Inference in Randomized Least Squares and PCA via Normality of Quadratic Forms

Abstract

Randomized sketching can make large-scale least-squares and PCA computations feasible, but it also introduces algorithmic uncertainty into the output. In this paper, we develop a framework to quantify this uncertainty via confidence intervals when using randomized sketching algorithms for two fundamental problems: least squares and principal component analysis (PCA). The dataset is treated as fixed, and the only randomness comes from the sketching algorithm. Our goal is perform statistical inference and construct confidence intervals for the full-data quantities—such as regression coefficients, PCA eigenvalues, and linear functionals of eigenvectors—from their randomized sketched analogues.

The framework applies to a broad class of sketches, including the subsampled randomized Hadamard transform (SRHT), sparse sign embeddings and CountSketch, sketching matrices with i.i.d. entries, Haar (partial uniform) sketching, and uniform subsampling. Its key step is to reduce inference to central limit theorems for quadratic forms of the form $a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n$ for fixed vectors $a_n, \tilde{a}_n$ and random sketching matrices $S_{m,n}$. We establish such central limit theorems for the sketching matrices studied, which may have broader interest. In particular, for the SRHT, our argument requires a dyadic expansion that exploits the recursive structure of the Hadamard transform.

Experiments support the theory and show that structured sketches can offer favorable computation–accuracy tradeoffs, in several regimes outperforming both Gaussian/i.i.d. sketching and recent optimal subsampling procedures.

Keywords: Random projection, subsampling, sketching algorithm, multivariate statistics, statistical inference.

Mathematics Subject Classification (2020): 62E20, 62F25, 68W20

1 Introduction

Randomized sketching can turn an otherwise infeasible PCA or least-squares computation into a routine one. The price is that the numerical answer itself becomes random: rerunning the same algorithm on the same data with a fresh sketch generally changes the output. This paper studies how to quantify that algorithmic uncertainty. More precisely, we work *data-conditionally*: the dataset is treated as fixed, and the only randomness comes from the sketching algorithm. Randomized sketching and random projections are now standard tools for accelerating large-scale linear algebra and data analysis (Halko et al., 2011a; Mahoney, 2011; Woodruff, 2014; Murray et al., 2023). Given a large data matrix $X_n \in \mathbb{R}^{n\times p}$, one draws a random sketching

matrix $S_{m,n} \in \mathbb{R}^{m\times n}$ with $m < n$ and works with the smaller matrix $\widetilde{X}_{m,n} = S_{m,n}X_n$ instead of X_n.

The inferential issue is especially transparent in principal component analysis (PCA). The eigenvalues of $X_n^\top X_n$ determine how much variation in the original dataset is explained by the principal components. If we run PCA on the sketch $\widetilde{X}_{m,n}$, then for each $i \in [p]$ we observe the *sketched eigenvalue* $\hat{\Lambda}_{m,i} = \lambda_i(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n}) = \lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n)$, whereas the quantity of real interest is the *full-data eigenvalue* $\Lambda_{n,i} = \lambda_i(X_n^\top X_n)$. The distinction matters. The numbers $\Lambda_{n,i}$ are the eigenvalues one would obtain by running PCA on the full dataset, but computing them may be expensive or impractical at scale. The numbers $\hat{\Lambda}_{m,i}$ are cheaper to compute from the sketch, but they are random because they depend on $S_{m,n}$. A central question is therefore whether the observed sketched eigenvalues can be used to construct confidence intervals for the full-data eigenvalues.

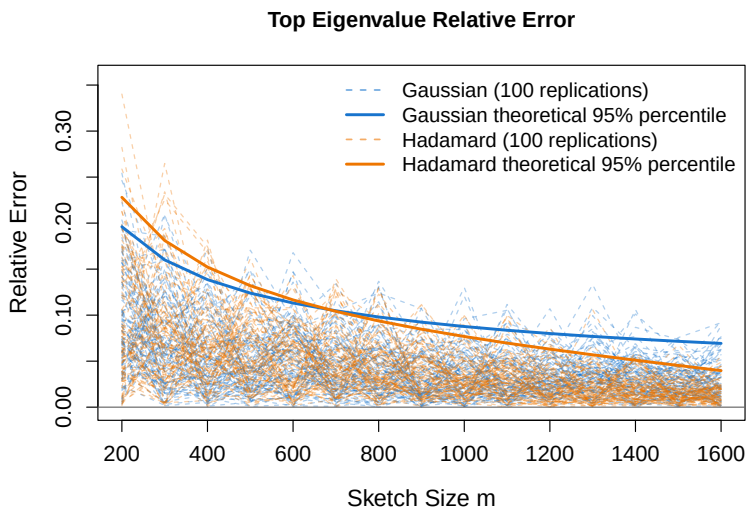


Figure 1: Relative error of the leading sketched eigenvalue under SRHT and Gaussian sketching for synthetic data generated as in Case 1 of Section 5 with $p = 15, n = 2048$. Dashed curves show 100 independent realizations of $|\hat{\Lambda}_{m,n,1} - \Lambda_{n,1}|/\Lambda_{n,1}$ for each sketch size m. Solid curves show the asymptotic 95% thresholds predicted by our theory: $z_{0.975}\sqrt{3(1-\gamma_n)/m}$ for SRHT and $z_{0.975}\sqrt{2/m}$ for Gaussian sketching, where $\gamma_n = m/n$. The curves intersect at $m = n/3$, illustrating the theoretical crossover: Gaussian sketching is more accurate for very small sketches, while SRHT becomes more accurate once m is moderately large.

The same phenomenon appears in least squares regression. For fixed data $X_n \in \mathbb{R}^{n\times p}$ and $y_n \in \mathbb{R}^n$, the target is the deterministic least-squares parameter $\beta_n = (X_n^\top X_n)^{-1} X_n^\top y_n$. Sketching replaces (X_n, y_n) by $(\widetilde{X}_{m,n}, \tilde{y}_{m,n}) = (S_{m,n}X_n, S_{m,n}y_n)$ and leads to randomized approximations such as the sketch-and-solve and partial-sketch estimators (e.g., Sarlos, 2006; Drineas et al., 2006)

$$\hat{\beta}_{m,n} = \left(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n}\right)^{-1} \widetilde{X}_{m,n}^\top \tilde{y}_{m,n} \qquad \text{and} \qquad \hat{\beta}_{m,n}^{(\mathrm{p})} = \left(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n}\right)^{-1} X_n^\top y_n. \tag{1}$$

These estimators are again random because a fresh sketch produces a fresh answer. Thus, in both PCA and least squares, the statistical task is the same: use the output of a randomized algorithm to perform inference on the deterministic quantity that would have been obtained

without sketching.

Recent work has developed inference results for particular sketching methods in least squares (Lopes et al., 2018, 2019, 2020; Wang et al., 2018a; Wang and Ma, 2021; Yu et al., 2022; Ahfock et al., 2021; Bartan and Pilanci, 2023; Lee and Ng, 2022; Zhang et al., 2023; Ma et al., 2022) and, to a lesser extent, in PCA (Lopes et al., 2020). So far, however, the theory has been largely case by case: each combination of statistical problem and sketching distribution has required its own analysis, and it has been unclear whether a unified perspective is possible. In particular, for the subsampled randomized Hadamard transform (SRHT), which is one of the most favorable sketching methods, comparable data-conditional inference results for PCA have not been available.

The SRHT is a natural test case because it is both practically important and mathematically subtle. Informally, it acts in three steps: it randomly flips signs of the coordinates of the input vector, applies a fast orthogonal Hadamard transform to mix the coordinates, and then keeps only m transformed coordinates after rescaling; Section 3.1 gives the formal definition. This sketch is attractive because it combines speed and accuracy: the Hadamard transform can be applied in $O(n \log n)$ operations, and SRHT has become a standard structured sketch in randomized numerical linear algebra (Ailon and Chazelle, 2006; Mahoney, 2011; Woodruff, 2014; Murray et al., 2023). At the same time, its strong dependence structure makes it much harder to analyze than dense Gaussian sketching. We show that this difficulty can nevertheless be overcome, and that one can obtain directly interpretable confidence intervals for full-data PCA eigenvalues under SRHT. We then show that this is in fact a special case of a much broader picture through a unified analysis that encompasses a variety of different sketching methods and both PCA and least squares.

1.1 A concrete result: sketched PCA with SRHT

We begin with a setting that aims to clearly illustrate both the promise and the mathematical challenges inherent in our approach: PCA under SRHT sketching. Recall that $X_n \in \mathbb{R}^{n\times p}$ is a fixed data matrix, let $S_{m,n}$ be an SRHT sketch, and recall $\Lambda_{n,i} = \lambda_i(X_n^\top X_n)$ and $\hat{\Lambda}_{m,i} = \lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n)$. Thus $\Lambda_{n,i}$ is the ith eigenvalue one would obtain from the full dataset (but cannot compute because X_n is not directly accessible), while $\hat{\Lambda}_{m,i}$ is the randomized eigenvalue computed from the sketch. While we expect that $\hat{\Lambda}_{m,i} - \lambda_i(X_n^\top X_n) \to_p 0$ as $m \to \infty$, knowing $\hat{\Lambda}_{m,i}$ alone does not tell us how far away it is from $\lambda_i(X_n^\top X_n)$, and how much error we are making.

Then the question we address is the following: how can we quantify the accuracy of the error of $\hat{\Lambda}_{m,i}$ for $\lambda_i(X_n^\top X_n)$? Specifically: How can we construct a confidence interval for $\lambda_i(X_n^\top X_n)$ based on $\hat{\Lambda}_{m,i}$? The next result provides an answer.

Theorem 1.1 (Informal)**.** *Let $\gamma_n = m/n < 1$, take $m, n \to \infty$ with p fixed, and choose $i \in [p]$. Under the conditions of Theorem 3.1, we have the asymptotic normality result*

$$\sqrt{\frac{m}{3(1-\gamma_n)}}\frac{\hat{\Lambda}_{m,i} - \Lambda_{n,i}}{\hat{\Lambda}_{m,i}} \Rightarrow \mathcal{N}(0,1). \tag{2}$$

Hence, if $z_{1-\alpha/2}$ denotes the standard normal $1-\alpha/2$ quantile, the set

$$\mathrm{CI}_{m,n}(\hat{\Lambda}_{m,i}) = \left[\hat{\Lambda}_{m,i}\left(1 - z_{1-\alpha/2}\sqrt{\frac{3(1-\gamma_n)}{m}}\right), \hat{\Lambda}_{m,i}\left(1 + z_{1-\alpha/2}\sqrt{\frac{3(1-\gamma_n)}{m}}\right)\right]$$

is an asymptotically valid $1-\alpha$ confidence interval for the full-data eigenvalue $\Lambda_{n,i}$, satisfying $P(\Lambda_{n,i} \in \mathrm{CI}_{m,n}(\hat{\Lambda}_{m,i})) \to_p 1-\alpha$.

Our confidence interval turns the output of a single sketched PCA computation into uncertainty quantification for the true and unobserved eigenvalue of the full dataset. Several points are worth mentioning:

- **Asymptotic exactness.** It is worth emphasizing that this confidence interval is asymptotically exact, in the sense that its coverage tends to $1-\alpha$ asymptotically as $m, n \to \infty$. This cannot be achieved by simple applications of concentration inequalities but rather requires establishing the asymptotic distribution of the randomized algorithm output.
- **No additional numerical linear algebra needed.** Remarkably, our confidence interval only depends on $\hat{\Lambda}_{m,i}$, and so it does not require any additional numerical linear algebra computation.
- **A special form for the asymptotic variance.** The asymptotic variance factor $3(1-\gamma_n)/m$ is special and only applies to SRHT. It is the result of a delicate calculation which is explained below. For other sketching distributions, for instance for Gaussian sketching or partial uniform sketching, we obtain different asymptotic variance factors and thus different intervals.
- **Approximate squared error.** The asymptotic limit of the probability distribution (2) informally also implies that the squared error is $(\hat{\Lambda}_{m,i} - \Lambda_{n,i})^2 \approx 3(1-\gamma_n)\hat{\Lambda}_{m,i}^2/m$. Under appropriate regularity conditions, the right hand side can be shown to provide a ratio-consistent estimator of the mean squared error $\mathbb{E}(\hat{\Lambda}_{m,i} - \Lambda_{n,i})^2$.

Since we view elements of the proof of Theorem 1.1 to be a key contribution, we provide a proof sketch below. The especially novel part of the proof is how we leverage the dyadic expansion of the Hadamard transform in order to find the asymptotic variance.

Proof sketch of Theorem 1.1

Write the singular value decomposition of the fixed data matrix as $X_n = U_n L_n V_n^\top$, and set $\Lambda_n = L_n^2 = \mathrm{diag}(\Lambda_{n,1}, \ldots, \Lambda_{n,p})$. Then

$$X_n^\top S_{m,n}^\top S_{m,n} X_n = V_n L_n (U_n^\top S_{m,n}^\top S_{m,n} U_n) L_n V_n^\top.$$

Since conjugation by V_n does not change eigenvalues, the random fluctuation of the sketched PCA eigenvalues is entirely encoded by the $p \times p$ matrix $U_n^\top S_{m,n}^\top S_{m,n} U_n$.

Now fix $i \in [p]$, and let $u_{n,i}$ denote the ith column of U_n. Because Condition 2.3 B gives a uniform eigengap, the map sending a symmetric matrix to its ith eigenvalue is smooth in a neighborhood of Λ_n. A first-order perturbation expansion therefore yields

$$\hat{\Lambda}_{m,i} - \Lambda_{n,i} = \Lambda_{n,i}\, u_{n,i}^\top (S_{m,n}^\top S_{m,n} - I_n) u_{n,i} + r_{m,n,i},$$

where $r_{m,n,i} = o_P(\sqrt{(1-\gamma_n)/m})$. Thus the leading fluctuation of the sketched eigenvalue is reduced to a single quadratic form of the sketch. This reduction is general and conceptually important: once quadratic forms of $S_{m,n}^\top S_{m,n}$ are understood, inference for PCA follows.

For SRHT, proving a central limit theorem for that quadratic form is the more nonstandard part. Set $a_n := u_{n,i}$, $R_{n,r} := (HDa_n)_r$, $r \in [n]$. Consider the Bernoulli random variables B_r, $r = 1, \ldots, n$, which determine whether or not we keep $R_{n,r}$ in the sub-sampling procedure of SRHT. Using these, one can rewrite

$$a_n^\top (S_{m,n}^\top S_{m,n} - I_n) a_n = \gamma_n^{-1} \sum_{r=1}^{n} (B_r - \gamma_n) R_{n,r}^2.$$

The variables B_r are independent Bernoulli random variables (independently of everything else), so the fluctuation of this triangular array with random weights $R_{n,r}^2$ can be characterized by a conditionally applied Lyapunov theorem. The challenge is to understand the conditional variance

$$\mathrm{Var}\Big(a_n^\top S_{m,n}^\top S_{m,n} a_n \mid D\Big) = \frac{1-\gamma_n}{\gamma_n} \sum_{r=1}^{n} R_{n,r}^4.$$

Since $\gamma_n = m/n$, to prove Theorem 1.1, it is enough to prove that $n \sum_{r=1}^{n} R_{n,r}^4 \to_P 3$. To see where this constant comes from, write the normalized Hadamard matrix as $H = (h_{rj})_{1 \leqslant r,j \leqslant n}$, so $R_{n,r} = \sum_{j=1}^{n} h_{rj} D_j a_{n,j}$. Because $h_{rj}^2 = 1/n$ and $\|a_n\|_2 = 1$, we may decompose

$$R_{n,r}^2 = \sum_{j=1}^{n} h_{rj}^2 a_{n,j}^2 + 2 \sum_{1 \leqslant j < k \leqslant n} h_{rj} h_{rk} D_j D_k a_{n,j} a_{n,k} = \frac{1}{n} + 2A_{n,r},$$

where $A_{n,r} := \sum_{1 \leqslant j < k \leqslant n} h_{rj} h_{rk} D_j D_k a_{n,j} a_{n,k}$. Hence

$$n \sum_{r=1}^{n} R_{n,r}^4 = n \sum_{r=1}^{n} \left(\frac{1}{n} + 2A_{n,r} \right)^2 = 1 + 4n \sum_{r=1}^{n} A_{n,r}^2,$$

because $\sum_{r=1}^{n} A_{n,r} = \sum_{1 \leqslant j < k \leqslant n} D_j D_k a_{n,j} a_{n,k} \sum_{r=1}^{n} h_{rj} h_{rk} = 0$ by the orthogonality of distinct Hadamard columns.

The nontrivial step is therefore to identify the limit of $4n \sum_r A_{n,r}^2$. Expanding the square gives

$$4n \sum_{r=1}^{n} A_{n,r}^2 = 4n \sum_{1 \leqslant j < k \leqslant n} \sum_{1 \leqslant \ell < m \leqslant n} a_{n,j} a_{n,k} a_{n,\ell} a_{n,m} D_j D_k D_\ell D_m \sum_{r=1}^{n} h_{rj} h_{rk} h_{r\ell} h_{rm}.$$

There are three types of contributions.

If $(j,k) = (\ell, m)$, then

$$4n \sum_{1 \leqslant j < k \leqslant n} a_{n,j}^2 a_{n,k}^2 \sum_{r=1}^{n} h_{rj}^2 h_{rk}^2 = 4 \sum_{1 \leqslant j < k \leqslant n} a_{n,j}^2 a_{n,k}^2 = 2 \left(1 - \sum_{j=1}^{n} a_{n,j}^4 \right).$$

A delocalization assumption implies $\|a_n\|_\infty \to 0$, hence $\sum_{j=1}^{n} a_{n,j}^4 \leqslant \|a_n\|_\infty^2 \sum_{j=1}^{n} a_{n,j}^2 = \|a_n\|_\infty^2 \to 0$, so the matched-pair contribution is $2 + o(1)$.

If exactly three distinct indices appear, then one Hadamard column is repeated and two

others are different, so the corresponding term vanishes by ordinary orthogonality. Thus only the fully distinct quadruples j, k, l, m remain. This is where the dyadic structure of the Hadamard matrix enters.

For notational simplicity, assume first that $n = 2^\ell$; the padded case is handled in exactly the same way. Let $\mathbb{F}_2$ be the finite field of size two, equipped with addition and multiplication. Identify $[n]$ with $\mathbb{F}_2^\ell$ by writing $x(j) \in \mathbb{F}_2^\ell$ for the binary expansion of $j-1$. Then the normalized Hadamard entries have the character representation $h_{rj} = n^{-1/2}(-1)^{\langle x(r), x(j)\rangle_2}$, where $\langle \cdot, \cdot \rangle_2$ denotes the inner product modulo 2. Consequently, for any four indices $j_1, j_2, j_3, j_4 \in [n]$,

$$t_{j_1,j_2,j_3,j_4} := \sum_{r=1}^{n} h_{rj_1} h_{rj_2} h_{rj_3} h_{rj_4} = \frac{1}{n^2} \sum_{x \in \mathbb{F}_2^\ell} (-1)^{\langle x,\, x(j_1) \oplus x(j_2) \oplus x(j_3) \oplus x(j_4)\rangle_2},$$

and the orthogonality of the Walsh characters yields the exact identity

$$t_{j_1,j_2,j_3,j_4} = \frac{1}{n} I\{x(j_1) \oplus x(j_2) \oplus x(j_3) \oplus x(j_4) = 0\}\,.$$

Equivalently, once three indices are fixed, there is a unique fourth index, $j_4 = j_1 \oplus j_2 \oplus j_3$, for which the fourfold Hadamard product survives. This is the key dyadic cancellation: the recursive Hadamard structure collapses a dense fourth-order interaction to a sparser family of additive quadruples on $\mathbb{F}_2^\ell$.

Let W_n denote the contribution of these fully distinct quadruples. By the preceding identity,

$$W_n = 4 \sum_{(j_1,j_2,j_3)} D_{j_1} D_{j_2} D_{j_3} D_{j_1 \oplus j_2 \oplus j_3}\, a_{n,j_1} a_{n,j_2} a_{n,j_3} a_{n,j_1 \oplus j_2 \oplus j_3},$$

where the sum is over triples for which the four displayed indices are pairwise distinct. This is a fourth-order Rademacher chaos. To show that it is negligible, it suffices to bound its second moment. Because the D_j are independent Rademacher variables, only terms in which every index appears an even number of times survive after squaring and taking expectation, so, again by a delocalization assumption,

$$\mathbb{E}_D[W_n^2] \leqslant C \sum_{j_1,j_2,j_3} a_{n,j_1}^2 a_{n,j_2}^2 a_{n,j_3}^2 a_{n,j_1 \oplus j_2 \oplus j_3}^2 \leqslant C \|a_n\|_\infty^2 \left(\sum_{j=1}^{n} a_{n,j}^2 \right)^3 = C \|a_n\|_\infty^2 \to 0$$

for a universal constant C. Hence $W_n = o_P(1)$.

Combining the matched-pair contribution with the negligible remainder gives $n \sum_{r=1}^{n} R_{n,r}^4 = 1 + 2 + o_P(1) = 3 + o_P(1)$. Once this variance limit is available, Hoeffding's inequality gives $\max_{1 \leqslant r \leqslant n} |R_{n,r}| = O_P(\sqrt{\log n / n})$, and this is enough to verify the conditional Lyapunov condition when $m / \log^2 n \to \infty$. Therefore

$$\mathrm{Var}\Big(a_n^\top S_{m,n}^\top S_{m,n} a_n \mid D \Big) = \frac{1-\gamma_n}{\gamma_n} \sum_{r=1}^{n} R_{n,r}^4 = \frac{1-\gamma_n}{m} \left(n \sum_{r=1}^{n} R_{n,r}^4 \right) = \frac{3(1-\gamma_n)}{m} + o_P(m^{-1}),$$

and the conditional central limit theorem yields

$$\sqrt{\frac{m}{3(1-\gamma_n)}}\left(a_n^\top S_{m,n}^\top S_{m,n} a_n - 1\right) \Rightarrow \mathcal{N}(0,1).$$

Retracing our steps, and using that $\hat{\Lambda}_{m,i}/\Lambda_{n,i} \to_P 1$, we obtain (2).

Comparison with Gaussian sketching. The Gaussian case provides a useful benchmark. Our framework also applies to Gaussian sketching, yielding $\sqrt{m/2}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1)$, so the limiting variance (scaled by m) is 2, compared with $3(1-\gamma_n)$ for SRHT. The reason for the difference is structural. A Gaussian sketch has independent entries and remains noisy even when m is close to n. By contrast, SRHT is a subsampled orthogonal transform: as m increases, it becomes closer to an exact geometry-preserving map, and at the endpoint $m = n$ it reduces to a signed Hadamard rotation that preserves the eigenvalues of $X_n^\top X_n$. This structural advantage appears in the factor $1 - \gamma_n$. The two asymptotic variances intersect at $m = n/3$: Gaussian sketching is more accurate for very small sketches, while SRHT becomes more accurate once the sketch is moderately large. Figure 1 shows this crossover in a simple synthetic example.

This statistical crossover is especially relevant because the two sketches also differ markedly in computational cost. For an $n \times p$ data matrix, an SRHT sketch can be applied implicitly by random sign flips, a fast Hadamard transform, and subsampling, which costs $O(np \log n)$ operations and avoids forming a dense $m \times n$ matrix. By contrast, applying a Gaussian sketch typically requires multiplying by a dense random matrix, with cost $O(mnp)$ and storage $O(mn)$. After the sketch is formed, the downstream computation of the top eigenvalue on the resulting $m \times p$ matrix has the same complexity for both methods.

This reasoning shows that the SRHT is a viable method in this problem: for lower sketch sizes, it can have a wider confidence interval than Gaussian sketching, while being faster—which leads to a user-specific trade-off. However, as we will see, Haar (partial uniform) sketches—which have the same order of complexity as Gaussian sketches—lead to more accurate inference, and our results provide the Variance formulas to use Haar sketching for asymptotic inference. For higher sketch sizes, SRHT leads to shorter confidence intervals than Gaussian sketching, while also being faster.

1.2 Main results and contributions

The above result is representative of our broader approach. Our analysis rests on studying quadratic forms $a_n^\top S_{m,n}^\top S_{m,n} \tilde{a}_n$, where a_n and $\tilde{a}_n$ are deterministic vectors determined by the data matrix and the parameter of interest. We show that once suitable central limit theorems are established for these quadratic forms, inference for both least squares and PCA follows systematically. In PCA, the quadratic-form limits yield asymptotic distributions for sketched eigenvalues and for linear functionals of sketched eigenvectors. In least squares, they yield asymptotic distributions for both sketch-and-solve and partial-sketch estimators.

Our general framework. We develop a framework that covers SRHT, CountSketch and sparse sign embeddings, uniform subsampling, sketching matrices with i.i.d. entries, Haar (partial uniform) sketching, and multisketching. As mentioned, the unifying perspective is that the least-squares and PCA problems discussed above, while different on the surface, are both con-

trolled by the same asymptotics of quadratic forms of the sketching matrix. We develop this framework in Section 2, where we formulate general conditions that make this principle precise. The case-by-case assumptions and asymptotic variance formulae for the main sketching methods are then summarized in Tables 1 and 2.

The framework applies to both sketch-and-solve (complete sketching) and partial sketching estimators in least squares (Sarlos, 2006; Drineas et al., 2006), and to inference for PCA eigenvalues and linear functionals of eigenvectors. Throughout, the theory is data-conditional: the dataset is treated as fixed, and the sketching matrix is the only source of randomness. We do not require the data to converge to a population limit or satisfy a probabilistic model; instead, we only impose mild conditions, such as a bounded condition number and, in PCA, an eigengap. Later in the paper we also explain how these data-conditional results can be combined with classical statistical models to make inference on population quantities, such as population regression coefficients and covariance eigenvalues. This approach broadens the scope of inference relative to earlier work that assumed convergence of the data (Lopes et al., 2020). At the same time, the unified perspective makes it possible to compare different sketches directly, clarifying the tradeoff between statistical accuracy and computational efficiency.

1.3 Related Work

Sketching and random projection methods have been studied extensively in multiple domains (Vempala, 2005; Li et al., 2006; Halko et al., 2011b; Mahoney, 2011; Woodruff, 2014; Drineas and Mahoney, 2016; Martinsson and Tropp, 2020; Cannings, 2021, etc). They have been applied to various problems, including PCA and SVD (e.g., Frieze et al., 2004; Halko et al., 2011b; Yang et al., 2021; Zhang and Tang, 2022, etc), linear and ridge regression (Lu et al., 2013; Chen et al., 2015; Wang et al., 2018b; Dobriban and Liu, 2019; Liu and Dobriban, 2019; Lacotte et al., 2020; Derezinski et al., 2021), two sample testing (Lopes et al., 2011; Srivastava et al., 2016), hypothesis testing in nonparametric regression (Liu et al., 2019) and in single-index models (Liu et al., 2022), convex optimization (Pilanci and Wainwright, 2015, 2016, 2017), etc. Works such as Lopes et al. (2018); Ahfock et al. (2021); Lee and Ng (2022); Bartan and Pilanci (2023) have developed methods for statistical inference in least squares regression. Ahfock et al. (2021) studies the statistical properties of complete, partial, and certain one-step corrected sketching estimators in least squares. They develop methods for finite-sample exact inference for Gaussian sketching. For Hadamard sketching and the CountSketch, they show the asymptotic normality of the sketched estimators when the sketching size is fixed, using a technical approach that is different from ours. Compared to Ahfock et al. (2021), we develop a unifying framework that includes *both PCA and least squares*, and consider a broader range of sketching distributions.

Bootstrap methods for inference using sketching have been studied separately for least squares (Lopes et al., 2018) and PCA (Lopes et al., 2020). The present work departs from these approaches in several key respects. We develop a unified framework that applies to both settings, whereas previous methods require separate treatments. Additionally, bootstrap methods provide posterior error analysis after running the sketching algorithm, whereas results such as Theorem 1.1 offer prior error analysis without execution. Prior approaches also rely on relatively strong convergence conditions on the unobserved data to ensure validity; we instead

require only weaker conditions such as a bounded condition number.

Zhang et al. (2023) proposes a framework for statistical inference via randomized algorithms. They consider arbitrary algorithms, only assuming that their output has a limiting error distribution around the true quantity or parameter of interest, without knowing what the distribution is. They develop a range of methods for statistical inference in this context, including sub-randomization, multi-run plug-in and multi-run aggregation. When the limiting distribution belongs to a known parametric family with unknown parameters, they consider inference based on estimating those parameters. Our setting belongs to this latter case, and our main contribution is to show how to derive the limiting distributions in a unifying way in sketching for linear regression and PCA.

Wang et al. (2018a); Wang and Ma (2021); Yu et al. (2022) investigate optimal subsampling for logistic regression, quantile regression models as well as for more general maximum quasi-likelihood estimators. They establish consistency and asymptotic normality of weighted subsampling estimators, and derive optimal subsampling probabilities that minimize the asymptotic mean squared error, as well as methods to estimate these probabilities. Yu et al. (2025) develops a sequential subsampling strategy for online estimation in multinomial logistic models. These approaches impose a regression model on the data, while we consider the data as fixed. Ma et al. (2022) studies weighted sampling estimators in a fixed-data setting, where the sample size is constant. In contrast, we consider a growing sample size.

1.4 Organization of the paper

Section 2 presents the unifying framework: we formalize the link between inference and quadratic forms, state the general conditions, and give the main theorem for sketch-and-solve least squares (Theorem 2.5) and the main theorem for PCA (Theorem 2.9). The appendix records the parallel abstract result for partial sketching. The assumptions are built up so that the concrete cases (SRHT, CountSketch, etc.) can be read off from Tables 1 and 2; the reader who prefers a case-by-case view can rely on those tables. Section 3 specializes to each sketching distribution and provides the proofs of the quadratic-form CLTs. Results related to SRHT are given in Section 3.1, while those for SSE and CountSketch are provided in Section 3.2. Section 3.3 studies uniform sampling. Further applications of our framework include multisketching and iterative Newton sketching, which are presented in Section 3.4 and Section 4, respectively. We then present numerical experiments in Section 5, including a comparison with certain optimal subsampling methods. Proofs and additional results are in the Appendix.

1.5 Notation and Definitions

We begin with the notation that is used throughout the paper. To keep the exposition lighter, we introduce only the basic objects here. Additional notation will be introduced as needed.

Basic notation. For a positive integer $p \geqslant 1$, we write $[p] = \{1, \ldots, p\}$. For any $n \geqslant 1$, we define $\mathbb{S}^{n-1}$ to be the unit sphere in $\mathbb{R}^n$. We write $a_n = \Theta(b_n)$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$.

Matrix notation. Let $\mathbb{S}_+^{p\times p}$ be the set of $p \times p$ positive definite matrices. For a vector $(v_1, \ldots, v_n)^\top$, we write $\operatorname{diag}(v_1, \ldots, v_n)$ for the diagonal matrix with diagonal entries v_i, $i \in [n]$.

For two matrices A, B of the same size $k \times l$, $A \odot B$ denotes the Hadamard product, with $(A \odot B)_{ij} = A_{ij}B_{ij}$ for all $(i,j) \in [k] \times [l]$.

2 Unifying Approach via Normality of Quadratic Forms

In this section we present the main abstract results behind our unified treatment of randomized least squares and randomized PCA. The common theme is that, after rewriting the sketched estimators in a suitable basis, the problem reduces to understanding the asymptotic behavior of quadratic forms involving $S_{m,n}^\top S_{m,n}$.

2.1 Unifying Conditions

We start with least squares, because it makes the basic mechanism especially transparent. Recall that the sketch-and-solve least squares estimator from (1) can be written as $\hat{\beta}_{m,n} = (X_n^\top S_{m,n}^\top S_{m,n} X_n)^{-1} X_n^\top S_{m,n}^\top S_{m,n} y_n$. Thus the key random objects are bilinear forms of the type $a_n^\top S_{m,n}^\top S_{m,n} \tilde{a}_n$ for appropriate deterministic vectors $a_n, \tilde{a}_n$. The same phenomenon appears for partial sketching and, in the PCA problem, for the singular values and singular vectors of the sketched data. Our analysis below formalizes this observation.

To isolate the relevant directions, let the SVD[1] of X_n be $X_n = U_n L_n V_n^\top$. Then

$$\hat{\beta}_{m,n} = V_n L_n^{-1} (U_n^\top S_{m,n}^\top S_{m,n} U_n)^{-1} U_n^\top S_{m,n}^\top S_{m,n} y_n. \tag{3}$$

Equation (3) shows that the left singular vectors U_n are a central geometric object: the sketch enters through $U_n^\top S_{m,n}^\top S_{m,n} U_n$ and $U_n^\top S_{m,n}^\top S_{m,n} y_n$. It is therefore natural to separate the signal subspace generated by U_n from its orthogonal complement.

Accordingly, we introduce an $n \times (n-p)$ orthogonal complement $U_{n,\perp}$ of U_n, so that $[U_n, U_{n,\perp}]$ is an $n \times n$ orthogonal matrix. We also let $\varepsilon_n = y_n - X_n \beta_n$ be the—deterministic—true residuals. Since β_n is the exact OLS parameter, ε_n is orthogonal to $\mathrm{col}(X_n) = \mathrm{col}(U_n)$. Whenever normalization is needed, we write $\bar{\varepsilon}_n = \varepsilon_n / \|\varepsilon_n\|$ and $\overline{X_n \beta_n} = X_n \beta_n / \|X_n \beta_n\|$, assuming these are well-defined when used. Finally, we write $\Lambda_n = L_n^2$.

