0)$,

$$m_2(m_1) = m_1 S^2(\bar{\alpha}, \bar{\gamma}) \left[\frac{\text{SE}(c^T \widetilde{\beta})}{c^T(\beta^0 - \beta_0)]} \right]^2,$$

where $S(\alpha, \gamma) := \Phi^{-1}(\gamma) + \Phi^{-1}(1-\alpha)$ and $\text{SE}(c^T\widetilde{\beta})$ is the standard error of $c^T\widetilde{\beta}$.

Alternatively, a data-oblivious sketch size for a pre-specified $\tau_2(\infty)$ is defined as

$$m_3 = n \frac{S^2(\bar{\alpha}, \bar{\gamma})}{\tau_2^2(\infty)}. \tag{2}$$

Note that m_3 only requires the choice of $\bar{\alpha}, \bar{\gamma}$, and $\tau_2(\infty)$ which, unlike m_2, can be computed without a preliminary sketch. The condition $m/n \to 0$ can be viewed as $\tau_2(\infty) \to \infty$ as $n \to \infty$.

6. Monte Carlo Experiments

In this section, we use Monte Carlo experiments to establish that when the errors are homoskedastic, estimates based on data sketched by random sampling or random projections will yield accurate inference. However, when the errors are heteroskedastic, sketching by random sampling will yield tests with size distortions, rejecting with much higher probability than the nominal size, unless robust standard errors are used.

6.1. When All the Regressors are Exogenous

We first consider the simulation design for which all the regressors are exogenous. The regressors $X_i = (1, X_{2,i}, \ldots, X_{p,i})^T$ consist of a constant term and a $(p-1)$-dimensional random vector $(X_{2,i}, \ldots, X_{p,i})^T$ generated from a multivariate normal distribution with mean zero vector and the variance covariance matrix Σ, whose (i,j) component is $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$. The dependent variable is generated by

$$y_i = X_i^T \beta_0 + \sigma(X_i) e_i,$$

where $\beta_0 = (0, 1, \ldots, 1)^T$, and e_i is generated from $N(0,1)$ independently from X_i. We consider two designs for $\sigma(X_i)$: (i) homoskedastic design $\sigma(X_i) = 1$ for all i and (ii) heteroskedastic design $\sigma(X_i) = \exp(X_{p,i})$, where $X_{p,i}$ is the p-th element of X_i. Throughout the Monte Carlo experiment, we set $n = 10^6$, $m = 500$, and $p = 6$. There were 5,000 replications for each experiment. Six sketching methods are considered: (i) Bernoulli sampling, (ii) uniform sampling, (iii) leverage score sampling and reweighted regression as in Ma et al. (2020); (iv) countsketch, (v) SRHT, (vi) subsampled randomized Fourier transforms using the real part of fast discrete Fourier transform (SRFT). Table 1 reports the empirical size and power of the t-test. The null and alternative hypotheses are that $H_0 : c^T\beta_0 = 1$ vs. $H_1 : c^T\beta_0 \neq 1$ with $c^T = (0, \ldots, 0, 1)$. Equivalently, the null hypothesis is $\beta_p = 1$. The power is obtained for the null value $c^T\beta_0 = 1.1$ for the homoskedastic design and $c^T\beta_0 = 1.4$ for the heteroskedastic design, respectively. The nominal size is 0.05.

Table 1. OLS based t test for $H_0 : \beta_p = 1$ vs $H_1 : \beta_p \neq 1$. S.E.0 and S.E.1 refer to homoskedasticity-only and heteroskedasticity-consistent standard errors, respectively.

	(1)	(2)	(3)	(4)
	Size		Power	
	S.E.0	S.E.1	S.E.0	S.E.1
(I) Homoskedastic Design				
Bernoulli	0.046	0.050	0.490	0.496
Uniform	0.047	0.052	0.489	0.490
Leverage	0.045	0.053	0.483	0.513
Countsketch	0.049	0.051	0.479	0.489
SRHT	0.056	0.061	0.492	0.498
SRFT	0.055	0.057	0.484	0.489
(II) Heteroskedastic Design				
Bernoulli	0.310	0.047	0.713	0.436
Uniform	0.301	0.053	0.719	0.435
Leverage	0.183	0.051	0.727	0.529
Countsketch	0.054	0.057	0.813	0.812
SRHT	0.054	0.056	0.804	0.809
SRFT	0.050	0.052	0.799	0.806

In column (1) in Table 1, we report the size of the test, namely, the probability of rejecting H_0 when the null value is true. In this column, the t-statistic is constructed using homoskedasticity-only standard errors S.E.0. Though many methods perform well, both Bernoulli and uniform sampling show substantial size distortions for the heteroskedastic design. Leverage score sampling combined with reweighted regression seems to account for heteroskedasticity to some extent, but not enough to remove all size distortions. In column (2) which reports results using robust standard errors S.E.1, all methods have satisfactory size. In column (3), we report the power of the test, i.e., the probability of rejecting H_0 when the null value is false. For the heteroskedastic design, the powers of the tests using homoskedastic standard errors S.E.0 are worse for Bernoulli, uniform and leverage samplings than those for countsketch, SRHT, and SRFT. The power loss of the RS schemes is much more pronounced when the robust standard errors are used in column (4). This efficiency loss is consistent with asymptotic theory developed in the paper because the squared regression error is positively correlated with one of the elements of squared X_i under the heteroskedastic design. All RP schemes perform similarly, hinting that even though a formal proof awaits future research, asymptotic normality may also hold for both SRHT and SRFT, and not just countsketch.

6.2. When One of the Regressors is Endogenous

We now move to the case when the regressors are $X_i = (1, X_{2,i}, \dots, X_{p-1,i}, X_{p,i})^T$, and y_i is generated by

$$y_i = X_i^T \beta_0 + \sigma_2(Z_i)(\eta_i + \epsilon_i), \tag{3}$$

where $\epsilon_i \sim N(0,1)$ is randomly drawn independently from X_i and η_i. The first $p-1$ regressors, including the intercept term, are exogenous, but

$$X_{p,i} = Z_i^T \zeta_0 + \sigma_1(Z_i)\eta_i, \tag{4}$$

where $\zeta_0 = (\zeta_{1,0}, \dots, \zeta_{q,0})^T$, $\eta_i \sim N(0,1)$ independently from $Z_i = (1, Z_{2,i}, \dots, Z_{q,i})^T$. The presence of η_i in both (3) and (4) induces endogeneity of $X_{p,i}$.

In each of the 1000 replications, $(X_{2,i}, \dots, X_{p-1,i})^T = (Z_{2,i}, \dots, Z_{p-1,i})^T$, while the $(q-1)$-dimensional Z_i is multivariate normal with mean zero and the variance Σ, whose (i,j) component is $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$. We consider two designs for $\sigma_1(Z_i)$: (i) homoskedastic design $\sigma_1(Z_i) = 1$ for all i and (ii) heteroskedastic design $\sigma_1(Z_i) = \exp\left(\frac{5}{q}\sum_{j=2}^{q}|Z_{j,i}|\right)/100$. As in the previous section, we set $n = 10^6$, $m = 500$, $p = 6$, and $q = 21$. We consider five sketching schemes and no longer include leverage score sampling since it is unclear how to implement it in the case of 2SLS. The nominal size is 0.05. Throughout, $(\zeta_{1,0}, \dots, \zeta_{p-1,0}) = (0, 0.1, \dots, 0.1)^T$, but values of $\zeta_{j,0}$ for $j \geq p$ depend on the context as explained below.

We first examine the so-called first-stage F-test for instrument relevance. In this case of a scalar endogenous regressor, the null hypothesis of irrelevant instruments amounts to a joint test of $H_0 : \zeta_{j,0} = 0$ for every $j = p, \dots, q$ in (4). The size of the test is evaluated at $\zeta_{j,0} = 0$ and the power at $\zeta_{j,0} = 0.1$ for $j = p, \dots, q$. The F-test statistic is constructed as

$$F = \frac{1}{q-p+1}\widehat{\zeta}_{-(p-1)}^T \left([\widehat{V}]_{-(p-1),-(p-1)}\right)^{-1} \widehat{\zeta}_{-(p-1)},$$

where $\widehat{\zeta}_{-(p-1)}$ is a $(q-p+1)$-dimensional vector of the OLS estimate $\widehat{\zeta}$ of regressing $X_{p,i}$ on Z_i, excluding the first $(p-1)$ elements, and $[\widehat{V}]_{-(p-1),-(p-1)}$ is the corresponding submatrix of $\widehat{V}$.

Table 2. F test for Instrument Relevance: V.0 and V.1 refer to homoskedasticity-only and heteroskedasticity-consistent asymptotic variance estimates, respectively.