The precise quadratic-form conditions depend on the sketching distribution. In general, more "random" sketches require weaker structural assumptions on the deterministic vectors involved. For instance, Gaussian sketches require weak conditions (but are computationally impractical), while uniform subsampling requires stronger conditions. To keep the theory unified, we formulate the needed assumptions abstractly through Conditions $\mathfrak{p}$, $\mathfrak{P}$, and $\mathfrak{P}'$, which will be instantiated for each sketch in Section 3. Condition $\mathfrak{p}$ refers to sequences of pairs of vectors on which we only need an appropriate form of tightness; analogously, Condition $\mathfrak{P}$ requires a central limit theorem. In contrast, Condition $\mathfrak{P}'$ is the corresponding condition on the sequence $(U_n)_{n \geqslant 1}$ that is needed for matrix-valued limits.

Condition 2.1 (Quadratic-form setup)**.** *Consider an asymptotic regime in which $n \to \infty$, $m = m_n \to \infty$ with $m < n$, while p is fixed. Let $(S_{m,n})_{m,n \geqslant 1}$ be a sequence of random $m \times n$*

[1] Here, U_n denotes a $n \times p$ partial orthogonal matrix satisfying $U_n^\top U_n = I_p$, $L_n = \mathrm{diag}(\ell_{n,1}, \ldots, \ell_{n,p})$, where $\ell_{n,1} \geqslant \ldots \geqslant \ell_{n,p} > 0$ are the singular values of X_n in non-increasing order, and V_n is a $p \times p$ orthogonal matrix containing the right singular vectors of X_n.

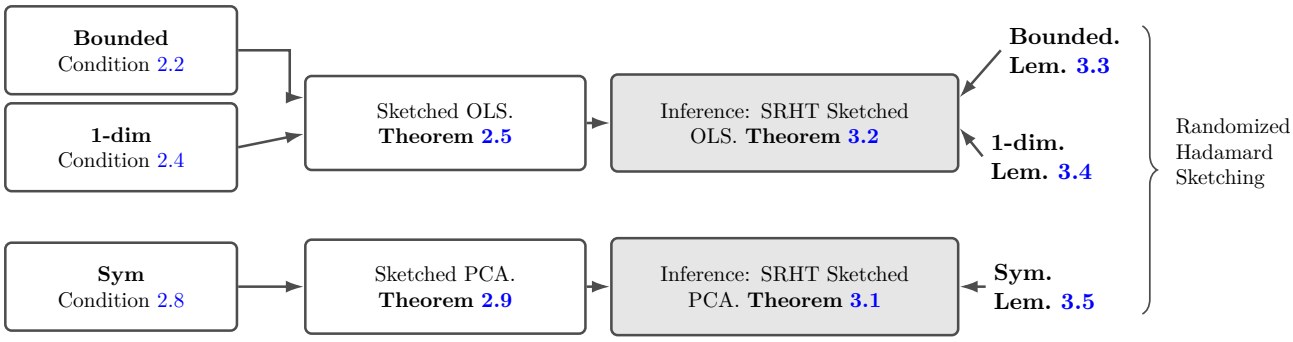


Figure 2: A flowchart illustrating how the abstract quadratic-form conditions and the SRHT lemmas combine to give the inference results.

sketching matrices, and let U_n be the $n \times p$ left singular matrix of the deterministic matrix X_n. Let $\bar{\varepsilon}_n = \varepsilon_n / \|\varepsilon_n\|$ and $\overline{X_n\beta_n} = X_n\beta_n / \|X_n\beta_n\|$, assuming they are well-defined when used. We also fix a sequence of positive scalars $(\tau_{m,n})_{m,n\geqslant 1}$ such that $\tau_{m,n}/m \to 0$. Consider Conditions $\mathfrak{p}$ and $\mathfrak{P}$, which refer to sequences of vector pairs $(a_n, \tilde{a}_n)_{n\geqslant 1}$ with $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$, and Condition $\mathfrak{P}'$ for the sequence of matrices $(U_n)_{n\geqslant 1}$.

As mentioned above, these conditions are kept abstract here and will be made concrete in the applications to specific sketching distributions. For instance, Condition $\mathfrak{p}$ can be taken to be all sequences $(a_n, \tilde{a}_n)_{n\geqslant 1}$ for Gaussian sketching matrices, while it requires that the entries of the vectors are sufficiently small for uniform subsampling.

We now describe a number of generic conditions used in our analysis. The first boundedness condition is used to show that appropriate error terms vanish.

Condition 2.2 (Boundedness under Condition $\mathfrak{p}$)**.** *For any sequence $(a_n, \tilde{a}_n)_{n\geqslant 1}$ satisfying Condition $\mathfrak{p}$, we have $\sqrt{m}\tau_{m,n}^{-1/2}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) = O_P(1)$.*

The factor $\tau_{m,n}$ records the sketch-dependent speed of convergence beyond the universal $\sqrt{m}$ scaling. The next mild regularity condition on the data has two components:

Condition 2.3 (Condition Number and Eigengap)**.** *For a sequence $(X_n)_{n\geqslant 1}$ of $n \times p$ data matrices, where $\ell_{n,1} \geqslant \ldots \geqslant \ell_{n,p} > 0$ are the singular values of X_n, consider the following conditions:*

- **A: Bounded Condition Number.** *We have $\limsup_{n\to\infty} \ell_{n,1}/\ell_{n,p} < \infty$.*
- **B: Global Relative Eigengap.** $\liminf_{n\to\infty} \min_{i\in[p-1]} \frac{\ell_{n,i}}{\ell_{n,i+1}} > 1$ *and* A *holds.*

Condition 2.3 A is a pure conditioning assumption: it keeps the singular directions well-scaled and is needed for tightness arguments in the proofs for sketched OLS. Condition B requires uniformly nonvanishing relative eigengaps. In particular, it implies that the singular values are distinct, which is what allows us to perform inference for individual eigenvalues and eigenvectors via sketched PCA.[2]

[2]One could relax this assumption and instead aim at inference for singular values associated with arbitrary singular spaces; we leave this to future work.

2.2 Asymptotic Distributions under Unifying Conditions

2.2.1 Sketched Least Squares

We first provide a result for sketched least squares. This requires a one-dimensional quadratic form condition.

Condition 2.4 (One-dimensional quadratic form under Condition $\mathfrak{P}$)**.** *For any sequence of vectors $(a_n, \tilde{a}_n)_{n\geqslant 1}$ satisfying Condition $\mathfrak{P}$, there is a sequence $\left(\sigma_n^2(a_n, \tilde{a}_n)\right)_{n\geqslant 1}$ uniformly bounded in $(0, \infty)$ for which*

$$\sqrt{m}\tau_{m,n}^{-1/2}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right)\sigma_n^{-1}(a_n, \tilde{a}_n) \Rightarrow \mathcal{N}(0, 1).$$

Moreover, for all $n \geqslant 1$, there is $M_n = M_n(\bar{\varepsilon}_n) \in \mathbb{R}^{p\times p}$, also possibly depending on U_n, such that for any $b_n \in \mathbb{S}^{p-1}$ with $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$ satisfying Condition $\mathfrak{P}$, σ_n^2 can be expressed in the form

$$\sigma_n^2(U_n b_n, \bar{\varepsilon}_n) = b_n^\top M_n(\bar{\varepsilon}_n) b_n. \tag{4}$$

The quantity $\sigma_n^2(a_n, \tilde{a}_n)$ is the asymptotic variance of a scalar quadratic form. As a special case, for $\psi \in \{0, 1\}$, we consider Condition `1-dim'`, in which $\sigma_n^2(a_n, \tilde{a}_n) = 1 + (\psi + 1)(a_n^\top \tilde{a}_n)^2$ for all $n \geqslant 1$ and $(a_n, \tilde{a}_n)_{n\geqslant 1}$ satisfying Condition $\mathfrak{P}$.

It is not hard to see that under the special case `1-dim'` in Condition 2.4, we can take $M_n = I_p$. In most cases—though not always—we set $\mathfrak{p} = \mathfrak{P}$ and derive Condition 2.2 from Condition 2.4. We state Condition 2.2 separately because it is strictly more general and is sometimes available under weaker assumptions than a full central limit theorem.

Recall that the deterministic target is the exact OLS parameter $\beta_n = (X_n^\top X_n)^{-1} X_n^\top y_n$. Under the quadratic-form conditions above and Condition 2.3, we obtain an asymptotic distribution for the sketch-and-solve estimator together with an empirical normalization that leads directly to statistical inference. The parallel abstract theorem for partial sketching is deferred to Appendix B. See Section G.1 for the proof of the next theorem.

Theorem 2.5 (Statistical Inference for Sketch-and-Solve OLS)**.** *Let $(X_n)_{n\geqslant 1}$ be a sequence of data matrices satisfying Condition 2.3* `A`*, and let $(S_{m,n})_{m,n\geqslant 1}$ be a sequence of sketching matrices. Suppose that for any $(b_n)_{n\geqslant 1}$ satisfying $b_n \in \mathbb{S}^{p-1}$ for all n, the pairs $(U_n b_n, U_n b_n)_{n\geqslant 1}$ and $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$ satisfy Condition $\mathfrak{P}$. Then we have the following two conclusions.*

Sketch-and-solve. *Under Conditions 2.2 and 2.4, the sketch-and-solve estimator $\hat{\beta}_{m,n}$ from (1) has the limiting distribution*

$$m^{1/2}\tau_{m,n}^{-1/2}\left(V_n L_n^{-1} M_n(\bar{\varepsilon}_n) L_n^{-1} V_n^\top\right)^{-1/2} \|\varepsilon_n\|^{-1}(\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p), \tag{5}$$

with M_n from (4).

Asymptotic inference under `1-dim'`. *If the special case `1-dim'` in Condition 2.4 holds for $\psi \in \{0, 1\}$, then we further obtain the following inferential result for the unobserved least-squares parameter β_n based on the observed $\widetilde{X}_{m,n}, \tilde{y}_{m,n}$, and $\tilde{\varepsilon}_{m,n} = \tilde{y}_{m,n} - \widetilde{X}_{m,n}\hat{\beta}_{m,n}$:*

$$m^{1/2}\tau_{m,n}^{-1/2}\|\tilde{\varepsilon}_{m,n}\|^{-1}(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{1/2}(\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p). \tag{6}$$

The theorem is best read in two layers. The first statement, (5), is the asymptotic normality result for sketch-and-solve least squares. It says that once the error is rescaled by $\|\varepsilon_n\|$ and by the covariance proxy $V_n L_n^{-1} M_n(\bar{\varepsilon}_n) L_n^{-1} V_n^\top$, the limiting law is standard Gaussian. The display (6) is the directly usable inferential version, because the normalization is written in terms of the sketched quantities themselves. Consequently, it can be used in the standard way to construct confidence regions for the unknown parameter β_n from the observed sketched data $\widetilde{X}_{m,n}$ and $\tilde{y}_{m,n}$.

The abstract form (5) is useful because, for any particular sketch, proving the theorem reduces to identifying the corresponding matrix M_n in (4). The parallel partial-sketching theorem is stated in Appendix B. To see this mechanism concretely, suppose that the special case **`1-dim'`** in Condition 2.4 holds. Then $M_n = I_p$, because $(U_n b_n)^\top \bar{\varepsilon}_n = 0$ for every $b_n \in \mathbb{S}^{p-1}$: the vector $U_n b_n$ lies in $\mathrm{col}(X_n)$, whereas $\bar{\varepsilon}_n$ is orthogonal to $\mathrm{col}(X_n)$. Therefore (5) reduces to

$$m^{1/2} \tau_{m,n}^{-1/2} \left(X_n^\top X_n \right)^{-1/2} \|y_n - X_n \beta_n\|^{-1} (\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p). \tag{7}$$

The asymptotic covariance matrix can then be estimated from the sketched data by a plug-in approach, as in (6).

The column for $\sigma_n^2(a_n, \tilde{a}_n)$ in Table 1 shows that the special case **`1-dim'`** in Condition 2.4 holds for all sketching methods studied here except uniform subsampling. A detailed case-by-case investigation leading to Table 1 is provided in Section 3. The last column of the table records the corresponding structural condition $\mathfrak{P}$, making precise the general principle that less "random" sketches require stronger regularity of the data.

Table 1: Summary of our results for quadratic forms $a_n^\top S_n^\top S_n \tilde{a}_n$. For various distributions of sketching matrices, we summarize the choices of $\tau_{m,n}$, $\sigma_n^2(a_n, \tilde{a}_n)$, and Condition $\mathfrak{P}$ used in Condition 2.4 for which we show Theorem 2.5; and G_n used in Condition 2.8 for which we show Theorem 2.9. Here $\gamma_n = m/n$, and as defined in Definition 2.7, $Q_p = \mathrm{vec}(I_p)\mathrm{vec}(I_p)^\top$, while P_p is the $p^2 \times p^2$ matrix representing matrix transposition in the vectorized matrix space. Also, $u_{n,i}$, $i \in [n]$ are the rows of the $n \times p$ left singular matrix U_n of X_n. See Section 3 for the details.

Sketching Distribution	$\tau_{m,n}$	$\sigma_n^2(a_n, \tilde{a}_n)$	G_n	Condition $\mathfrak{P}$
Hadamard	$1-\gamma_n$	$1 + 2(a_n^\top \tilde{a}_n)^2$	$I_{p^2} + P_p + Q_p$	ℓ_∞-delocalization
CountSketch/SSE	1	$1 + (a_n^\top \tilde{a}_n)^2$	$I_{p^2} + P_p$	ℓ_4-delocalization
i.i.d., $\kappa_{n,4} - 3 = o(1)$	1	$1 + (a_n^\top \tilde{a}_n)^2$	$I_{p^2} + P_p$	
Haar	$1-\gamma_n$	$1 + (a_n^\top \tilde{a}_n)^2$	$I_{p^2} + P_p$	
Uniform Subsampling	$1-\gamma_n$	$n \sum_{i=1}^n a_{n,i}^2 \tilde{a}_{n,i}^2$	$n \sum_{i=1}^n \left((u_{n,i} u_{n,i}^\top) \otimes (u_{n,i} u_{n,i}^\top) \right)$	See Section 3.3

Our theorem applies to *arbitrary fixed data* X_n, y_n and imposes no distributional assumptions on them. In particular, the inference is conditional on the observed data and randomness comes only from the sketch. The sketched solution can therefore be used as a basis for inference on the population regression coefficient when the observed responses arise from a noisy model; see Section C for details.

2.2.2 Sketched PCA

We now turn to sketch-and-solve principal component analysis. It turns out that this analysis requires a joint CLT for multiple quadratic forms (written in a matrix).

To describe such limiting distributions of symmetric matrices, it is convenient to vectorize them and then keep only the nonredundant upper-triangular coordinates. Hence, we define $\mathrm{vec} : \mathbb{R}^{k\times l} \to \mathbb{R}^{kl}$ as the vectorization operator that stacks the columns of a matrix from left to right. We shall also use the Kronecker delta δ_{ij}, which equals one if $i = j$ and zero otherwise.

Definition 2.6 (Upper Triangulation Operator)**.** *For every $p \in \mathbb{N}_{>0}$, we consider the upper triangulation operator $D_p : \mathbb{R}^{p^2} \to \mathbb{R}^{p(p+1)/2}$, which in matrix form is given by $(D_p)_{(ij),(kl)} = \delta_{ik}\delta_{jl}$, $i, j, k, l \in [p]$, and $k \leqslant l$.*

We also need the following two special matrices, which allow us to capture the dependencies between the entries of the vectorized quadratic form. These arise due to reusing the same randomness multiple times, as well as the same directions in U_n, in the quadratic form.

Definition 2.7 (Special Matrices Q_p and P_p)**.** *For a positive integer p, we denote $Q_p = \mathrm{vec}(I_p)\mathrm{vec}(I_p)^\top$. Moreover, we denote by P_p the $p^2 \times p^2$ matrix representing matrix transposition in the vectorized matrix space, so that $(P_p)_{(ij),(kl)} = \delta_{il}\delta_{jk}$ for all $i, j, k, l \in [p]$.*

The condition we require is then as follows:

Condition 2.8 (Matrix quadratic form under Condition $\mathfrak{P}'$)**.** *For $(U_n)_{n\geqslant 1}$ satisfying Condition $\mathfrak{P}'$, there exists a sequence of $p^2 \times p^2$ symmetric matrices $(G_n)_{n\geqslant 1}$ such that there are $0 < c < C < \infty$ for which the upper triangulation operator D_p from Definition 2.6 satisfies $cI_{p(p+1)/2} \preceq D_p^\top G_n D_p \preceq CI_{p(p+1)/2}$, and*

$$\sqrt{m}(\tau_{m,n} D_p^\top G_n D_p)^{-1/2} D_p^\top \mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) \Rightarrow \mathcal{N}(0, I_{p(p+1)/2}). \tag{8}$$

As a special case, we consider Condition `Sym-1`*, under which there exists a $p^2 \times p^2$ symmetric matrix G such that*

$$\sqrt{m}\tau_{m,n}^{-1/2}\mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) \Rightarrow \mathcal{N}(0, G). \tag{9}$$

As an even more special case, for $\psi \in \{0, 1\}$, we consider Condition `Sym-1'`*, in which, for P_p and Q_p from Definition 2.7, $G = I_{p^2} + P_p + \psi Q_p$.*

The matrix G_n is the asymptotic covariance of the matrix quadratic form in the `Sym` condition. Table 1 summarizes how these objects are instantiated for the sketching distributions considered in this paper.

For a $p\times p$ symmetric matrix Ξ and each $i \in [p]$, we denote by $\lambda_i(\Xi)$ the i-th largest eigenvalue of Ξ, and by $v_i(\Xi)$ an associated eigenvector whose first nonzero coordinate is positive. We may regard λ_i and v_i as continuous functions on the space of symmetric $p\times p$ matrices, and when Ξ has distinct eigenvalues they are differentiable in a neighborhood of Ξ; see, for example, Magnus and Neudecker (2019). In particular, we define $\Lambda_{n,i} = \lambda_i(X_n^\top X_n)$ and $\hat{\Lambda}_{m,i} = \lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n)$ for $i \in [p]$. We also write $v_{n,i}$ for an eigenvector of $X_n^\top X_n$ associated with $\Lambda_{n,i}$, and $\hat{v}_{m,n,i}$ for an eigenvector of $X_n^\top S_{m,n}^\top S_{m,n} X_n$ associated with $\hat{\Lambda}_{m,i}$.

To state the eigenvector limit, we need one further object. For each $i \in [p]$, the matrix Δ_i below captures the asymptotic covariance structure of the i-th eigenvector after first-order

perturbation. For $\Xi \in \mathbb{S}_+^{p\times p}$ with distinct eigenvalues and a $p^2 \times p^2$ symmetric matrix G, define

$$\Delta_i(\Xi, G) := \sum_{k\neq i, l\neq i} \frac{\lambda_i(\Xi)\sqrt{\lambda_k(\Xi)\lambda_l(\Xi)}}{(\lambda_i(\Xi) - \lambda_k(\Xi))\,(\lambda_i(\Xi) - \lambda_l(\Xi))} G_{(ik),(il)} v_k(\Xi) v_l(\Xi)^\top. \tag{10}$$

If Ξ has repeated eigenvalues, we formally set $\Delta_i(\Xi, G) = \infty \cdot I_p$, so that for all nonzero $c \in \mathbb{R}^p$, $(c^\top \Delta_i(\Xi, G)c)^{-1/2} = 0$. This convention simply encodes that a nondegenerate one-dimensional normalization is unavailable without eigenvalue separation. We further define $\Delta_{n,i} := \Delta_i(X_n^\top X_n, G_n)$ and the plug-in estimator $\hat{\Delta}_{m,n,i} := \Delta_i(X_n^\top S_{m,n}^\top S_{m,n} X_n, G_n)$.

Our main abstract result for PCA is the following; see Section G.3 for the proof.

Theorem 2.9 (Statistical Inference in Sketched PCA)**.** *Let $(X_n)_{n\geqslant 1}$ satisfy Condition 2.3* `B`*, $(U_n)_{n\geqslant 1}$ satisfy Condition $\mathfrak{P}'$, and $(S_{m,n})_{m,n\geqslant 1}$ satisfy Condition 2.8 for some $(G_n)_{n\geqslant 1}$. For any $i \in [p]$, we have the following conclusions.*

Eigenvalues. *For the unobserved data eigenvalue $\Lambda_{n,i} = \lambda_i(X_n^\top X_n)$ and the observed sketched eigenvalue $\hat{\Lambda}_{m,i} = \lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n)$, we have*

$$\sqrt{\frac{m}{\tau_{m,n}(G_n)_{(ii),(ii)}}}\,\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1). \tag{11}$$

Eigenvectors. *For Δ_i from (10) and $\Delta_{n,i} := \Delta_i(X_n^\top X_n, G_n)$, and for any vector $c \in \mathbb{S}^{p-1}$ with $\liminf_{n\to\infty} c^\top \Delta_{n,i} c > 0$, we have the following inferential result for the unobserved linear combination $c^\top v_{n,i}$ of the eigenvector $v_{n,i}$ of $X_n^\top X_n$ associated with the eigenvalue $\Lambda_{n,i}$, based on the observed eigenvector $\hat{v}_{m,n,i}$ of $X_n^\top S_{m,n}^\top S_{m,n} X_n$ associated with the eigenvalue $\hat{\Lambda}_{m,i}$ and the plug-in estimator $\hat{\Delta}_{m,n,i} := \Delta_i(X_n^\top S_{m,n}^\top S_{m,n} X_n, G_n)$:*

$$\sqrt{m}\tau_{m,n}^{-1/2}(c^\top \hat{\Delta}_{m,n,i} c)^{-1/2} c^\top\left(\hat{v}_{m,n,i} - v_{n,i}\right) \Rightarrow \mathcal{N}(0,1). \tag{12}$$

Special case `Sym-1'`***.*** *If the special case* `Sym-1'` *in Condition 2.8 holds, then*

$$\sqrt{\frac{m}{(2+\psi)\tau_{m,n}}}\,\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1), \tag{13}$$

and for any vector $c \in \mathbb{S}^{p-1}$ satisfying $\limsup_{n\to\infty}(c^\top v_{n,i})^2 < 1$,

$$\sqrt{\frac{m}{\tau_{m,n}}}\left(\sum_{k\neq i} \frac{\hat{\Lambda}_{m,i}\hat{\Lambda}_{m,n,k}}{(\hat{\Lambda}_{m,i} - \hat{\Lambda}_{m,n,k})^2}(c^\top \hat{v}_{m,n,k})^2\right)^{-1/2} c^\top\left(\hat{v}_{m,n,i} - v_{n,i}\right) \Rightarrow \mathcal{N}(0,1). \tag{14}$$

If $C \in \mathbb{R}^{p\times k}$ with $\liminf_n \lambda_{\min}(C^\top \Delta_{n,i} C) > 0$, the same argument yields inference for the k-dimensional functional $C^\top v_{n,i}$, namely $\sqrt{m}\tau_{m,n}^{-1/2}(C^\top \hat{\Delta}_{m,n,i} C)^{-1/2} C^\top\left(\hat{v}_{m,n,i} - v_{n,i}\right) \Rightarrow \mathcal{N}(0, I_k)$.

Again, the theorem is best read one statement at a time. Equation (11) gives asymptotic normality for the individual sketched eigenvalue. The normalization is multiplicative in the eigenvalue scale, through $\hat{\Lambda}_{m,i}^{-1}$, and the variance factor is read directly from the covariance matrix G_n of the underlying matrix quadratic form. Equation (12) then turns to eigenvectors. Here the relevant target is not the full vector $v_{n,i}$ by itself, but an arbitrary linear functional

$c^\top v_{n,i}$. The matrix $\Delta_{n,i}$ is the asymptotic covariance from first-order eigenvector perturbation, and its plug-in version $\hat{\Delta}_{m,n,i}$ makes the normalization observable.

The special case `Sym-1'` in Condition 2.8 is particularly transparent. In (13), the variance factor simplifies to $2+\psi$. In (14), the normalization depends only on the observed sketched eigenvalues and eigenvectors, so the limiting distribution is again directly usable for inference.

A key feature of the theorem is that it does *not* assume any convergence of the data matrices. By using a subsequence argument, we obtain the stated limits without requiring X_n to approach a deterministic population matrix. This substantially broadens the scope of the result compared with, for example, Lopes et al. (2020). Table 2 summarizes several concrete instances obtained in Section 3 by applying Theorem 2.9.

Table 2: Summary of some of our results for sketching in PCA. For various distributions of sketching matrices, and for any $i \in [p]$, the table shows the asymptotic variances $\sigma^2(T_\lambda)$ of $T_\lambda := \sqrt{m}\hat{\Lambda}_{m,i}^{-1}(\hat{\Lambda}_{m,i} - \Lambda_{n,i})$ and $\sigma^2(T_{c^\top v})$ of $T_{c^\top v} := \sqrt{m}\left(\sum_{k\neq i}\hat{\Lambda}_{m,i}\hat{\Lambda}_{m,n,k}/(\hat{\Lambda}_{m,i}-\hat{\Lambda}_{m,n,k})^2(c^\top\hat{v}_{m,n,k})^2\right)^{-1/2} c^\top(\hat{v}_{m,n,i}-v_{n,i})$. We also show Condition $\mathfrak{P}$ used in the quadratic-form framework above for which we prove Theorem 2.9. We recall that $\gamma_n = m/n$. See Section 3 for the details.

Sketching Distribution	$\sigma^2(T_\lambda)$	$\sigma^2(T_{c^\top v})$	Condition $\mathfrak{P}$
Hadamard	$3(1-\gamma_n)$	$1-\gamma_n$	ℓ_∞-delocalization
CountSketch/SSE	2	1	ℓ_4-delocalization
i.i.d., $\kappa_{n,4}-3=o(1)$	2	1	
Haar	$2(1-\gamma_n)$	$1-\gamma_n$	
Uniform subsampling	$(n-m)\sum_{k=1}^n (U_n)_{ki}^4$	$1-\gamma_n$	See Section 3.3

Table 2 also highlights an important statistical point. It shows that sketching can have a smaller asymptotic variance than simple subsampling, providing a theoretical explanation for an empirical phenomenon that, to our knowledge, has not previously been justified in general. For example, if there exists an index k with $(U_n)_{k1} \geqslant (3/n)^{1/4}$, then the asymptotic variance of the largest eigenvalue under uniform subsampling can exceed the corresponding variance under the other sketching methods in the table.

When the leverage scores become small, the delocalization conditions needed for Hadamard sketching can still hold, and our theory then yields valid inference. The quantity $\sum_{k=1}^n (U_n)_{ki}^4$, known as the inverse participation ratio (Livan et al., 2018), measures the sparsity of vectors. Under suitable delocalization assumptions, the results of Section 3.3 imply that this quantity can itself be estimated by a plug-in procedure.

For completeness, we also provide an inference result for the setting in which the data are i.i.d. from an underlying statistical model; see Section C for details.

3 Statistical Inference for Specific Sketching Distributions

We now derive inference results in sketched LS and PCA for a variety of sketching distributions. We start with the Subsampled Randomized Hadamard Transform or SRHT (Ailon and Chazelle, 2006), which is computationally efficient and applies under broad conditions on the data; but is highly nontrivial to analyze (Section 3.1). We also provide results for the CountSketch (Charikar

et al., 2002) and Sparse Sign Embeddings (Achlioptas, 2001), which have similar advantages (Section 3.2). We then study uniform sampling (Ma et al., 2015) (Section 3.3). While this method is fast, it requires stringent assumptions on the data. We further analyze sketching matrices with i.i.d. entries (Section D.1) and uniform orthogonal (Haar) sketching (Section D.2). In these sections, we present the results for PCA first, as they are simpler to state. Section E summarizes the computational cost of all the sketching methods studied. The proofs of all results are provided in Section G.

3.1 Subsampled Randomized Hadamard Transform

Hadamard matrices H_t are defined for non-negative integers t inductively, starting with $H_0 = 1$ and letting for all $t \geqslant 0$, $H_{t+1} = \frac{1}{\sqrt{2}} \begin{pmatrix} H_t & H_t \\ H_t & -H_t \end{pmatrix}$. When $n = 2^l$ is a power of two, for some non-negative integer l, an $m \times n$, $m \leqslant n$, subsampled randomized Hadamard transform (SRHT) sketching matrix (Ailon and Chazelle, 2006) is defined as $S_{m,n} = \sqrt{n/m} B H_l D$, where B is an $n \times n$ diagonal matrix with i.i.d. Bernoulli random variables with a success probability of m/n. Furthermore, D is an $n \times n$ diagonal matrix with i.i.d. ± 1 Rademacher random variables on the diagonal.[3]

For other n, one can use the SRHT for the smallest power of two greater than n, padding the data with zeroes, see Section A. Thus, we can assume without loss of generality that n is a power of two because this does not affect the validity of the conditions below. Matrix-vector multiplication via SRHT can be efficiently performed in $O(n \log n)$ floating-point operations, leveraging the Fast Fourier Transform (Ailon and Chazelle, 2006).

Our results rely on the following conditions. Let $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ and U_n be the $n \times p$ left singular matrix of X_n. Denoting for $i \in [p]$ the i-th row of U_n by $U_{i:}$.

$$\textbf{Condition } \mathfrak{P} : \|a_n\|_\infty \to 0, \text{ or } \|\tilde{a}_n\|_\infty \to 0; \tag{15}$$

$$\textbf{Condition } \mathfrak{P}' : \lim_{n\to\infty} \max_{i=1,\dots,n} \max_{i=1,\dots,n} \frac{\|U_{i:}\|}{\|U_n\|_F} = 0. \tag{16}$$

Condition $\mathfrak{P}'$ requires that all *leverage scores* tend to zero; which can be interpreted as there being no X_n-outliers. Condition $\mathfrak{P}$ is used implicitly in our inference results, and explicitly in the supporting results (Lemmas 3.3 and 3.4). Moreover, Condition $\mathfrak{P}'$ is used explicitly in our inference results.

PCA. We have the following inferential result for the unobserved data eigenvalues as well as eigenvectors, using the observed sketched eigenvalues $\hat{\Lambda}_{m,i} = \lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n)$ and eigenvectors $\hat{v}_{m,n,i} = v_i(X_n^\top S_{m,n}^\top S_{m,n} X_n)$, $i \in [p]$. Its proof and the proofs of all other theoretical results in this section are given in Section G.5.

Theorem 3.1 (Stat. Inference for Eigenvalues/vectors with SRHT)**.** *For all $(X_n)_{n\geqslant 1}$ with SVD $U_n L_n V_n^\top$ satisfying Condition 2.3* B*, suppose $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$ from (16). Let*

[3] We remark that this matrix has size $n \times n$, unlike the $m \times n$ matrices considered in the introduction. However only a number of its rows are non-zero and the expected value of the number of those rows is equal to m. We could equivalently replace the matrix with its sub-matrix consisting of the non-zero rows; however we prefer to keep it in the current form due to notational simplicity.

$\gamma_n = m/n$. *Then as* $m, n \to \infty$ *with* p *fixed, such that* $m/\log^2 n \to \infty$ *and* $\limsup_n \gamma_n < 1$, *for any* $i \in [p]$, *we have for the unobserved data eigenvalues* $\Lambda_{n,i} = \lambda_i(X_n^\top X_n)$ *and the observed sketched eigenvalues that* (2) *holds. Further, we have for the unobserved linear combinations* $c^\top v_{n,i}$ *of the eigenvector* $v_{n,i}$ *of* $X_n^\top X_n$ *associated with the eigenvalue* $\Lambda_{n,i}$, *and the observed eigenvector* $\hat{v}_{m,n,i}$ *of* $X_n^\top S_{m,n}^\top S_{m,n} X_n$ *associated with the eigenvalue* $\hat{\Lambda}_{m,i}$ *that for any vector* $c \in \mathbb{S}^{p-1}$ *with* $\limsup_{n\to\infty} (c^\top v_{n,i})^2 < 1$:

$$\sqrt{\frac{m}{1-\gamma_n}} \left(\sum_{k \neq i} \frac{\hat{\Lambda}_{m,i}\hat{\Lambda}_{m,n,k}}{(\hat{\Lambda}_{m,i} - \hat{\Lambda}_{m,n,k})^2} (c^\top \hat{v}_{m,n,k})^2 \right)^{-1/2} c^\top (\hat{v}_{m,n,i} - v_{n,i}) \Rightarrow \mathcal{N}(0,1). \tag{17}$$

LS. The following result shows how to make inferences on least squares parameters based on sketched LS estimators when using SRHT.

Theorem 3.2 (Statistical Inference in OLS with SRHT)**.** *Let* $(S_{m,n})_{m,n\geqslant 1}$ *be a sequence of subsampled randomized Hadamard sketching matrices, and let* X_n *satisfy Condition 2.3* `A`. *Denoting for* $i \in [p]$ *the* i*-th row of* U_n *by* $U_{i:}$, *if either*

$$(U_n)_{n\geqslant 1} \text{ satisfies Condition } \mathfrak{P}' \text{ from (16)} \quad \textit{or} \quad \lim_{n\to\infty} \max_{i=1,\dots,n} \frac{|\varepsilon_{n,i}|}{\|\varepsilon_n\|} = 0, \tag{18}$$

then, as $m \to \infty$, $m/\log^2 n \to \infty$, *and for* $\gamma_n = m/n$, $\limsup \gamma_n < 1$, *we have for the sketch-and-solve LS estimator* $\hat{\beta}_{m,n}$ *from* (1) *that*

$$(m/(1-\gamma_n))^{1/2} \|\tilde{\varepsilon}_{m,n}\|^{-1} (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{1/2} (\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p).$$

If either

$$(U_n)_{n\geqslant 1} \text{ satisfies Condition } \mathfrak{P}' \text{ from (16)} \quad \textit{or} \quad \lim_{n\to\infty} \max_{i=1,\dots,n} \frac{|y_{n,i} - \varepsilon_{n,i}|}{\|y_n - \varepsilon_n\|} = 0, \tag{19}$$

then, in the same asymptotic regime, we have for the partial sketching estimator $\hat{\beta}_{m,n}^{(\mathrm{p})}$ *from* (1) *that*

$$(m/(1-\gamma_n))^{1/2} \left(\|\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})}\|^2 \cdot (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} + 2\hat{\beta}_{m,n}^{(\mathrm{p})}\hat{\beta}_{m,n}^{(\mathrm{p})\top} \right)^{-1/2} (\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p).$$

The second condition in (18) ensures that the normalized residuals are small, which can be viewed as there being no outliers in the outcome vector y_n.