	(1)	(2)	(3)	(4)
	SIZE		POWER	
	V.0	V.1	V.0	V.1
(I) HOMOSKEDASTIC DESIGN				
BERNOULLI	0.047	0.063	1.000	0.999
UNIFORM	0.049	0.063	0.997	0.999
COUNTSKETCH	0.040	0.058	1.000	0.999
SRHT	0.048	0.051	0.999	0.998
SRFT	0.050	0.052	1.000	0.999
(II) HETEROSKEDASTIC DESIGN				
BERNOULLI	0.350	0.033	0.914	0.843
UNIFORM	0.338	0.024	0.900	0.828
COUNTSKETCH	0.045	0.060	0.879	0.883
SRHT	0.038	0.052	0.897	0.895
SRFT	0.050	0.059	0.890	0.888

Table 3. 2SLS based t test for $H_0 : \beta_p = 1, H_1 : \beta_p \neq 1$

	(1)	(2)	(3)	(4)
	SIZE		POWER	
	S.E.0	S.E.1	S.E.0	S.E.1
(I) HOMOSKEDASTIC DESIGN				
BERNOULLI	0.065	0.067	0.687	0.695
UNIFORM	0.056	0.057	0.686	0.693
COUNTSKETCH	0.055	0.060	0.698	0.705
SRHT	0.043	0.046	0.710	0.714
FFT	0.061	0.068	0.704	0.703
(II) HETEROSKEDASTIC DESIGN				
BERNOULLI	0.274	0.050	0.844	0.648
UNIF	0.291	0.046	0.864	0.654
COUNTSKETCH	0.042	0.047	0.930	0.930
SRHT	0.052	0.056	0.941	0.944
FFT	0.055	0.055	0.933	0.942

In Table 2, we report the size and power of the F-test for $H_0 : \zeta_{p,0} = \zeta_{p+1,0} = \ldots = \zeta_{q,0} = 0$ using homoskedasticity-only (V.0) and heteroskedasticity-consistent (V.1) asymptotic variance estimates, respectively. As in the previous subsection, Bernoulli and uniform sampling sketches suffer from size distortions in the heteroskedastic design but V.0 is used. Tests based on V.1 have good size without sacrificing much power when the F test is constructed from data sketched by RP.

We now turn to 2SLS estimation of β_0. To ensure that the instruments are powerful enough to estimate β_0 well, we now set $\zeta_{j,0} = 0.5$ for $j = p, \dots, q$ with $\sigma_1(Z_i) = 1$ for all i. We set $\beta_0 = (0, 1, \dots, 1)^T$ and consider two designs for $\sigma_2(Z_i)$: (i) homoskedastic design $\sigma_2(Z_i) = 1$ for all i and (ii) heteroskedastic design $\sigma_2(Z_i) = \exp\left(\frac{5}{q}\sum_{j=2}^{q}|Z_{j,i}|\right)/100$.

Table 4. OLS in the empirical illustration: S.E.0 and S.E.1 refer to homoskedasticity-only and heteroskedasticity-consistent standard errors, respectively ($n = 247,199$, $m = 15,283$)

	ESTIMATE	S.E.0	S.E.1
FULL SAMPLE	0.08016	0.00036	0.00039
BERNOULLI	0.07989	0.00142	0.00158
UNIFORM	0.07931	0.00146	0.00163
LEVERAGE	0.07779	0.00144	0.00149
COUNTSKETCH	0.08105	0.00143	0.00147
SRHT	0.07975	0.00142	0.00143
SRFT	0.08296	0.00143	0.00143

Table 5. 2SLS in the empirical illustration ($n = 247,199$, $m = 61,132$)

	ESTIMATE	S.E.0	S.E.1
FULL SAMPLE	0.077	0.015	0.015
BERNOULLI	0.053	0.027	0.028
UNIFORM	0.094	0.021	0.021
COUNTSKETCH	0.076	0.021	0.023
SRHT	0.115	0.018	0.018
SRFT	0.081	0.022	0.022

As in the previous subsection, we test $H_0 : \beta_p = 1$ against $H_1 : \beta_p \neq 1$, or equivalently, $c^T = (0, \dots, 0, 1)$. The power is obtained for $\beta_p = 1.05$ in the homoskedastic design and $\beta_p = 1.10$ for the heteroskedastic design, respectively. Table 3 reports results for nominal size of 0.05. Basically, the same patterns are observed as in the previous section. Thus, simulations support the theoretical result that robust standard errors are not needed for inference when estimation is based on sketched data using sketching schemes in the RP class.

7. An Empirical Illustration

An exemplary application of the 2SLS in economics is causal inference, such as to estimate the return to education. Suppose that y_i is the wages for worker i (typically in logs) and X_i contains educational attainment $\texttt{edu}_i$ (say, years of schooling completed). Here, the unobserved random variable e_i includes worker i's unobserved ability among other things. Then, $\texttt{edu}_i$ will be correlated with e_i if workers with higher ability tends to attain higher levels of education. The least-squares estimator may not provide a consistent estimate of the return to schooling. To overcome this problem, economists use an instrumental variable that is uncorrelated with e_i but correlated with $\texttt{edu}_i$.

We now look at the OLS and 2SLS estimates of return to education in columns (1) and (2) of Table IV in Angrist & Krueger (1991). The dependent variable y is the log weekly wages, the covariates X include years of education, the intercept term and 9 year-of-birth dummies ($p = 11$). Following Angrist & Krueger (1991), the instruments Z are the exogenous regressors (i.e., the intercept and year-of-birth dummies) and a full set of quarter-of-birth (one quarter omitted) times year-of-birth interactions ($q = 1 + 9 + 3 \times 10 = 40$). Their idea was that season of birth is unlikely to be correlated with workers' ability but can affect educational attainment because of compulsory schooling laws. The full sample size is $n = 247,199$.

We start with how to choose m in this application. We focus on the data-oblivious sketch size m_3 defined in (2), as it is simpler to use. We set the target size $\alpha = 0.05$ and the target power $\gamma = 0.8$. Then, $S^2(\bar{\alpha}, \bar{\gamma}) = 6.18$. It remains to specify $\tau_2(\infty)$, which can be interpreted as the value of t-statistic when the sample size is really large.

For OLS, we take $\tau_2(\infty) = 10$ and the integer part of m_3, resulting in $m = 15,283$ (about 6% of n). Table 4 reports empirical results for the OLS estimates. For each sketching scheme, only one random sketch is drawn; hence, the results can change if we redraw sketches. Remarkably, all sketched estimates are 0.08, reproducing the full sample estimate up to the second digit. The sketched homoskedasticity-only standard errors are also very much the same across different methods. The Eicker-Huber-White standard error S.E.1 is a bit larger than the homoskedastic standard error S.E.0 with the full sample. As expected, the same pattern is observed for Bernoulli and uniform sampling, as these sampling schemes preserve conditional heteroskedasticity.

For 2SLS, as it is more demanding to achieve good precision, we take $\tau_2(\infty) = 5$ and the integer part of m_3, resulting in $m = 61,132$ (about 25% of n). Table 5 reports empirical results for the 2SLS estimates. The sketched estimates vary from 0.053 to 0.115, reflecting that the 2SLS estimates are less precisely estimated than the OLS estimates. Both types of standard errors are almost identical across all sketches for 2SLS. There are multiple topics for 2SLS unstudied in this paper. For example, we may consider the two-sample 2SLS estimator analyzed in Angrist & Krueger (1992; 1995) and Inoue & Solon (2010). This is the topic for future research.

Acknowledgements

We would like to thank anonymous reviewers for helpful comments. The first author would like to thank the European Research Council for financial support (ERC-2014-CoG646917-ROMIA) and the UK Economic and Social Research Council for research grant (ES/P008909/1) to the CeMMAP. The second author would like to thank the National Science Foundation for financial support (SES: 1558623).

References

Ahfock, D. C., Astle, W. J., and Richardson, S. Statistical properties of sketching algorithms. *Biometrika*, 108(2): 283–297, 2020.

Angrist, J. D. and Krueger, A. B. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Angrist, J. D. and Krueger, A. B. The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418): 328–336, 1992.

Angrist, J. D. and Krueger, A. B. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235, 1995.

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

Clarkson, K. L. and Woodruff, D. P. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pp. 81–90, 2013.

Clarkson, K. L. and Woodruff, D. P. Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, 63(6):1–45, 2017.

Cohen, M. B., Nelson, J., and Woodruff, D. P. Optimal Approximate Matrix Product in Terms of Stable Rank. In Chatzigiannakis, I., Mitzenmacher, M., Rabani, Y., and Sangiorgi, D. (eds.), *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 11:1–11:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-013-2. doi: 10.4230/LIPIcs.ICALP.2016.11. URL http://drops.dagstuhl.de/opus/volltexte/2016/6278.

Dobriban, E. and Liu, S. Asymptotics for sketching in least squares. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 3675–3685, 2019.

Drineas, P. and Mahoney, M. W. Lectures on randomized numerical linear algebra. In Mahoney, M. W., Duchi, J. C., and Gilbert, A. C. (eds.), *The Mathematics of Data*, pp. 1–48. AMS/IAS/SIAM, 2018.

Drineas, P., Mahoney, M., and Muthukrishnan, S. Sampling algorithms for l2 regression and applications. *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1127–1136, 2006.

Drineas, P., Mahoney, M., Muthukrishnan, S., and Sarlos, T. Faster least squares approximation. *Numerical Mathematics*, 117:219–249, 2011.

Geppert, L., Ickstadt, K., Munteanu, A., Qudedenfeld, J., and Sohler, C. Random projections for Bayesian regressions. *Statistical Computing*, 27::79–101, 2017.

Hall, P. Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of multivariate analysis*, 14(1):1–16, 1984.

Inoue, A. and Solon, G. Two-sample instrumental variables estimators. *Review of Economics and Statistics*, 92(3): 557–561, 2010.

Kane, D. M. and Nelson, J. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM*, 61(1):1–23, 2014.

Lee, S. and Ng, S. An econometric perspective on algorithmic subsampling. *Annual Review of Economics*, 12(1): 45–80, 2020.

Liu, S. and Dobriban, E. Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklRwaEKwB.

Ma, P., Mahoney, M. W., and Yu, B. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911, 2015.

Ma, P., Zhang, X., Xing, X., Ma, J., and Mahoney, M. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1026–1035. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/ma20b.html.

Martinsson, P.-G. and Tropp, J. A. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

Nelson, J. and Nguyên, H. L. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 117–126, 2013.

Raskutti, G. and Mahoney, M. W. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

Sarlos, T. Improved approximation algorithms for large matrices via random projections. *Proceedings of the 47*

IEEE Symposium on Foundations of Computer Science, 2006.

Wang, H. More efficient estimation for logistic regression with optimal subsamples. *Journal of Machine Learning Research*, 20(132):1–59, 2019.

Wang, S., Gittens, A., and Mahoney, M. W. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18(218):1–50, 2018.

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

Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

A. Appendix: Proofs for OLS

Recall that $\widehat{\beta}_{OLS} - \beta_0 = (X^T X)^{-1} X^T e$ and $\widetilde{\beta}_{OLS} - \beta_0 = (\widetilde{X}^T \widetilde{X})^{-1} \widetilde{X}^T \widetilde{e}$. Thus

$$\begin{aligned}\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS} &= \Big((\widetilde{X}^T \widetilde{X})^{-1} - (X^T X)^{-1} \Big) X^T e + (\widetilde{X}^T \widetilde{X})^{-1} \Big(\widetilde{X}^T \widetilde{e} - X^T e \Big) \\ &\quad + \Big((\widetilde{X}^T \widetilde{X})^{-1} - (X^T X)^{-1} \Big) \Big(\widetilde{X}^T \widetilde{e} - X^T e \Big) \\ &= (\widetilde{A}_n - \widehat{A}_n) \widehat{g}_n + \widehat{A}_n (\widetilde{g}_n - \widehat{g}_n) + (\widetilde{A}_n - \widehat{A}_n)(\widetilde{g}_n - \widehat{g}_n),\end{aligned}$$

where $\widetilde{g}_n := \widetilde{X}^T \widetilde{e}/n$, $\widehat{g}_n := X^T e/n$, $\widetilde{A}_n := (\widetilde{X}^T \widetilde{X}/n)^{-1}$, and $\widehat{A}_n := (X^T X/n)^{-1}$. By the law of large numbers and the continuous mapping theorem, $\widehat{A}_n - A = o_p(1)$ and by the central limit theorem, $\widehat{g}_n = O_p(n^{-1/2})$. Furthermore, by repeated applications of Theorem 3.1,

$$\text{MSE}[(\widetilde{X}^T \widetilde{e} - X^T e)/n] = O(m^{-1}) \quad \text{and} \quad \text{MSE}[(\widetilde{X}^T \widetilde{X} - X^T X)/n] = O(m^{-1}),$$

and by Chebyshev's inequality, $(\widetilde{X}^T \widetilde{e} - X^T e)/n = O_p(m^{-1/2})$ and $(\widetilde{X}^T \widetilde{X} - X^T X)/n = O_p(m^{-1/2})$. The latter combined with the continuous mapping theorem yields that $\widetilde{A}_n - \widehat{A}_n = O_p(m^{-1/2})$. Thus,

$$\begin{aligned}\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS} &= A(\widetilde{g}_n - \widehat{g}_n) + (\widehat{A}_n - A)(\widetilde{g}_n - \widehat{g}_n) + O_p(m^{-1/2} n^{-1/2} + m^{-1}) \\ &= A(\widetilde{g}_n - \widehat{g}_n) + o_p(m^{-1/2}).\end{aligned}$$

We start with asymptotic normality for Bernoulli sampling.

Proof of Theorem 2.2(i). In view of the Cramer-Wold device, it suffices to show that for any nonzero constant vector $c \in \mathbb{R}^p$,

$$m^{1/2} \left[c^T \mathbb{E}(e_i^2 X_i X_i^T) c \right]^{-1/2} c^T (\widetilde{g}_n - \widehat{g}_n) \to_d N(0,1).$$

Write

$$c^T (\widetilde{g}_n - \widehat{g}_n) = n^{-1} \sum_{i=1}^n \left(\frac{n}{m} B_{ii} - 1 \right) e_i X_i^T c.$$

Because the summands are i.i.d. with mean zero and finite variance, the central limit theorem yields the desired result immediately. □

Proof of Theorem 2.2(ii). This result is a special case of Theorem 3.3(ii) and we prove Theorem 3.3(ii) below. □

In what follows, we focus on the instance that r.dim$(\Pi) = m$. Recall that

$$n^{-1} \left(U^T \Pi^T \Pi V - U^T V \right) = n^{-1} \sum_{i=1}^n \psi_i U_i V_i + n^{-1} \sum_{i=1}^n \sum_{j=1, j \neq i}^n U_i \varphi_{ij} V_j =: T_{n1} + T_{n2},$$

where $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^n$ are vectors of certain i.i.d. random variables $(U_i, V_i) \in \mathbb{R}^2$ that are independent of Π,

$$\psi_i := \sum_{k=1}^m \Pi_{ki}^2 - 1, \; \varphi_{ij} := \sum_{k=1}^m \Pi_{ki} \Pi_{kj}.$$

There are two important cases. In case (i), T_{n1} is the leading term and T_{n2} is identically zero. The latter is true if $\varphi_{ij} = 0$ for all $i \neq j$. Methods in this class generate sketches by sampling from the full data matrix using deterministic or data dependent probabilities. The case includes random sampling with replacement (RS).

In case (ii), T_{n2} is the leading term and T_{n1} is identically zero or asymptotically negligible. Recall that for $w = (u, v, \pi_1, \dots, \pi_m)^T$,

$$\widetilde{H}(w_1, w_2) := \sum_{k=1}^{m} u_1 \pi_{k1} \pi_{k2} v_2$$

and $H(w_1, w_2) := \widetilde{H}(w_1, w_2) + \widetilde{H}(w_2, w_1)$. Then,

$$T_{n2} = n^{-1} \sum\sum_{1 \le i < j \le n} H(W_i, W_j). \tag{5}$$

Note that $H(W_i, W_j)$ is symmetric, i.e., $H(W_i, W_j) = H(W_j, W_i)$. The canonical form of Π we consider for case (ii) is random projection whose properties are given in the main text.

Before proving Theorems 3.1 and 3.3, we first prove Lemma A.1 and establish some useful lemmas.

Lemma A.1. *GP, CS, and SRHT satisfy the conditions for RP.*

Proof of Lemma A.1. For GP, it is straightforward to check all the conditions as all elements of Π are i.i.d. We omit the details. For CS, note that the columns of Π are i.i.d. and Π has only one non-zero entry in each column, hence implying that $\Pi_{ki}\Pi_{kj} = 0$ for all $k, i \neq j$. Then, it is easy to see that all the conditions are satisfied. It is more involving to check the conditions for SRHT. To do so, write

$$\Pi_{ki} = \sqrt{\frac{n}{m}} \sum_{j=1}^{n} S_{kj} H_{ji} D_{ii}.$$

Using the fact that for each k, $S_{k\ell_1} S_{k\ell_2} = 0$ whenever $\ell_1 \neq \ell_2$ (the property of uniform sampling), further write

$$\begin{aligned}
\Pi_{ki}\Pi_{kj} &= \frac{n}{m} \sum_{\ell_1=1}^{n} \sum_{\ell_2=1}^{n} S_{k\ell_1} H_{\ell_1 i} D_{ii} S_{k\ell_2} H_{\ell_2 j} D_{jj} \\
&= \frac{n}{m} \sum_{\ell=1}^{n} S_{k\ell} H_{\ell i} D_{ii} H_{\ell j} D_{jj}
\end{aligned}$$

and

$$\Pi_{ki}\Pi_{kj}\Pi_{\ell p}\Pi_{\ell q} = \frac{n^2}{m^2} \sum_{t_1=1}^{n} \sum_{t_2=1}^{n} S_{kt_1} H_{t_1 i} D_{ii} H_{t_1 j} D_{jj} S_{\ell t_2} H_{t_2 p} D_{pp} H_{t_2 q} D_{qq}.$$

Using the facts that $\mathbb{E}(S_{kj}) = n^{-1}$, $\mathbb{E}(D_{ii}) = 0$, $\sum_{j=1}^{n} H_{ji}^2 = 1$, and $|H_{ji}| = n^{-1/2}$, we have

$$\begin{aligned}
\mathbb{E}(\Pi_{ki}) &= \sqrt{\frac{n}{m}} \sum_{j=1}^{n} \mathbb{E}(S_{kj}) H_{ji} \mathbb{E}(D_{ii}) = 0, \\
\mathbb{E}(\Pi_{ki}^2) &= \frac{n}{m} \sum_{j=1}^{n} \mathbb{E}(S_{kj}) H_{ji}^2 = \frac{1}{m} \sum_{j=1}^{n} H_{ji}^2 = \frac{1}{m}, \\
\mathbb{E}(\Pi_{ki}^2 \Pi_{kj}^2) &= \frac{n^2}{m^2} \sum_{\ell=1}^{n} \mathbb{E}(S_{k\ell}) H_{\ell i}^2 H_{\ell j}^2 = \frac{n}{m^2} \sum_{\ell=1}^{n} H_{\ell i}^2 H_{\ell j}^2 = \frac{1}{m^2}, \\
\mathbb{E}(\Pi_{ki}^4) &= \frac{n^2}{m^2} \sum_{\ell=1}^{n} \mathbb{E}(S_{k\ell}) H_{\ell i}^4 = \frac{n}{m^2} \sum_{\ell=1}^{n} H_{\ell i}^4 = \frac{1}{m^2}.
\end{aligned}$$