Our result for the sketch-and-solve estimator $\hat{\beta}_{m,n}$ is related to Theorem 3 (i) of Ahfock et al. (2021). They obtain the asymptotic distribution of the sketched data $S_{m,n}X_n$ for a finite m. In contrast, we focus on the case where the sketch size m tends to infinity, which leads to the additional $1 - \gamma_n$ factor in the variance. In contrast to the work by Ahfock et al. (2021), our approach allows for X_n-outliers, as long as there are no y_n-outliers. Moreover, we only require a bounded condition number for X_n, while their results require certain convergence conditions on X_n and y_n.

CLT. The following lemmas play a fundamental role in the proofs of the above theorems by verifying Conditions 2.2, 2.4, and 2.8 for SRHT sketching. We show that normality holds if

Condition $\mathfrak{P}$ is satisfied, but does not need to hold otherwise.

Lemma 3.3 (Bounded Quadratic Forms for SRHT Sketching)**.** *For any sequences of vectors $(a_n, \tilde{a}_n)_{n\geqslant 1}$, such that $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ for all $n \geqslant 1$, we have*

$$\sqrt{\frac{m}{1-\gamma_n}}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) = O_P(1).$$

The next lemmas establish the normality of quadratic forms arising in SRHT. In the proof, finding the asymptotic variances requires intricate and novel calculations, using a dyadic representation for the Hadamard matrix elements. See Section G.5.4.

Lemma 3.4 (Distributions of 1D Quadratic Forms for SRHT Sketching)**.** *For any sequences of vectors $(a_n)_{n\geqslant 1}$, $(\tilde{a}_n)_{n\geqslant 1}$ satisfying Condition $\mathfrak{P}$ defined in (15) such that $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ for all $n \geqslant 1$, if $m/\log^2 n \to \infty$ and $\limsup m/n < 1$, as $m, n \to \infty$, we have*

$$\sqrt{\frac{m}{1-\gamma_n}}\sqrt{\frac{1}{1+2(a_n^\top \tilde{a}_n)^2}}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) \Rightarrow \mathcal{N}(0,1). \tag{20}$$

Moreover, if Condition $\mathfrak{P}$ fails, $\sqrt{m}a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n$ does not need to be asymptotically normal even if m/n converges.

Lemma 3.5 (Normality of Symmetric QF for SRHT)**.** *If $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, and $m/\log^2 n \to \infty$ with $\limsup m/n < 1$ as $m, n \to \infty$, then for the $p^2 \times p^2$ positive definite matrix $G = I_{p^2} + P_p + Q_p$,*

$$\sqrt{\frac{m}{1-\gamma_n}}\mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n}U_n - I_p) \Rightarrow \mathcal{N}(0, G).$$

3.2 Sparse Sign Embeddings and CountSketch

Sparse Sign Embeddings are a type of sparse dimension reduction maps (e.g., Achlioptas, 2001; Meng and Mahoney, 2013, etc). For a given sparsity $\zeta_m \in \{1, 2, \ldots\}$, a sparse sign embedding is a random matrix of the form $S_{m,n} = (R_{m,1}, \ldots, R_{m,n})$, where for all $i \in [n]$, $R_{m,i}$ are i.i.d. uniformly distributed over the set of m-vectors with ζ_m nonzero elements equal to $\pm 1/\sqrt{\zeta_m}$, i.e., $\{f/\sqrt{\zeta_m} \in \mathbb{R}^m : f_1, \ldots, f_m \in \{\pm 1, 0\}$ and $\|f\|^2 = \zeta_m\}$. Due to their tunable sparsity, sparse sign embeddings allow the user to control the computational cost of sketching, see Section E. Recently, they have been suggested as a "sensible default" sketching method (Epperly, 2023). For $\zeta_m = 1$, sparse sign embeddings are also known as CountSketch (Charikar et al., 2002; Clarkson and Woodruff, 2017).

Our results rely on the following conditions, where $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ and U_n is the $n \times p$ left singular matrix of X_n for all $n \geqslant 1$:

$$\textbf{Condition } \mathfrak{P} : (\sqrt{m}/\zeta_m) \cdot \sum_{i=1}^n a_{n,i}^4 \to 0, \text{ and } (\sqrt{m}/\zeta_m) \cdot \sum_{i=1}^n \tilde{a}_{n,i}^4 \to 0; \tag{21}$$

$$\textbf{Condition } \mathfrak{P}' : (\sqrt{m}/\zeta_m) \cdot \|\mathrm{vec}(U_n)\|_4^4 \to 0. \tag{22}$$

As before, these will be used explicitly or implicitly in our results. Further, in this section, we use $\sum_{i\neq j}$ to abbreviate the summation over $(i,j) \in [n] \times [n]$ subject to $i \neq j$.

PCA. We have the following result on inference for eigenvalues and eigenvectors with sparse sign embeddings. Its proof, along with the proofs of all other theoretical results in this section, is given in Section G.6. This result allows the per-column sparsity ζ_m to be $o(\sqrt{m})$; which in particular includes the CountSketch.

Theorem 3.6 (Statistical Inference for Eigenvalues/vectors with SSE)**.** *For $(X_n)_{n\geqslant 1}$ satisfying Condition 2.3* B*, suppose $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, let $m, n \to \infty$, and suppose the sparsity satisfies $\zeta_m^2/m \to 0$. Then for the unobserved data eigenvalues $\Lambda_{n,i} = \lambda_i(X_n^\top X_n)$ and the observed sketched eigenvalues $\hat{\Lambda}_{m,i} = \lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n)$, we have that $\sqrt{\frac{m}{2}}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1)$. Further, we have for the unobserved linear combinations $c^\top v_{n,i}$ of the eigenvector $v_{n,i}$ of $X_n^\top X_n$ associated with the eigenvalue $\Lambda_{n,i}$, and the observed eigenvector $\hat{v}_{m,n,i}$ of $X_n^\top S_{m,n}^\top S_{m,n} X_n$ associated with the eigenvalue $\hat{\Lambda}_{m,i}$ that for any vector $c \in \mathbb{S}^{p-1}$ with $\limsup_{n\to\infty}(c^\top v_{n,i})^2 < 1$,*

$$\sqrt{m}\left(\sum_{k\neq i}\frac{\hat{\Lambda}_{m,i}\hat{\Lambda}_{m,n,k}}{(\hat{\Lambda}_{m,i}-\hat{\Lambda}_{m,n,k})^2}(c^\top \hat{v}_{m,n,k})^2\right)^{-1/2} c^\top(\hat{v}_{m,n,i} - v_{n,i}) \Rightarrow \mathcal{N}(0,1).$$

LS. We also have the following result for inference about the least squares parameters for sketching with sparse sign embeddings.

Theorem 3.7 (Statistical Inference in OLS with Sparse Sign Embeddings)**.** *Suppose X_n satisfies Condition 2.3* A*, $m, n \to \infty$, and $\zeta_m^2/m \to 0$. If*

$$(U_n)_{n\geqslant 1} \text{ satisfies Condition } \mathfrak{P}' \text{ from (22)} \quad \textit{and} \quad (\sqrt{m}/\zeta_m)\cdot\frac{\|\varepsilon_n\|_4^4}{\|\varepsilon_n\|_2^4} \to 0, \tag{23}$$

then for the sketch-and-solve LS estimator $\hat{\beta}_{m,n}$ from (1), we have with $\hat{\Sigma}_{m,n} = \|\tilde{\varepsilon}_{m,n}\|^2 \cdot (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1}$ that $\sqrt{m}\hat{\Sigma}_{m,n}^{-1/2}(\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$. Similarly, if

$$(U_n)_{n\geqslant 1} \text{ satisfies Condition } \mathfrak{P}' \text{ from (22)} \quad \textit{and} \quad (\sqrt{m}/\zeta_m)\cdot\frac{\|y_n - \varepsilon_n\|_4^4}{\|y_n - \varepsilon_n\|_2^4} \to 0, \tag{24}$$

then for the partial sketching LS estimator $\hat{\beta}_{m,n}^{(\mathrm{p})}$ from (1), with $\hat{\Sigma}_{m,n}^{(\mathrm{p})} = \hat{\beta}_{m,n}^{(\mathrm{p})\top}\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})} \cdot (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} + \hat{\beta}_{m,n}^{(\mathrm{p})}\hat{\beta}_{m,n}^{(\mathrm{p})\top}$, we have that $\sqrt{m}(\hat{\Sigma}_{m,n}^{(\mathrm{p})})^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$.

CLT. The inference results mentioned above are based on the following central limit theorems for quadratic forms underlying Conditions 2.2, 2.4, and 2.8 for Sparse Sign Embeddings. First, we have a result on one-dimensional quadratic forms.

Lemma 3.8 (Limiting Distributions of 1D Quadratic Forms for SSE)**.** *For a fixed n, and any sequences $(a_n)_{n\geqslant 1}$, $(\tilde{a}_n)_{n\geqslant 1}$ such that $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ for all $n \geqslant 1$, if Condition $\mathfrak{P}$ from (21) holds as $m \to \infty$ and $\zeta_m^2/m \to 0$, then we have*

$$\sqrt{\frac{m}{1+(a_n^\top \tilde{a}_n)^2}}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) \Rightarrow \mathcal{N}(0,1). \tag{25}$$

This lemma represents a novel result about CountSketch and Sparse Sign Embeddings. Our proof treats the quadratic forms as degenerate U-statistics and uses the martingale central limit theorem. We next have a result on vectorized symmetric quadratic forms.

Lemma 3.9 (Asymptotic Normality for Sparse Signed Embeddings)**.** *If as $m, n \to \infty$, $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$ from (22) while $\zeta_m^2/m \to 0$, then*

$$\sqrt{m}\mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) \Rightarrow \mathcal{N}\left(0, I_{p^2} + P_p\right). \tag{26}$$

3.3 Uniform Subsampling

We assume[4] $S_{m,n} = \sqrt{\frac{n}{m}}\bar{B}_n$, where $\bar{B}_n = \mathrm{diag}(B_1, \ldots, B_n)$ with i.i.d. $B_i \sim \mathrm{Bernoulli}(\gamma_n)$, $\gamma_n = m/n$ for all $i \in [n]$. Thus, $\bar{B}_n$ represents a uniform random subsampling procedure. Our results rely on the following conditions, where $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ for all $n \geqslant 1$, $u_{n,i}$, $i \in [n]$ are the rows of the $n \times p$ left singular matrix U_n of X_n, and $G_n := n\sum_{i=1}^n (u_{n,i}u_{n,i}^\top) \otimes (u_{n,i}u_{n,i}^\top)$:

$$\textbf{Cond. } \mathfrak{p} : \text{for } 0 < c < C,\ c \leqslant n\sum_{i=1}^n a_{n,i}^2 \tilde{a}_{n,i}^2 \leqslant C;\ \textbf{Cond.}\mathfrak{P} : \max_{i\in[n]} \frac{(a_{n,i}\tilde{a}_{n,i})^2}{\sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2} \to 0; \tag{27}$$

$$\textbf{Cond. } \mathfrak{P}' : \text{for some } 0 < c < C,\ cI_{p(p+1)/2} \preceq D_p^\top G_n D_p \preceq CI_{p(p+1)/2}; \tag{28}$$

$$\text{for all upper triangular } \Phi \in \mathbb{R}^{p\times p},\ \frac{\max_{i\in[n]}(u_{n,i}^\top \Phi u_{n,i})^2}{\sum_{i=1}^n (u_{n,i}^\top \Phi u_{n,i})^2} \to 0.$$

Again, these results appear explicitly in our asymptotic normality results, and sometimes only implicitly in our inference results.

PCA. If the data X_n has very heterogeneous rows, then subsampling can introduce a great deal of variability in the resulting estimates. Therefore, obtaining inference and asymptotic normality requires stronger delocalization conditions on U_n, and leads to an asymptotic variance of a delicate form. We have the following inferential result; its proof and all other results related to uniform subsampling are given in Section G.7.

Theorem 3.10 (Inference for Eigenvalues/vectors with Subsampling)**.** *Suppose $(X_n)_{n\geqslant 1}$ satisfies Condition 2.3* B*, $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, and $(S_{m,n})_{m,n\geqslant 1}$ satisfies Condition 2.8. Then we have for any $i \in [p]$ that*

$$\sqrt{\frac{m}{(1-\gamma_n)\cdot n\sum_{k=1}^n (U_n)_{ki}^4}}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1). \tag{29}$$

Let $\widetilde{X}_{m,n} = \widetilde{U}_{m,n}\hat{L}_{m,n}\hat{V}_{m,n}^\top$ be the SVD of $\widetilde{X}_{m,n}$. If $\liminf \gamma_n > 0$, we have

$$\frac{1}{\sqrt{(1-\gamma_n)\cdot m\sum_{j=1}^m \tilde{U}_{ji}^4}}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1). \tag{30}$$

For Δ_i defined in (10), and any vector $c \in \mathbb{S}^{p-1}$ satisfying $\liminf_{n\to\infty} c^\top \Delta_{n,i} c > 0$,

$$\sqrt{m/(1-\gamma_n)}(c^\top \hat{\Delta}_{m,n,i} c)^{-1/2} c^\top (\hat{v}_{m,n,i} - v_{n,i}) \Rightarrow \mathcal{N}(0,1).$$

[4]Here as for the SRHT we prefer to keep the zero rows of the sketching matrix for notational simplicity.

LS. Additionally, we have the following theorem for inference about the least squares parameters based on uniform subsampling.

Theorem 3.11 (Statistical Inference in OLS with Random Sampling)**.** *Let p be fixed and $m, n \to \infty$ with $0 < \liminf m/n \leqslant \limsup m/n < 1$. Suppose that X_n satisfies Condition 2.3* A*, and $\|\mathrm{vec}(U_n)\|_\infty = O(1/\sqrt{n})$. Also, suppose there exist some constants $0 < c < C$ such that*

$$cI_p \preceq nU_n^\top \operatorname{diag}\left(\frac{|\varepsilon_n|}{\|\varepsilon_n\|} \odot \frac{|\varepsilon_n|}{\|\varepsilon_n\|}\right) U_n \preceq CI_p. \tag{31}$$

For

$$\hat{\Sigma}_{m,n} = m(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} \widetilde{X}_{m,n}^\top \operatorname{diag}(\tilde{\varepsilon}_{m,n} \odot \tilde{\varepsilon}_{m,n}) \widetilde{X}_{m,n} (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1},$$

we have $m^{1/2}(1-\gamma_n)^{-1/2}\hat{\Sigma}_{m,n}^{-1/2}(\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$. Furthermore, assume there exist $0 < c < C$ such that

$$cI_p \preceq nU_n^\top \operatorname{diag}\left(\frac{|y_n - \varepsilon_n|}{\|y_n - \varepsilon_n\|} \odot \frac{|y_n - \varepsilon_n|}{\|y_n - \varepsilon_n\|}\right) U_n \preceq CI_p. \tag{32}$$

Then with

$$\hat{\Sigma}_{m,n}^{(\mathrm{p})} = m(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} \widetilde{X}_{m,n}^\top \operatorname{diag}\left((\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})}) \odot (\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})})\right) \widetilde{X}_{m,n} (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1},$$

we have $m^{1/2}(1-\gamma_n)^{-1/2}(\hat{\Sigma}_{m,n}^{(\mathrm{p})})^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$.

In fact, (31) holds if there are constants $0 < c < C$ such that $c/\sqrt{n} \leqslant \min_{i\in[n]} |\varepsilon_{n,i}|/\|\varepsilon_n\|$ and $\max_{i\in[n]} |\varepsilon_{n,i}|/\|\varepsilon_n\| \leqslant C/\sqrt{n}$, and similarly (32) holds if $c/\sqrt{n} \leqslant \min_{i\in[n]} |y_{n,i} - \varepsilon_{n,i}|/\|y_n - \varepsilon_n\|$ and $\max_{i\in[n]} |y_{n,i} - \varepsilon_{n,i}|/\|y_n - \varepsilon_n\| \leqslant C/\sqrt{n}$.

CLT. The following lemmas play a significant role in the proofs of the above theorems by establishing the boundedness and normality statements in Conditions 2.2, 2.4, and 2.8 for uniform subsampling.

Lemma 3.12 (Asymptotic Normality of QF for Random Sampling)**.** *For any sequences of vectors $(a_n, \tilde{a}_n)_{n\geqslant 1}$ such that $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ for all $n \geqslant 1$ satisfying Condition* $\mathfrak{p}$ *from (27), we have*

$$\sqrt{\frac{m}{1-\gamma_n}}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) = O_P(1). \tag{33}$$

Further, if Condition $\mathfrak{P}$ from (27) holds, then

$$\sqrt{\frac{m}{1-\gamma_n}}\sqrt{\frac{1}{n\sum_{i=1}^n a_{n,i}^2 \tilde{a}_{n,i}^2}}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) \Rightarrow \mathcal{N}(0,1). \tag{34}$$

Moreover, there exist sequences $(a_n)_{n\geqslant 1}$, $(\tilde{a}_n)_{n\geqslant 1}$ of unit norm vectors with $a_n, \tilde{a}_n \in \mathbb{R}^n$ that do not satisfy Condition $\mathfrak{P}$, such that $\sqrt{m}a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n$ is not asymptotically normal.

Lemma 3.13 (Asymptotic Normality of Symmetric QF for Uniform Sampling)**.** *If U_n satisfies Condition $\mathfrak{P}'$ from (28), then we have*

$$\sqrt{\frac{m}{1-\gamma_n}}(D_p^\top G_n D_p)^{-1/2} D_p^\top \mathrm{vec}\left(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p\right) \Rightarrow \mathcal{N}(0, I_{p(p+1)/2}).$$

3.4 Multisketching

Several sketching matrices can be combined; a typical approach is multisketching (Kapralov et al., 2016; Higgins et al., 2025). Its general form is

$$S_{m,n}X_n = S_{2,m,\tilde{m}}(S_{1,\tilde{m},n}X_n)$$

where the first sketching via $S_{1,\tilde{m},n}$ quickly reduces the size of the data X_n from n to $\tilde{m}$, and the second sketching $S_{2,m,\tilde{m}}$ further reduces the size using a denser sketch from $\tilde{m}$ to m. The following result characterizes the asymptotic normality of the CountGauss sketch Kapralov et al. (2016), where $S_{1,\tilde{m},n}$ is a CountSketch and $S_{2,m,\tilde{m}}$ is a Gaussian sketching matrix. See Section G.9 for its proof.

Lemma 3.14 (Asymptotic Normality of QF for CountGauss sketch)**.** *Assume $m/\tilde{m} \to 0$ and $\tilde{m}/n \to 0$, then for any sequences of vectors $(a_n, \tilde{a}_n)_{n\geqslant 1}$ such that $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ for all $n \geqslant 1$ satisfying Condition $\mathfrak{P}$ from (21) with m replaced by $\tilde{m}$, we have*

$$\sqrt{m}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) = O_P(1),$$

and

$$\sqrt{\frac{m}{1+(a_n^\top \tilde{a}_n)^2}}\left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) \Rightarrow \mathcal{N}(0,1).$$

A direct consequence of this lemma is the statistical inference result for least squares problem with CountGauss sketch. The same conclusions as in Theorem 3.7 hold, with the conditions involving m replaced by $\tilde{m}$. We remark that Condition $\mathfrak{P}$ from (21) used in Lemma 3.14 can be weakened due to the additional Gaussian sketching. Here we state such a preliminary result to illustrate the application of our unified inference framework using quadratic forms.

4 Statistical Inference for Iterative Newton Sketching

Iterative sketching algorithms offer alternatives to the sketch-and-solve algorithms and often achieve more precise solutions. For instance, iterative Newton sketching (Pilanci and Wainwright, 2017) has many uses, including in generalized linear models (GLMs). Our theoretical framework can also be applied to this algorithm.

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a smooth strongly convex function with a global minimizer $\theta_* := \operatorname{argmin} f(\theta)$. The standard Newton's algorithm starts at some θ_0, and after $t \geqslant 0$, it performs the iteration $\theta_{t+1} = \theta_t - (\nabla^2 f(\theta_t))^{-1}\nabla f(\theta_t)$. However, evaluating the Hessian matrix $\nabla^2 f$ can be a bottleneck. For instance, in GLMs, if we let $X \in \mathbb{R}^{n\times p}$, $n \geqslant p$, be the data matrix, where x_i is its i-th row, then the cost per iteration is $O(np^2)$. The cost mainly comes from computing $\nabla^2 f(\theta_t) = X^\top W_t X$ where W_t is a diagonal matrix of weights $w_i(t) = \psi(\theta_t^\top x_i)$ for some function ψ that depends on the model. For instance, for logistic regression, $\psi : u \mapsto \exp(u)/[1+\exp(u)]^2$.

Iterative Newton sketching uses a sketched approximation of the Hessian in each iteration. Specifically, let $H(\theta)$ be the Hessian evaluated at θ and denote $H_t = H(\theta_t)$. Assume that $H(\theta_*)$ can be decomposed as the multiplication of $H_{l,*} \in \mathbb{R}^{p\times n}$ and $H_{r,*} \in \mathbb{R}^{n\times p}$, where $H_{l,*}$, $H_{r,*}$ can be interpreted as matrix square roots such that $H_{r,*} = H_{l,*}^\top$. Similarly, let H_t be decomposed

as $H_t = H_{t,l}H_{t,r}$. In many cases, computing $H_{t,l}$ is efficient. For instance, in GLMs, we can define $H_{t,l} = X^\top W_t^{1/2}$ and $H_{t,r} = W_t^{1/2}X$. Iterative Newton sketching starts from some θ_0 and performs the update $\theta_{t+1} = \theta_t - \tilde{H}_t^{-1}\nabla f(\theta_t)$, where $\tilde{H}_t := (H_{t,l})S_{t,m,n}^\top S_{t,m,n}(H_{t,r})$, and where $\{S_{t,m,n}\}_{t\geqslant 0}$ are independent sketching matrices for different t.

We assume the following condition for the distribution results of the Newton sketch.

Condition 4.1 (Regularity conditions for f)**.** *Assume that each Hessian H_l is Lipschitz continuous, meaning that there exists a universal constant L satisfying $\|H_l(\theta+\Delta) - H_l(\theta)\| \leqslant L\|\Delta\|$ for all θ, Δ. Moreover, let $\max_{\alpha_i \geqslant 0: \sum_{i=1}^p \alpha_i = 3} \left| \frac{\partial^3 f}{\partial\theta_1^{\alpha_1}\cdots\partial\theta_p^{\alpha_p}} \right|$ be uniformly bounded over θ.*

For two sequences of random vectors or variables X_n and Y_n, if they converge to the same limiting distribution, we denote this as $X_n \overset{d}{\approx} Y_n$. The following theorem, whose proof is given in Section G.9, establishes the limiting distribution of the outputs of iterative Newton sketching, characterized through the asymptotic normality of quadratic forms as per Condition 2.8. Denote the SVD of $H_{*,r} \in \mathbb{R}^{n,p}$ by $H_{*,r} = U_* L_* V_*^\top$.

Theorem 4.2. *Consider the asymptotic regime that $m = m_n \to \infty$ with $m < n$. Let $\Delta_t = \theta_t - \theta_*$, where θ_* is the minimizer of f. Let θ_{t_0} be a starting point that is independent of $S_{t,m,n}$, and $\|\Delta_{t_0}\| = o_P(m^{-1/2})$. Denote $\Omega_t = U_*^\top(S_{t,m,n}^\top S_{t,m,n} - I_n)U_*$. Conditional on Δ_{t_0}, we have for any fixed $t > t_0$,*

$$\frac{m^{(t-t_0)/2}\Delta_t}{\|\Delta_{t_0}\|} \overset{d}{\approx} m^{(t-t_0)/2} V_* L_*^{-1} \left\{ \prod_{i=t-1}^{t_0} \Omega_i \right\} L_* V_*^\top \frac{\Delta_{t_0}}{\|\Delta_{t_0}\|}. \tag{35}$$

Letting $\|x\|_A := \sqrt{x^\top A x}$, this further implies that for Hadamard sketching and i.i.d. sketching, we have

$$\frac{m^{t-t_0}\|\Delta_t\|_{H_*}^2}{\|\Delta_{t_0}\|_{H_*}^2} \overset{d}{\approx} m^{t-t_0} e_1^\top \left\{ \prod_{i=t_0}^{t-1} \Omega_i \right\} \left\{ \prod_{i=t-1}^{t_0} \Omega_i \right\} e_1.$$

An approximate initial point θ_{t_0} for sketched Newton satisfying $\|\Delta_{t_0}\| = o(m^{-1/2})$ can be obtained via SGD with $t_0 \gg m$ steps. This requires some mild conditions on the data points, see for instance Lemma 2 of Rakhlin et al. (2012).

Under Condition 2.8, the asymptotic distribution of $m^{1/2}\Omega_i$ can be characterized depending on the types of sketching. This result also underscores the role of the loss function's curvature and its interaction with the initial error $\theta_{t_0} - \theta_*$ in the uncertainty of the outputs. Theorem 4.2 implies that sub-randomization Zhang et al. (2023) can be used for statistical inference for θ_*, viewing Δ_{t_0} as fixed. However, Theorem 4.2 does not lead to valid pivotal inference. Even though V_*, and L_* can be estimated consistently, the asymptotic distribution depends on Δ_{t_0}, which cannot be easily estimated. It is also possible to investigate the bootstrap Lopes et al. (2018). We do not pursue the details in this work.

5 Numerical Experiments

We conduct experiments on both simulated and empirical data to evaluate statistical performance and computational efficiency across different sketching methods (Sections 5.1 and F).

The inference targets include the least squares coefficient using full data $\beta_n = (X_n^\top X_n)^{-1} X_n^\top y_n$ as well as the eigenvalues and eigenvectors of $X_n^\top X_n$. For the simulated data, we consider the following three cases:

Case 1: For both LS and PCA, we generate $X_n = U_n L_n V_n^\top$, where U_n is sampled uniformly from the set of $n \times p$ partial orthogonal matrices, $V_n^\top$ is sampled uniformly from the set of $p \times p$ orthogonal matrices, and $L_n = \mathrm{diag}(1, 1/2, \ldots, 1/p)$. Additionally, for LS we generate $y_n \in \mathbb{R}^n$ with i.i.d. Uniform $(0, 1)$ entries.

Case 2: As Lopes et al. (2018) in a slightly different setting, we let A_n be an $n \times p$ matrix with i.i.d. rows distributed according to the multivariate t-distribution $t_2(0, C)$ where $C = (c_{ij}) = (2 \cdot 0.5^{|i-j|})$. Let U_n be the left singular matrix of A_n. We construct L_n with entries equally spaced in the interval $[0.1, 1]$ and generate $V_n^\top$ as the right singular matrix of a matrix with i.i.d. $\mathcal{N}(0, 1)$ entries. Finally, we set $X_n = U_n L_n V_n^\top$ for both LS and PCA. To generate y_n in LS, we set $b = (1_{0.2p}, t1_{0.6p}, 1_{0.2p})^\top$ with $t = 0.1$, ensuring that $0.2p$ is an integer. Moreover, we generate $\mathcal{E}_n \in \mathbb{R}^n$ with i.i.d. normal entries having a standard deviation of 0.01, and then let $y_n = X_n b + \mathcal{E}_n$.

Case 3: For both LS and PCA, we let X_n be an $n \times p$ Gaussian random matrix such that for all $j \in [p]$, $(X_n)_{i,j} \sim \mathcal{N}(0, 1)$ when $i \leqslant n/2$, and $(X_n)_{i,j} \sim \mathcal{N}(5, 1)$ otherwise. After that, we standardize each column to have zero mean and unit variance. For linear regression, we also generate $y_n \in \mathbb{R}^n$ with i.i.d. Uniform$(0, 1)$ entries.

We also compare our sketched least squares methods with the optimal subsampling algorithm from Yu et al. (2022).

5.1 Statistical Inference from Sketch-and-Solve Estimators

We perform experiments to compare several sketching methods discussed in the paper.

Simulated Data. In this section, we compare the performance of various sketching methods for statistical inference in least squares and PCA. We generate simulated data with $n = 2,048$ and $p = 15$ for each case. To evaluate the methods, we generate 500 independent sketches for each value of m ranging from 200 to 1600 with gaps of 200. We use three sketching methods: i.i.d. sketching, Hadamard sketching, and CountSketch. For Case 1, Section 3 presents the coverage of 95% intervals for the first coordinate of β_n, the largest eigenvalue, and the first coordinate of the corresponding eigenvector of $X_n^\top X_n$. For detailed cases, we show the corresponding simulation results in Section F.

Empirical Data Example. As an empirical illustration of sketch-and-solve methods, we use the New York flights (nycflights13) dataset (Wickham, 2018) and the Human Genome Diversity Project (HGDP) dataset, see Section F.2 for details. We also perform sketched PCA using the Million Song Dataset (MSD, Bertin-Mahieux et al. (2011)), as an example of a larger dataset, with $n = 515,344$ and $p = 90$. We show coverage results in Section F.2.

Running Time. To empirically analyze the running time of our methods, we conduct experiments with various types of sketching methods for PCA in Cases 1 and 3. For a relative small sample size, see Section F.4 for details. Here, we analyze a larger data set in Case 1, with $n = 2^{21}, p = 100$, and $k = 1$. This leads to a dataset that is large enough to be somewhat challenging to be analyzed on a PC, but still realistic to occur in applications. We use SRHT as

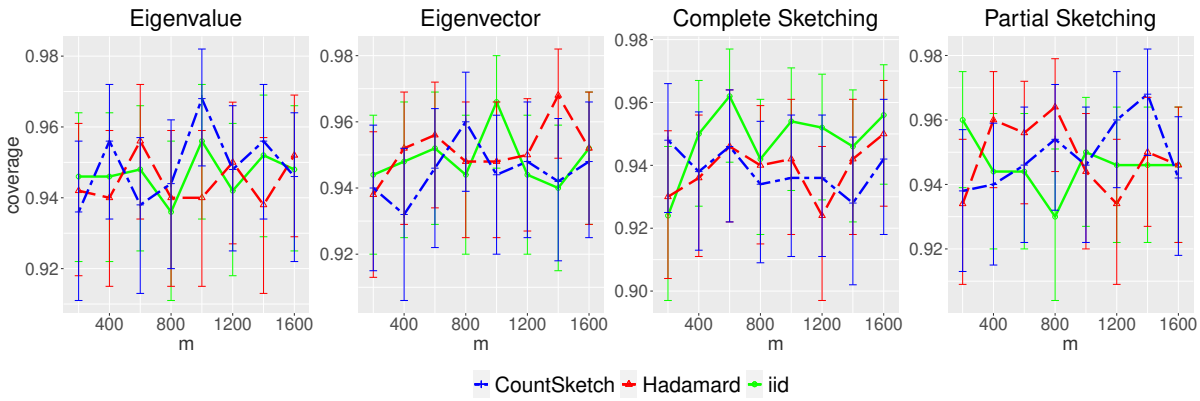


Figure 3: Coverage of nominal 95% intervals, along with 95% Clopper-Pearson intervals for the coverage, for λ_k, $c^\top v_k$, and the first coordinates of two least squares solutions in Case 1, with $p = 15, n = 2048$, and 500 Monte Carlo trials for each setting.

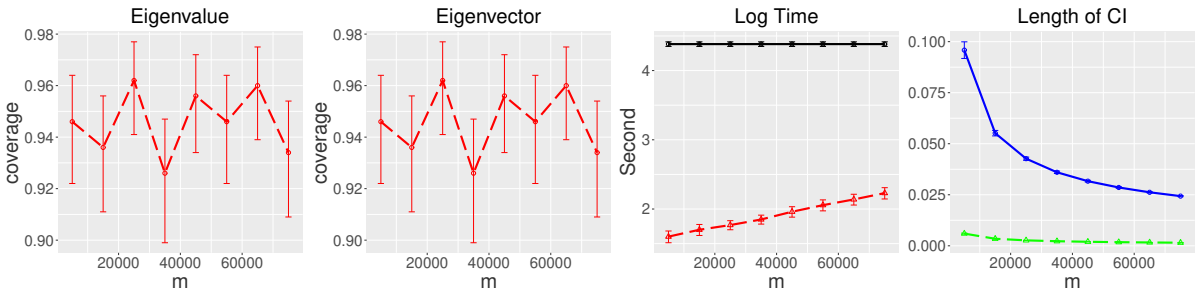


Figure 4: The first two subfigures show the coverage of nominal 95% intervals, along with 95% Clopper-Pearson intervals for the coverage, for λ_1 and $c^\top v_1$ as a function of the sketch size m; see Section 5.1. The results cover Case 1, with $p = 100, n = 2^{21}$, and 500 trials. The third subfigure displays the computational time for the SRHT sketched PCA (red dashed line) and the original PCA solution (black solid line). The fourth subfigure shows the length of the confidence intervals of SRHT sketched eigenvalues (blue solid line) and eigenvectors (green dashed line).

the primary sketching method for PCA. We vary the value of the sketch size m from 5,000 to 75,000 with increments of 10,000. The coverage and computation time results are presented in Section 4. Our solution achieves the target coverage even when the sketch dimension is small. Additionally, as the sketch size increases, the computational time required for SRHT sketched PCA grows slowly, but remains significantly lower than that of PCA on the full data. At the same time, as the sketching dimension increases, the length of the confidence intervals for the SRHT sketched solution decreases, as expected. Here the values of the target parameters of interest are $\lambda_1 = 1$ and $c^\top v_1 = 0.087$; hence the confidence intervals are short and informative. These experiments illustrate the trade-off between statistical accuracy and computational costs.