Furthermore, note that the diagonal elements of D are i.i.d. and the rows of S are i.i.d. Then, we have that $\mathbb{E}[\Pi_{ki}\Pi_{kj}] = 0$ for all $k, i \neq j$ and $\mathbb{E}[\Pi_{ki}\Pi_{kj}\Pi_{\ell p}\Pi_{\ell q}] = 0$ for all $k \neq \ell, i \neq j, p \neq q$. Therefore, we have verified all the required conditions. □

Lemma A.2. *If* Π *is a random matrix satisfying RS, then,*

$$\mathbb{E}\left[n^{-1}\left(U^T\Pi^T\Pi V - U^TV\right)\right] = 0,$$
$$\operatorname{Var}\left[n^{-1}\left(U^T\Pi^T\Pi V - U^TV\right)\right] = \left\{\frac{1}{m} - \frac{1}{n} + \left(1 - \frac{1}{m}\right)\sum_{i=1}^{n} p_i^2\right\}\operatorname{Var}(U_iV_i).$$

In particular, when $p_i = n^{-1}$, the variance is reduced to $\frac{n-1}{n}\frac{1}{m}\operatorname{Var}(U_iV_i)$.

Proof of Lemma A.2. Using the property of RS, we have that

$$\begin{aligned}
\mathbb{E}[\psi_i] &= \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] - 1 = np_i - 1,\\
\mathbb{E}[\psi_i^2] &= \sum_{k=1}^{m}\sum_{\ell=1}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{\ell i}^2] - 2\sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] + 1\\
&= \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^4] + \sum_{k=1}^{m}\sum_{\ell=1,\ell\neq k}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{\ell i}^2] - 2\sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] + 1\\
&= \frac{n^2}{m}p_i + \frac{n^2 m(m-1)}{m^2}p_i^2 - 2np_i + 1,
\end{aligned}$$

and for $i \neq j$, using the fact that $\Pi_{ki}\Pi_{kj} = 0$ whenever $i \neq j$,

$$\begin{aligned}
\mathbb{E}[\psi_i\psi_j] &= \sum_{k=1}^{m}\sum_{\ell=1}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{\ell j}^2] - \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] - \sum_{\ell=1}^{m}\mathbb{E}[\Pi_{\ell j}^2] + 1\\
&= \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] + \sum_{k=1}^{m}\sum_{\ell=1,\ell\neq k}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{\ell j}^2] - \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] - \sum_{\ell=1}^{m}\mathbb{E}[\Pi_{\ell j}^2] + 1\\
&= \frac{n^2 m(m-1)}{m^2}p_ip_j - np_i - np_j + 1.
\end{aligned}$$

Note that $T_{n2} = 0$ because $\varphi_{ij} = 0$. Hence, it suffices to compute the mean and variance of T_{n1}. Write

$$\begin{aligned}
\mathbb{E}(T_{n1}) &= n^{-1}\sum_{i=1}^{n}\mathbb{E}(\psi_i)\mathbb{E}(U_iV_i) = n^{-1}\sum_{i=1}^{n}(np_i - 1)\mathbb{E}(U_iV_i) = 0,\\
\operatorname{Var}(T_{n1}) &= n^{-2}\sum_{i=1}^{n}\mathbb{E}(\psi_i^2)\mathbb{E}(U_i^2V_i^2) + n^{-2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}(\psi_i\psi_j)\mathbb{E}(U_iV_i)\mathbb{E}(U_jV_j)\\
&= n^{-2}\sum_{i=1}^{n}\left(\frac{n^2}{m}p_i + \frac{n^2m(m-1)}{m^2}p_i^2 - 2np_i + 1\right)\mathbb{E}(U_i^2V_i^2)\\
&+ n^{-2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(\frac{n^2m(m-1)}{m^2}p_ip_j - np_i - np_j + 1\right)\mathbb{E}(U_iV_i)\mathbb{E}(U_jV_j)\\
&= \left\{\frac{1}{m} + \left(1 - \frac{1}{m}\right)\sum_{i=1}^{n}p_i^2 - \frac{1}{n}\right\}\mathbb{E}(U_i^2V_i^2)\\
&+ \left\{\left(1 - \frac{1}{m}\right)\left(1 - \sum_{i=1}^{n}p_i^2\right) - \frac{n-1}{n}\right\}\mathbb{E}(U_iV_i)\mathbb{E}(U_jV_j)\\
&= \left\{\frac{1}{m} - \frac{1}{n} + \left(1 - \frac{1}{m}\right)\sum_{i=1}^{n}p_i^2\right\}\left\{\mathbb{E}(U_i^2V_i^2) - [\mathbb{E}(U_iV_i)]^2\right\}.
\end{aligned}$$

Therefore, we have proved the lemma. □

Lemma A.3. *If* Π *is a random matrix satisfying RP, then,*

$$\mathbb{E}\left(T_{n1}\right) = \mathbb{E}\left(T_{n2}\right) = 0, \operatorname{Var}\left(T_{n1}\right) = O(n^{-1}), \quad \textit{and} \quad \operatorname{Var}\left(T_{n2}\right) = \frac{n-1}{n}\frac{1}{m}\{\mathbb{E}(U_i^2)\mathbb{E}(V_i^2) + [\mathbb{E}(U_iV_i)]^2\}.$$

Proof of Lemma A.3. As in the proof of Lemma A.2, we have that

$$\begin{aligned}
\mathbb{E}[\psi_i] &= \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] - 1 = 0,\\
\mathbb{E}[\psi_i^2] &= \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^4] + \sum_{k=1}^{m}\sum_{\ell=1,\ell\neq k}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{\ell i}^2] - 2\sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] + 1\\
&= O(1) \quad \text{using the assumption that } \mathbb{E}[\Pi_{ki}^4] = O(m^{-1}),
\end{aligned}$$

and for $i \neq j$,

$$\begin{aligned}
\mathbb{E}[\psi_i\psi_j] &= \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] + \sum_{k=1}^{m}\sum_{\ell=1,\ell\neq k}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{\ell j}^2] - \sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^2] - \sum_{\ell=1}^{m}\mathbb{E}[\Pi_{\ell j}^2] + 1\\
&= \frac{1}{m} + \frac{m(m-1)}{m^2} - 2 + 1\\
&= 0.
\end{aligned}$$

Now write

$$\begin{aligned}
\mathbb{E}(T_{n1}) &= n^{-1}\sum_{i=1}^{n}\mathbb{E}(\psi_i)\mathbb{E}(U_iV_i) = 0,\\
\operatorname{Var}(T_{n1}) &= n^{-2}\sum_{i=1}^{n}\mathbb{E}(\psi_i^2)\mathbb{E}(U_i^2V_i^2) + n^{-2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}(\psi_i\psi_j)\mathbb{E}(U_iV_i)\mathbb{E}(U_jV_j)\\
&= O(n^{-1}).
\end{aligned}$$

Furthermore, $\mathbb{E}[H(W_i,W_j)] = \mathbb{E}[\mathbb{E}[H(W_i,W_j)|U_i,U_j,V_i,V_j]] = 0$ by RP(ii), specifically, $\mathbb{E}[\Pi_{ki}\Pi_{kj}] = 0$. Hence, $\mathbb{E}(T_{n2}) = 0$.

For the final result, write

$$\mathbb{E}[\{H(W_i,W_j)\}^2] = \mathbb{E}\left\{\left[\widetilde{H}(W_i,W_j) + \widetilde{H}(W_j,W_i)\right]\left[\widetilde{H}(W_i,W_j) + \widetilde{H}(W_j,W_i)\right]\right\}.$$

Write

$$\{H(W_i,W_j)\}^2 = T_{ij1} + T_{ij2} + 2T_{ij3}, \tag{6}$$

where

$$\begin{aligned}
T_{ij1} &:= \widetilde{H}(W_i,W_j)\widetilde{H}(W_i,W_j),\\
T_{ij2} &:= \widetilde{H}(W_j,W_i)\widetilde{H}(W_j,W_i),\\
T_{ij3} &:= \widetilde{H}(W_i,W_j)\widetilde{H}(W_j,W_i).
\end{aligned}$$

Use RP(ii)-(iii), in particular, $\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] = m^{-2}$ and $\mathbb{E}[\Pi_{ki}\Pi_{kj}\Pi_{\ell i}\Pi_{\ell j}] = 0$ whenever $k \neq \ell$, $i \neq j$, to obtain

$$\begin{aligned}
\mathbb{E}[T_{ij1}|U_i,V_i,U_j,V_j] &= \sum_{k=1}^{m}U_i^2V_j^2\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] = \frac{1}{m}U_i^2V_j^2,\\
\mathbb{E}[T_{ij2}|U_i,V_i,U_j,V_j] &= \sum_{k=1}^{m}U_j^2V_i^2\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] = \frac{1}{m}U_j^2V_i^2,\\
\mathbb{E}[T_{ij3}|U_i,V_i,U_j,V_j] &= \sum_{k=1}^{m}U_iU_jV_iV_j\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] = \frac{1}{m}U_iU_jV_iV_j.
\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[\{H(W_i,W_j)\}^2] &= \mathbb{E}[\mathbb{E}[\{H(W_i,W_j)\}^2|U_i,V_i,U_j,V_j]]\\ &= \frac{1}{m}\mathbb{E}\left(U_i^2V_j^2+U_j^2V_i^2+U_iU_jV_iV_j\right)\\ &= \frac{2}{m}\{\mathbb{E}(U_i^2)\mathbb{E}(V_i^2)+[\mathbb{E}(U_iV_i)]^2\}.\end{aligned}$$