5.2 Comparison with Optimal Subsampling

We compare our sketched least squares methods with the optimal subsampling algorithm from Yu et al. (2022). Later, we also compare various sketching methods in Section F.3. We compare the asymptotic MSE of the least squares solutions in Figure 5. Details can be found in Section F.5. For Case 2, all subsampling methods are less accurate than Hadamard sketching. This shows the benefits of using projection methods beyond subsampling. In Case 2, where the

data matrices differ significantly from i.i.d. models, subsampling methods tend to be less stable. On the other hand, sketching methods like Hadamard sketching *mix the data* before subsampling, which helps mitigate the impact of "outlier" data points. Figure 5 shows that Hadamard sketching achieves the "empirical Pareto frontier", which means that among the methods compared, empirically Hadamard sketching has the best tradeoff in balancing computational cost and statistical accuracy.

Figure 5: Comparison of subsampling methods for Case 2, using $p = 25, n = 4096$ with 2000 Monte Carlo trials for each setting. Here "optsubsamp_800" is the optimal subsampling method from Algorithm 2 in Yu et al. (2022) with pilot subsample size $r_0 = 800$, while "optsubsamp_1200" and "optsubsamp_1400" refer to the same algorithm using $r_0 = 1200$ and $r_0 = 1400$.

6 Summary and Discussion

In this work, we propose a unified methodology for statistical inference in sketched PCA and LS built on the asymptotic normality of quadratic forms. This framework unlocks inference for a variety of sketching methods, also reveals their distinct behaviors. Some novel asymptotic distributions for quadratic forms involving sketching are derived, including those for Sparse Sign Embeddings and Subsampled Randomized Hadamard Transform (SRHT) sketching. Our simulations show that in certain scenarios, SRHT sketching achieves a better computation-accuracy tradeoff compared to subsampling-type methods.

We focus on the fixed-p regime, which is widely examined in the sketching literature, particularly in the context of the sketch-and-solve paradigm. When p is large, other sketching algorithms are typically more suitable, including iterative methods as well as sketch-and-precondition. Investigating the similarities and distinctions between various sketching techniques and sampling strategies, especially within frameworks such as sketch-and-precondition, remains an important direction for future research.

References

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

Daniel C Ahfock, William J Astle, and Sylvia Richardson. Statistical properties of sketching algorithms. *Biometrika*, 108(2):283–297, 2021.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

Theodore Wilbur Anderson. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, pages 122–148, 1963.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Burak Bartan and Mert Pilanci. Distributed sketching for randomized optimization: Exact characterization, concentration, and lower bounds. *IEEE Transactions on Information Theory*, 69(6):3850–3879, 2023.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Howard M Cann, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, Walter F Bodmer, Batsheva Bonne-Tamir, Anne Cambon-Thomsen, et al. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

Timothy I Cannings. Random projections: Data perturbation for classification problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(1):e1499, 2021.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

Shouyuan Chen, Yang Liu, Michael R Lyu, Irwin King, and Shengyu Zhang. Fast relative-error approximation algorithm for ridge regression. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 201–210, 2015.

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

Michal Derezinski, Zhenyu Liao, Edgar Dobriban, and Michael Mahoney. Sparse sketches with small inversion bias. In *Conference on Learning Theory*, pages 1467–1510. PMLR, 2021.

Edgar Dobriban and Sifan Liu. Asymptotics for sketching in least squares regression. In *Advances in Neural Information Processing Systems*, pages 3675–3685, 2019.

Edgar Dobriban and Art B Owen. Deterministic parallel analysis: an improved method for selecting factors and principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):163–183, 2019.

Petros Drineas and Michael W Mahoney. RandNLA: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for l_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Ethan N. Epperly. Which sketch should I-use?, 2023. URL https://www.ethanepperly.com/index.php/2023/11/27/which-sketch-should-i-use/.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

Gene H Golub and Charles F Van Loan. Matriz computations. *The Johns Hopkins University Press, Baltimore, USA*, 1989.

AK Gupta and DK Nagar. *Matrix Variate Distributions*, volume 104. CRC Press, 1999.

Nathan Halko, Per-Gunnar Martinsson, Yoel Shkolnisky, and Mark Tygert. An algorithm for the principal component analysis of large data sets. *SIAM Journal on Scientific computing*, 33(5):2580–2594, 2011a.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011b.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Andrew James Higgins, Erik Boman, and Ichitaro Yamazaki. A high performance gpu countsketch implementation and its application to multisketching and least squares problems. In *Proceedings of the SC'25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1808–1815, 2025.

Thomas Holgersson and Jolanta Pielaszkiewicz. A collection of moments of the wishart distribution. *Recent Developments in Multivariate and Random Matrix Analysis: Festschrift in Honour of Dietrich von Rosen*, pages 147–162, 2020.

Michael Kapralov, Vamsi Potluru, and David Woodruff. How to fake multiply by a gaussian matrix. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2016.

Jonathan Lacotte, Sifan Liu, Edgar Dobriban, and Mert Pilanci. Limiting spectrum of randomized hadamard transform and optimal iterative sketching methods. In *Conference on Neural Information Processing Systems*, 2020.

Sokbae Lee and Serena Ng. Least squares estimation using sketched data with heteroskedastic errors. In *International Conference on Machine Learning*, pages 12498–12520. PMLR, 2022.

Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza,

and Richard M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, 2008.

Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296. ACM, 2006.

Changyu Liu, Xingqiu Zhao, and Jian Huang. A random projection approach to hypothesis tests in high-dimensional single-index models. *Journal of the American Statistical Association*, pages 1–21, 2022.

Meimei Liu, Zuofeng Shang, and Guang Cheng. Sharp theoretical analysis for nonparametric testing under random projection. In *Conference on Learning Theory*, pages 2175–2209. PMLR, 2019.

Sifan Liu and Edgar Dobriban. Ridge regression: Structure, cross-validation, and sketching. *arXiv preprint arXiv:1910.02373, International Conference on Learning Representations (ICLR) 2020*, 2019.

Giacomo Livan, Marcel Novaes, and Pierpaolo Vivo. Introduction to random matrices theory and practice. *Monograph Award*, 63(54):914, 2018.

Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

Miles Lopes, Shusen Wang, and Michael Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. In *International Conference on Machine Learning*, pages 3217–3226. PMLR, 2018.

Miles Lopes, N Benjamin Erichson, and Michael Mahoney. Error estimation for sketched svd via the bootstrap. In *International Conference on Machine Learning*, pages 6382–6392. PMLR, 2020.

Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

Yichao Lu, Paramveer Dhillon, Dean P Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems*, pages 369–377, 2013.

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

Ping Ma, Yongkai Chen, Xinlian Zhang, Xin Xing, Jingyi Ma, and Michael W. Mahoney. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. *Journal of Machine Learning Research*, 23(177):1–45, 2022.

Jan R Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Per-Gunnar Martinsson and Joel A Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100, 2013.

Riley Murray, James Demmel, Michael W Mahoney, N Benjamin Erichson, Maksim Melnichenko, Osman Asif Malik, Laura Grigori, Piotr Luszczek, Michał Dereziński, Miles E Lopes, et al. Randomized numerical linear algebra: A perspective on the field with an eye to software. *arXiv preprint arXiv:2302.11474*, 2023.

Jolanta Pielaszkiewicz and Thomas Holgersson. Mixtures of traces of wishart and inverse wishart matrices. *Communications in Statistics-Theory and Methods*, 50(21):5084–5100, 2020.

Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17 (1):1842–1879, 2016.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Symposium on Foundations of Computer Science, 2006(FOCS'06)*., pages 143–152. IEEE, 2006.

Radhendushka Srivastava, Ping Li, and David Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970, 2016.

Aad W Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

Santosh S Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

Haiying Wang and Yanyuan Ma. Optimal subsampling for quantile regression in big data. *Biometrika*, 108(1):99–112, 2021.

HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018a.

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

Christine M Waternaux. Asymptotic distribution of the sample roots for a nonnormal population. *Biometrika*, 63(3):639–645, 1976.

Hadley Wickham. *nycflights13: Flights that Departed NYC in 2013*, 2018. URL https://CRAN.R-project.org/package=nycflights13. R package version 1.0.0.

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

Fan Yang, Sifan Liu, Edgar Dobriban, and David P Woodruff. How to reduce dimension with pca and random projections? *IEEE Transactions on Information Theory*, 67(12):8154–8189, 2021.

Jianfeng Yao, Zhidong Bai, and Shurong Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, New York, 2015.

Jun Yu, HaiYing Wang, Mingyao Ai, and Huiming Zhang. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. *Journal of the American Statistical Association*, 117(537):265–276, 2022.

Jun Yu, Zhiqiang Ye, Mingyao Ai, and Ping Ma. Optimal subsampling for data streams with measurement constrained categorical responses. *Journal of Computational and Graphical Statistics*, 34(3):994–1004, 2025.

Yichi Zhang and Minh Tang. Perturbation analysis of randomized svd and its applications to high-dimensional statistics. *arXiv preprint arXiv:2203.10262*, 2022.

Zhixiang Zhang, Sokbae Lee, and Edgar Dobriban. A framework for statistical inference via randomized algorithms. *arXiv preprint arXiv:2307.11255*, 2023.

Appendix

A Technical Reminders

We collect additional notation used throughout the appendix, besides those in Section 1.5. For any real numbers x and y, we define $\lceil x \rceil$ as the smallest integer greater than or equal to x, and $x \wedge y = \min\{x, y\}$. We use the symbol " d" to denote the differential of a function. For an event A, $I\{A\}$ denotes its indicator function, which equals to unity if A happens and to zero otherwise. We use $\sum_{i<j}$ and $\sum_{i\neq j}$ to abbreviate the summation over $(i,j) \in [n] \times [n]$ subject to $i < j$ and $i \neq j$, respectively. We define $1_m = (1, \ldots, 1)^\top \in \mathbb{R}^m$, and $\{e_i\}_{i=1}^m$ be the canonical orthonormal basis of $\mathbb{R}^m$. For $i \in [p]$, we let E_{ii} be a $p \times p$ matrix with the (i,i)-th entry equal to unity, and other entries equal to zero. For positive integers k, l, we often denote the columns of a $k \times l$ matrix A by $(a_j)_{j\in[l]}$. We will sometimes denote the Euclidean inner product by $\langle \cdot, \cdot \rangle$.

Denote by $A =_d B$ that two random vectors A, B have the same distribution. For any positive integer $p \geqslant 1$, we denote the probability distribution of a random vector $X \in \mathbb{R}^p$ by $L(X)$. Since the space of probability measures on $\mathbb{R}^p$ is metrizable, there is a metric d_p on this space, such that for a sequence of random vectors $(X_n)_{n\geqslant 1}$ and a random vector X, the weak convergence $X_n \Rightarrow X$ is equivalent to $d_p(L(X_n), L(X)) \to 0$.

Tensor Operations. Next, recall that for any matrices A, B and C of conformable sizes, we have that $\mathrm{vec}(ABC) = (C^\top \otimes A)\mathrm{vec}(B)$; see for example Theorem 2.4.2 in Magnus and Neudecker (2019). Additionally, for two vectors $a, b \in \mathbb{R}^p$, we have that

$$b \otimes a = (b \otimes I_p)a, \tag{36}$$

as well as $b \otimes a = \mathrm{vec}(ab^\top)$. Further, for the matrix P_p from Definition 2.7, representing matrix transposition in the vectorized matrix space,

$$(a \otimes b)^\top P_p (c \otimes d) = (a \otimes b)^\top (d \otimes c) = (a^\top d)(b^\top c). \tag{37}$$

In particular, for any $a \in \mathbb{R}^p$, $\Big((a \otimes I_p)^\top P_p (a \otimes I_p)\Big)_{ij} = (a \otimes e_i)^\top P_p (a \otimes e_j) = (a^\top e_i)(a^\top e_j) = e_i^\top a a^\top e_j = (aa^\top)_{ij}$, for all $i, j \in [p]$, so

$$(a \otimes I_p)^\top P_p (a \otimes I_p) = aa^\top. \tag{38}$$

Moreover, for the matrix $Q_p = \mathrm{vec}(I_p)\mathrm{vec}(I_p)^\top$ from Definition 2.7, we have $(Q_p)_{ij,kl} = \delta_{ij}\delta_{kl}$. Thus, for all $p \times p$ matrices Φ, Φ, we have

$$\mathrm{vec}(\Phi)^\top Q_p \mathrm{vec}(\Psi) = \sum_{(ij),(kl)} \Phi_{ij}\Psi_{kl}\delta_{ij}\delta_{kl} = \sum_{i,k} \Phi_{ii}\Psi_{kk} = \left(\sum_{i=1}^p \Phi_{ii}\right)\left(\sum_{k=1}^p \Psi_{kk}\right) = (\mathrm{tr}\,\Phi)(\mathrm{tr}\,\Psi). \tag{39}$$

Thus, for any $a \in \mathbb{R}^p$, $\Big((a \otimes I_p)^\top Q_p (a_n \otimes I_p)\Big)_{ij} = (a \otimes e_i)^\top Q_p (a \otimes e_j) = (a^\top e_i)(a^\top e_j) = e_i^\top a a^\top e_j = (aa^\top)_{ij}$, for all $i, j \in [p]$, so

$$(a \otimes I_p)^\top Q_p (a \otimes I_p) = aa^\top. \tag{40}$$

Subsampled randomized Hadamard transform (SRHT). For a positive integer n, we let $l = \lceil \log_2 n \rceil$ and $n' := 2^l \geqslant n$. An SRHT sketching matrix (Ailon and Chazelle, 2006) can be expressed as $S_{m,n} = \sqrt{n'/m} B H_l D \Pi_n$, where B is an $n' \times n'$ diagonal matrix with i.i.d. Bernoulli random variables with a success probability of m/n'.

Furthermore, D is an $n' \times n'$ diagonal matrix with i.i.d. ± 1 Rademacher random variables on the diagonal. Finally, Π_n is an $n' \times n$ with $\Pi_n^\top = [I_n, 0_{n'-n,n}]^\top$, where $0_{n'-n,n}$ is a $(n'-n) \times n$ zero matrix.

B Partial Sketching under the Unifying Conditions

For completeness we record the partial-sketching analogue of Theorem 2.5, which we defer from Section 2. A corresponding population-level PCA statement appears later in Corollary C.2.

Condition B.1 (Additional variance structure for partial sketching)**.** *Assume Condition 2.4. For all $n \geqslant 1$, there is $M_n^{(\mathrm{p})} = M_n^{(\mathrm{p})}(\overline{X_n \beta_n}) \in \mathbb{R}^{p \times p}$, also possibly depending on U_n, such that for any $b_n \in \mathbb{S}^{p-1}$ with $(U_n b_n, \overline{X_n \beta_n})_{n \geqslant 1}$ satisfying Condition $\mathfrak{P}$, σ_n^2 can be expressed in the form*

$$\sigma_n^2(U_n b_n, \overline{X_n \beta_n}) = b_n^\top M_n^{(\mathrm{p})}(\overline{X_n \beta_n}) b_n. \tag{41}$$

Theorem B.2 (Statistical Inference for the Partial Sketching LS Estimator)**.** *Let $(X_n)_{n \geqslant 1}$ be a sequence of data matrices satisfying Condition 2.3* `A`*, and let $(S_{m,n})_{m,n \geqslant 1}$ be a sequence of sketching matrices. Suppose that for any $(b_n)_{n \geqslant 1}$ satisfying $b_n \in \mathbb{S}^{p-1}$ for all n, the pairs $(U_n b_n, U_n b_n)_{n \geqslant 1}$ and $(U_n b_n, \overline{X_n \beta_n})_{n \geqslant 1}$ satisfy Condition $\mathfrak{P}$. Then we have the following two conclusions.*

Partial sketching. *Under Conditions 2.2, 2.4, and B.1, the partial sketching estimator $\hat{\beta}_{m,n}^{(\mathrm{p})}$ from (1) has the limiting distribution*

$$m^{1/2} \tau_{m,n}^{-1/2} \left(V_n L_n^{-1} M_n^{(\mathrm{p})}(\overline{X_n \beta_n}) L_n^{-1} V_n^\top \right)^{-1/2} \|X_n \beta_n\|^{-1} (\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p). \tag{42}$$

Observable normalization under* `1-dim'`*. *If the special case* `1-dim'` *in Condition 2.4 holds for $\psi \in \{0, 1\}$, then we further obtain the following inferential result for the unobserved least-squares parameter β_n based on the observed $\widetilde{X}_{m,n}, \tilde{y}_{m,n}$:*

$$m^{1/2} \tau_{m,n}^{-1/2} \left(\|\widetilde{X}_{m,n} \hat{\beta}_{m,n}^{(\mathrm{p})}\|^2 (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} + (\psi+1) \hat{\beta}_{m,n}^{(\mathrm{p})} \hat{\beta}_{m,n}^{(\mathrm{p})\top} \right)^{-1/2} (\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p). \tag{43}$$

Under the special case `1-dim'` in Condition 2.4, Condition B.1 is automatic, with $M_n^{(\mathrm{p})}(\overline{X_n \beta_n}) = I_p + (\psi + 1) U_n^\top \overline{X_n \beta_n} \overline{X_n \beta_n}^\top U_n$.

C Detailed results in the framework

While our results are data-conditional and do not assume that the data was generated by a statistical model, they can also be used for inference on population regression parameters when the data follows a population linear regression model. The following corollary illustrates this. See Appendix G.2 for the proof.

Corollary C.1 (Inference for regression parameters in a linear model)**.** *Assume the data follows a standard linear model with true parameter β^*, that is, $y_n = X_n\beta^* + \varepsilon_n$, where the error vector ε_n has i.i.d. components with mean zero and variance σ_o^2, and X_n can be either random or deterministic. Under the conditions of Theorem 2.5, let Condition 2.4* `1-dim'` *and the Lindeberg-Feller condition as stated in Example 2.28 of Van der Vaart (1998) hold. If in addition $m/(n\tau_{m,n}) \to 0$, we have*

$$m^{1/2}\tau_{m,n}^{-1/2}\|\tilde{\varepsilon}_{m,n}\|^{-1}(\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n})^{1/2}(\hat{\beta}_{m,n} - \beta^*) \Rightarrow \mathcal{N}(0, I_p). \tag{44}$$

Under the conditions of Theorem B.2, if instead $m/(\tau_{m,n}\|X_n\beta^*\|^2) \to_P 0$, then

$$m^{1/2}\tau_{m,n}^{-1/2}\left(\|\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})}\|^2 \cdot (\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n})^{-1} + (\psi+1)\hat{\beta}_{m,n}^{(\mathrm{p})}\hat{\beta}_{m,n}^{(\mathrm{p})\top}\right)^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta^*) \Rightarrow \mathcal{N}(0, I_p). \tag{45}$$

For various types of sketching methods discussed in this paper, $m/(n\tau_{m,n}) \to 0$ reduces to $m/n \to 0$. Further, $m/(\tau_{m,n}\|X_n\beta^*\|^2) \to_P 0$ has a similar interpretation if $\|X_n\beta^*\|^2 = \Theta(n)$, which is a genericity condition that holds if e.g., the matrix X_n has i.i.d. entries with zero mean and unit variance, and if $\|\beta^*\|^2 = \Theta(1)$. Therefore, given data sampled i.i.d. from a population, if the sketched sample size grows slower than the original sample size, we can directly use the sketched solution for inference about the population regression coefficient. However, we emphasize that our general theory is data-conditional and is thus stronger and more broadly applicable, as it does not assume any specific statistical model generating the data. For the following result, see Appendix G.4 for the proof.

Corollary C.2 (Inference for eigenvalues of the covariance matrix with i.i.d. data)**.** *Let the rows of X_n be sampled i.i.d. from a p-dimensional distribution having finite fourth cumulant with a covariance matrix $\Sigma \in \mathbb{S}_+^{p\times p}$. Denote $\lambda_i^* := \lambda_i(\Sigma)$ for $i \in [p]$. Under the conditions of Theorem 2.9, if in addition $m \cdot \left(n\tau_{m,n}(G_n)_{(ii),(ii)}\right)^{-1} \to 0$, we have for any $i \in [p]$ that*

$$\sqrt{\frac{m}{\tau_{m,n}(G_n)_{(ii),(ii)}}}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \lambda_i^*\right) \Rightarrow \mathcal{N}(0, 1). \tag{46}$$

Similar to the inference in a linear model as in Corollary C.1, $m \cdot \left(n\tau_{m,n}(G_n)_{(ii),(ii)}\right)^{-1} \to 0$ usually reduces to $m/n \to 0$. Thus, we can use sketching for inference about the population eigenvalues if $m \ll n$. Results for eigenvectors are similar and thus omitted.

D Inference Results for Other Sketching Distributions

D.1 Sketching Matrices with i.i.d. Entries

We now let $S_{m,n}$ be an $m \times n$ random sketching matrix with i.i.d. entries for all $n \geqslant 1$. Moreover, we denote the entries of the scaled sketching matrix as $\sqrt{m}S_{m,n} = (T_{ij})_{i\in[m],j\in[n]}$, where $\mathbb{E}T_{ij} = 0$ and $\mathbb{E}T_{ij}^2 = 1$, so that $\mathbb{E}\left[S_{m,n}\right] = 0$ and $\mathbb{E}\left[S_{m,n}^\top S_{m,n}\right] = I_n$. Additionally, for some positive δ', we assume $\mathbb{E}T_{ij}^4 = \kappa_{n,4} > 1 + \delta'$, which excludes distributions whose kurtosis is very close to the theoretical minimum of unity, such as the uniform distribution over $\{-1, 1\}$. See the

discussion after Theorem D.2 for more details. Moreover, for $\delta \geqslant 0$, we denote the moments $\kappa_{n,4+\delta} := \mathbb{E}T_{ij}^{4+\delta} < +\infty$.

Our results rely on the following conditions, where $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$, and $u_{n,i}$, $i \in [n]$ are the rows of the $n \times p$ left singular matrix U_n of X_n:

$$\textbf{Condition } \mathfrak{P} : \sup_{n\geqslant 1} \kappa_{n,4+\delta} < +\infty; \text{ Or } \kappa_{n,4} \to \infty, \kappa_{n,4+\delta}^{4/(4+\delta)}/\kappa_{n,4} = o(m^{\delta/(4+\delta)}),$$
$$\text{and for some } 0 < c < C, c \leqslant \sum_{i=1}^{n} (a_{n,i}\tilde{a}_{n,i})^2 \leqslant C; \tag{47}$$

$$\textbf{Condition } \mathfrak{P}' : \sup_{n\geqslant 1} \kappa_{n,4+\delta} < +\infty; \text{ Or } \kappa_{n,4} \to \infty, \kappa_{n,4+\delta}^{4/(4+\delta)}/\kappa_{n,4} = o(m^{\delta/(4+\delta)}),$$
$$\text{and there exists } c > 0 \text{ with } cI_{p(p+1)/2} \preceq D_p^\top \sum_{i=1}^{n} \left((u_{n,i}u_{n,i}^\top) \otimes (u_{n,i}u_{n,i}^\top)\right) D_p. \tag{48}$$

Both conditions can be satisfied *either* if the $4+\delta$-th moments $\kappa_{n,4+\delta}$ are uniformly bounded, or if they grow sufficiently slowly, but the data satisfies certain additional regularity conditions. As before, these conditions are used implicitly or explicitly in our results.

PCA. We have the following result for inference in PCA for sketching matrices with i.i.d. entries.

Theorem D.1 (Inference for Eigenvalues/vectors with i.i.d. Sketching)**.** *Consider the SVD of X_n given by $U_nL_nV_n^\top$. Let the entries of the scaled i.i.d. sketching matrix $\sqrt{m}S_{m,n} = (T_{ij})_{i,j}$ satisfy $\mathbb{E}T_{ij} = 0, \mathbb{E}T_{ij}^2 = 1$, $\mathbb{E}T_{ij}^4 = \kappa_{n,4} > 1+\delta'$ for some $\delta' > 0$, and $\mathbb{E}T_{ij}^{4+\delta} < +\infty$ for some $\delta > 0$. Suppose that $(X_n)_{n\geqslant 1}$ satisfies Condition 2.3* B *and $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$. For all $n \geqslant 1$, define the $p^2 \times p^2$ matrix Γ_n, where $(\Gamma_n)_{(k_1k_2),(k_3k_4)} = (\kappa_{n,4} - 3)\sum_{h=1}^{n}(U_n)_{hk_1}(U_n)_{hk_2}(U_n)_{hk_3}(U_n)_{hk_4}$ for every $k_1, k_2, k_3, k_4 \in [p]$. Then*

$$\sqrt{\frac{m}{2 + (\Gamma_n)_{(ii),(ii)}}}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1); \tag{49}$$

and letting

$$\hat{\Delta}_{m,n,i} = \sum_{k\neq i, l\neq i} \frac{\hat{\Lambda}_{m,i}\sqrt{\hat{\Lambda}_{m,n,k}\hat{\Lambda}_{m,n,l}}}{(\hat{\Lambda}_{m,i} - \hat{\Lambda}_{m,n,k})(\hat{\Lambda}_{m,i} - \hat{\Lambda}_{m,n,l})}\left(\delta_{kl} + (\Gamma_n)_{(ik),(il)}\right)\hat{v}_{m,n,k}\hat{v}_{m,n,l}^\top,$$

for any vector $c \in \mathbb{S}^{p-1}$ satisfying $\liminf_{n\to\infty} c^\top\Delta_{n,i}c > 0$, with $\Delta_{n,i} = \Delta_i(X_n^\top X_n, G_n)$ from Section 2.2.2,

$$\sqrt{m}(c^\top\hat{\Delta}_{m,n,i}c)^{-1/2}c^\top(\hat{v}_{m,n,i} - v_{n,i}) \Rightarrow \mathcal{N}(0,1).$$

In (49), the excess kurtosis term Γ_n vanishes if $\kappa_{n,4} = 3$, which holds for Gaussian distributions. In that case (49) depends only on the empirical data. For $\kappa_{n,4} \neq 3$, (49) only leads to inference if $(\Gamma_n)_{(ii),(ii)}$ vanishes, which can be viewed as a delocalization condition on the data. Theorem D.1 is related to the results of Anderson (1963) for Gaussian sketching matrices and to those of Waternaux (1976). Those results obtain the joint limiting distribution of all eigenvalues, while our results focus on individual eigenvalues but (1) establish inference methods, and (2) broaden the conditions, not requiring the convergence of the data; see Section H for more

discussion.

LS. We consider the problem of inference for β_n based on sketching matrices with i.i.d. entries. The result here is consistent with the special case of Theorem 3.2 in Zhang et al. (2023) for fixed p, but our conclusion holds for a broader class of sketching distributions, only assuming that a $4+\delta$-th moment (for any small $\delta > 0$) grows sufficiently slowly, whereas they assume all moments are bounded. In particular, our work allows sparse i.i.d. sketches, which requires the $4+\delta$-th moment to grow, whereas Zhang et al. (2023) does not allow this.

Theorem D.2 (Inference in OLS with i.i.d. Sketching)**.** *Let X_n satisfy Condition 2.3* `A`*, let p be fixed and $m, n \to \infty$. Also, let the entries of $S_{m,n}$ be i.i.d. copies of $S_{1,1}/m^{1/2}$, where $S_{1,1}$ has zero mean, unit variance and kurtosis $\kappa_{n,4} = \mathbb{E}S_{1,1}^4 > 1+\delta'$ for some $\delta' > 0$. Suppose either that*

(i) for some fixed $\delta > 0$, $\kappa_{n,4+\delta} = \mathbb{E}|S_{1,1}|^{4+\delta}$ satisfies $\sup_{n\geqslant 1} \kappa_{n,4+\delta} < +\infty$; or

(ii) we have $\kappa_{n,4} \to \infty$, $\kappa_{n,4+\delta}^{4/(4+\delta)}/\kappa_{n,4} = o(m^{\delta/(4+\delta)})$ and also there exist some constants $0 < c < C$ such that

$$cI_p \preceq U_n^\top \operatorname{diag}\left(\frac{|\varepsilon_n|}{\|\varepsilon_n\|} \odot \frac{|\varepsilon_n|}{\|\varepsilon_n\|}\right) U_n \preceq CI_p. \tag{50}$$

With

$$\hat{\Sigma}_{m,n} = m(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} \widetilde{X}_{m,n}^\top \operatorname{diag}(\tilde{\varepsilon}_{m,n} \odot \tilde{\varepsilon}_{m,n}) \widetilde{X}_{m,n} (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1}, \tag{51}$$

we have for the sketch-and-solve LS estimator $\hat{\beta}_{m,n}$ from (1) that $m^{1/2}\hat{\Sigma}_{m,n}^{-1/2}(\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$. Furthermore, with

$$\hat{\Sigma}_{m,n}^{(\mathrm{p})} = m(\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} \widetilde{X}_{m,n}^\top \operatorname{diag}\left((\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})}) \odot (\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})})\right) \widetilde{X}_{m,n} (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1},$$

if assuming either (i); or assuming (ii) with

$$cI_p \preceq U_n^\top \operatorname{diag}\left(\frac{|y_n - \varepsilon_n|}{\|y_n - \varepsilon_n\|} \odot \frac{|y_n - \varepsilon_n|}{\|y_n - \varepsilon_n\|}\right) U_n \preceq CI_p$$

instead of (50), then we have for the partial sketching LS estimator $\hat{\beta}_{m,n}^{(\mathrm{p})}$ from (1) that $m^{1/2}(\hat{\Sigma}_{m,n}^{(\mathrm{p})})^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$.

We know that $\kappa_{n,4} \geqslant 1$, and the minimum value is achieved for a Radamacher variable. This minimum is excluded from our analysis, because if $\kappa_{n,4} = 1$, some projections of $\hat{\beta}_{m,n} - \beta_n$ in certain directions may be deterministically zero, allowing us to exactly recover some linear functionals of β_n. Thus, one will need a more complex result to describe the behavior of $\hat{\beta}_{m,n} - \beta_n$. This can be inferred from equation (95) in the proofs by choosing specific a_n and $\tilde{a}_n$. For example, when $a_n = \tilde{a}_n = (1, 0, \ldots, 0)^\top$ or $a_n = (1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)^\top$ and $\tilde{a}_n = (1/\sqrt{2}, -1/\sqrt{2}, 0, \ldots, 0)^\top$, the right-hand side of (95) is zero. Since sketches whose entries are i.i.d. Rademacher variables are somewhat rarely used, we leave investigating this phenomenon to future work.

For Gaussian sketching, where $S_{m,n} = m^{-1/2}Z_{m,n}$ and $Z_{m,n}$ contains independent standard Gaussian entries, Theorem 1 of Ahfock et al. (2021) shows that the distribution of $\hat{\beta}_{m,n}$ given $\widetilde{X}_{m,n}$ is normal, with a known mean and covariance matrix. Hence, exact finite-sample statistical inference—conditional on $\widetilde{X}_{m,n}$—can be performed. Our result corresponds to taking the limit

as $m \to \infty$ in Theorem 1, part (ii) of Ahfock et al. (2021), for the marginal distribution of $\hat{\beta}_{m,n}$, where our m is their k and our $\sum_{i=1}^n \varepsilon_{n,i}^2 = \|y_n - X_n\beta_n\|^2$ is their RSS_F. An equivalent result for Gaussian sketching also appears in equation (5) of Bartan and Pilanci (2023).

CLT. The previous results are based on the following central limit theorems for quadratic forms underlying Conditions 2.2, 2.4, and 2.8 for i.i.d. sketching.

Lemma D.3 (Limiting Distributions of 1D Quadratic Forms for i.i.d. Sketching)**.** *Suppose that the entries of $S_{m,n}$ are i.i.d. copies of $S_{1,1}/m^{1/2}$, where $S_{1,1}$ has zero mean, unit variance, kurtosis $\kappa_{n,4} = \mathbb{E}S_{1,1}^4 > 1 + \delta'$ for some $\delta' > 0$, and that for some fixed $\delta > 0$, Condition $\mathfrak{P}$ from (47) holds. Then we have*

$$\sqrt{\frac{m}{1 + (a_n^\top \tilde{a}_n)^2 + (\kappa_{n,4} - 3)\sum_{i=1}^n (a_{n,i}\tilde{a}_{n,i})^2}} \left(a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\right) \Rightarrow \mathcal{N}(0,1).$$

Lemma D.4 (Asymptotic Normality of Symmetric QF for i.i.d. Sketching)**.** *Suppose that the entries of $S_{m,n}$ are i.i.d. copies of $S_{1,1}/m^{1/2}$, where $S_{1,1}$ has zero mean, unit variance, kurtosis $\kappa_{n,4} = \mathbb{E}S_{1,1}^4 > 1 + \delta'$ for some $\delta' > 0$. Let $(U_n)_{n\geqslant 1}$ satisfy Condition $\mathfrak{P}'$. Then for $G_n = I_{p^2} + P_p + \Gamma_n$ with Γ_n from Theorem D.1, we have*

$$\sqrt{m}(D_p^\top G_n D_p)^{-1/2} D_p^\top \operatorname{vec}\left(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p\right) \Rightarrow \mathcal{N}(0, I_{p(p+1)/2}).$$

D.2 Uniform Orthogonal Sketching

Now, we let $\gamma_n = m/n$ and introduce a sketching matrix $S_{m,n} = \gamma_n^{-1/2} S_0$, where S_0 is an $m \times n$ matrix drawn from the uniform measure over $m \times n$ partial orthogonal matrices; so that $\mathbb{E}S_{m,n} = 0$ and $\mathbb{E}S_{m,n}^\top S_{m,n} = I_n$. Here, Condition $\mathfrak{P}$ and $\mathfrak{P}'$ do not impose any restriction.

PCA. We have the following result for inference in PCA when using sketching matrices uniformly distributed on the space of all $m \times n$ partial orthogonal matrices.