Then, Var (T_{n2}) can be obtained by combining the U statistic formula given in (5) with RP(iii). □

Proof of Theorem 3.1. The theorem for RS follows immediately from Lemma A.2 under the condition that $\sum_{i=1}^n p_i^2 = o(m^{-1})$. The theorem for RP follows straightforwardly from Lemma A.3 using Cauchy–Schwarz inequality. We now consider BS. Recall that

$$n^{-1}\left(U^T\Pi^T\Pi V - U^TV\right) = n^{-1}\sum_{i=1}^n\left(\frac{n}{m}B_{ii}-1\right)U_iV_i$$

and that the summands are i.i.d.,

$$\mathbb{E}\left[\left(\frac{n}{m}B_{ii}-1\right)U_iV_i\right]=0, \quad \text{and} \quad \mathrm{Var}\left[\left(\frac{n}{m}B_{ii}-1\right)U_iV_i\right]=\left(\frac{n}{m}-1\right)\mathbb{E}\left(U_i^2V_i^2\right).$$

Then, the desired result follows immediately. □

In order to prove Theorem 3.3(ii), we use the central limit theorem for degenerate U-statistics of Hall (1984). For the sake of easy referencing, we reproduce it below.

Lemma A.4 (Theorem 1 of Hall (1984))**.** *Assume that* $\{W_1,\dots,W_n\}$ *are independent and identically distributed random vectors. Define*

$$\mathbb{U}_n := \sum\sum_{1\le i<j\le n} H_n(W_i,W_j).$$

Assume H_n *is symmetric,* $\mathbb{E}[H_n(W_1,W_2)|W_1]=0$ *almost surely and* $\mathbb{E}[H_n^2(W_1,W_2)]<\infty$ *for each* n*. Let* $G_n(w_1,w_2) := \mathbb{E}[H_n(W_1,w_1)H_n(W_1,w_2)]$*. If*

$$\frac{\mathbb{E}[G_n^2(W_1,W_2)]+n^{-1}\mathbb{E}[H_n^4(W_1,W_2)]}{\{\mathbb{E}[H_n^2(W_1,W_2)]\}^2}\to 0$$

as $n\to\infty$*, then*

$$\mathbb{V}_n^{-1/2}\frac{\mathbb{U}_n}{n}\to_d N(0,1),$$

where $\mathbb{V}_n := \frac{1}{2}\mathbb{E}[H_n^2(W_1,W_2)]$.

Lemma A.5. *Let* Π *be a random matrix satisfying RP. Let* $G(w_1,w_2) := \mathbb{E}[H(W_i,w_1)H(W_i,w_2)]$*. Then,*

$$\mathbb{E}[G^2(W_i,W_j)] = O(m^{-3}).$$

Proof. First, write

$$G(w_1,w_2) = \mathbb{E}\left\{\left[\widetilde{H}(W_i,w_1)+\widetilde{H}(w_1,W_i)\right]\left[\widetilde{H}(W_i,w_2)+\widetilde{H}(w_2,W_i)\right]\right\}.$$

Because $\mathbb{E}[\Pi_{ki}^2]=m^{-1}$ and $\mathbb{E}[\Pi_{ki}\Pi_{\ell i}]=0$ whenever $k\neq\ell$ for each i, we have that

$$\begin{aligned}\mathbb{E}[\widetilde{H}(W_i,w_1)\widetilde{H}(W_i,w_2)|U_i,V_i,U_j,V_j] &= \sum_{k=1}^m\sum_{j=1}^m U_i^2v_1v_2\mathbb{E}[\Pi_{ki}\Pi_{ji}]\pi_{k1}\pi_{j2}\\ &= \sum_{k=1}^m U_i^2v_1v_2\mathbb{E}[\Pi_{ki}^2]\pi_{k1}\pi_{k2}\\ &= m^{-1}\sum_{k=1}^m U_i^2v_1v_2\pi_{k1}\pi_{k2}.\end{aligned}$$

Similarly,

$$\mathbb{E}[\widetilde{H}(w_1, W_i)\widetilde{H}(w_2, W_i)|U_i, V_i, U_j, V_j] = m^{-1}\sum_{k=1}^{m} u_1 u_2 V_i^2 \pi_{k1}\pi_{k2},$$

$$\mathbb{E}[\widetilde{H}(W_i, w_1)\widetilde{H}(w_2, W_i)|U_i, V_i, U_j, V_j] = m^{-1}\sum_{k=1}^{m} U_i u_2 v_1 V_i \pi_{k1}\pi_{k2},$$

$$\mathbb{E}[\widetilde{H}(w_1, W_i)\widetilde{H}(W_i, w_2)|U_i, V_i, U_j, V_j] = m^{-1}\sum_{k=1}^{m} u_1 U_i V_i v_2 \pi_{k1}\pi_{k2}.$$

Then, by simple algebra,

$$G(W_i, W_j) = m^{-1}\sum_{k=1}^{m}\left\{\mathbb{E}[U_i^2]V_iV_j + \mathbb{E}[V_i^2]U_iU_j + \mathbb{E}[U_iV_i]U_iV_j + \mathbb{E}[U_iV_i]U_jV_i\right\}\Pi_{ki}\Pi_{kj}.$$

Using RP(ii)-(iii), write

$$\begin{aligned}
\mathbb{E}[G^2(W_i, W_j)] &= \mathbb{E}[\mathbb{E}[G^2(W_i, W_j)|U_i, V_i, U_j, V_j]] \\
&= m^{-2}\sum_{k=1}^{m}\mathbb{E}\left[\left\{\mathbb{E}[U_i^2]V_iV_j + \mathbb{E}[V_i^2]U_iU_j + \mathbb{E}[U_iV_i]U_iV_j + \mathbb{E}[U_iV_i]U_jV_i\right\}^2\right]\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2] \\
&= m^{-4}\sum_{k=1}^{m}\mathbb{E}\left[\left\{\mathbb{E}[U_i^2]V_iV_j + \mathbb{E}[V_i^2]U_iU_j + \mathbb{E}[U_iV_i]U_iV_j + \mathbb{E}[U_iV_i]U_jV_i\right\}^2\right] \\
&= O(m^{-3}),
\end{aligned}$$

which proves the lemma. □

Lemma A.6. *Let* Π *be a random matrix satisfying RP. Furthermore, assume that the columns of* Π *are i.i.d. Then, for* $i \neq j$*,* $\mathbb{E}[\{H(W_i, W_j)\}^4] = O(m^{-1})$*.*

Proof of Lemma A.6. Using (6), write

$$\{H(W_i, W_j)\}^4 = (T_{ij1} + T_{ij2} + 2T_{ij3})(T_{ij1} + T_{ij2} + 2T_{ij3}).$$

We expand the right-hand side of the equation above. The first term has the form

$$T_{ij1}T_{ij1} = \sum_{k_1=1}^{m}\sum_{k_2=1}^{m}\sum_{k_3=1}^{m}\sum_{k_4=1}^{m} U_i^4 V_j^4 \Pi_{k_1i}\Pi_{k_1j}\Pi_{k_2i}\Pi_{k_2j}\Pi_{k_3i}\Pi_{k_3j}\Pi_{k_4i}\Pi_{k_4j}.$$

Combining RP with the additional assumption that the columns of Π are i.id., we have that $\mathbb{E}[\Pi_{ki}^4\Pi_{kj}^4] = O(m^{-2})$ uniformly. Also, $\mathbb{E}[\Pi_{k_1i}\Pi_{k_1j}\Pi_{k_2i}\Pi_{k_2j}\Pi_{k_3i}\Pi_{k_3j}\Pi_{k_4i}\Pi_{k_4j}]$ is nonzero only if all four indices are the same ($k_1 = k_2 = k_3 = k_4 = k$) or two pairs of the indices are the same (e.g., $k_1 = k_2$ and $k_3 = k_4$). This implies that

$$\mathbb{E}[T_{ij1}T_{ij1}] = \mathbb{E}\left(U_i^4 V_j^4\right)\left\{\sum_{k=1}^{m}\mathbb{E}[\Pi_{ki}^4\Pi_{kj}^4] + 6\sum_{k=1}^{m}\sum_{\ell=1,\ell\neq k}^{m}\mathbb{E}[\Pi_{ki}^2\Pi_{kj}^2\Pi_{\ell i}^2\Pi_{\ell j}^2]\right\} = O(m^{-1}).$$

Moreover, using similar arguments, we can show that all other terms $\mathbb{E}[T_{ijk}T_{ij\ell}] = O(m^{-1})$, where $k, \ell \in \{1, 2, 3\}$. Therefore, we have proved the lemma. □

Lemma A.7. *Let* Π *be a random matrix satisfying RP. Furthermore, assume that the columns of* Π *are i.i.d. Then, as* $n \to \infty$*,*

$$\sqrt{m}\, T_{n2} \to_d N[0, \{\mathbb{E}(U_i^2)\mathbb{E}(V_i^2) + \mathbb{E}(U_iV_i)^2\}],$$