Theorem D.5 (Inference for Eigenvalues/vectors with Haar Sketching)**.** *If $(X_n)_{n\geqslant 1}$ satisfies Condition 2.3* B *and $m/n \to \gamma \in (0,1)$, then*

$$\sqrt{\frac{m}{2(1-\gamma_n)}} \hat{\Lambda}_{m,i}^{-1} \left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}(0,1),$$

and for any vector $c \in \mathbb{S}^{p-1}$ satisfying $\limsup_{n\to\infty}(c^\top v_{n,i})^2 < 1$,

$$\sqrt{\frac{m}{(1-\gamma_n)}} \left(\sum_{k\neq i} \frac{\hat{\Lambda}_{m,i}\hat{\Lambda}_{m,n,k}}{(\hat{\Lambda}_{m,i} - \hat{\Lambda}_{m,n,k})^2} (c^\top \hat{v}_{m,n,k})^2\right)^{-1/2} c^\top(\hat{v}_{m,n,i} - v_{n,i}) \Rightarrow \mathcal{N}(0,1).$$

LS. Theorem D.6 provides a method to construct confidence regions for β_n for uniform orthogonal sketching. This result is a special case of Theorem 3.8 in Zhang et al. (2023) for fixed p, and is presented only for completeness. Moreover, we also present a shorter proof following our unified framework.

Theorem D.6 (Inference in OLS with Haar Sketching, Zhang et al. (2023))**.** *Let p be fixed and $m, n - m \to \infty$. Suppose that $\gamma_n^{1/2} S_{m,n}$ is uniformly distributed over the space of all partial*

$m \times n$ orthogonal matrices. Additionally, we assume that X_n satisfies Condition 2.3 A. *For the estimator $\hat{\Sigma}_{m,n}$ defined as $\hat{\Sigma}_{m,n} = \|\tilde{\varepsilon}_{m,n}\|^2 \cdot (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1}$, we have $(m/(1-\gamma_n))^{1/2}\, \hat{\Sigma}_{m,n}^{-1/2}(\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$. Furthermore, for the estimator $\hat{\Sigma}_{m,n}^{(\mathrm{p})}$ defined as*

$$\hat{\Sigma}_{m,n}^{(\mathrm{p})} = \hat{\beta}_{m,n}^{(\mathrm{p})\top} \widetilde{X}_{m,n}^\top \widetilde{X}_{m,n} \hat{\beta}_{m,n}^{(\mathrm{p})} \cdot (\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n})^{-1} + \hat{\beta}_{m,n}^{(\mathrm{p})} \hat{\beta}_{m,n}^{(\mathrm{p})\top},$$

we have $(m/(1-\gamma_n))^{1/2}\, (\hat{\Sigma}_{m,n}^{(\mathrm{p})})^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$.

CLT. The results of the inference mentioned earlier are based on the following central limit theorems for quadratic forms arising from Haar sketching matrices.

Lemma D.7 (Limiting Distributions of 1D Quadratic Forms for Haar Sketching)**.** *For any sequences $(a_n)_{n\geqslant 1}$, $(\tilde{a}_n)_{n\geqslant 1}$ such that $a_n, \tilde{a}_n \in \mathbb{S}^{n-1}$ for all $n \geqslant 1$, as $m, n-m \to \infty$ we have*

$$\sqrt{\frac{m}{1-\gamma_n}} \sqrt{\frac{1}{1+(a_n^\top \tilde{a}_n)^2}} \left(a_n^\top S_{m,n}^\top S_{m,n} \tilde{a}_n - a_n^\top \tilde{a}_n \right) \Rightarrow \mathcal{N}(0,1).$$

Lemma D.8 (Asymptotic Normality of Symmetric QF for Haar Sketching)**.** *With the same notation as above and $G = I_{p^2} + P_p$, $\gamma_n \to \gamma \in (0,1)$, we have*

$$\sqrt{m}(1-\gamma_n)^{-1/2} \mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) \Rightarrow \mathcal{N}(0, G).$$

E Computational Considerations

We now examine the computational complexity of performing statistical inference using various sketching methods for both least squares and PCA. We consider computation over a single machine, in which case communication cost is not a bottleneck. We perform floating-point operation (flop) count calculations based on standard matrix computation algorithms.

The computational cost consists of two primary components. The first part involves obtaining the sketched data $(\widetilde{X}_{m,n}, \tilde{y}_{m,n})$ for least squares, or only $\widetilde{X}_{m,n}$ for PCA. The second part involves solving the normal equations for the least squares problem, or the computation of eigenvalues and eigenvectors for $\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n}$. The cost of the second part is $O(mp^2)$, using standard matrix computation methods (Golub and Van Loan, 1989) irrespective of the chosen sketching method. Therefore, we focus on the cost incurred in the initial step of obtaining the sketched data For least squares, the original dataset (X_n, y_n) is of size $n \times (p+1)$, while the sketched dataset $(\widetilde{X}_{m,n}, \tilde{y}_{m,n})$ has size $m \times (p+1)$; for PCA, we only have the $n \times p$ matrix X_n and the $m \times p$ matrix $\widetilde{X}_{m,n}$. The following table summarizes the computational cost.

Table 3: Computational cost of performing sketched LS and PCA using various types of sketching matrices. Here, the sample size is n, the dimension of the data is $p \leqslant n$, and the sketched sample size is $p \leqslant m \leqslant n$; while $\mathrm{nnz}(X_n)$ denotes the number of nonzeros of the matrix X_n.

	i.i.d.	Haar	Hadamard	Sparse Sign Embeddings
LS	$O(mnp)$	$O(m^2 n)$	$O(np\log_2 n + mp^2)$	$O(\zeta_m \mathrm{nnz}(X_n) + mp^2)$
PCA	$O(mnp)$	$O(m^2 n)$	$O(np\log_2 n + mp^2)$	$O(\zeta_m \mathrm{nnz}(X_n) + mp^2)$

The conclusions of this analysis in general align with the existing common beliefs that Hadamard sketching and Sparse Sign Embeddings are two of the best sketching methods. Their

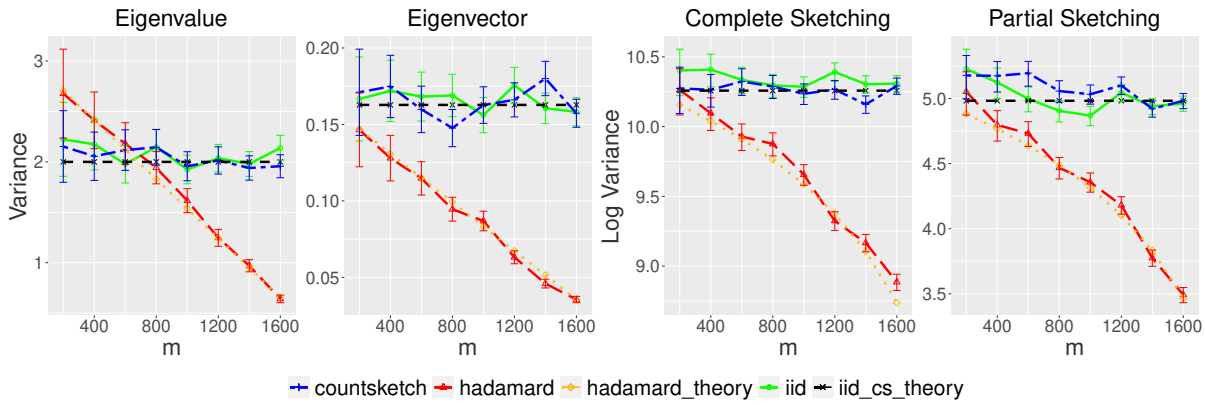


Figure 6: Variances of $\sqrt{m}\hat{\Lambda}_{m,n,k}$ and $\sqrt{m}c^\top \hat{v}_{m,n,k}$, and logarithm of variances of $\sqrt{m}c^\top \hat{\beta}_{m,n}$ and $\sqrt{m}c^\top \hat{\beta}^{(\mathrm{p})}_{m,n}$ for Case 1, using $p = 15, n = 2048, k = 1$, and $c = (1, 0, \ldots, 0)^\top$. Here "iid_theory" is obtained from Theorem H.1 and Theorem D.2, "hadamard_theory" from Theorem 3.1, "CountSketch_theory" from Theorem 3.6 and Theorem 3.7.

computational complexity is lower than that of the other approaches considered such as i.i.d. and Haar sketching. Moreover, we show in experiments (see Figure 5) that Hadamard sketching can have a better time accuracy trade-off than subsampling methods. However we emphasize that to achieve their full potential, each of these methods requires optimized numerical implementations.

F Details for Numerical Experiments

F.1 Illustrations for synthetic datasets

In Case 1, the left singular matrix meets all delocalization conditions required by our results, due to the isotropy of the normal distribution. The design in Case 2 may not satisfy the delocalization conditions required for Hadamard sketching and CountSketch. This case illustrates that those conditions are in fact needed for asymptotic normality. The delocalization is governed by the distribution from which the rows are drawn. If we sample each row from a heavy-tailed distribution, we expect that more outliers—observations with a large norm—may be present in the matrix. Consequently, the singular vectors are more likely to be localized. Simulations with several degrees of freedom for the t-distributions and with several covariance matrices show that our formulas are more accurate for larger degrees of freedom. For a given degree of freedom, our formulas have roughly the same level of accuracy for various covariance matrices.

In Case 3, before normalization, X_n has one spiked singular value of order $\sqrt{n}$ and $p-1$ singular values of order $\sqrt{n}$. The relative eigengap condition holds for the first singular value.

F.2 Empirical Data Example

As an empirical illustration of sketched least squares, we consider the New York flights (nycflights13) dataset (Wickham, 2018). This has 60,448 observations and 21 features, with the response being the arrival delay. We show coverage results for the least squares solution in Figure 9. We observe that all methods have coverage close to the nominal level.

To illustrate sketch-and-solve PCA, we use the Human Genome Diversity Project (HGDP)

dataset, previously analyzed in studies such as Cann et al. (2002); Li et al. (2008). We use the Centre d'Etude du Polymorphisme Humain panel, which contains Single Nucleotide Polymorphism (SNP) data for 1,043 observations, representing 51 populations from Africa, Europe, Asia, Oceania, and the Americas. Focusing on the SNPs located on chromosome 22, we create the data matrix X_n by selecting the first 20 features. In this $n \times p$ matrix, X_{ij} represents the count of the minor allele of SNP j in the genome of individual i and takes values in $\{0, 1, 2\}$. To preprocess the data, we perform SNP-wise standardization by centering each SNP using its mean and scaling by its standard error. Subsequently, we impute the missing values as zeros, as in Dobriban and Owen (2019). Figures 8 and 10 shows the coverage results as well as the asymptotic variances for different sketching methods based on the HGDP dataset. We observe that the coverage probability of the considered methods is close to the nominal level.

F.3 Comparison of Various Forms of Sketching

We proceed to compare the asymptotic variances across various sketching methods. Figures 6 and 10 compare various sketching methods in principal component analysis with respect to the ratio m/n of the sketch size to the sample size. These figures display the variance of $\sqrt{m}c^\top \hat{\Lambda}_{m,n,k}$ for simulated data in Cases 1, 2, and 3, respectively, where $n = 2048$, $p = 10$, and 500 replications are considered for each m. The asymptotic variances of i.i.d. sketching estimators for least squares are provided in equations (88) and (89) for the sketch-and-solve and partial sketching cases, respectively. Similarly, for PCA, the theoretical results can be derived from Table 2. Notably, the asymptotic properties of CountSketch closely resemble those of i.i.d. Gaussian sketching, denoted as "iid_cs_theory". Furthermore, we use "hadamard_theory" to denote the asymptotic results for SRHT sketching.

For sketch-and-solve least squares in Case 1, when the signal-to-noise ratio R_n^2 approaches zero, partial sketching estimators have smaller variances than sketch-and-solve estimators, see Figure 6. This finding aligns with the comparison of the relative efficiency of complete and partial sketching from Ahfock et al. (2021). The empirical variances closely approximate the theoretical asymptotic variances.

In Case 1 for PCA, the eigengap is large for all $k \in [p]$, ensuring that the variances are close to the predicted values, see Figure 6. We also observe good performance in other cases, see Figure 10. Furthermore, when employing Hadamard sketching, the variances of both $\sqrt{m}\hat{\Lambda}_{m,n,k}$ and $\sqrt{m}c^\top \hat{v}_{m,n,k}$ show a linear decrease in m, and tend to zero when m approaches n. In contrast, there is no significant decreasing trend in the asymptotic variance of i.i.d. sketching and CountSketch when p remains fixed. Instead, these variances remain relatively stable as m increases. These observations are consistent with the theoretical results summarized in Table 2.

F.4 Running Time

To empirically analyze the running time of our methods, we conduct experiments with various types of sketching methods for PCA in Cases 1 and 3. We choose m ranging from 200 to 1600 in increments of 200. For a fixed m, we repeat the process 500 times.

From the results in Figure 7, Table 4 and 5, CountSketch is the fastest method, while sparse sign embeddings with $\zeta_m > 1$ are slower but still perform very well. In terms of computational

cost, SRHT is slower than CountSketch and sparse sign embeddings. Sketching matrices with i.i.d. entries are slower, and Haar sketching matrices are the slowest. The results from Figure 7 qualitatively align with our theoretical analysis of computational complexity from Appendix E. We emphasize that specific wall clock times depend on the numerical implementation used, and our experiments should be viewed as giving only a rough idea of the relative performance of the various methods.

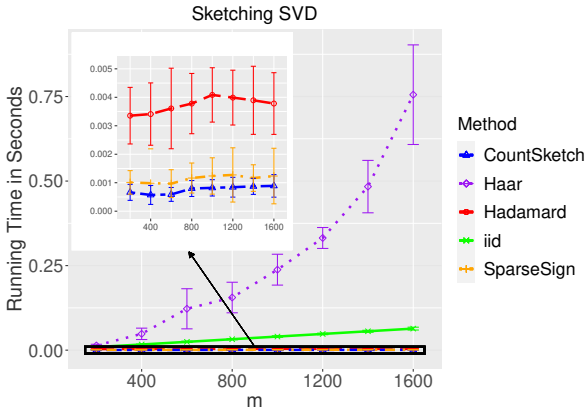


Figure 7: Computational cost experiment in Case 1 with $p = 15$, $n = 2048$, where m ranges from 200 to 1600 in increments of 200. For each sketching method, 500 simulations are conducted, and the mean time of a single step is calculated and plotted. Standard errors over 500 simulations are shown. Here we use Sparse Sign Embeddings with $\zeta_m = 8$.

Table 4: Computational cost (in seconds) of the experiments for Case 1 with 500 simulations, $p = 15$, $n = 2048$.

m	200	400	600	800	1000	1200	1400	1600
i.i.d. Gaussian Sketching	4.54	8.13	11.96	15.73	19.52	23.28	27.27	31.31
SRHT Sketching	2.18	2.08	2.06	2.26	2.52	2.25	2.14	2.19
CountSketch	0.32	0.30	0.33	0.38	0.39	0.46	0.48	0.51
SSE, $\zeta_m = 8$	0.67	0.71	0.72	0.83	0.85	0.84	0.90	0.96

Table 5: Computational costs (in seconds) of the experiments for Case 3, using the same protocol as in Appendix F.4.

m	200	400	600	800	1000	1200	1400	1600
i.i.d. Gaussian Sketching	4.48	8.22	12.13	16.31	19.93	24.78	29.98	33.26
SRHT Sketching	2.23	2.08	2.23	2.22	3.07	2.41	2.74	2.43
CountSketch	0.39	0.37	0.37	0.49	0.51	0.55	0.59	0.67
SSE, $\zeta_m = 8$	0.85	0.76	1.04	1.14	1.23	1.19	1.24	1.25

F.5 Comparison with Optimal Subsampling

Their Algorithm 2 involves three steps. First, a pilot subsample of r_0 datapoints is chosen uniformly, and is used to form a rough estimate of the parameter. Then, this estimate is used to further estimate the optimal subsample weights, and the new datapoints are added by weighted subsampling. The final subset is used to solve the maximum likelihood equation and obtain the final estimate. We focus on Hadamard sketching due to the availability of an efficient

implementation in the same language—R—for which the implementation of the methods from Yu et al. (2022) are available.

G Proofs of Results

G.1 Proofs of Theorem 2.5 and Theorem B.2

Sketch-and-solve

We first present the proof for $\hat{\beta}_{m,n}$. Condition 2.2 implies that for any sequence of nonzero vectors $(a_n, \tilde{a}_n)_{n\geqslant 1}$ of unit norm satisfying Condition $\mathfrak{P}$, $|a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n| = O_P(\tau_{m,n}^{1/2}/\sqrt{m})$. By assumption, for any $(b_n)_{n\geqslant 1}$ satisfying $b_n \in \mathbb{S}^{p-1}$ for all n, $(U_n b_n, U_n b_n)_{n\geqslant 1}$ and $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$ satisfy Condition $\mathfrak{P}$. Therefore, taking $a_n = U_n b_n$ and $\tilde{a}_n = \bar{\varepsilon}_n$, and noticing that $\|U_n b_n\|^2 = b_n^\top U_n^\top U_n b_n = b_n^\top b_n = 1$ and $\|\bar{\varepsilon}_n\| = 1$, we find that

$$|b_n^\top (U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) b_n| = O_P\left(\frac{\tau_{m,n}^{1/2}}{\sqrt{m}}\right) \quad \text{and} \quad |b_n^\top U_n^\top S_{m,n}^\top S_{m,n}\bar{\varepsilon}_n| = O_P\left(\frac{\tau_{m,n}^{1/2}}{\sqrt{m}}\right).$$

Therefore, since p is fixed, for any small $\varepsilon > 0$, we can take a union bound over b_n in a fixed ε-cover of $\mathbb{S}_+^{p\times p}$ to conclude that

$$\|U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p\| = O_P\left(\frac{\tau_{m,n}^{1/2}}{\sqrt{m}}\right), \quad \|U_n^\top S_{m,n}^\top S_{m,n}\bar{\varepsilon}_n\| = O_P\left(\frac{\tau_{m,n}^{1/2}}{\sqrt{m}}\right). \tag{52}$$

Using that $\tau_{m,n}/m \to 0$, it follows that $\|(U_n^\top S_{m,n}^\top S_{m,n} U_n)^{-1} - I_p\| = O_P(\tau_{m,n}^{1/2}/\sqrt{m})$. From equation (3), we have that

$$\hat{\beta}_{m,n} = \beta_n + \|\varepsilon_n\| \cdot V_n L_n^{-1} (U_n^\top S_{m,n}^\top S_{m,n} U_n)^{-1} U_n^\top S_{m,n}^\top S_{m,n}\bar{\varepsilon}_n. \tag{53}$$

Therefore, from (52), for any $c \in \mathbb{R}^n$ such that $X_n^{\dagger\top} c \neq 0$, using also that $U_n^\top U_{n,\perp} = 0$ so that $U_n^\top \bar{\varepsilon}_n = 0$, and that $X_n^{\dagger\top} = U_n L_n^{-1} V_n^\top$, we have that if $\varepsilon_n \neq 0$,

$$\begin{aligned}
&\|\varepsilon_n\|^{-1}\|X_n^{\dagger\top} c\|^{-1} c^\top(\hat{\beta}_{m,n} - \beta_n) \\
&= \|X_n^{\dagger\top} c\|^{-1} c^\top V_n L_n^{-1}[(U_n^\top S_{m,n}^\top S_{m,n} U_n)^{-1} - I_p] U_n^\top S_{m,n}^\top S_{m,n}\bar{\varepsilon}_n \\
&\quad + \|X_n^{\dagger\top} c\|^{-1} c^\top V_n L_n^{-1} U_n^\top S_{m,n}^\top S_{m,n}\bar{\varepsilon}_n \\
&= O_P\left(\frac{\tau_{m,n}}{m}\right) + \|X_n^{\dagger\top} c\|^{-1} c^\top V_n L_n^{-1} U_n^\top (S_{m,n}^\top S_{m,n} - I_n)\bar{\varepsilon}_n + \|X_n^{\dagger\top} c\|^{-1} c^\top V_n L_n^{-1} U_n^\top \bar{\varepsilon}_n \\
&= \|X_n^{\dagger\top} c\|^{-1} c^\top V_n L_n^{-1} U_n^\top (S_{m,n}^\top S_{m,n} - I_n)\bar{\varepsilon}_n + O_P\left(\frac{\tau_{m,n}}{m}\right).
\end{aligned} \tag{54}$$

We will use this approximation to study the distribution of $\hat{\beta}_{m,n}$. The distribution of the first term of the last line of (54) can be studied via Condition 2.4. In that condition, taking $a_n = U_n b_n$ and $\tilde{a}_n = \bar{\varepsilon}_n$ for every $n \geqslant 1$, where $b_n \in \mathbb{S}^{p-1}$, we find that with M_n defined in (4),

$$m^{1/2}\tau_{m,n}^{-1/2}\left(b_n^\top M_n b_n\right)^{-1/2} b_n^\top U_n^\top \left(S_{m,n}^\top S_{m,n} - I_n\right)\bar{\varepsilon}_n \Rightarrow \mathcal{N}(0,1).$$

Since $X_n^{\dagger\top} = U_n L_n^{-1} V_n^\top$, by taking $b_n = L_n^{-1} V_n^\top c / \|X_n^{\dagger\top} c\| \in \mathbb{S}^{p-1}$ and cancelling out $\|X_n^{\dagger\top} c\|$, we can rewrite this as

$$m^{1/2} \tau_{m,n}^{-1/2} \left(c^\top V_n L_n^{-1} M_n L_n^{-1} V_n^\top c \right)^{-1/2} c^\top V_n L_n^{-1} U_n^\top (S_{m,n}^\top S_{m,n} - I_n) \bar{\varepsilon}_n \Rightarrow \mathcal{N}(0,1). \tag{55}$$

Noting that $L_n^{-1} = \mathrm{diag}(\ell_{n,1}^{-1}, \ldots, \ell_{n,p}^{-1})$, with $0 < \ell_{n,1}^{-1} \leqslant \ldots \leqslant \ell_{n,p}^{-1}$, let

$$T_n := \frac{V_n L_n^{-1} M_n L_n^{-1} V_n^\top}{\|L_n^{-2}\|} \text{ and } W_n := m^{1/2} \tau_{m,n}^{-1/2} T_n^{-1/2} \|\varepsilon_n\|^{-1} \|L_n^{-1}\|^{-1} (\hat{\beta}_{m,n} - \beta_n).$$

where recall that $\|\cdot\|$ is the operator norm. Then, recalling the metric d_p metrizing weak convergence from Appendix A, our desired claim (5) amounts to $d_p(L(W_n), \mathcal{N}(0, I_p)) \to 0$.

We will prove this by a subsequence argument. First, we argue that we can find a compact set $B \subset \mathbb{S}_+^{p\times p}$ such that for all $n \geqslant 1$, $T_n \in B$. To see this, first, we have $\|T_n\| \leqslant \|M_n\|$; also, $\lambda_{\min}(T_n) \geqslant \lambda_{\min}(L_n^{-1})^2 \cdot \lambda_{\min}(M_n) / \|L_n^{-2}\| = \ell_{n,p}^2 / \ell_{n,1}^2 \lambda_{\min}(M_n)$. Since by Condition 2.4, for all $n \geqslant 1$, $a^\top M_n a = \sigma_n^2(U_n a, \bar{\varepsilon}_n)$ are bounded uniformly bounded away from zero and infinity for any $a \in \mathbb{S}^{p-1}$, the spectra of all $(M_n)_{n\geqslant 1}$ are uniformly bounded away from zero and infinity, and thus using that $(X_n)_{n\geqslant 1}$ satisfies Condition 2.3 `A`, so are the the spectra of all $(T_n)_{n\geqslant 1}$. This shows that there is a compact set $B \subset \mathbb{S}_+^{p\times p}$ such that for all $n \geqslant 1$, $T_n \in B$.

Next, we continue with the subsequence argument. Consider any subsequence $(n_k)_{k\geqslant 1}$ of $\mathbb{N}$. Since $T_{n_k} \in B$ for all $k \geqslant 1$ and B is bounded, there exists a further sub-subsequence $(n'_k)_{k\geqslant 1}$ of $(n_k)_{k\geqslant 1}$ such that $T_{n'_k} \to T$, for some $T \in \mathbb{S}_+^{p\times p}$. We denote by $(m'_k)_{k\geqslant 1}$ the sequence of sketch sizes associated with $(n'_k)_{k\geqslant 1}$. Using that $(c^\top V_n L_n^{-1} M_n L_n^{-1} V_n^\top c)^{1/2} = \|T_n^{1/2} c\| \cdot \|L_n^{-1}\|$, it follows from (55) that

$$(m'_k)^{1/2} \tau_{m'_k,n'_k}^{-1/2} \|L_{n'_k}^{-1}\|^{-1} c^\top V_{n'_k} L_{n'_k}^{-1} U_{n'_k}^\top (S_{m'_k,n'_k}^\top S_{m'_k,n'_k} - I_{n'_k}) \bar{\varepsilon}_{n'_k} \Rightarrow \mathcal{N}(0, c^\top T c).$$

Then, by (54), since $\|X_{n'_k}^{\dagger\top} c\| / \|L_{n'_k}^{-1}\| = O(1)$ and $\tau_{m,n}/m \to 0$, it follows that

$$(m'_k)^{1/2} \tau_{m'_k,n'_k}^{-1/2} \|\varepsilon_{n'_k}\|^{-1} \|L_{n'_k}^{-1}\|^{-1} c^\top \left(\hat{\beta}_{m'_k,n'_k}^{(\mathrm{s})} - \beta_{n'_k} \right) \Rightarrow \mathcal{N}(0, c^\top T c).$$

Using the Cramer-Wold device, we conclude that

$$(m'_k)^{1/2} \tau_{m'_k,n'_k}^{-1/2} \|\varepsilon_{n'_k}\|^{-1} \|L_{n'_k}^{-1}\|^{-1} \left(\hat{\beta}_{m'_k,n'_k}^{(\mathrm{s})} - \beta_{n'_k} \right) \Rightarrow \mathcal{N}(0, T).$$

Consequently, using again that $T_{n'_k} \to T$, $d_p\left(L(W_{n'_k}), \mathcal{N}(0, I_p)\right) \to 0$ holds along the subsequence $(n'_k)_{k\geqslant 1}$. Thus, for any subsequence $(n_k)_{k\geqslant 1}$, there exists a further sub-subsequence $(n'_k)_{k\geqslant 1}$ of $(n_k)_{k\geqslant 1}$, such that $d_p\left(L(W_{n'_k}), \mathcal{N}(0, I_p)\right) \to 0$. Thus, $d_p\left(L(W_n), \mathcal{N}(0, I_p)\right)$ tends to zero, finishing the proof of (5).

Partial sketching

Continuing with partial sketching, if $X_n\beta_n \neq 0$,

$$\begin{aligned}\hat{\beta}^{(\mathrm{p})}_{m,n} &= V_n L_n^{-1}(U_n^\top S_{m,n}^\top S_{m,n} U_n)^{-1} U_n^\top y_n \\ &= \beta_n + \|X_n\beta_n\| \cdot V_n L_n^{-1}\left[(U_n^\top S_{m,n}^\top S_{m,n} U_n)^{-1} - I_p\right] \overline{X_n\beta_n}.\end{aligned} \tag{56}$$

Thus, using (52), we have for any $c \in \mathbb{R}^n$ such that $X_n^{\dagger\top} c \neq 0$ that

$$\begin{aligned}&\|X_n\beta_n\|^{-1}\|X_n^{\dagger\top}c\|^{-1} c^\top(\hat{\beta}^{(\mathrm{p})}_{m,n} - \beta_n) \\ &= -\|X_n^{\dagger\top}c\|^{-1} c^\top V_n L_n^{-1} U_n^\top (S_{m,n}^\top S_{m,n} - I_n)\overline{X_n\beta_n} + O_P\left(\frac{\tau_{m,n}}{m}\right).\end{aligned} \tag{57}$$

By Conditions 2.4 and B.1, for $a_n = U_n b_n$ and $\tilde{a}_n = \overline{X_n\beta_n}$, for every $n \geqslant 1$, where $b_n \in \mathbb{S}^{p-1}$, we have

$$m^{1/2}\tau_{m,n}^{-1/2}(b_n^\top M_n^{(\mathrm{p})} b_n)^{-1/2} b_n^\top U_n^\top \left(S_{m,n}^\top S_{m,n} - I_n\right)\overline{X_n\beta_n} \Rightarrow \mathcal{N}(0,1),$$

where $M_n^{(\mathrm{p})}$ is defined in (41). Choosing $b_n = L_n^{-1}V_n^\top c/\|X_n^{\dagger\top}c\|$, it follows that

$$m^{1/2}\tau_{m,n}^{-1/2}\left(c^\top V_n L_n^{-1} M_n^{(\mathrm{p})} L_n^{-1} V_n^\top c\right)^{-1/2} c^\top V_n L_n^{-1} U_n^\top (S_{m,n}^\top S_{m,n} - I_n)\overline{X_n\beta_n} \Rightarrow \mathcal{N}(0,1).$$

Similar to the aforementioned proof for sketch-and-solve least squares, we define $T_n' := V_n L_n^{-1} M_n^{(\mathrm{p})} L_n^{-1} V_n^\top / \|L_n^{-2}\|$, $W_n' := m^{1/2}\tau_{m,n}^{-1/2} \cdot (T_n')^{-1/2}\|X_n\beta_n\|^{-1}\|L_n^{-1}\|^{-1}(\hat{\beta}^{(\mathrm{p})}_{m,n} - \beta_n)$, and need to show that $d_p(L(W_n'), \mathcal{N}(0, I_p)) \to 0$. This is followed by a subsequence argument essentially identical to the one used above in the proof for sketch-and-solve least squares, and we omit the proof to avoid repetition.

Inference

In particular, if Condition 2.4 `1-dim'` holds, then $M_n = I_p$, and so (5) reduces to (7). Further, we can verify that Condition 2.4 `1-dim'` with $\psi \in \{0,1\}$ implies that $M_n^{(\mathrm{p})}(\overline{X_n\beta_n}) = I_p + (\psi + 1)U_n^\top \overline{X_n\beta_n}\,\overline{X_n\beta_n}^\top U_n$ for $M_n^{(\mathrm{p})}$ from (41). Hence, with $\Sigma_n^{(\mathrm{p})} = \|y_n - \varepsilon_n\|^2(X_n^\top X_n)^{-1} + (\psi+1)\beta_n\beta_n^\top$, we have

$$m^{1/2}\tau_{m,n}^{-1/2}\left(\Sigma_n^{(\mathrm{p})}\right)^{-1/2}(\hat{\beta}^{(\mathrm{p})}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p). \tag{58}$$

Without loss of generality, we can consider $y_n \in \mathbb{S}^{n-1}$ and X_n of unit spectral radius. Indeed, for general y_n and X_n, we let $y_n' = y_n/\|y_n\|$ and $X_n' = X_n/\|X_n\|$, and we consider the observed data $S_{m,n}(X_n', y_n')$. By applying a similar argument and corresponding rescaled estimators, $\|X_n\|$ and $\|y_n\|$ cancel out, leading to the desired result for general X_n and y_n.

Next, since $\|U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p\| = O_P(\tau_{m,n}^{1/2}/\sqrt{m})$ and $\|X_n\| = 1$, we have

$$\|\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n} - X_n^\top X_n\| = O_P(\tau_{m,n}^{1/2}/\sqrt{m}). \tag{59}$$

Moreover, we can write

$$\begin{aligned}
&\|\tilde{\varepsilon}_{m,n}\|^2 = \|S_{m,n}(y_n - X_n\hat{\beta}_{m,n})\|^2 \\
&= \|S_{m,n}(y_n - X_n\beta_n)\|^2 + \|S_{m,n}X_n(\beta_n - \hat{\beta}_{m,n})\|^2 + 2\langle S_{m,n}(y_n - X_n\beta_n), S_{m,n}X_n(\beta_n - \hat{\beta}_{m,n})\rangle.
\end{aligned} \tag{60}$$

By using Condition 2.2, we have $\|S_{m,n}(y_n - X_n\beta_n)\|^2/\|\varepsilon_n\|^2 - 1 = O_P(\tau_{m,n}^{1/2}/\sqrt{m}) = o_P(1)$. From (59),

$$\begin{aligned}
\|S_{m,n}X_n(\beta_n - \hat{\beta}_{m,n})\|^2 &= \|X_n(\beta_n - \hat{\beta}_{m,n})\|^2 + O_P(\|\beta_n - \hat{\beta}_{m,n}\|^2 \cdot \tau_{m,n}^{1/2}/\sqrt{m}) \\
&\leqslant \|\beta_n - \hat{\beta}_{m,n}\|^2(1 + o_P(1)).
\end{aligned} \tag{61}$$

Also, from (7), since $\|X_n\| = 1$ and the condition number of X_n is bounded above from Condition 2.3 A, $\|(X_n^\top X_n)^{-1}\| = O(1)$ and

$$\|\hat{\beta}_{m,n} - \beta_n\| = O_P\left(\frac{\tau_{m,n}^{1/2}\|\varepsilon_n\| \cdot \|(X_n^\top X_n)^{-1}\|^{1/2}}{\sqrt{m}}\right) = O_P\left(\frac{\tau_{m,n}^{1/2}\|\varepsilon_n\|}{\sqrt{m}}\right).$$

Therefore, the second term on the second line of (60) is $O_P\left(m^{-1}\tau_{m,n}\|\varepsilon_n\|^2\right)$. Further, the third term is $O_P\left(m^{-1/2}\tau_{m,n}^{1/2}\|\varepsilon_n\|^2\right)$ by the Cauchy-Schwarz inequality. Therefore, we can conclude that $\|\tilde{\varepsilon}_{m,n}\|^2/\|\varepsilon_n\|^2 \to_P 1$. This verifies (6).