Proof of Lemma A.7. Note that $H(w_1, w_2) = H(w_2, w_1)$ and $\mathbb{E}(H(W_1, W_2)|W_1) = \mathbb{E}(H(W_1, W_2)|W_2) = 0$. Thus, T_{n2} is a degenerate U-statistic. By Lemmas A.3, A.5 and A.6, we have that

$$\frac{\mathbb{E}[G^2(W_1, W_2)] + n^{-1}\mathbb{E}[H^4(W_1, W_2)]}{\{\mathbb{E}[H^2(W_1, W_2)]\}^2} = O(m^{-1} + n^{-1}m) = o(1).$$

Then, the conclusion of Lemma A.7 follows directly by applying Lemma A.4 along with Lemma A.3. □

Proof of Theorem 3.3(ii). Recall that

$$\widetilde{\beta}_{OLS} - \widehat{\beta}_{OLS} = A(\widetilde{g}_n - \widehat{g}_n) + o_p(m^{-1/2}).$$

Then, the theorem follows immediately by applying Lemma A.7 with $U_i = c^T X_i$ and $V_i = e_i$ for each constant vector $c \in \mathbb{R}^p$. □

B. Appendix: Proofs for 2SLS

B.1. Proof of Theorem 4.2

Recall that using the singular value decomposition of X and Z, we write $X = U_X \Sigma_X V_X^T$ and $Z = U_Z \Sigma_Z V_Z^T$. Define

$$\widehat{\theta} := \left(U_X^T U_Z U_Z^T U_X\right)^{-1} U_X^T U_Z U_Z^T y \tag{7}$$

and

$$\widetilde{\theta} := \left(U_X^T \Pi^T \Pi U_Z \left(U_Z^T \Pi^T \Pi U_Z\right)^{-1} U_Z^T \Pi^T \Pi U_X\right)^{-1} U_X^T \Pi^T \Pi U_Z \left(U_Z^T \Pi^T \Pi U_Z\right)^{-1} U_Z^T \Pi^T \Pi y. \tag{8}$$

It would be convenient to work with $U_X\widehat{\theta}$ and $U_X\widetilde{\theta}$ in order to analyze algorithmic properties of sketched 2SLS estimators because U_X is an orthonormal matrix. The following lemma establishes the equivalence between $X\widehat{\beta}$ and $U_X\widehat{\theta}$.

Lemma B.1. *Let Assumption 4.1 hold. Then, $X\widehat{\beta} = U_X\widehat{\theta}$.*

Proof. By the singular value decomposition of X and Z, we have that

$$\begin{aligned}
Z^T Z &= V_Z \Sigma_Z^2 V_Z^T, \\
Z(Z^T Z)^{-1} Z^T &= U_Z U_Z^T, \\
X^T Z(Z^T Z)^{-1} Z^T X &= V_X \Sigma_X U_X^T U_Z U_Z^T U_X \Sigma_X V_X^T, \\
\left(X^T Z(Z^T Z)^{-1} Z^T X\right)^{-1} &= V_X \Sigma_X^{-1} \left(U_X^T U_Z U_Z^T U_X\right)^{-1} \Sigma_X^{-1} V_X^T, \\
X^T Z(Z^T Z)^{-1} Z^T y &= V_X \Sigma_X U_X^T U_Z U_Z^T y.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widehat{\beta} &= V_X \Sigma_X^{-1} \left(U_X^T U_Z U_Z^T U_X\right)^{-1} U_X^T U_Z U_Z^T y, \\
X\widehat{\beta} &= U_X \left(U_X^T U_Z U_Z^T U_X\right)^{-1} U_X^T U_Z U_Z^T y,
\end{aligned}$$

which in turn implies the conclusion in view of the definition of $\widehat{\theta}$ in (7). □

As in Lemma B.1, the equivalence between $X\widetilde{\beta}$ and $U_X\widetilde{\theta}$ holds.

Lemma B.2. *Assume that (i) $\widetilde{Z}^T\widetilde{Z}$ is non-singular and (ii) $\widetilde{X}^T\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T\widetilde{X}$ is non-singular. Then, $X\widetilde{\beta} = U_X\widetilde{\theta}$.*

Proof. As in the proof of Lemma B.1, we have that

$$\begin{aligned}
\widetilde{Z}^T\widetilde{Z} &= V_Z\Sigma_Z U_Z^T\Pi^T\Pi U_Z\Sigma V_Z^T,\\
\left(\widetilde{Z}^T\widetilde{Z}\right)^{-1} &= V_Z\Sigma_Z^{-1}\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}\Sigma_Z^{-1}V_Z^T,\\
\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T &= \Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T,\\
\widetilde{X}^T\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T\widetilde{X} &= V_X\Sigma_X U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi U_X\Sigma_X V_X^T,\\
\left(\widetilde{X}^T\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T\widetilde{X}\right)^{-1} &= V_X\Sigma_X^{-1}\left(U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi U_X\right)^{-1}\Sigma_X^{-1}V_X^T,\\
\widetilde{X}^T\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T\widetilde{y} &= V_X\Sigma_X U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi y.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\widetilde{\beta} &= V_X\Sigma_X^{-1}\left(U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi U_X\right)^{-1}U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi y,\\
X\widetilde{\beta} &= U_X\left(U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi U_X\right)^{-1}U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi y,
\end{aligned}$$

which again implies the conclusion in view of the definition of $\widetilde{\theta}$ in (8). □

Abusing the notation a bit, define now

$$\begin{aligned}
\widetilde{A} &:= U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi U_X,\\
\widehat{A} &:= U_X^T U_Z U_Z^T U_X,\\
\widetilde{B} &:= U_X^T\Pi^T\Pi U_Z\left(U_Z^T\Pi^T\Pi U_Z\right)^{-1}U_Z^T\Pi^T\Pi\widehat{e},\\
\widehat{B} &:= U_X^T U_Z U_Z^T\widehat{e}.
\end{aligned}$$

Lemma B.3. *Let Assumption 4.1 hold. Then,* $\widehat{A}^{-1}\widehat{B} = 0$.

Proof. Note that

$$\begin{aligned}
\widehat{A}^{-1}\widehat{B} &= \left(U_X^T U_Z U_Z^T U_X\right)^{-1}U_X^T U_Z U_Z^T\widehat{e}\\
&= \left(U_X^T U_Z U_Z^T U_X\right)^{-1}U_X^T U_Z U_Z^T y - \left(U_X^T U_Z U_Z^T U_X\right)^{-1}U_X^T U_Z U_Z^T X\widehat{\beta}\\
&= 0,
\end{aligned}$$

since $X\widehat{\beta} = U_X\left(U_X^T U_Z U_Z^T U_X\right)^{-1}U_X^T U_Z U_Z^T y$. □

Under Assumption 4.1, we first obtain the following lemma.

Lemma B.4. *Let Assumption 4.1 hold. Then, the following holds jointly with probability at least* $1-\delta$:

$$\begin{aligned}
\left\|\widetilde{A}-\widehat{A}\right\|_2 &\le f_1(\varepsilon_1,\varepsilon_2),\\
\left\|\widetilde{B}-\widehat{B}\right\|_2 &\le \varepsilon_3\left\|\widehat{e}\right\| + f_2(\varepsilon_1,\varepsilon_2)\left[1+\varepsilon_3\left\|\widehat{e}\right\|\right].
\end{aligned}$$

Proof. Let $\widetilde{A}_1 := U_Z^T \Pi^T \Pi U_X$, $\widetilde{A}_2 := \left(U_Z^T \Pi^T \Pi U_Z\right)^{-1}$, $\widehat{A}_1 := U_Z^T U_X$, and $\widehat{A}_2 := I$. Then we have that

$$
\begin{aligned}
& \widetilde{A} - \widehat{A} \\
& = \widetilde{A}_1^T \widetilde{A}_2 \widetilde{A}_1 - \widehat{A}_1^T \widehat{A}_2 \widehat{A}_1 \\
& = (\widetilde{A}_1 - \widehat{A}_1)^T \widetilde{A}_2 (\widetilde{A}_1 - \widehat{A}_1) + \widehat{A}_1^T \widetilde{A}_2 (\widetilde{A}_1 - \widehat{A}_1) + (\widetilde{A}_1 - \widehat{A}_1)^T \widetilde{A}_2 \widehat{A}_1 + \widehat{A}_1^T (\widetilde{A}_2 - \widehat{A}_2) \widehat{A}_1 \\
& = (\widetilde{A}_1 - \widehat{A}_1)^T \widehat{A}_2 (\widetilde{A}_1 - \widehat{A}_1) + (\widetilde{A}_1 - \widehat{A}_1)^T (\widetilde{A}_2 - \widehat{A}_2) (\widetilde{A}_1 - \widehat{A}_1) \\
& + \widehat{A}_1^T \widehat{A}_2 (\widetilde{A}_1 - \widehat{A}_1) + \widehat{A}_1^T (\widetilde{A}_2 - \widehat{A}_2) (\widetilde{A}_1 - \widehat{A}_1) \\
& + (\widetilde{A}_1 - \widehat{A}_1)^T \widehat{A}_2 \widehat{A}_1 + (\widetilde{A}_1 - \widehat{A}_1)^T (\widetilde{A}_2 - \widehat{A}_2) \widehat{A}_1 \\
& + \widehat{A}_1^T (\widetilde{A}_2 - \widehat{A}_2) \widehat{A}_1.
\end{aligned}
$$

It is straightforward to show that $\left\| \widetilde{A}_2 - \widehat{A}_2 \right\|_2 \leq \varepsilon_1/(1-\varepsilon_1)$ using Assumption 4.1(i). Since $\left\| \widehat{A}_1 \right\|_2 \leq \|U_Z\|_2 \|U_X\|_2 = 1$ and $\left\| \widehat{A}_2 \right\|_2 = 1$, we have that

$$
\begin{aligned}
\left\| \widetilde{A} - \widehat{A} \right\|_2 & \leq \varepsilon_2^2 + \varepsilon_2^2 \varepsilon_1/(1-\varepsilon_1) + 2\varepsilon_2 + 2\varepsilon_2\varepsilon_1/(1-\varepsilon_1) + \varepsilon_1/(1-\varepsilon_1) \\
& = \frac{\varepsilon_1 + \varepsilon_2(\varepsilon_2 + 2)}{1 - \varepsilon_1} = f_1(\varepsilon_1, \varepsilon_2),
\end{aligned}
$$

using Assumption 4.1. This proves the first desired result.

Now let $\widetilde{B}_1 := U_Z^T \Pi^T \Pi \widehat{e}$ and $\widehat{B}_1 := U_Z^T \widehat{e}$. Consider

$$
\begin{aligned}
\widetilde{B} - \widehat{B} & = U_X^T \Pi^T \Pi U_Z \left(U_Z^T \Pi^T \Pi U_Z\right)^{-1} U_Z^T \Pi^T \Pi \widehat{e} - U_X^T U_Z U_Z^T \widehat{e} \\
& = \widetilde{A}_1^T \widetilde{A}_2 \widetilde{B}_1 - \widehat{A}_1^T \widehat{A}_2 \widehat{B}_1 \\
& = (\widetilde{A}_1 - \widehat{A}_1)^T \widetilde{A}_2 (\widetilde{B}_1 - \widehat{B}_1) + \widehat{A}_1^T \widetilde{A}_2 (\widetilde{B}_1 - \widehat{B}_1) + (\widetilde{A}_1 - \widehat{A}_1)^T \widetilde{A}_2 \widehat{B}_1 + \widehat{A}_1^T (\widetilde{A}_2 - \widehat{A}_2) \widehat{B}_1 \\
& = (\widetilde{A}_1 - \widehat{A}_1)^T \widehat{A}_2 (\widetilde{B}_1 - \widehat{B}_1) + (\widetilde{A}_1 - \widehat{A}_1)^T (\widetilde{A}_2 - \widehat{A}_2) (\widetilde{B}_1 - \widehat{B}_1) \\
& + \widehat{A}_1^T \widehat{A}_2 (\widetilde{B}_1 - \widehat{B}_1) + \widehat{A}_1^T (\widetilde{A}_2 - \widehat{A}_2) (\widetilde{B}_1 - \widehat{B}_1) \\
& + (\widetilde{A}_1 - \widehat{A}_1)^T \widehat{A}_2 \widehat{B}_1 + (\widetilde{A}_1 - \widehat{A}_1)^T (\widetilde{A}_2 - \widehat{A}_2) \widehat{B}_1 \\
& + \widehat{A}_1^T (\widetilde{A}_2 - \widehat{A}_2) \widehat{B}_1.
\end{aligned}
$$

Since $\left\| \widehat{B}_1 \right\|_2 = \left\| U_Z^T \widehat{e} \right\|_2 \leq \|U_Z\|_2 \|\widehat{e}\| \leq \|\widehat{e}\|$, we have that

$$
\begin{aligned}
\left\| \widetilde{B} - \widehat{B} \right\|_2 & \leq \varepsilon_3 \|\widehat{e}\| + \left[\varepsilon_2 + \varepsilon_1/(1-\varepsilon_1) + \varepsilon_2\varepsilon_1/(1-\varepsilon_1)\right] \left[1 + \varepsilon_3 \|\widehat{e}\|\right] \\
& = \varepsilon_3 \|\widehat{e}\| + f_2(\varepsilon_1, \varepsilon_2) \left[1 + \varepsilon_3 \|\widehat{e}\|\right],
\end{aligned}
$$

again using Assumption 4.1. This proves the second desired result. □

Lemma B.5. *Let Assumptions 4.1 hold. Then, the following holds with probability at least* $1 - \delta$:

$$
\sigma_{min}(\widetilde{A}) \geq \frac{1}{2} \sigma_{min}^2 (U_Z^T U_X).
$$

Proof. Use the fact that for real matrices C and D,

$$
\sigma_{\min}(C + D) \geq \sigma_{\min}(C) - \sigma_{\max}(D)
$$

to obtain

$$
\sigma_{\min}(\widetilde{A}) \geq \sigma_{\min}(\widehat{A}) - \sigma_{\max}(\widetilde{A} - \widehat{A}).
$$

Then the desired result follows from the first conclusion of Lemma B.4, since

$$\sigma_{\min}(\widehat{A}) = \sigma_{\min}(U_X^T U_Z U_Z^T U_X) = \sigma^2_{\min}(U_Z^T U_X) \quad \text{and} \quad \sigma_{\max}(\widetilde{A} - \widehat{A}) \leq \left\|\widetilde{A} - \widehat{A}\right\|_2 .$$

□

Lemma B.5 implies that $\widetilde{A}^{-1}$ is well defined with probability at least $1 - \delta$.

Lemma B.6. *Let Assumptions 4.1 hold. Then, the following holds with probability at least* $1 - \delta$:

$$\left\|\widetilde{A}^{-1} - \widehat{A}^{-1}\right\|_2 \leq \frac{2 f_1(\varepsilon_1, \varepsilon_2)}{\sigma^4_{min}(U_Z^T U_X)}.$$

Proof. Write

$$\widetilde{A}^{-1} - \widehat{A}^{-1} = \widehat{A}^{-1}\left(\widehat{A} - \widetilde{A}\right)\widetilde{A}^{-1}.$$

Thus,

$$\begin{aligned}
\left\|\widetilde{A}^{-1} - \widehat{A}^{-1}\right\|_2 &\leq \left\|\widehat{A}^{-1}\right\|_2 \left\|\widehat{A} - \widetilde{A}\right\|_2 \left\|\widetilde{A}^{-1}\right\|_2 \\
&\leq \frac{2 f_1(\varepsilon_1, \varepsilon_2)}{\sigma^4_{\min}(U_Z^T U_X)}
\end{aligned}$$

since $\left\|\widehat{A}^{-1}\right\|_2 = \left[\sigma^2_{\min}(U_Z^T U_X)\right]^{-1}$, by Lemma B.4, $\left\|\widehat{A} - \widetilde{A}\right\|_2 \leq f_1(\varepsilon_1, \varepsilon_2)$ and, by Lemma B.5, $\left\|\widetilde{A}^{-1}\right\|_2 \leq 2\left[\sigma^2_{\min}(U_Z^T U_X)\right]^{-1}$ with probability at least $1 - \delta$. □

Proof of Theorem 4.2. By Lemmas B.1 and B.2,

$$X(\widetilde{\beta} - \widehat{\beta}) = U_X(\widetilde{\theta} - \widehat{\theta}),$$

so that

$$\sigma_{\min}(X)\left\|\widetilde{\beta} - \widehat{\beta}\right\| \leq \left\|\widetilde{\theta} - \widehat{\theta}\right\| .$$

Thus, it suffices to bound $\left\|\widetilde{\theta} - \widehat{\theta}\right\|$. To do so, write

$$\widetilde{y} = \Pi\left(X\widehat{\beta} + \widehat{e}\right) = \widetilde{X}\widehat{\beta} + \widetilde{e} = \Pi U_X \widehat{\theta} + \widetilde{e}, \tag{9}$$

where $\widetilde{e} = \Pi\widehat{e}$. Plugging (9) into (8) yields

$$\widetilde{\theta} - \widehat{\theta} = \widetilde{A}^{-1}\widetilde{B}.$$

Then, by Lemma B.3, we have that $\widetilde{\theta} - \widehat{\theta} = \widetilde{A}^{-1}\widetilde{B} = \widetilde{A}^{-1}\widetilde{B} - \widehat{A}^{-1}\widehat{B}$. Further, write

$$\widetilde{\theta} - \widehat{\theta} = \left(\widetilde{A}^{-1} - \widehat{A}^{-1}\right)\widehat{B} + \widehat{A}^{-1}\left(\widetilde{B} - \widehat{B}\right) + \left(\widetilde{A}^{-1} - \widehat{A}^{-1}\right)\left(\widetilde{B} - \widehat{B}\right).$$

Thus,

$$\begin{aligned}
\|\widetilde{\theta} - \widehat{\theta}\| &= \|\widetilde{A}^{-1}\widetilde{B} - \widehat{A}^{-1}\widehat{B}\|_2 \\
&\leq \left\|\widetilde{A}^{-1} - \widehat{A}^{-1}\right\|_2 \left\|\widehat{B}\right\|_2 + \left\|\widehat{A}^{-1}\right\| \left\|\widetilde{B} - \widehat{B}\right\|_2 + \left\|\widetilde{A}^{-1} - \widehat{A}^{-1}\right\| \left\|\widetilde{B} - \widehat{B}\right\|_2 \\
&\leq \frac{2 f_1(\varepsilon_1, \varepsilon_2)}{\sigma^4_{\min}(U_Z^T U_X)}\|\widehat{e}\| + \frac{\varepsilon_3\|\widehat{e}\| + f_2(\varepsilon_1, \varepsilon_2)\left[1 + \varepsilon_3\|\widehat{e}\|\right]}{\sigma^2_{\min}(U_Z^T U_X)} \\
&\quad + \frac{2 f_1(\varepsilon_1, \varepsilon_2)}{\sigma^4_{\min}(U_Z^T U_X)}\left\{\varepsilon_3\|\widehat{e}\| + f_2(\varepsilon_1, \varepsilon_2)\left[1 + \varepsilon_3\|\widehat{e}\|\right]\right\} \\
&= \frac{f_2(\varepsilon_1, \varepsilon_2) + \varepsilon_3\|\widehat{e}\|\left[1 + f_2(\varepsilon_1, \varepsilon_2)\right]}{\sigma^2_{\min}(U_Z^T U_X)}\left[1 + \frac{2 f_1(\varepsilon_1, \varepsilon_2)}{\sigma^2_{\min}(U_Z^T U_X)}\right],
\end{aligned}$$

where the last inequality follows from Assumption 4.1. □

We now specialize Theorem 4.2 for countsketch.