Continuing with partial sketching, we have from (58) that

$$\|\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n\| = O_P\left(\frac{\tau_{m,n}^{1/2}\|\Sigma_n^{(\mathrm{p})}\|^{1/2}}{\sqrt{m}}\right) = O_P\left(\frac{\tau_{m,n}^{1/2}\|\beta_n\|}{\sqrt{m}}\right). \tag{62}$$

Thus, $\|\hat{\beta}_{m,n}^{(\mathrm{p})}\hat{\beta}_{m,n}^{(\mathrm{p})\top} - \beta_n\beta_n^\top\| = o_P(\|\beta_n\|^2)$. Further, since we can write $\|y_n - \varepsilon_n\|^2 = \|X_n\beta_n\|^2$, using (59) and (62), we have

$$\begin{aligned}
&\|\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})}\|^2 - \|y_n - \varepsilon_n\|^2 = (\|\widetilde{X}_{m,n}\hat{\beta}_{m,n}^{(\mathrm{p})}\|^2 - \|\widetilde{X}_{m,n}\beta_n\|^2) + (\|\widetilde{X}_{m,n}\beta_n\|^2 - \|y_n - \varepsilon_n\|^2) \\
&= \mathrm{tr}\left(\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n}(\hat{\beta}_{m,n}^{(\mathrm{p})}\hat{\beta}_{m,n}^{(\mathrm{p})\top} - \beta_n\beta_n^\top)\right) + \beta_n^\top(\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n} - X_n^\top X_n)\beta_n \\
&\leqslant \|\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n}\|\|\hat{\beta}_{m,n}^{(\mathrm{p})}\hat{\beta}_{m,n}^{(\mathrm{p})\top} - \beta_n\beta_n^\top\| + \|\beta_n\|^2\|\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n} - X_n^\top X_n\| = o_P(\|\beta_n\|^2).
\end{aligned}$$

Therefore,

$$\|\Sigma_n^{(\mathrm{p})} - \hat{\Sigma}_{m,n}^{(\mathrm{p})}\|_{\mathrm{Fr}} = o_P(\|\beta_n\|^2).$$

Now, $\min_{a\in\mathbb{S}^{p-1}} a^\top\left(\|X_n\beta_n\|^2(X_n^\top X_n)^{-1} + (\psi+1)\beta_n\beta_n^\top\right)a = \lambda_{\min}(\Sigma_n^{(\mathrm{p})})$, and we have for any $a \in \mathbb{S}^{p-1}$ that

$$\|X_n\beta_n\|^2 a^\top(X_n^\top X_n)^{-1}a \geqslant \lambda_{\min}(X_n^\top X_n)\|\beta_n\|^2\lambda_{\min}\left((X_n^\top X_n)^{-1}\right) = \lambda_{\min}(X_n^\top X_n)\|\beta_n\|^2.$$

Combined with Condition 2.3 A, it follows that $\|(\Sigma_n^{(\mathrm{p})})^{-1}\| = \lambda_{\min}(\Sigma_n^{(\mathrm{p})})^{-1} = O_P(\|\beta_n\|^{-2})$. Combined with $\|\Sigma_n^{(\mathrm{p})} - \hat{\Sigma}_{m,n}^{(\mathrm{p})}\|_{\mathrm{Fr}} = o_P(\|\beta_n\|^2)$, this leads to $\|(\Sigma_n^{(\mathrm{p})})^{-1}\hat{\Sigma}_{m,n}^{(\mathrm{p})} - I_p\| = o_P(1)$, and (43) follows.

G.2 Proof of Corollary C.1

Recall from Theorem 2.5 that conditional on (X_n, y_n), (6) holds. This implies that we also have the unconditional result

$$m^{1/2}\tau_{m,n}^{-1/2}\|\tilde{\varepsilon}_{m,n}\|^{-1}(\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n})^{1/2}(\hat{\beta}_{m,n}-\beta_n) \Rightarrow \mathcal{N}(0, I_p). \tag{63}$$

Moreover, we have $\sigma_o^{-1}(X_n^\top X_n)^{1/2}(\beta_n - \beta^*) \Rightarrow \mathcal{N}(0,1)$, see Example 2.28 in Van der Vaart (1998). From (59) and (60), we have $(X_n^\top X_n)^{-1}(\widetilde{X}_{m,n}^\top\widetilde{X}_{m,n}) \to_P I_p$ and $\|\tilde{\varepsilon}_{m,n}\|^{-1}\|\varepsilon_n\| \to_P 1$. Due to the law of large numbers, we have $\|\varepsilon_n\|^2/n \to_P \sigma_o^2$. From Slutsky's theorem, since $m/(n\tau_{m,n}) \to 0$, we conclude (44).

For partial sketching we have from (58) in Appendix G.1 that unconditionally,

$$m^{1/2}\tau_{m,n}^{-1/2}\left(\Sigma_n^{(\mathrm{p})}\right)^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})}-\beta_n) \Rightarrow \mathcal{N}(0, I_p). \tag{64}$$

Now, $\sigma_o^{-1}(X_n^\top X_n)^{1/2}(\beta_n - \beta^*) \Rightarrow \mathcal{N}(0,1)$ is equivalent to

$$\sqrt{\frac{\tau_{m,n}}{m}}\sigma_o^{-1}(X_n^\top X_n)^{1/2}\left(\Sigma_n^{(\mathrm{p})}\right)^{1/2}\cdot\sqrt{\frac{m}{\tau_{m,n}}}\left(\Sigma_n^{(\mathrm{p})}\right)^{-1/2}(\beta_n-\beta^*) \Rightarrow \mathcal{N}(0,1).$$

Observe that

$$\begin{aligned}
\left\|\left(\Sigma_n^{(\mathrm{p})}\right)^{-1/2}(X_n^\top X_n)^{-1/2}\right\| &= \lambda_{\min}^{-1}\left((X_n^\top X_n)^{1/2}\left(\Sigma_n^{(\mathrm{p})}\right)^{1/2}\right) \\
&= \lambda_{\min}^{-1/2}\left(\|y_n-\varepsilon_n\|^2+(\psi+1)X_n^\top X_n\beta_n\beta_n^\top\right) \\
&\le \left(\|y_n-\varepsilon_n\|^2+(\psi+1)\lambda_{\min}(X_n^\top X_n)\lambda_{\min}(\beta_n\beta_n^\top)\right)^{-1/2} \\
&\le \|y_n-\varepsilon_n\|^{-1}.
\end{aligned}$$

From our assumptions, it can be checked that $m^{1/2}/(\tau_{m,n}^{1/2}\|X_n\beta^*\|) \to_P 0$. Therefore,

$$m^{1/2}\tau_{m,n}^{-1/2}\left\|\left(\Sigma_n^{(\mathrm{p})}\right)^{-1/2}(X_n^\top X_n)^{-1/2}\right\| \to_P 0.$$

Combining with Slutsky's theorem, we have $m^{1/2}\tau_{m,n}^{-1/2}\left(\Sigma_n^{(\mathrm{p})}\right)^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})}-\beta^*) \Rightarrow \mathcal{N}(0, I_p)$. Finally, as in the proof of (43), we can replace $\Sigma_n^{(\mathrm{p})}$ to its consistent estimator $\hat{\Sigma}_{m,n}^{(\mathrm{p})}$. This completes the proof of (45).

G.3 Proof of Theorem 2.9

The proof relies on the following result about asymptotic distributions for the sketched PCA solutions, obtained leveraging the delta method.

Proposition G.1 (Asymptotic Normality of Eigenvectors and Eigenvalues)**.** *Suppose that the sequence $(X_n^\top X_n)_{n\geqslant 1}$ of matrices converges to $\Sigma \in \mathbb{S}_+^{p\times p}$, and $(G_n)_{n\geqslant 1}$ converges to a positive semi-definite matrix G. Assume Σ has the spectral decomposition $V^\top L^2 V$, where $V =$*

$(\nu_1, \nu_2, \ldots, \nu_p)$ *and* $L = \mathrm{diag}(\lambda_{1,0}, \ldots, \lambda_{p,0})^{1/2}$. *Under Condition 2.8, we have for all* $i \in [p]$ *that*

$$\sqrt{m}\tau_{m,n}^{-1/2}\lambda_{i,0}^{-1}\left(\hat{\Lambda}_{m,i} - \Lambda_{n,i}\right) \Rightarrow \mathcal{N}\left(0, G_{(ii),(ii)}\right) \tag{65}$$

and with Δ_i *from* (10),

$$\sqrt{m}\tau_{m,n}^{-1/2}\left(\hat{v}_{m,n,i} - v_i\right) \Rightarrow \mathcal{N}\left(0, \Delta_i(\Sigma, G)\right). \tag{66}$$

Proof. From (8) in Condition 2.8 and $G_n \to G$, we conclude that (9) holds. Also, we can consider λ_i and v_i, $i \in [p]$, as functions on $\mathbb{R}^{p^2}$; and since all eigenvalues of Σ are distinct, λ_i, v_i, $i \in [p]$, can be viewed as smooth functions in a neighborhood $N \subset \mathbb{R}^{p^2}$ of $\mathrm{vec}(\Sigma)$. Given the assumption that $X_n^\top X_n \to \Sigma$, it follows that for some $n_0 \in \mathbb{N}$, $\mathrm{vec}(X_n^\top X_n) \in N$ for all $n \geqslant n_0$. Further, (9) implies that $\mathrm{vec}\left(X_n^\top S_{m,n}^\top S_{m,n} X_n - X_n^\top X_n\right) \to_P 0$, so the events $\Omega_n = \{\mathrm{vec}(X_n^\top S_{m,n}^\top S_{m,n} X_n) \in N\}$ have $P(\Omega_n) \to 1$. Now fix $i \in [p]$. By the mean value theorem, for some $\theta_{m,n} \in [0,1]$,

$$\begin{aligned}\lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n) - \lambda_i(X_n^\top X_n) &= \nabla\lambda_i\left(X_n^\top X_n + \theta_{m,n}(X_n^\top S_{m,n}^\top S_{m,n} X_n - X_n^\top X_n)\right)^\top \\ &\quad \cdot \mathrm{vec}(X_n^\top S_{m,n}^\top S_{m,n} X_n - X_n^\top X_n) I(\Omega_n) + \left(\lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n) - \lambda_i(X_n^\top X_n)\right) I(\Omega_n^c).\end{aligned}$$

Since $P(\Omega_n^c) \to 0$, by Slutsky's theorem, the limiting distribution of $\lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n) - \lambda_i(X_n^\top X_n)$ is determined by the first term on the event Ω_n. Applying the continuous mapping theorem on Ω_n, we have

$$\nabla\lambda_i\left(X_n^\top X_n + \theta_{m,n}(X_n^\top S_{m,n}^\top S_{m,n} X_n - X_n^\top X_n)\right) \to_P \nabla\lambda_i(\Sigma).$$

Now, defining $\bar{X}_n = X_n(X_n^\top X_n)^{-1/2}\Sigma^{1/2}$, we can write

$$\sqrt{m}(X_n^\top S_{m,n}^\top S_{m,n} X_n - X_n^\top X_n) = \sqrt{m}(X_n^\top X_n)^{1/2}\Sigma^{-1/2}(\bar{X}_n^\top S_{m,n}^\top S_{m,n}\bar{X}_n - \Sigma)\Sigma^{-1/2}(X_n^\top X_n)^{1/2}.$$

According to Slutsky's theorem, it is sufficient to characterize the limiting distribution of $\sqrt{m}(\bar{X}_n^\top S_{m,n}^\top S_{m,n}\bar{X}_n - \Sigma)$.

Using the SVD $X_n = U_n L_n V_n^\top$ and that $\Sigma = V^\top L^2 V$, we can write $\bar{X}_n = U_n V_n^\top V L V^\top$. Denoting $\bar{U}_n = U_n V_n^\top V$, the SVD of $\bar{X}_n$ is therefore $\bar{X}_n = \bar{U}_n L V^\top$. Thus, $\bar{X}_n^\top S_{m,n}^\top S_{m,n}\bar{X}_n - \Sigma = VL(\bar{U}_n^\top S_{m,n}^\top S_{m,n}\bar{U}_n - I_p)LV^\top$, and consequently, recalling the discussion from Appendix A,

$$\mathrm{vec}(\bar{X}_n^\top S_{m,n}^\top S_{m,n}\bar{X}_n - \Sigma) = ((VL)\otimes(VL))\mathrm{vec}(\bar{U}_n^\top S_{m,n}^\top S_{m,n}\bar{U}_n - I_p). \tag{67}$$

Additionally, since $X_n^\top X_n \to \Sigma$ and the eigenvalues of Σ are distinct, and as our conventions resolve the sign-ambiguity of eigenvectors, we have that $V_n \to V$. Since $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, by Condition 2.8 and applying Slutsky's theorem to $U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p = V_n^\top V(\bar{U}_n^\top S_{m,n}^\top S_{m,n}\bar{U}_n - I_p)V^\top V_n$, we have $\sqrt{m}\tau_{m,n}^{-1/2}\mathrm{vec}(\bar{U}_n^\top S_{m,n}^\top S_{m,n}\bar{U}_n - I_p) \Rightarrow \mathcal{N}(0, G)$. Consequently, by (67), with $\rho := ((VL)\otimes(VL))G((LV^\top)\otimes(LV^\top))$, we have

$$\sqrt{m}\tau_{m,n}^{-1/2}\mathrm{vec}(\bar{X}_n^\top S_{m,n}^\top S_{m,n}\bar{X}_n - \Sigma) \Rightarrow \mathcal{N}(0, \rho).$$

Next, we recall the gradient $\nabla\lambda_i(\Sigma)$ and the Jacobian $Jv_i(\Sigma)$, in a form appropriate for our use. Based on Theorem 8.9 in Magnus and Neudecker (2019), we know that $\mathrm{d}\lambda_i = v_i^\top \mathrm{d}(X) v_i$. This implies that $\mathrm{d}\lambda_i = (v_i \otimes v_i)^\top \mathrm{d}(\mathrm{vec}(X))$. Therefore, we can conclude that $\nabla\lambda_i(\Sigma) = \nu_i \otimes \nu_i$. Similarly, $\mathrm{d}v_i = (\lambda_{i,0} I_p - \Sigma)^\dagger \mathrm{d}(X) v_i$, and thus $\mathrm{d}v_i = \Big(v_i \otimes (\lambda_{i,0} I_p - \Sigma)^\dagger\Big)^\top \mathrm{d}(\mathrm{vec}(X))$. Consequently, we have $Jv_i(\Sigma) = \nu_i \otimes (\lambda_{i,0} I_p - \Sigma)^\dagger$.

Applying Slutsky's theorem and combining terms, we find that

$$\sqrt{m}\tau_{m,n}^{-1/2}\Big(\lambda_i(X_n^\top S_{m,n}^\top S_{m,n} X_n) - \lambda_i(X_n^\top X_n)\Big) \Rightarrow \mathcal{N}\Big(0, (\nu_i \otimes \nu_i)^\top \rho(\nu_i \otimes \nu_i)\Big),$$

as well as

$$\sqrt{m}\tau_{m,n}^{-1/2}\Big(v_i(X_n^\top S_{m,n}^\top S_{m,n} X_n) - v_i(X_n^\top X_n)\Big) \Rightarrow \mathcal{N}\Big(0, (\nu_i \otimes (\lambda_{i,0} I_p - \Sigma)^\dagger)^\top \rho(\nu_i \otimes (\lambda_{i,0} I_p - \Sigma)^\dagger)\Big).$$

Since $V = (\nu_1, \nu_2, \ldots, \nu_p)$, we have $LV^\top \nu_i = Le_i = \lambda_{i,0} e_i$ for each $i \in [p]$. Therefore, $((LV^\top) \otimes (LV^\top))(\nu_i \otimes \nu_i) = (LV^\top \nu_i) \otimes (LV^\top \nu_i) = \lambda_{i,0}^2 e_i \otimes e_i$. Thus,

$$(\nu_i \otimes \nu_i)^\top \rho(\nu_i \otimes \nu_i) = \lambda_{i,0}^2 (e_i \otimes e_i)^\top G (e_i \otimes e_i) = \lambda_{i,0}^2 G_{(ii),(ii)}.$$

This shows (65). Moreover, $(\lambda_{i,0} I_p - \Sigma)^\dagger = V^\top \mathrm{diag}((\lambda_{i,0} - \lambda_{k,0})^{-1}(1 - \delta_{ki}))_{k=1}^p V$, where we interpret $0/0 := 0$. Also, $LV^\top(\lambda_{i,0} I_p - \Sigma)^\dagger = \mathrm{diag}(\sqrt{\lambda_{k,0}}(\lambda_{i,0} - \lambda_{k,0})^{-1}(1 - \delta_{ki}))_{k=1}^p V^\top$. Denote $\alpha_k = \sqrt{\lambda_{k,0}}/(\lambda_{i,0} - \lambda_{k,0})$ for all $k \neq i$ and $\alpha_i = 0$. Also let $\tilde{V} = \mathrm{diag}(\sqrt{\lambda_{k,0}}(\lambda_{i,0} - \lambda_{k,0})^{-1}(1 - \delta_{ki}))_{k=1}^p V^\top = (\alpha_1 \nu_1, \ldots, \alpha_p \nu_p)^\top$. Then,

$$(\nu_i \otimes (\lambda_{i,0} I_p - \Sigma)^\dagger)^\top \rho(\nu_i \otimes (\lambda_{i,0} I_p - \Sigma)^\dagger) = \lambda_{i,0}(e_i \otimes \tilde{V})^\top G (e_i \otimes \tilde{V}).$$

We denote by $\tilde{G}^i$ the $p \times p$ principal submatrix of G, for which $\tilde{G}^i_{kl} = G_{(ik),(il)}$ for all $k, l \in [p]$. Then

$$(e_i \otimes \tilde{V})^\top G (e_i \otimes \tilde{V}) = \tilde{V}^\top \tilde{G}^i \tilde{V} = (\alpha_1 \nu_1, \ldots, \alpha_p \nu_p) \tilde{G}^i (\alpha_1 \nu_1, \ldots, \alpha_p \nu_p)^\top = \sum_{k,l} \alpha_k \alpha_l (\tilde{G}^i)_{kl} \nu_k \nu_l^\top.$$

Hence,

$$\lambda_{i,0}(e_i \otimes \tilde{V})^\top G (e_i \otimes \tilde{V}) = \sum_{k \neq i, l \neq i} \frac{\lambda_{i,0}\sqrt{\lambda_{k,0}\lambda_{l,0}}}{(\lambda_{i,0} - \lambda_{k,0})(\lambda_{i,0} - \lambda_{l,0})} G_{(ik),(il)} \nu_k \nu_l^\top.$$

Due to the definition of Δ_i from (10), this proves (66).

□

We first show Theorem 2.9 for a sequence of data matrices $(X_n)_{n \geqslant 1}$ and covariance matrices $(G_n)_{n \geqslant 1}$ satisfying the conditions of Proposition G.1. In that case, by using that (65) holds due to Proposition G.1, and from Slutsky's theorem, we conclude (11). Next, recall that from the continuous mapping theorem, since $X_n^\top X_n \to \Sigma$, $G_n \to G$, and $X_n^\top S_{m,n}^\top S_{m,n} X_n - X_n^\top X_n \to_P 0$, we have $\Lambda_k \to \lambda_{k,0}$, $\hat{\Lambda}_{m,n,k} \to_P \lambda_{k,0}$, $v_k \to \nu_k$, and $\hat{v}_{m,n,k} \to_P \nu_k$ for all $k \in [p]$. This shows that $\hat{\Delta}_{m,n,i}$ is well-defined with probability tending to unity, and $\hat{\Delta}_{m,n,i} \to_P \Delta_i(\Sigma, G)$, $\Delta_{n,i} \to \Delta_i(\Sigma, G)$. Since $\liminf_{n\to\infty} c^\top \Delta_{n,i} c > 0$, there exists $\varepsilon > 0$ such that $c^\top \Delta_{n,i} c > \varepsilon$ for all

sufficiently large $n \in \mathbb{N}$, and as a result, $c^\top \Delta_i(\Sigma, G)c \geqslant \varepsilon > 0$. Thus, the left-hand side in (12) is well-defined for sufficiently large $n \in \mathbb{N}$. Moreover, using (66) and Slutsky's theorem, (12) holds under the conditions of Proposition G.1.

Finally, we use a subsequence argument to show Theorem 2.9 under the conditions from its statement. Without loss of generality we can assume that $\|X_n\| = 1$ by working with $X_n/\|X_n\|$. Indeed, note that Δ_i from (10) is a homogeneous function with respect to Ξ, i.e., for any $c > 0$, any $G \in \mathbb{R}^{p^2 \times p^2}$, and all feasible choices of $\Xi \in \mathbb{S}_+^{p\times p}$, $\Delta_i(\Xi, G) = \Delta_i(c\Xi, G)$. We observe that the left-hand side of (11) and (12) remains invariant when we scale the data X_n to $X_n' = X_n/g_n$ for some nonzero scaling sequence $(g_n)_{n\geqslant 1}$. In this case, we have $\hat{\Lambda}'_{m,n,i} = \hat{\Lambda}_{m,i}/g_n^2$ and $\Lambda'_{n,i} = \Lambda_{n,i}/g_n^2$, leading to $\hat{\Lambda}_{m,i}^{-1}(\hat{\Lambda}_{m,i} - \Lambda_{n,i}) = (\hat{\Lambda}'_{m,i})^{-1}(\hat{\Lambda}'_{m,n,i} - \Lambda'_{n,i})$, and due to the homogeneity of the function Δ_i, we have $(c^\top \hat{\Delta}_{m,n,i}c)^{-1/2}c^\top(\hat{v}_{m,n,i} - v_{n,i}) = (c^\top \hat{\Delta}'_{m,n,i}c)^{-1/2}c^\top(\hat{v}'_{m,n,i} - v'_{n,i})$. Hence, from now on, we assume without loss of generality that $\|X_n\| = 1$.

Let

$$
\begin{aligned}
&T_n := X_n^\top X_n, \qquad W_n := m^{1/2}\tau_{m,n}^{-1/2}(G_n)_{(ii),(ii)}^{-1/2}\hat{\Lambda}_{m,i}^{-1}(\hat{\Lambda}_{m,i} - \Lambda_{n,i}),\\
&\chi_n := m^{1/2}\tau_{m,n}^{-1/2}(c^\top \Delta_{n,i}c)^{-1/2}c^\top(\hat{v}_{m,n,i} - v_{n,i}),\\
&\tilde{\chi}_n := m^{1/2}\tau_{m,n}^{-1/2}(c^\top \hat{\Delta}_{m,n,i}c)^{-1/2}c^\top(\hat{v}_{m,n,i} - v_{n,i}).
\end{aligned}
\tag{68}
$$

We have that $\|T_n\| = 1$, and the singular values of T_n are $(\ell_{n,i}^2/\ell_{n,1}^2)_{i\in[p]} = (\ell_{n,i}^2)_{i\in[p]}$. Thus, due to Condition 2.3 B, there is a compact set $K \subset \mathbb{S}_+^{p\times p}$ such that for all $n \geqslant 1$, $T_n \in K$. Since there exists $0 < c \leqslant C$, such that $cI_{p(p+1)/2} \preceq D_p^\top G_n D_p \preceq CI_{p(p+1)/2}$, there is a compact set $\tilde{K} \subset \mathbb{R}^{p(p+1)/2\times p(p+1)/2}$ such that for all $n \geqslant 1$, $D_p^\top G_n D_p \in \tilde{K}$.

Moreover, (11) is equivalent to $d_1(L(W_n), \mathcal{N}(0,1)) \to 0$. Now, consider any subsequence $(n_k)_{k\geqslant 1}$ of $\mathbb{N}$. Since $T_{n_k} \in K, D_p^\top G_{n_k} D_p \in \tilde{K}$ for all $k \geqslant 1$ and $K, \tilde{K}$ are compact, there exists a further sub-subsequence $(n'_k)_{k\geqslant 1}$ of $(n_k)_{k\geqslant 1}$ such that $T_{n'_k} \to \Sigma$ and $D_p^\top G_{n'_k} D_p \to D_p^\top G D_p$, for some $\Sigma \in \mathbb{S}_+^{p\times p}$ and absolutely symmetric $p^2 \times p^2$ matrix G. Since all eigenvalues of T_n are uniformly separated by Condition 2.3 B, all eigenvalues of Σ are distinct. Then $d_1(L(W_{n'_k}), \mathcal{N}(0,1)) \to 0$ holds along the subsequence $(n'_k)_{k\geqslant 1}$. Thus, for any subsequence $(n_k)_{k\geqslant 1}$, there exists a further sub-subsequence $(n'_k)_{k\geqslant 1}$ of $(n_k)_{k\geqslant 1}$ such that $d_1(L(W_{n'_k}), \mathcal{N}(0,1)) \to 0$. Therefore, $d_1(L(W_n), \mathcal{N}(0,1)) \to 0$, proving (11).

The proof of (12) is similar, with a few differences. We prove the result for χ_n first. By $c^\top \Delta_{n,i}c > \varepsilon$ for all $n \geqslant 1$, for every possible sub-subsequence limit Σ, $c^\top \Delta_i(\Sigma, G)c \geqslant \varepsilon > 0$. This shows that χ_n and $\tilde{\chi}_n$ are well defined for all sufficiently large $n \in \mathbb{N}$. As in the previous proof, $d_1(L(\chi_n), \mathcal{N}(0,1)) \to 0$. Since $(c^\top \hat{\Delta}_{m,n,i}c)^{-1}c^\top \Delta_{n,i}c \to_P 1$, we have $d_1(L(\tilde{\chi}_n), \mathcal{N}(0,1)) \to 0$ by Slutsky's theorem. This finishes the proof.

If Condition 2.8 `Sym-1'` holds, then $G_{(ii),(ii)} = 2 + \psi$ and $\Delta_{n,i} = \sum_{k\neq i} \Lambda_{n,i}\Lambda_{n,k}/(\Lambda_{n,i} - \Lambda_{n,k})^2 v_{n,k}v_{n,k}^\top$. Since $\Lambda_{n,i}\Lambda_{n,k}/(\Lambda_{n,i} - \Lambda_{n,k})^2 \geqslant \Lambda_{n,p}^2/\Lambda_{n,1}^2 > c_0^2$, for some $c_0 > 0$, we can conclude $c^\top \Delta_{n,i}c > c_0^2\varepsilon$ if $\sum_{k\neq i}(c^\top v_{n,k})^2 > \varepsilon$, which is equivalent to $(c^\top v_{n,i})^2 < 1 - \varepsilon$. Based on the proof above, (13) and (14) hold.

G.4 Proof of Corollary C.2

Recall from Theorem 2.9 that conditional on $(X_n)_{n\geqslant 1}$, (11) holds. This implies that we also have the unconditional result

$$\sqrt{\frac{m}{\tau_{m,n}(G_n)_{(ii),(ii)}}}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i}-\Lambda_{n,i}\right)\Rightarrow\mathcal{N}(0,1). \tag{69}$$

Moreover, from Waternaux (1976) on inference for one eigenvalue, if we define κ_4^i as the kurtosis of the marginal distribution of the i-th dimension, we have

$$\sqrt{\frac{n}{2+\kappa_4^i/\lambda_i^{*2}}}\lambda_i^{*-1}\left(\Lambda_{n,i}-\lambda_i^*\right)\Rightarrow\mathcal{N}(0,1). \tag{70}$$

From (69) and (70), we have $\Lambda_{n,i}\to_P\lambda_i^*$ and $\hat{\Lambda}_{m,i}^{-1}\to_P\lambda_i^*$. If $m/\left(n\tau_{m,n}(G_n)_{(ii),(ii)}\right)\to 0$, from Slutsky's theorem, we can conclude (46).

G.5 Proofs of Results in Section 3.1

G.5.1 Proof of Theorem 3.1

By Lemma 3.5, if $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, Condition 2.8 `Sym-1` holds with $G=I_{p^2}+P_p+Q_p$ and $\tau_{m,n}=1-\gamma_n$. Therefore, Condition 2.8 `Sym-1'` holds with $\psi=1$. Then since Condition 2.3 `B` is satisfied, we conclude that (2) and (17) hold due to (13) and (14) from Theorem 2.9, respectively.

G.5.2 Proof of Theorem 3.2

From Lemma 3.3, we conclude that for any $(b_n)_{n\geqslant 1}$ such that $b_n\in\mathbb{S}^{p-1}$ for all $n\geqslant 1$, $(U_nb_n,U_nb_n)_{n\geqslant 1}$ satisfies Condition 2.2. If (18) holds, $(U_nb_n,\bar{\varepsilon}_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}$. Thus, from Lemma 3.4 applied to $(a_n,\tilde{a}_n)_{n\geqslant 1}=(U_nb_n,\bar{\varepsilon}_n)_{n\geqslant 1}$, Condition 2.4 `1-dim'` holds for $(U_nb_n,\bar{\varepsilon}_n)_{n\geqslant 1}$ with $\psi=1$. Then by Theorem 2.5, we conclude the first result.

Similarly, if (19) holds, then from Lemma 3.4, it follows that $(U_nb_n,\overline{X_n\beta}_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}$. Then the second claim follows from Theorem B.2 by the same argument.

G.5.3 Proof of Lemma 3.3

Recall that the sketching matrix is defined as $S_{m,n}=\sqrt{n/m}BHD$, where $H=H_l$ is the Hadamard matrix. Denote $R_i:=(HDa_n)_i$ and $\tilde{R}_i:=(HD\tilde{a}_n)_i$ for $i\in[n]$. Recall $\gamma_n=m/n$ and $B=\mathrm{diag}(B_1,\ldots,B_n)$, where B_i are i.i.d. Bernoulli random variables with a success probability of γ_n. Let h_{kl}, $k,l\in[n]$ be the entries of H, and $D=\mathrm{diag}(D_1,\ldots,D_n)$. Then, letting $a_n=(a_{n,1},\ldots,a_{n,n})^\top$ and $\tilde{a}_n=(\tilde{a}_{n,1},\ldots,\tilde{a}_{n,n})^\top$, we have $R_i=\sum_{j\in[n]}h_{ij}D_ja_{n,j}$. Moreover, we can write $a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n=\sum_{i=1}^n R_i\tilde{R}_i\gamma_n^{-1}B_i$ and thus find that with

$$s_n^2:=\mathrm{Var}\left[a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n\mid D\right]=\frac{1-\gamma_n}{\gamma_n}\sum_{i=1}^n R_i^2\tilde{R}_i^2, \tag{71}$$

we have

$$\mathbb{E}\left[s_n^2\right] = \frac{1-\gamma_n}{\gamma_n}\mathbb{E}\sum_{i=1}^n \sum_{j_1,j_2,j_3,j_4=1}^n h_{ij_1}h_{ij_2}h_{ij_3}h_{ij_4}D_{j_1}D_{j_2}D_{j_3}D_{j_4}a_{n,j_1}a_{n,j_2}\tilde{a}_{n,j_3}\tilde{a}_{n,j_4}$$
$$= \frac{1-\gamma_n}{\gamma_n}n^{-1}\left(\sum_{j=1}^n a_{n,j}^2\tilde{a}_{n,j}^2 + \sum_{j_1\neq j_2} a_{n,j_1}^2\tilde{a}_{n,j_2}^2 + 2\sum_{j_1\neq j_2} a_{n,j_1}\tilde{a}_{n,j_1}a_{n,j_2}\tilde{a}_{n,j_2}\right)$$
$$= \frac{1-\gamma_n}{\gamma_n}n^{-1}\left(1 + 2\left(a_n^\top\tilde{a}_n\right)^2 - 2\sum_{j=1}^n a_{n,j}^2\tilde{a}_{n,j}^2\right),$$

where the first step uses that B_i are i.i.d. Bernoulli(γ_n) variables, the third step uses that D_i are i.i.d. variables equal to ± 1 with probability $1/2$, and the last step uses that a_n and $\tilde{a}_n$ are of unit norm. We also have

$$\mathrm{Var}\left[\mathbb{E}\left[a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n \mid D\right]\right] = \mathrm{Var}\left[\sum_{i=1}^n R_i\tilde{R}_i\right] = \mathrm{Var}\left[a_n^\top\tilde{a}_n\right] = 0.$$

Combining the above two equations, we have $\mathrm{Var}\left[a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n\right] = O\left((1-\gamma_n)/m\right)$, and Condition 2.2 follows.

G.5.4 Proof of Lemma 3.4

Proof. For all $i \in [n]$, recall the notations $R_i = (HDa_n)_i$ and $\tilde{R}_i = (HD\tilde{a}_n)_i$ from the proof of Lemma 3.3 in Appendix G.5.3. Recall that $a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n = \sum_{i=1}^n R_i\tilde{R}_i\gamma_n^{-1}B_i$, is a weighted sum of the n independent Bernoulli variables B_i, $i \in [n]$, given D. With s_n^2 from (71), we will show that for a small $\delta > 0$,

$$\frac{\sum_{i=1}^n |R_i\tilde{R}_i|^{2+\delta}\gamma_n^{-2-\delta}\mathbb{E}|B_i - \mathbb{E}B_i|^{2+\delta}}{s_n^{2+\delta}} = o_P(1), \tag{72}$$

and then, by the Lyapunov central limit theorem and Lemma G.2 below, the result will follow. To verify (72), we will show the following bound:

$$\mathbb{P}\left\{\max_{1\leqslant i\leqslant n}\{|R_i| \vee |\tilde{R}_i|\} \leqslant \sqrt{\frac{\log n}{n}}\right\} \geqslant 1 - \frac{8}{n}. \tag{73}$$

By writing $R_i = \sum_{j=1}^n h_{ij}D_ja_{n,j}$, we can apply Hoeffding's inequality to obtain $\mathbb{P}(|R_i| > t) \leqslant 2\exp\left(-2nt^2\right)$ for any $t > 0$. Then (73) holds by taking $t = n^{-1/2}\log^{1/2} n$ and further taking a union bound over $i \in [n]$. Define the event

$$\Xi := \left\{\max_{1\leqslant i\leqslant n}\{|R_i| \vee |\tilde{R}_i|\} \leqslant (n^{-1}\log n)^{1/2}\right\} \bigcap \left\{1/2 < n\sum_{i=1}^n R_i^2\tilde{R}_i^2 < 4\right\}.$$

By (73) and Lemma G.2 below, the event Ξ holds with probability tending to one. On Ξ, we have $\sum_{i=1}^n (R_i\tilde{R}_i)^{2+\delta} \leqslant 4(n^{-1}\log n)^\delta/n$ and $s_n^{2+\delta} > [2n\gamma_n/(1-\gamma_n)]^{-1-\delta/2}$. We also find $\mathbb{E}|B_i - \mathbb{E}B_i|^{2+\delta} = \gamma_n(1-\gamma_n)^{2+\delta} + \gamma_n^{2+\delta}(1-\gamma_n) \leqslant \gamma_n(1-\gamma_n)$. Combining these bounds, the left side of (72) is bounded by $C(m^{-1/2}\log n)^\delta/(1-\gamma_n)^{\delta/2}$ for some constant $C > 0$. Therefore,

since $m/\log^2 n \to \infty$ and $\limsup m/n < 1$, we conclude (72).

Lemma G.2. *We have*

$$n\sum_{i=1}^{n} R_i^2 \tilde{R}_i^2 - [1 + 2(a_n^\top \tilde{a}_n)^2] = o_P(1). \tag{74}$$

Proof. Below, all sums over indices such as $j_1, j_2, \cdots$ will range over the index set $[n]$, and we recall $H = (h_{ij})_{i,j\in[n]}$. We start by considering

$$\begin{aligned} n\sum_{i=1}^{n} R_i^2 \tilde{R}_i^2 &= n\sum_{i=1}^{n}\left(\sum_{j=1}^{n} h_{ij} D_j a_{n,j}\right)^2 \left(\sum_{j=1}^{n} h_{ij} D_j \tilde{a}_{n,j}\right)^2 \\ &= n\sum_{i=1}^{n}\left(\sum_{j=1}^{n} h_{ij}^2 a_{n,j}^2 + \sum_{j_1\neq j_2} h_{ij_1} h_{ij_2} D_{j_1} D_{j_2} a_{n,j_1} a_{n,j_2}\right) \times \\ &\qquad \left(\sum_{j=1}^{n} h_{ij}^2 \tilde{a}_{n,j}^2 + \sum_{j_1\neq j_2} h_{ij_1} h_{ij_2} D_{j_1} D_{j_2} \tilde{a}_{n,j_1} \tilde{a}_{n,j_2}\right). \end{aligned}$$

We can thus decompose $n\sum_{i=1}^{n} R_i^2 \tilde{R}_i^2$ into the sum of the following terms:

$$\begin{aligned} I_1 &= n\sum_{i=1}^{n}\sum_{j_1,j_2} h_{ij_1}^2 h_{ij_2}^2 a_{n,j_1}^2 \tilde{a}_{n,j_2}^2; \qquad I_2 = n\sum_{i=1}^{n}\left(\sum_{j_1\neq j_2} h_{ij_1} h_{ij_2} D_{j_1} D_{j_2} a_{n,j_1} a_{n,j_2}\right)\left(\sum_{j=1}^{n} \tilde{a}_{n,j}^2\right); \\ I_3 &= n\sum_{i=1}^{n}\left(\sum_{j_1\neq j_2} h_{ij_1} h_{ij_2} D_{j_1} D_{j_2} \tilde{a}_{n,j_1} \tilde{a}_{n,j_2}\right)\left(\sum_{j=1}^{n} a_{n,j}^2\right); \end{aligned}$$

and

$$\begin{aligned} I_4 &= n\sum_{i=1}^{n}\left(\sum_{j_1\neq j_2}\sum_{j_3\neq j_4} h_{ij_1} h_{ij_2} h_{ij_3} h_{ij_4} D_{j_1} D_{j_2} D_{j_3} D_{j_4} a_{n,j_1} a_{n,j_2} \tilde{a}_{n,j_3} \tilde{a}_{n,j_4}\right) \\ &= n\sum_{i=1}^{n}\left(\sum_{(j_1,j_2,j_3,j_4)} \star + 2\sum_{j_1=j_3, j_2=j_4, j_1\neq j_2} \star + 2\sum_{j_1=j_3, j_2\neq j_4} \star\right) =: I_{41} + I_{42} + I_{43}, \end{aligned}$$

where $\star$ stands for the term inside the sum on the first line, and (j_1, j_2, j_3, j_4) shows that these four indices must take different values.

To calculate I_1, we observe that $h_{ij} = \pm 1/\sqrt{n}$ for all i, j, so that $I_1 = \sum_{j_1,j_2} a_{n,j_1}^2 \tilde{a}_{n,j_2}^2 = 1$. To calculate I_2 and I_3, we change the order of summation and use that $\sum_{i=1}^{n} h_{ij_1} h_{ij_2} = 0$ when $j_1 \neq j_2$, to find that $I_2 = I_3 = 0$. Similarly, we find that $I_{43} = 0$.

Next, $I_{42} = 2\sum_{j_1\neq j_2} a_{n,j_1} a_{n,j_2} \tilde{a}_{n,j_1} \tilde{a}_{n,j_2}$, since $h_{ij}^2 = 1/n$ and $D_j^2 = 1$ for any $i, j \in [n]$. If $a_n \neq \pm\tilde{a}_n$, define $a_n' = (I - a_n a_n^\top)\tilde{a}_n / \|(I - a_n a_n^\top)\tilde{a}_n\|$, otherwise, define a_n' as an arbitrary unit length vector orthogonal to a_n, so that we have

$$\tilde{a}_n = a_n a_n^\top \cdot \tilde{a}_n + (I_n - a_n a_n^\top) \cdot \tilde{a}_n = (a_n^\top \tilde{a}_n) a_n + (1 - (a_n^\top \tilde{a}_n)^2)^{1/2} a_n'. \tag{75}$$

Moreover, denote $t_1 = a_n^\top \tilde{a}_n, t_2 = (1-(a_n^\top \tilde{a}_n)^2)^{1/2}$. Then, we can express I_{42} as

$$\begin{aligned} I_{42} &= 2\sum_{j_1\neq j_2} a_{n,j_1}a_{n,j_2}(t_1 a_{n,j_1} + t_2 a'_{n,j_1})(t_1 a_{n,j_2} + t_2 a'_{n,j_2}) \\ &= 2\sum_{j_1\neq j_2} t_1^2 a_{n,j_1}^2 a_{n,j_2}^2 + 2\sum_{j_1\neq j_2} t_1 t_2 a_{n,j_1}a_{n,j_2}(a_{n,j_1}a'_{n,j_2} + a'_{n,j_1}a_{n,j_2}) \\ &\qquad + 2\sum_{j_1\neq j_2} t_2^2 a_{n,j_1}a_{n,j_2}a'_{n,j_1}a'_{n,j_2} \\ &=: I_{42}^{(1)} + I_{42}^{(2)} + I_{42}^{(3)}. \end{aligned}$$

Due to Condition $\mathfrak{P}$ defined in (15), without loss of generality, we can assume that $\max_{1\leqslant i\leqslant n} |a_{n,i}| \to 0$. Hence, we have

$$I_{42}^{(1)} = 2t_1^2\left[\left(\sum_{j=1}^n a_{n,j}^2\right)^2 - \sum_{j=1}^n a_{n,j}^4\right] = 2t_1^2\left(1+o(1)\right);$$

and since a'_n is orthogonal to a_n,

$$I_{42}^{(3)} = 2t_2^2\left[\left(\sum_{j=1}^n a_{n,j}a'_{n,j}\right)^2 - \sum_{j=1}^n a_{n,j}^2(a'_{n,j})^2\right] = o(t_2^2).$$

Moreover, we find

$$\left|I_{42}^{(2)}\right| = \left|-4\sum_{j=1}^n t_1 t_2 a_{n,j}^3 a'_{n,j}\right| \leqslant 4t_1t_2 \max_{1\leqslant j\leqslant n}|a_{n,j}a'_{n,j}| \cdot \sum_{j=1}^n a_{n,j}^2 = o(t_1t_2),$$

Therefore, $I_{42} = 2(a_n^\top \tilde{a}_n)^2 + o(1)$.

To analyze I_{41}, we further introduce $t_{j_1,j_2,j_3,j_4} = \sum_{i=1}^n h_{ij_1}h_{ij_2}h_{ij_3}h_{ij_4}$ for any indices $j_1,j_2,j_3,j_4 \in [n]$, and express

$$I_{41} = n\sum_{(j_1,j_2,j_3,j_4)} t_{j_1,j_2,j_3,j_4} D_{j_1}D_{j_2}D_{j_3}D_{j_4} a_{n,j_1}a_{n,j_2}\tilde{a}_{n,j_3}\tilde{a}_{n,j_4}.$$

Then, we consider

$$\mathbb{E}I_{41}^2 = n^2\mathbb{E}\sum_{(j_1,\ldots,j_4),(j_5,\ldots,j_8)} t_{j_1,\ldots,j_4}t_{j_5,\ldots,j_8} D_{j_1}\ldots D_{j_8} a_{n,j_1}a_{n,j_2}\tilde{a}_{n,j_3}\tilde{a}_{n,j_4}a_{n,j_5}a_{n,j_6}\tilde{a}_{n,j_7}\tilde{a}_{n,j_8}.$$

Now, only terms such that $\{j_1,j_2,j_3,j_4\} = \{j_5,j_6,j_7,j_8\}$ have a non-zero expectation. Thus after decomposing $\tilde{a}_n$ into two orthogonal vectors as in handling I_{42} above, using Lemma G.3 and the assumption that $\max_{i\in[n]}|a_{n,i}| \to 0$, we conclude that the above expectation is $o(1)$. Thus $I_{41} = o_P(1)$ and we conclude (74) by summing up the above bounds.

□

Lemma G.3. *For any $j_1,j_2,j_3 \in [n]$ there is a unique j_4 such that $t_{j_1,j_2,j_3,j_4} \neq 0$, and then $t_{j_1,j_2,j_3,j_4} = 1/n$.*

Proof. With $l = \log_2 n \in \mathbb{N}$, we define the binary representation of $j_i - 1$ by $j_{i,l-1}2^{l-1} + \ldots + j_{i,2}2^2 + j_{i,1}2 + j_{i,0}1$, and define the binary representation $(\ell_{l-1}, \ldots, \ell_0)$ of $\ell - 1$ similarly. Then $h_{j_i,\ell} = n^{-1/2}(-1)^{\sum_{k=0}^{l-1} j_{i,k}\ell_k}$, and it follows that

$$t_{j_1,j_2,j_3,j_4} = \frac{1}{n^2}\sum_{\ell=1}^{n}(-1)^{\sum_{k=0}^{l-1}\sum_{i=1}^{4} j_{i,k}\ell_k}.$$

We claim the following two facts: First, if $\sum_{i=1}^{4} j_{i,k}$ is even for any $k \in [0:l] := \{0, 1, \ldots, l\}$, then $t_{j_1,j_2,j_3,j_4} = 1/n$. Second, if $\sum_{i=1}^{4} j_{i,k_0}$ is odd for some $k_0 \in [0:l]$, then $t_{j_1,j_2,j_3,j_4} = 0$. These two claims imply Lemma G.3.

Next, we prove the two claims. The first claim holds by noting that $\sum_{k=0}^{l-1}\sum_{i=1}^{4} j_{i,k}\ell_k$ is even for any $\ell \in [n]$. To conclude the second claim, we write

$$t_{j_1,j_2,j_3,j_4} = \frac{1}{n^2}\sum_{\ell=1}^{n}(-1)^{\ell_{k_0}} \cdot (-1)^{\sum_{k=0,k\neq k_0}^{l-1}\sum_{i=1}^{4} j_{i,k}\ell_k} =: \frac{1}{n^2}\sum_{\ell=1}^{n} g_\ell.$$

For any $\ell_1 \in [n]$, we can find a unique $\ell_2 \in [n]$ such that $\ell_{1,k_0} + \ell_{2,k_0} = 1$ and the other coefficients of their binary representations match. For such a pair of ℓ_1 and ℓ_2, we have

$$\begin{aligned} g_{\ell_1} + g_{\ell_2} &= (-1)^{\ell_{1,k_0}} \cdot (-1)^{\sum_{k=0,k\neq k_0}^{l-1}\sum_{i=1}^{4} j_{i,k}\ell_{1,k}} + (-1)^{\ell_{2,k_0}} \cdot (-1)^{\sum_{k=0,k\neq k_0}^{l-1}\sum_{i=1}^{4} j_{i,k}\ell_{2,k}} \\ &= \left[(-1)^{\ell_{1,k_0}} + (-1)^{\ell_{2,k_0}}\right] \cdot (-1)^{\sum_{k=0,k\neq k_0}^{l-1}\sum_{i=1}^{4} j_{i,k}\ell_{1,k}} = 0. \end{aligned}$$

Moreover, we can find $n/2$ such pairs that run through $[n]$, thus $t_{j_1,j_2,j_3,j_4} = n^{-2}\sum_{\ell=1}^{n} g_\ell = 0$. □

Finally we provide an example to show that (20) is generally not true if Condition $\mathfrak{P}$ define(15) is not satisfied. Let $n \geqslant 4$ and consider $a_n = (1/2, 1/2, -1/2, -1/2, 0, \ldots, 0)^\top$ and $\tilde{a}_n = (-1/2, 1/2, -1/2, 1/2, 0 \ldots 0)^\top$. For all $i \in [n]$, denote the i-th column of H by h_i. By the structure of the Hadamard matrix, for any diagonal matrix B, we have $h_1^\top B h_2 = h_3^\top B h_4$, $h_1^\top B h_3 = h_2^\top B h_4$, and $h_1^\top B h_4 = h_2^\top B h_3$. Additionally, we have $h_i^\top B h_i = \sum_{i=1}^{n} B_i/n$. Then, we can calculate:

$$\begin{aligned} a_n^\top DHBHD\tilde{a}_n &= \sum_{i=1}^{4} a_{n,i}\tilde{a}_{n,i} h_i^\top B h_i \\ &+ [(a_{n,1}\tilde{a}_{n,2} + a_{n,2}\tilde{a}_{n,1})D_1D_2 + (a_{n,3}\tilde{a}_{n,4} + a_{n,4}\tilde{a}_{n,3})D_3D_4]h_1^\top B h_2 \\ &+ [(a_{n,1}\tilde{a}_{n,3} + a_{n,3}\tilde{a}_{n,1})D_1D_3 + (a_{n,2}\tilde{a}_{n,4} + a_{n,4}\tilde{a}_{n,2})D_2D_4]h_1^\top B h_3 \\ &+ [(a_{n,1}\tilde{a}_{n,4} + a_{n,4}\tilde{a}_{n,1})D_1D_4 + (a_{n,2}\tilde{a}_{n,3} + a_{n,3}\tilde{a}_{n,2})D_2D_3)h_1^\top B h_4 \\ &= \frac{1}{2}(D_1D_4 - D_2D_3)h_1^\top B h_4. \end{aligned}$$

Since D_i are uniformly distributed over ± 1 for all $i \in [4]$, the above random variable places a non-vanishing point mass at zero, and is thus not asymptotically normal. □

G.5.5 Proof of Lemma 3.5

With the Cramer-Wold device, since $U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p$ is a symmetric matrix, it suffices to show that for any symmetric $\Phi \in \mathbb{R}^{p\times p}$,

$$\sqrt{m}\,\mathrm{tr}\left(\Phi(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p)\right) \Rightarrow \mathcal{N}\left(0, \mathrm{vec}(\Phi)^\top G\ \mathrm{vec}(\Phi)\right). \tag{76}$$

By the spectral decomposition $\Phi = \sum_{i=1}^p \mu_i w_i w_i^\top$ of Φ, where $(w_i)_{i\in[p]}$ forms an orthonormal basis of $\mathbb{R}^p$, we have

$$\mathrm{tr}\left(\Phi(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p)\right) = \sum_{i=1}^p \mu_i (w_i^\top U_n^\top S_{m,n}^\top S_{m,n} U_n w_i - 1).$$

Now $U_n w_i$, $i \in [p]$ constitutes an orthogonal basis of the p-dimensional linear subspace of $\mathbb{R}^n$ given by the span of the columns of U_n. Since $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, we have $\|U_n w_i\|_\infty \leqslant \max_{i=1,\dots,n} \|U_{i:}\|\|w_i\| \to 0$ for every $i \in [p]$, due to Condition $\mathfrak{P}'$ from (16). We now denote $R_i^k := (HDU_n w_k)_i$ for $i \in [n], k \in [p]$, so that

$$\sum_{i=1}^p \mu_i w_i^\top U_n^\top S_{m,n}^\top S_{m,n} U_n w_i = \sum_{j=1}^n \left(\sum_{i=1}^p \mu_i \cdot \left(R_j^i\right)^2\right) \gamma_n^{-1} B_j.$$

As in the proof of Lemma 3.4, conditional on D, this is a weighted sum of n independent Bernoulli variables, with D-conditional variance $s_n^2 = \sum_{j=1}^n \frac{1-\gamma_n}{\gamma_n}\left(\sum_{i=1}^p \mu_i \cdot \left(R_j^i\right)^2\right)^2$. We write $n\sum_{j=1}^n \left(\sum_{i=1}^p \mu_i \cdot \left(R_j^i\right)^2\right)^2 = n\sum_{i_1,i_2}\sum_{j=1}^n \mu_{i_1}\mu_{i_2} \cdot \left(R_j^{i_1}\right)^2 \left(R_j^{i_2}\right)^2$. Since $(w_j)_{j\in[p]}$ are orthonormal, according to (74) that for $i_1 \neq i_2 \in [p]$, it holds that $\sum_{j=1}^n (R_j^{i_1})^2 (R_j^{i_2})^2 \to_P 1$, and for $i_1 = i_2 \in [p]$, it holds that $\sum_{j=1}^n (R_j^{i_1})^2 (R_j^{i_2})^2 \to_P 3$. Therefore,

$$n\sum_{j=1}^n \left(\sum_{i=1}^p \mu_i \cdot \left(R_j^i\right)^2\right)^2 - \left(2\|\mu\|^2 + \left(\sum_{i=1}^p \mu_i\right)^2\right) = o_P(1),$$

Now, since μ_i are the eigenvalues of Φ, $2\|\mu\|^2 = 2\|\Phi\|_F^2 = \mathrm{vec}(\Phi)^\top (I_p + P_p)\mathrm{vec}(\Phi)$. Also,

$$\mathrm{vec}(\Phi)^\top Q_p \mathrm{vec}(\Phi) = \sum_{(ij),(kl)} \Phi_{ij}\Phi_{kl}\delta_{ij}\delta_{kl} = \sum_{i,k} \Phi_{ii}\Phi_{kk} = (\mathrm{tr}\,\Phi)^2 = \left(\sum_{i=1}^p \mu_i\right)^2.$$

Thus, $ms_n^2/(1-\gamma_n)$ tends to $\mathrm{vec}(\Phi)^\top (I_{p^2} + P_p + Q_p)\mathrm{vec}(\Phi)$.

Moreover, we can verify similarly to the proof of (72) that for a small $\delta > 0$

$$\sum_{j=1}^n \left(\sum_{i=1}^p \mu_i \cdot \left(R_j^i\right)\right)^{2+\delta} \gamma_n^{-2-\delta} \mathbb{E}|B_i - \mathbb{E}B_i|^{2+\delta} s_n^{-(2+\delta)} = o_P(1),$$

assuming that $m/\log^2 n \to \infty$. This can be verified by Hoeffding's inequality similar to how it is used above in the proof for (72). Thus, we conclude (76).

G.6 Proofs of Results in Section 3.2

G.6.1 Proof of Theorem 3.6

From Lemma 3.9, if $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, Condition 2.8 `Sym-1` holds with $G = I_{p^2} + P_p$ and $\tau_{m,n} = 1$. Therefore, Condition 2.8 `Sym-1'` holds with $\psi = 0$. Then, since Condition 2.3 `B` holds by assumption, we conclude the desired results based on Theorem 2.9.

G.6.2 Proof of Theorem 3.7

Due to (23), we conclude that for any $b_n \in \mathbb{S}^{p-1}$, $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$ and $(U_n b_n, U_n b_n)_{n\geqslant 1}$ satisfy Condition $\mathfrak{P}$. Thus, from Lemma 3.8 Condition 2.4 `1-dim'` holds for $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$ with $\psi = 0$. Then by Theorem 2.5, we conclude the first result. Similarly, if (24) holds, $(U_n b_n, \overline{X_n \beta_n})_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}$. Then the second claim follows from Theorem B.2 together with Lemma 3.8.

G.6.3 Proof of Lemma 3.8

For all $i \in [n]$, we denote the i-th column of $S_{m,n}$ as $R_{m,i}$. Then, we have $R_{m,i} =_d \bar{D}_i G_{m,i}$, where $\bar{D}_i = \mathrm{diag}(D_{1,i}, \ldots, D_{m,i})$, $D_{m,i}$ are i.i.d. Rademacher random variables, and $G_{m,i}$ are uniformly distributed on the set $\{f/\sqrt{\zeta_m} \in \mathbb{R}^m : f_1, \ldots, f_m \in \{0,1\} \text{ and } \|f\|^2 = \zeta_m\}$, independently across i. We can express $a_n^\top S_{m,n}^\top S_{m,n} \tilde{a}_n = \sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}$
$+ \sum_{i\neq j} a_{n,i}\tilde{a}_{n,j} R_{m,i}^\top R_{m,j}$.

Defining $h_n(x,y) = x^\top y$ for $x, y \in \mathbb{R}^n$, we have

$$\sum_{i\neq j} a_{n,i}\tilde{a}_{n,j} R_{m,i}^\top R_{m,j} = \sum_{i<j} (a_{n,i}\tilde{a}_{n,j} + a_{n,j}\tilde{a}_{n,i}) h_n(R_{m,i}, R_{m,j}).$$

The variables $T_i := R_{m,i}$ are i.i.d. across $i \in [n]$, making the sum a weighted U-statistic of order two. Moreover, for $i \neq j$, $\mathbb{E}[h_n(T_i, T_j)|T_i] = (R_{m,i})^\top \mathbb{E}[R_{m,j}] = 0$ and $\mathbb{E}[h_n(T_i, T_j)|T_j] = 0$. This shows that the U-statistic is degenerate.

We use the martingale central limit theorem, see e.g., Corollary 3.1 from Hall and Heyde (2014), to derive our result. In particular, we focus on analyzing the sum $\sum_{i<j} w_{ij} h_n(T_i, T_j)$, where $w_{ij} = a_{n,i}\tilde{a}_{n,j} + a_{n,j}\tilde{a}_{n,i}$ for all i, j. To analyze this sum, we construct the sequence

$$\Upsilon_k = \sum_{j=1}^{k-1} w_{jk} h_n(T_j, T_k),\ k \in [n]. \tag{77}$$

Due to the independence between T_j and T_k for all $j \neq k$, we have $\mathbb{E}[\Upsilon_k | T_1, \ldots, T_{k-1}] = 0$ for all $k \in [n]$. Consequently, the sequence $\{S_i = \sum_{l=2}^i \Upsilon_l, 2 \leqslant i \leqslant n\}$ forms a martingale with respect to the σ-field generated by $T_1, \ldots, T_i$, $i \in [n]$. Moreover, we can write $S_n = \sum_{i<j} w_{ij} h_n(T_i, T_j)$.

To show that $S_n/\sqrt{\mathrm{Var}\,[S_n]}$ tends to a standard normal distribution, it is enough to show the Lyapunov condition $s_n^{-4} \sum_{i=2}^n \mathbb{E}\Upsilon_i^4 \to 0$. Additionally, we need to show that $s_n^{-2} q_n^2 \to_P 1$, where $q_n^2 := \sum_{i=2}^n \mathbb{E}[\Upsilon_i^2 | T_1, \ldots, T_{i-1}]$.

Proof of the Lyapunov condition: We show the Lyapunov condition $s_n^{-4} \sum_{i=2}^n \mathbb{E}\Upsilon_i^4 \to 0$

first. Using the properties of square-integrable martingales, we have $s_n^2 = \sum_{i=2}^n \mathbb{E}\Upsilon_i^2$, and

$$\begin{aligned}\mathbb{E}\Upsilon_k^2 &= \mathbb{E}\left(\sum_{j=1}^{k-1} w_{jk} h_n(T_j, T_k)\right)^2 = \sum_{j=1}^{k-1}\sum_{i=1}^{k-1} w_{ik} w_{jk} \mathbb{E}[h_n(T_j, T_k) h_n(T_i, T_k)] \\ &= \sum_{j=1}^{k-1}\sum_{i=1}^{k-1} w_{ik} w_{jk} \mathbb{E}[R_{m,j}^\top R_{m,k} R_{m,i}^\top R_{m,k}] = \sum_{i=1}^{k-1} w_{ik}^2 \mathbb{E}[(R_{m,i}^\top R_{m,k})^2].\end{aligned}$$

Here, for $i \neq j$, we use the independence of $R_{m,i}$, $R_{m,j}$ and $R_{m,k}$. Also, since $D_{m,i}$ are i.i.d. Rademacher random variables, and in particular $D_{k,i}^2 = 1$ for all k, i

$$\begin{aligned}&\mathbb{E}[(R_{m,i}^\top R_{m,k})^2] = \mathbb{E}[(G_{m,i}^\top \bar{D}_i \bar{D}_k G_{m,k})^2] = \mathbb{E}[G_{m,i}^\top \bar{D}_i \bar{D}_k G_{m,k} G_{m,k}^\top \bar{D}_i \bar{D}_k G_{m,i}] \\ &= \mathbb{E}[G_{m,i}^\top \operatorname{diag}(G_{m,k} \odot G_{m,k}) G_{m,i}].\end{aligned}$$

Since $G_{m,i}$, $G_{m,j}$ are independently and uniformly distributed on the set $\{f/\sqrt{\zeta_m} \in \mathbb{R}^m : f_1, \ldots, f_m \in \{0,1\} \text{ and } \|f\|^2 = \zeta_m\}$, every component of $G_{m,i}$ equals to zero or $1/\sqrt{\zeta_m}$, and hence $G_{m,i} \odot G_{m,i} = G_{m,i}/\sqrt{\zeta_m}$ for all i. Further, $\mathbb{E}[G_{m,k}] = \sqrt{\zeta_m} 1_m/m$, and we have for $i \neq k$ that

$$\begin{aligned}\mathbb{E}[G_{m,i}^\top \operatorname{diag}(G_{m,k} \odot G_{m,k}) G_{m,i}] &= \mathbb{E}[(G_{m,i} \odot G_{m,i})^\top (G_{m,k} \odot G_{m,k})] \\ &= \mathbb{E}[G_{m,i}^\top]\mathbb{E}[G_{m,k}]/\zeta_m = 1/m.\end{aligned}$$

Therefore,

$$\begin{aligned}s_n^2 &= \sum_{k=2}^n \mathbb{E}\Upsilon_k^2 = \sum_{k=2}^n \frac{1}{m}\sum_{i=1}^{k-1} w_{ik}^2 = \frac{1}{m}\sum_{i<k} w_{ik}^2 = \frac{1}{m}\sum_{i<k}(a_{n,i}\tilde{a}_{n,k} + a_{n,k}\tilde{a}_{n,i})^2 \\ &= \frac{1}{m}\sum_{i<k}\left(a_{n,i}^2\tilde{a}_{n,k}^2 + a_{n,k}^2\tilde{a}_{n,i}^2 + 2a_{n,i}\tilde{a}_{n,i}a_{n,k}\tilde{a}_{n,k}\right) = \frac{1}{m}\sum_{i\neq k}\left(a_{n,i}^2\tilde{a}_{n,k}^2 + a_{n,i}\tilde{a}_{n,i}a_{n,k}\tilde{a}_{n,k}\right) \\ &= \frac{1}{m}\left(\left(\sum_{i=1}^n a_{n,i}^2\right)\left(\sum_{i=1}^n \tilde{a}_{n,i}^2\right) - \sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2 + \left(\sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}\right)^2 - \sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2\right) \\ &= \frac{1}{m}\left(1 + (a_n^\top \tilde{a}_n)^2 - 2\sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2\right). \qquad (78)\end{aligned}$$

Additionally, $\mathbb{E}\Upsilon_k^4$ equals

$$\begin{aligned}&\mathbb{E}\left[\sum_{j=1}^{k-1} w_{jk} h_n(T_j, T_k)\right]^4 \\ &= \sum_{i_1,i_2,i_3,i_4=1}^{k-1} w_{i_1k} w_{i_2k} w_{i_3k} w_{i_4k} \mathbb{E}\left[h_n(T_{i_1}, T_k) h_n(T_{i_2}, T_k) h_n(T_{i_3}, T_k) h_n(T_{i_4}, T_k)\right] \\ &= \sum_{i_1,i_2,i_3,i_4=1}^{k-1} w_{i_1k} w_{i_2k} w_{i_3k} w_{i_4k} \mathbb{E}\left[R_{m,i_1}^\top R_{m,k} R_{m,i_2}^\top R_{m,k} R_{m,i_3}^\top R_{m,k} R_{m,i_4}^\top R_{m,k}\right].\end{aligned}$$

Due to independence and as $\mathbb{E}[R_{m,i}] = 0$ for all $i \in [k-1]$, if an index occurs only once in

(i_1, i_2, i_3, i_4), the corresponding term is zero. Therefore, the sum becomes

$$\sum_{i=1}^{k-1} w_{ik}^4 \mathbb{E}\left[\left(R_{m,i}^\top R_{m,k}\right)^4\right] + 3\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(1-\delta_{ij}) w_{ik}^2 w_{jk}^2 \mathbb{E}\left[\left(R_{m,i}^\top R_{m,k}\right)^2 \left(R_{m,j}^\top R_{m,k}\right)^2\right].$$

Define $M_{\zeta_m} \in \mathbb{R}^{m\times m}$, with $(M_{\zeta_m})_{ii} = 1$ for $i \in [m]$ and $(M_{\zeta_m})_{ij} = (\zeta_m - 1)/(m-1)$ for $i \neq j, i, j \in [m]$. Then one can verify that for all k, $\mathbb{E}[G_{m,k} G_{m,k}^\top] = M_{\zeta_m}/m$. Therefore, since $D_{k,i}^2 = 1$ for all k, i, and $G_{m,i} \odot G_{m,i} = G_{m,i}/\sqrt{\zeta_m}$ for all i we have for all distinct i, j, k that

$$\begin{aligned}
&\mathbb{E}\left[(R_{m,i}^\top R_{m,k})^2 (R_{m,j}^\top R_{m,k})^2\right] = \mathbb{E}\left[(G_{m,i}^\top \bar{D}_i \bar{D}_k G_{m,k})^2 (G_{m,j}^\top \bar{D}_j \bar{D}_k G_{m,k})^2\right] \\
&= \frac{1}{\zeta_m^2}\mathbb{E}\left[G_{m,i}^\top G_{m,k} G_{m,k}^\top G_{m,j}\right] = \frac{1}{\zeta_m^2}\mathbb{E}\left[G_{m,i}\right]^\top \mathbb{E}\left[G_{m,k} G_{m,k}^\top\right] \mathbb{E}\left[G_{m,j}\right] \\
&= \frac{1}{\zeta_m^3}\left(\frac{\zeta_m}{m}\right)^2 \frac{1_m^\top M_{\zeta_m} 1_m}{m} = \frac{1}{m^2}.
\end{aligned}$$

In addition, for all $i \neq k$, again since $G_{m,i} \odot G_{m,i} = G_{m,i}/\sqrt{\zeta_m}$ for all i

$$\begin{aligned}
&\mathbb{E}\left[\left(R_{m,i}^\top R_{m,k}\right)^4\right] \\
&= \mathbb{E}\left[\left(G_{m,i}^\top \bar{D}_i \bar{D}_k G_{m,k}\right)^4\right] = \mathbb{E}\left[\left(\sum_{l=1}^m D_{l,i} D_{l,k} \left(G_{m,i}\right)_l \left(G_{m,k}\right)_l\right)^4\right] \\
&= \mathbb{E}\left[\sum_{l=1}^m \left(G_{m,i}\right)_l^4 \left(G_{m,k}\right)_l^4\right] + 6\mathbb{E}\left[\sum_{l_1<l_2} \left(G_{m,i}\right)_{l_1}^2 \left(G_{m,k}\right)_{l_1}^2 \left(G_{m,i}\right)_{l_2}^2 \left(G_{m,k}\right)_{l_2}^2\right] \\
&= \frac{1}{\zeta_m^3}\mathbb{E}\left[\sum_{l=1}^m \left(G_{m,i}\right)_l \left(G_{m,k}\right)_l\right] + \frac{3}{\zeta_m^2}\mathbb{E}\left[\sum_{l_1\neq l_2} \left(G_{m,i}\right)_{l_1} \left(G_{m,k}\right)_{l_1} \left(G_{m,i}\right)_{l_2} \left(G_{m,k}\right)_{l_2}\right] \\
&= \frac{1}{\zeta_m^3}\mathbb{E}\left[G_{m,i}^\top G_{m,k}\right] + \frac{3}{\zeta_m^2}\mathbb{E}\left[\left(\sum_{l=1}^m \left(G_{m,i}\right)_l \left(G_{m,k}\right)_l\right)^2 - \sum_{l=1}^m \left(G_{m,i}\right)_l^2 \left(G_{m,k}\right)_l^2\right].
\end{aligned}$$