Theorem B.7. *Let data $\mathcal{D}_n$ be fixed, Z^TZ and X^TP_ZX are non-singular. Let $\Pi \in \mathbb{R}^{m\times n}$ be counsketch with $m \geq \max\{q(q+1), 2pq\}/(\varepsilon^2\delta)$ for some $\varepsilon \in (0, 1/3]$. Suppose that $\sigma^2_{min}(U_Z^TU_X) \geq \frac{16\varepsilon(1+\varepsilon)}{1-\varepsilon}$. and let $\underline{\sigma}* = \left[\sigma_{min}(X)\sigma^2_{min}(U_Z^TU_X)\right]^{-1}$. Then, the following holds with probability at least $1-\delta$:*

$$\left\|\widetilde{\beta}_{2SLS} - \widehat{\beta}_{2SLS}\right\| \leq \frac{4\varepsilon}{1-\varepsilon}\left[2 + 3\frac{\|\widehat{e}\|}{\sqrt{p}}\right]\underline{\sigma}^*$$

To establish Lemma B.10 given below, we first state the following known results in the literature.

Lemma B.8 (Theorem 6.2 of Kane & Nelson (2014))**.** *Distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ is defined to have $(\varepsilon, \delta, 2)$-JL (Johnson-Lindenstrauss) moments if for any $x \in \mathbb{R}^n$ such that $\|x\| = 1$,*

$$\mathbb{E}_{\Pi\sim\mathcal{D}}\left[\left|\|\Pi x\|^2 - 1\right|^2\right] \leq \varepsilon^2\delta.$$

Given $\varepsilon, \delta \in (0, 1/2)$, let $\mathcal{D}$ be any distribution over matrices with n columns with the $(\varepsilon, \delta, 2)$-JL moment property. Then, for any A and B real matrices each with n rows,

$$\mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|A^T\Pi^T\Pi B - A^TB\|_F > 3\varepsilon\|A\|_F\|B\|_F\right) < \delta.$$

Lemma B.9 (Theorem 2.9 of Woodruff (2014))**.** *Let $\Pi \in \mathbb{R}^{m\times n}$ be countsketch with $m \geq 2/(\varepsilon^2\delta)$. Then, Π satisfies the $(\varepsilon, \delta, 2)$-JL moment property.*

Lemma B.10. *Let $\Pi \in \mathbb{R}^{m\times n}$ be countsketch with $m \geq \max\{q(q+1), 2pq\}/(\varepsilon^2\delta)$ for some $\varepsilon \in (0, 1/2)$. Then, Assumption 4.1 holds with $\varepsilon_1 = \varepsilon, \varepsilon_2 = 3\varepsilon, \varepsilon_3 = 3\varepsilon p^{-1/2}$.*

Proof of Lemma B.10. As shown in the proof of Theorem 2 of Nelson & Nguyên (2013),

$$\mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|U_Z^T\Pi^T\Pi U_Z - I_q\|_2 > \varepsilon\right) < \delta,$$

provided that $m \geq q(q+1)/(\varepsilon^2\delta)$. This verifies the first condition of Assumption 4.1.

Now to verify conditions (ii) and (iii) of Assumption 4.1, note that since countsketch with $m \geq 2/(\varepsilon^2\delta)$ satisfies the $(\varepsilon, \delta, 2)$-JL moment property, we have, for any any A and B real matrices each with n rows,

$$\begin{aligned}&\mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|A^T\Pi^T\Pi B - A^TB\|_2 > 3\varepsilon\|A\|_F\|B\|_F\right)\\ &\leq \mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|A^T\Pi^T\Pi B - A^TB\|_F > 3\varepsilon\|A\|_F\|B\|_F\right) < \delta.\end{aligned}$$

Since $\|U_X\|_F^2 = p$, $\|U_Z\|_F^2 = q$ and $\|\widehat{e}\|_F = \|\widehat{e}\|$, we have that

$$\begin{aligned}\mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|U_Z^T\Pi^T\Pi U_X - U_Z^TU_X\|_2 > 3\varepsilon\sqrt{pq}\right) < \delta,\\ \mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|U_Z^T\Pi^T\Pi\widehat{e} - U_Z^T\widehat{e}\| > 3\varepsilon\sqrt{q}\|\widehat{e}\|\right) < \delta,\end{aligned}$$

provided that $m \geq 2/(\varepsilon^2\delta)$. Replacing ε with $\varepsilon/\sqrt{pq}$ yields that

$$\begin{aligned}\mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|U_Z^T\Pi^T\Pi U_X - U_Z^TU_X\|_2 > 3\varepsilon\right) < \delta,\\ \mathbb{P}_{\Pi\sim\mathcal{D}}\left(\|U_Z^T\Pi^T\Pi\widehat{e} - U_Z^T\widehat{e}\| > 3\varepsilon p^{-1/2}\|\widehat{e}\|\right) < \delta,\end{aligned}$$

provided that $m \geq 2pq/(\varepsilon^2\delta)$. Thus, we have proved Lemma B.10. □

Proof of Theorem B.7. In view of Lemma B.10, this theorem follows directly from applying Theorem 4.2 to the case when Π is a countsketch. □

B.2. Proof of Theorem 4.4

Proof of Theorem 4.4. It follows from the definition of the estimator that

$$\begin{aligned}\widetilde{\beta} &= \left(\widetilde{X}^T\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T\widetilde{X}\right)^{-1}\widetilde{X}^T\widetilde{Z}(\widetilde{Z}^T\widetilde{Z})^{-1}\widetilde{Z}^T\left(\widetilde{X}\beta_0+\widetilde{e}\right)\\ &= \beta_0+\left\{(\widetilde{X}^T\widetilde{Z}/n)(\widetilde{Z}^T\widetilde{Z}/n)^{-1}(\widetilde{Z}^T\widetilde{X}/n)\right\}^{-1}(\widetilde{X}^T\widetilde{Z}/n)(\widetilde{Z}^T\widetilde{Z}/n)^{-1}(\widetilde{Z}^T\widetilde{e}/n).\end{aligned}$$

Thus,,

$$\widehat{\beta} = \beta_0 + \left[(X^TZ/n)(Z^TZ/n)^{-1}(Z^TX/n)\right]^{-1}(X^TZ/n(Z^TZ/n)^{-1}Z^Te/n.$$

Write

$$\widetilde{\beta}-\widehat{\beta} = (\widetilde{A}_n - A)g_n + A(\widetilde{g}_n - g_n) + (\widetilde{A}_n - A)(\widetilde{g}_n - g_n),$$

where $\widetilde{g}_n = \widetilde{Z}^T\widetilde{e}/n$, $g_n = Z^Te/n$,

$$\begin{aligned}\widetilde{A}_n &= \left\{(\widetilde{X}^T\widetilde{Z}/n)(\widetilde{Z}^T\widetilde{Z}/n)^{-1}(\widetilde{Z}^T\widetilde{X}/n)\right\}^{-1}(\widetilde{X}^T\widetilde{Z}/n)(\widetilde{Z}^T\widetilde{Z}/n)^{-1},\\ A &= \left[\mathbb{E}(X_iZ_i^T)\left[\mathbb{E}(Z_iZ_i^T)\right]^{-1}\mathbb{E}(Z_iX_i^T)\right]^{-1}\mathbb{E}(X_iZ_i^T)\left[\mathbb{E}(Z_iZ_i^T)\right]^{-1}.\end{aligned}$$

As in the proof for OLS, $\widetilde{A}_n - A = o_p(1)$ and $\widetilde{g}_n - g_n = O_p(m^{-1/2})$. By the central limit theorem, $g_n = O_p(n^{-1/2})$. Hence,

$$\widetilde{\beta}-\widehat{\beta} = A(\widetilde{g}_n - g_n) + o_p\left(n^{-1/2}+m^{-1/2}\right).$$

Moreover, by Lemma A.7 that

$$m^{1/2}(\widetilde{g}_n - g_n) \to_d N[0, \mathbb{E}(e_i^2Z_iZ_i^T)].$$

Combining all the arguments above yields

$$m^{1/2}(\widetilde{\beta}-\widehat{\beta}) \to_d N[0, A\mathbb{E}(e_i^2Z_iZ_i^T)A^T],$$

which gives the conclusion of the theorem. □