This further equals

$$\begin{aligned}
&\frac{1}{\zeta_m^3}\mathbb{E}\left[G_{m,i}^\top G_{m,k}\right] + \frac{3}{\zeta_m^2}\mathbb{E}\left[\left(\sum_{l=1}^m \left(G_{m,i}\right)_l \left(G_{m,k}\right)_l\right)^2\right] - \frac{3}{\zeta_m^3}\mathbb{E}\left[\sum_{l=1}^m (G_{m,i})_l (G_{m,k})_l\right] \\
&= -\frac{2}{m\zeta_m^2} + \frac{3}{\zeta_m^2}\mathbb{E}\left[\mathbb{E}\left[G_{m,i}^\top G_{m,k} G_{m,k}^\top G_{m,i} | G_{m,i}\right]\right] \\
&= -\frac{2}{m\zeta_m^2} + \frac{3}{\zeta_m^2}\mathbb{E}\left[\sum_{l=1}^m \left(G_{m,i}\right)_l^2 + \frac{\zeta_m - 1}{m-1}\sum_{l_1\neq l_2} \left(G_{m,i}\right)_{l_1} \left(G_{m,i}\right)_{l_2}\right] \\
&= -\frac{2}{m\zeta_m^2} + \frac{3}{m\zeta_m^2}\left(\frac{m}{\zeta_m}\cdot\frac{\zeta_m}{m} + \frac{\zeta_m - 1}{m-1}\cdot\frac{m(m-1)}{\zeta_m}\cdot\frac{\zeta_m(\zeta_m-1)}{m(m-1)}\right) = \frac{m - 1 + 3(\zeta_m - 1)^2}{m(m-1)\zeta_m^2}.
\end{aligned}$$

Hence, we can express $\mathbb{E}\Upsilon_k^4$ as

$$\mathbb{E}\Upsilon_k^4 = \frac{m - 1 + 3(\zeta_m - 1)^2}{m(m-1)\zeta_m^2}\sum_{i=1}^{k-1} w_{ik}^4 + \frac{3}{m^2}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(1-\delta_{ij}) w_{ik}^2 w_{jk}^2.$$

Therefore,

$$\frac{\sum_{k=2}^n \mathbb{E}\Upsilon_k^4}{s_n^4} = \frac{\frac{m(m-1+3(\zeta_m-1)^2)}{(m-1)\zeta_m^2}\sum_{i<k} w_{ik}^4 + 3\sum_{k=2}^n \sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(1-\delta_{ij})w_{ik}^2 w_{jk}^2}{\left(1+(a_n^\top \tilde{a}_n)^2 - 2\sum_{i=1}^n a_{n,i}^2 \tilde{a}_{n,i}^2\right)^2}.$$

Thus, we only need to show that

$$\frac{m(m-1+3(\zeta_m-1)^2)}{(m-1)\zeta_m^2}\sum_{i<k} w_{ik}^4 \to 0, \quad \text{and} \quad \sum_{k=2}^n \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} w_{ik}^2 w_{jk}^2 \to 0, \tag{79}$$

where recall that $w_{ij} = a_{n,i}\tilde{a}_{n,j} + a_{n,j}\tilde{a}_{n,i}$ for all i, j. Using the Cauchy-Schwarz inequality, we find $w_{ij}^2 = (a_{n,i}\tilde{a}_{n,j} + a_{n,j}\tilde{a}_{n,i})^2 \leqslant (a_{n,i}^2 + \tilde{a}_{n,i}^2)(a_{n,j}^2 + \tilde{a}_{n,j}^2)$. Consequently,

$$\sum_{i<k} w_{ik}^4 \leqslant \sum_{i,k}(a_{n,i}^2 + \tilde{a}_{n,i}^2)^2(a_{n,k}^2 + \tilde{a}_{n,k}^2)^2 = \left(\sum_{i=1}^n (a_{n,i}^2 + \tilde{a}_{n,i}^2)^2\right)^2 \leqslant \left(2\sum_{i=1}^n (a_{n,i}^4 + \tilde{a}_{n,i}^4)\right)^2.$$

Due to Condition $\mathfrak{P}$ from (21) and $\zeta_m^2/m \to 0$, we find $\sqrt{(m-1+3(\zeta_m-1)^2)/\zeta_m^2}\sum_{i=1}^n a_{n,i}^4 \to 0$ and $\sqrt{(m-1+3(\zeta_m-1)^2)/\zeta_m^2}\cdot\sum_{i=1}^n \tilde{a}_{n,i}^4 \to 0$, thus

$$m\left(m-1+3(\zeta_m-1)^2\right)/\left((m-1)\zeta_m^2\right)\cdot\sum_{i<k} w_{ik}^4 \to 0.$$

Similarly, since $\zeta_m^2/m \to 0$, $\sum_{k=1}^n (a_{n,i}^4 + \tilde{a}_{n,i}^4) = o(1)$. Thus,

$$\begin{aligned}\sum_{k=2}^n \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} w_{ik}^2 w_{jk}^2 &= \sum_{k=2}^n \left(\sum_{i=1}^{k-1} w_{ik}^2\right)^2 \leqslant \sum_{k=1}^n \left(\sum_{i=1}^{k-1}(a_{n,i}^2 + \tilde{a}_{n,i}^2)(a_{n,k}^2 + \tilde{a}_{n,k}^2)\right)^2 \\ &\leqslant \sum_{k=1}^n \left(\sum_{i=1}^n (a_{n,i}^2 + \tilde{a}_{n,i}^2)(a_{n,k}^2 + \tilde{a}_{n,k}^2)\right)^2 = 4\sum_{k=1}^n (a_{n,k}^2 + \tilde{a}_{n,k}^2)^2 \leqslant 8\sum_{k=1}^n (a_{n,i}^4 + \tilde{a}_{n,i}^4) \to 0.\end{aligned}$$

In the last equation, we have used that for all i, $\sum_{i=1}^n a_{n,i}^2 = \sum_{i=1}^n \tilde{a}_{n,i}^2 = 1$.

Additionally, $2\sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2 \leqslant \sum_{i=1}^n (a_{n,i}^4 + \tilde{a}_{n,i}^4) \to 0$. In summary, we have shown that $\sum_{k=2}^n \mathbb{E}\Upsilon_k^4/s_n^4 \to 0$, which verifies the Lyapunov condition.

Proof of ratio-consistency of variance: Next, recalling $q_n^2 = \sum_{k=2}^n \mathbb{E}[\Upsilon_k^2|T_1,\ldots,T_{k-1}]$, and s_n^2 from (78), we aim to prove that $s_n^{-2}q_n^2 \to_P 1$. For this, we calculate, for any $k \in \{2,\ldots,n\}$

$$\begin{aligned}\mathbb{E}[\Upsilon_k^2|T_1,\ldots,T_{k-1}] &= \sum_{j=1}^{k-1}\sum_{i=1}^{k-1} w_{ik}w_{jk}\mathbb{E}[h_n(T_j,T_k)h_n(T_i,T_k)|T_1,\ldots,T_{k-1}] \\ &= \sum_{j=1}^{k-1}\sum_{i=1}^{k-1} w_{ik}w_{jk}R_{m,j}^\top \mathbb{E}[R_{m,k}R_{m,k}^\top]R_{m,i} = \sum_{j=1}^{k-1}\sum_{i=1}^{k-1} w_{ik}w_{jk}R_{m,j}^\top M_{\zeta_m}R_{m,i}/m.\end{aligned}$$

Therefore, we have

$$
\begin{aligned}
q_n^2 &= \frac{1}{m}\sum_{k=1}^{n}\sum_{j=1}^{k-1}\sum_{i=1}^{k-1} w_{ik}w_{jk}R_{m,j}^\top M_{\zeta_m}R_{m,i} \\
&= \frac{1}{m}\sum_{k=1}^{n}\left(\sum_{i=1}^{k-1} w_{ik}^2 + \sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(1-\delta_{ij})w_{ik}w_{jk}R_{m,j}^\top M_{\zeta_m}R_{m,i}\right).
\end{aligned}
$$

and thus

$$
q_n^2 - s_n^2 = \frac{1}{m}\sum_{k=1}^{n}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(1-\delta_{ij})w_{ik}w_{jk}R_{m,j}^\top M_{\zeta_m}R_{m,i}.
$$

As a result,

$$
\begin{aligned}
\mathbb{E}[(q_n^2 - s_n^2)^2] &= \frac{1}{m^2}\mathbb{E}\left[\left(\sum_{k=1}^{n}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(1-\delta_{ij})w_{ik}w_{jk}R_{m,j}^\top M_{\zeta_m}R_{m,i}\right)^2\right] \qquad (80) \\
&= \frac{1}{m^2}\sum_{i_1\neq j_1<k_1}\sum_{i_2\neq j_2<k_2} w_{i_1k_1}w_{j_1k_1}w_{i_2k_2}w_{j_2k_2}\mathbb{E}\left[R_{m,i_1}^\top M_{\zeta_m}R_{m,j_1}R_{m,i_2}^\top M_{\zeta_m}R_{m,j_2}\right],
\end{aligned}
$$

where $\sum_{i_1\neq j_1<k_1}\star$ is an abbreviation for $\sum_{k_1=1}^{n}\sum_{i_1=1}^{k_1-1}\sum_{j_1=1}^{k_1-1}(1-\delta_{i_1j_1})\star$. From the computation of the fourth moment of Υ_k, we know that the terms above are nonzero only when either $i_2 = i_1, j_2 = j_1$ or $i_2 = j_1, j_2 = i_1$. Due to the symmetry of the inner product, $R_{m,i_2}^\top M_{\zeta_m}R_{m,j_2} = R_{m,i_1}^\top M_{\zeta_m}R_{m,j_1}$ in both cases. Moreover, based on our previous results, for $i_1 \neq j_1$,

$$
\begin{aligned}
\mathbb{E}\left[\left(R_{m,i_1}^\top M_{\zeta_m}R_{m,j_1}\right)^2\right] &= \mathbb{E}\left[R_{m,i_1}^\top M_{\zeta_m}\mathbb{E}[R_{m,j_1}R_{m,j_1}^\top]M_{\zeta_m}R_{m,i_1}\right] \\
&= m^{-1}\mathbb{E}\left[\mathrm{tr}(M_{\zeta_m}^2 R_{m,i_1}R_{m,i_1}^\top)\right] = 1/m^2\cdot\mathrm{tr}(M_{\zeta_m}^2) \\
&= 1/m\cdot\left(1+(\zeta_m-1)^2/(m-1)\right).
\end{aligned}
$$

This yields that (80) equals

$$
\frac{1+(\zeta_m-1)^2/(m-1)}{m^3}\sum_{k_1=2}^{n}\sum_{k_2=2}^{n}\sum_{i_1=1}^{k_1\wedge k_2-1}\sum_{j_1=1}^{k_1\wedge k_2-1}(1-\delta_{i_1j_1})w_{i_1k_1}w_{j_1k_1}w_{i_1k_2}w_{j_1k_2}. \qquad (81)
$$

Now, recalling that $w_{ij} = a_{n,i}\tilde{a}_{n,j} + a_{n,j}\tilde{a}_{n,i}$ for all i, j, and applying the Cauchy-Schwarz inequality, we have

$$
|w_{i_1k_1}w_{j_1k_1}w_{i_1k_2}w_{j_1k_2}| \leqslant \left(a_{n,i_1}^2+\tilde{a}_{n,i_1}^2\right)\left(a_{n,j_1}^2+\tilde{a}_{n,j_1}^2\right)\left(a_{n,k_1}^2+\tilde{a}_{n,k_1}^2\right)\left(a_{n,k_2}^2+\tilde{a}_{n,k_2}^2\right).
$$

Thus, using that for all i, $\sum_{i=1}^n a_{n,i}^2 = \sum_{i=1}^n \tilde{a}_{n,i}^2 = 1$, (81) can be bounded by

$$\begin{aligned}
&\frac{1+\frac{(\zeta_m-1)^2}{m-1}}{m^3}\sum_{k_1=2}^n\sum_{k_2=2}^n\sum_{i_1,j_1=1}^{k_1\wedge k_2-1}\left(a_{n,i_1}^2+\tilde{a}_{n,i_1}^2\right)\left(a_{n,j_1}^2+\tilde{a}_{n,j_1}^2\right)\left(a_{n,k_1}^2+\tilde{a}_{n,k_1}^2\right)\left(a_{n,k_2}^2+\tilde{a}_{n,k_2}^2\right)\\
&\leqslant \frac{1+\frac{(\zeta_m-1)^2}{m-1}}{m^3}\sum_{i_1,j_1,k_1,k_2}\left(a_{n,i_1}^2+\tilde{a}_{n,i_1}^2\right)\left(a_{n,j_1}^2+\tilde{a}_{n,j_1}^2\right)\left(a_{n,k_1}^2+\tilde{a}_{n,k_1}^2\right)\left(a_{n,k_2}^2+\tilde{a}_{n,k_2}^2\right)\\
&= \frac{16\left(1+\frac{(\zeta_m-1)^2}{m-1}\right)}{m}.
\end{aligned}$$

Therefore, since $\zeta_m^2/m \to 0$, we have established that

$$s_n^{-4}\mathbb{E}[(q_n^2-s_n^2)^2] \leqslant \frac{16/m^3\cdot(1+(\zeta_m-1)^2/(m-1))}{\left(1+(a_n^\top\tilde{a}_n)^2-2\sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2\right)^2/m^2} \to 0.$$

This shows $s_n^{-2}q_n^2 \xrightarrow{\ell_2} 1$, and thus $s_n^{-2}q_n^2 \to_P 1$. Since $s_n^{-1}S_n \Rightarrow \mathcal{N}(0,1)$ and $ms_n^2/\left(1+(a_n^\top\tilde{a}_n)^2\right) \to_P 1$, (25) follows.

G.6.4 Proof of Lemma 3.9

We will use a martingale central limit theorem argument, as in the proof of Lemma 3.8. By the Cramer-Wold device as shown in (76), we can focus on

$$\operatorname{tr}\left(\Phi(U_n^\top S_{m,n}^\top S_{m,n}U_n - I_p)\right) = \sum_{i=1}^p \mu_i(w_i^\top U_n^\top S_{m,n}^\top S_{m,n}U_n w_i - 1).$$

We denote $t_i^k := (U_n w_i)_k$ for $i\in[p], i\in[n]$. Then, we can rewrite the above as

$$\sum_{i=1}^p \mu_i\left(\sum_{k\neq j} t_k^i t_j^i R_{m,k}^\top R_{m,j}\right) = \sum_{k<j} 2\left(\sum_{i=1}^p \mu_i t_k^i t_j^i\right) R_{m,k}^\top R_{m,j}.$$

As in (77), this is a weighted degenerate U-statistic, where $w_{kj} = 2\left(\sum_{i=1}^p \mu_i t_k^i t_j^i\right)$ for all k,j. Using notations from Lemma 3.8, as well as $t^k = (t_i^k)_{i\in[n]}$ for $k\in[p]$, we find

$$\begin{aligned}
ms_n^2 &= \sum_{i<j} w_{ij}^2 = 4\sum_{i<j}\left(\sum_{k=1}^p \mu_k t_i^k t_j^k\right)^2 = 4\sum_{k_1,k_2}\mu_{k_1}\mu_{k_2}\sum_{i<j} t_i^{k_1}t_j^{k_1}t_i^{k_2}t_j^{k_2}\\
&= 2\sum_{k_1,k_2}\mu_{k_1}\mu_{k_2}\left(\left(t^{k_1,\top}t^{k_2}\right)^2 - \sum_{i=1}^n\left(t_i^{k_1}\right)^2\left(t_i^{k_2}\right)^2\right).
\end{aligned}$$

By the orthogonality of U_n, $t^{k_1,\top}t^{k_2} = w_{k_1}^\top w_{k_2} = \delta_{k_1k_2}$, and so this equals

$$2\sum_{k_1,k_2}\mu_{k_1}\mu_{k_2}\left(\delta_{k_1k_2} - \sum_{i=1}^n\left(t_i^{k_1}\right)^2\left(t_i^{k_2}\right)^2\right) = 2\sum_{k=1}^p\mu_k^2 - 2\sum_{k_1,k_2}\mu_{k_1}\mu_{k_2}\sum_{i=1}^n\left(t_i^{k_1}\right)^2\left(t_i^{k_2}\right)^2.$$

Since $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$ from (22), we have $\sum_{i=1}^n (U_n)_{ki}^4 = o\left(\zeta_m/\sqrt{m}\right)$ for every $k \in [p]$, and so recalling that $t_i^k = (U_n w_k)_i$ for $k \in [p], i \in [n]$, by the Cauchy-Schwarz inequality, we can derive that $\sum_{i=1}^n \left(t_i^{k_1}\right)^2 \left(t_i^{k_2}\right)^2 = o\left(\zeta_m/\sqrt{m}\right)$ for every $k_1, k_2 \in [p]$. Thus, $ms_n^2 - 2\sum_{k=1}^p \mu_k^2 \to 0$. By applying the Cauchy-Schwarz inequality once more, we find that

$$w_{ij}^2 = 4\left(\sum_{k=1}^p \mu_k t_i^k t_j^k\right)^2 \leqslant 4\left(\sum_{k=1}^p \mu_k^2 \left(t_i^k\right)^2\right)\left(\sum_{k=1}^p \left(t_j^k\right)^2\right) \tag{82}$$

$$\leqslant 4\left(\sum_{k=1}^p \mu_k^2\right)\left(\sum_{k=1}^p \left(t_i^k\right)^2\right)\left(\sum_{k=1}^p \left(t_j^k\right)^2\right). \tag{83}$$

Denoting $C := 4\left(\sum_{k=1}^p \mu_k^2\right)$, we arrive at

$$\sum_{i<j} w_{ij}^4 \leqslant C^2 \sum_{i<j}\left(\sum_{k=1}^p \left(t_i^k\right)^2\right)^2\left(\sum_{k=1}^p \left(t_j^k\right)^2\right)^2 \leqslant C^2 \sum_{i,j}\left(\sum_{k=1}^p \left(t_i^k\right)^2\right)^2\left(\sum_{k=1}^p \left(t_j^k\right)^2\right)^2$$
$$= \left(C\sum_{i=1}^n\left(\sum_{k=1}^p \left(t_i^k\right)^2\right)^2\right)^2 = \left(C\sum_{k_1,k_2}\sum_{i=1}^n \left(t_i^{k_1}\right)^2\left(t_i^{k_2}\right)^2\right)^2.$$

Since $\sum_{i=1}^n \left(t_i^{k_1}\right)^2\left(t_i^{k_2}\right)^2 = o(\zeta_m/\sqrt{m})$, we conclude that $\frac{m\left(m-1+3(\zeta_m-1)^2\right)}{((m-1)\zeta_m^2)} \cdot \sum_{i<j} w_{ij}^4 \to 0$, which verifies the first part of (79). Additionally, since for all $k \in [p]$, $\sum_{i=1}^n (t_i^k)^2 = \sum_{i=1}^n (U_n w_k)_i^2 = \|w_k\|^2 = 1$,

$$\sum_{j=1}^n\left(\sum_{i=1}^{j-1} w_{ij}^2\right)^2 \leqslant C^2\sum_{j=1}^n\left[\sum_{i=1}^{j-1}\left(\sum_{k=1}^p \left(t_i^k\right)^2\right)\left(\sum_{k=1}^p \left(t_j^k\right)^2\right)\right]^2$$
$$\leqslant C^2\sum_{j=1}^n\left[\sum_{i=1}^{n}\left(\sum_{k=1}^p \left(t_i^k\right)^2\right)\left(\sum_{k=1}^p \left(t_j^k\right)^2\right)\right]^2$$
$$= C^2\sum_{j=1}^n\left(p\sum_{k=1}^p \left(t_j^k\right)^2\right)^2 = p^2C^2\sum_{k_1,k_2}\sum_{i=1}^n \left(t_i^{k_1}\right)^2\left(t_i^{k_2}\right)^2 \to 0,$$

which verifies the second part of (79). Finally, we have

$$\frac{1+(\zeta_m-1)^2/(m-1)}{m}\sum_{k_1=2}^n\sum_{k_2=2}^n\sum_{i_1=1}^{k_1\wedge k_2-1}\sum_{j_1=1}^{k_1\wedge k_2-1} w_{i_1k_1}w_{j_1k_1}w_{i_1k_2}w_{j_1k_2}$$
$$\leqslant \frac{C^2\left(1+\frac{(\zeta_m-1)^2}{(m-1}\right)}{m}\sum_{i_1,j_1,k_1,k_2}\left(\sum_{l=1}^p \left(r_{i_1}^l\right)^2\right)\left(\sum_{l=1}^p \left(r_{j_1}^l\right)^2\right)\left(\sum_{l=1}^p \left(r_{k_1}^l\right)^2\right)\left(\sum_{l=1}^p \left(r_{k_2}^l\right)^2\right)$$
$$= \frac{C^2\left(1+(\zeta_m-1)^2/(m-1)\right)}{m}\left(\sum_{i=1}^n\sum_{l=1}^p \left(r_i^l\right)^2\right)^4 = \frac{C^2p^4\left(1+(\zeta_m-1)^2/(m-1)\right)}{m} \to 0.$$

Thus, the conditions needed for the martingale central limit theorem, as used previously in the proof of Lemma 3.8, are satisfied. Since $2\sum_{k=1}^p \mu_k^2 = \mathrm{vec}(\Phi)^\top(I_{p^2}+P_p)\mathrm{vec}(\Phi)$, we conclude that

$$\sqrt{m}\,\mathrm{tr}\left(\Phi(U_n^\top S_{m,n}^\top S_{m,n}U_n - I_p)\right) \Rightarrow \mathcal{N}\left(0, \mathrm{vec}(\Phi)^\top(I_{p^2}+P_p)\mathrm{vec}(\Phi)\right),$$

and (26) follows.

G.7 Proofs in Section 3.3

G.7.1 Proof of Theorem 3.10

The proofs of all claims other than (30) can be directly obtained from Theorem 2.9 by letting $G_n := n\sum_{i=1}^n \left((u_{n,i}u_{n,i}^\top) \otimes (u_{n,i}u_{n,i}^\top)\right)$, as $(G_n)_{(ii),(ii)} = n\sum_{k=1}^n (U_n)_{ki}^4$. Now, to conclude (30) from (29), let $u_{n,i}$ be the i-th column of U_n. By calculating variances, we argue that

$$\frac{m\sum_{j=1}^m (s_j^\top u_{n,i})^4}{n\sum_{i=1}^n (U_n)_{ki}^4} \to_P 1. \tag{84}$$

Since $\sum_{j=1}^m (s_j^\top u_{n,i})^4 = n^2/m^2 \cdot \sum_{l=1}^n (b_l (U_n)_{li})^4 = n^2/m^2 \cdot \sum_{l=1}^n (U_n)_{li}^4 b_l$, by Condition $\mathfrak{P}'$, the variance of the left hand side of (84) is

$$\frac{1-\gamma_n}{\gamma_n}\frac{\sum_{l=1}^n (U_n)_{li}^8}{(\sum_{l=1}^n (U_n)_{li}^4)^2} \leqslant \frac{1-\gamma_n}{\gamma_n}\frac{\max_{l\in[n]}(U_n)_{li}^4}{\sum_{l=1}^n (U_n)_{li}^4} \to_P 0.$$

Now, recall that $\widetilde{X}_{m,n} = \widetilde{U}_{m,n}\hat{L}_{m,n}\hat{V}_{m,n}^\top$ is the SVD of $\widetilde{X}_{m,n}$. By Condition $\mathfrak{P}'$ defined in (28), $n\sum_{k=1}^n (U_n)_{ki}^4/m = (G_n)_{(ii),(ii)}/m \to 0$; combining this with Theorem 3.10, we have $L_n^{-1}\hat{L}_{m,n} \to_P I_p$ and $V_n^{-1}\hat{V}_{m,n} \to_P I_p$. Note that $\widetilde{U}_{m,n} = \widetilde{X}_{m,n}\hat{V}_{m,n}\hat{L}_{m,n}^{-1}$, which further equals $S_{m,n}U_nL_nV_n^\top\hat{V}_{m,n}\hat{L}_{m,n}^{-1}$, and $L_nV_n^\top\hat{V}_{m,n}\hat{L}_{m,n}^{-1} \to_P I_p$. Expanding the summation in the numerator according to the rules of matrix element multiplication, we conclude that

$$n^{-1}m\sum_{j=1}^m \tilde{U}_{ji}^4 / \left[n^{-1}m\sum_{j=1}^m (s_j^\top u_{n,i})^4\right] \to_P 1.$$

Hence, (30) follows from (29).

G.7.2 Proof of Theorem 3.11

The proof is based on the following asymptotic distribution and consistency properties.

Lemma G.4 (Asymptotic Normality in OLS with Uniform Sampling)**.** *Under the conditions of Theorem 3.11, with*

$$\Sigma_n := n(X_n^\top X_n)^{-1}\left(\sum_{i=1}^n \varepsilon_{n,i}^2 x_{n,i}x_{n,i}^\top\right)(X_n^\top X_n)^{-1}, \tag{85}$$

we have $m^{1/2}(1-\gamma_n)^{-1/2}\Sigma_n^{-1/2}(\hat{\beta}_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$. *Similarly, defining*

$$\Sigma_n^{(\mathrm{p})} := n(X_n^\top X_n)^{-1}\left(\sum_{i=1}^n (y_{n,i} - \varepsilon_{n,i})^2 x_{n,i}x_{n,i}^\top\right)(X_n^\top X_n)^{-1}, \tag{86}$$

we have $m^{1/2}(1-\gamma_n)^{-1/2}(\Sigma_n^{(\mathrm{p})})^{-1/2}(\hat{\beta}_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$.

The asymptotic result (85) coincides with Theorem 2 of Yu et al. (2022) by taking $\Psi(x) = x$

and $p_i = r/n$ therein. Their result is assumes that X_n has i.i.d. rows and y_n has i.i.d. entries, and then conditions on X_n and y_n. In contrast, we do not make this assumption on X_n and y_n.

Proof. For any $b_n \in \mathbb{S}^{p-1}$, $n \geqslant 1$, since $\|\mathrm{vec}(U_n)\|_\infty = O(1/\sqrt{n})$, we have $\|U_n b_n\|_\infty = O(1/\sqrt{n})$. Then $(U_n b_n, U_n b_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{p}$, and due to Lemma 3.12, Condition 2.2 holds for the sequences $(U_n b_n, U_n b_n)_{n\geqslant 1}$.

From Lemma 3.12, $\sigma_n^2(a_n, \tilde{a}_n) = n\sum_{i=1}^n (a_{n,i}\tilde{a}_{n,i})^2$. By (31), $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}$ and the sequence of variances $\sigma_n^2(U_n b_n, \bar{\varepsilon}_n) = n\sum_{i=1}^n (U_n b_n)_i^2 (\bar{\varepsilon}_n)_i^2$ is uniformly bounded away from zero and infinity. Therefore, Condition 2.4 holds for the sequences $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$, and $M_n = nU_n^\top \operatorname{diag}(\bar{\varepsilon}_n \odot \bar{\varepsilon}_n) U_n$. Similarly to the proof in Lemma G.6, combining this with Lemma 3.12, we conclude the desired claim for $\hat{\beta}_{m,n}$.

Similarly, Condition 2.4 holds for $(U_n b_n, \overline{X_n\beta_n})_{n\geqslant 1}$, and using Lemma 3.12, we have $M_n^{(\mathrm{p})} = n\|X_n\beta_n\|^{-2} U_n^\top \operatorname{diag}((X_n\beta_n)\odot(X_n\beta_n))U_n \in \mathbb{S}_+^{p\times p}$. Therefore, if we take $b_n = L_n^{-1}V_n^\top c/\|X_n^{\dagger\top}c\|$, the claim for $\hat{\beta}_{m,n}^{(\mathrm{p})}$ follows as in Lemma G.6. □

Proposition G.5 (Consistency via Sampling for Covariance Matrices)**.** *Under the conditions of Theorem 3.11, we have*

$$\|\Sigma_n^{-1}\hat{\Sigma}_{m,n} - I_p\|_{\mathrm{Fr}} = o_P(1) \text{ and } \|(\Sigma_n^{(\mathrm{p})})^{-1}\hat{\Sigma}_{m,n}^{(\mathrm{p})} - I_p\|_{\mathrm{Fr}} = o_P(1).$$

Proof. We consider $y_n \in \mathbb{S}^{n-1}$ and X_n of a unit spectral radius first. From (52), $\|\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n} - X_n^\top X_n\| = o_P(1)$. Thus, it suffices to show that

$$\left\| \frac{m}{n}\widetilde{X}_{m,n}^\top \operatorname{diag}(\tilde{\varepsilon}_{m,n}\odot\tilde{\varepsilon}_{m,n})\widetilde{X}_{m,n} - \sum_{i=1}^n \varepsilon_{n,i}^2 x_{n,i}x_{n,i}^\top \right\| = o_P(1).$$

Since $S_{m,n} = \operatorname{diag}(\sqrt{n/m}B_1, \ldots, \sqrt{n/m}B_n)$, we have

$$\begin{aligned}
&\widetilde{X}_{m,n}^\top \operatorname{diag}(\tilde{\varepsilon}_{m,n}\odot\tilde{\varepsilon}_{m,n})\widetilde{X}_{m,n} \\
&= \sqrt{n/m}X_n^\top S_{m,n}\operatorname{diag}(B_1^2(y_n - X_n\hat{\beta}_{m,n})_1^2, \ldots, B_n^2(y_n - X_n\hat{\beta}_{m,n})_n^2)S_{m,n}X_n \\
&= (n/m)\cdot\sum_{i=1}^n (n/m)B_i(\varepsilon_n + X_n(\beta_n - \hat{\beta}_{m,n}))_i^2 x_{n,i}x_{n,i}^\top.
\end{aligned}$$

Here, we can check that $\|\beta_n - \hat{\beta}_{m,n}\| = O_P(\sqrt{1-\gamma_n)/m})$ and $\sum_{i=1}^n [(n/m)B_i - 1]\varepsilon_{n,i}^2 (x_{n,i}^\top c)^2 = o_P(1)$. Therefore,

$$\sum_{i=1}^n (n/m)B_i \cdot 2\varepsilon_{n,i}x_{n,i}^\top(\beta_n - \hat{\beta}_{m,n})(x_{n,i}^\top c)^2 = O_P(\sqrt{(1-\gamma_n)/m}\cdot n/m)$$

and $\sum_{i=1}^n (n/m)B_i \cdot (x_{n,i}^\top(\beta_n - \hat{\beta}_{m,n}))^2(x_{n,i}^\top c)^2 = O_P((1-\gamma_n)/m\cdot n/m)$.

Thus, the first claim of Proposition G.5 holds when $\|y_n\| = 1$ and $\|X_n\| = 1$. For general y_n and X_n, the proof is similar to that of Proposition G.7. For the second conclusion, the proof is almost identical to the above argument. □

G.7.3 Proof of Lemma 3.12

To show (33), notice that $\mathrm{Var}\left[a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n\right] = n\sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2\cdot(1-\gamma_n)/m$. Then (33) follows by Condition $\mathfrak{p}$ from (27).

To see (34), we know that $a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n = \frac{n}{m}\sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}B_i$ and $\mathbb{E}B_i = m/n$. Denoting $J_i = (1/\gamma_n - 1)^{-1/2}(\frac{n}{m}B_i - 1)$, we have that J_i are i.i.d. with $\mathbb{E}J_i = 0$ and $\mathbb{E}J_i^2 = 1$. As a consequence, $a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top\tilde{a}_n = (1/\gamma_n - 1)^{1/2}\sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}J_i$. We check the Lindeberg condition

$$\sum_{i=1}^n \mathbb{E}[(a_{n,i}\tilde{a}_{n,i})^2 J_i^2 I\{(a_{n,i}\tilde{a}_{n,i})^2 J_i^2 > \varepsilon s_n^2\}]/s_n^2 \to 0 \tag{87}$$

for all $\varepsilon > 0$, where $s_n^2 = \mathrm{Var}\left[\sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}J_i\right] = \sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2\,\mathrm{Var}\left[J_i\right] = \sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2$. We have

$$\mathbb{E}[J_i^2 I\{(a_{n,i}\tilde{a}_{n,i})^2 J_i^2 > \varepsilon s_n^2\}] \leqslant \mathbb{E}\left[J_i^2 I\left\{\max_{i\in[n]}\frac{(a_{n,i}\tilde{a}_{n,i})^2}{s_n^2}J_i^2 > \varepsilon\right\}\right],$$

so the left hand side of (87) is upper bounded by

$$\sum_{i=1}^n \frac{(a_{n,i}\tilde{a}_{n,i})^2}{s_n^2}\mathbb{E}\left[J_i^2 I\left\{\max_{i\in[n]}\frac{(a_{n,i}\tilde{a}_{n,i})^2}{s_n^2}J_i^2 > \varepsilon\right\}\right] = \mathbb{E}\left[J_i^2 I\left\{\max_{i\in[n]}\frac{(a_{n,i}\tilde{a}_{n,i})^2}{s_n^2}J_i^2 > \varepsilon\right\}\right] \to 0,$$

where the convergence to zero holds due to Condition $\mathfrak{P}$. Then (34) follows.

A counterexample that does not satisfy Condition $\mathfrak{P}$, such that $\sqrt{m}a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n$ is not asymptotically normal is $a_n = \tilde{a}_n = (1, 0, \ldots, 0)$. Then,

$$\sqrt{\frac{m}{1-\gamma_n}}\sqrt{\frac{1}{n\sum_{i=1}^n a_{n,i}^2\tilde{a}_{n,i}^2}}\sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}J_i = \sqrt{\frac{m}{n-m}}J_1.$$

This random variable takes two values, so it is not asymptotically normal.

G.7.4 Proof of Lemma 3.13

By the subsequence argument and Cramer-Wold device as in the beginning of the proof of Lemma D.4, it suffices to show that for any upper triangular $\Phi \in \mathbb{R}^{p\times p}$,

$$\sqrt{\frac{m}{1-\gamma_n}}\left(\mathrm{vec}(\Phi)^\top G_n \mathrm{vec}(\Phi)\right)^{-1/2}\mathrm{vec}(\Phi)^\top \mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n}U_n - I_p) \Rightarrow \mathcal{N}(0,1).$$

Denote $U_n = (u_{n,1}, \ldots, u_{n,n})^\top$, where $u_{n,i}^\top$ is the i-th row of U_n. Then

$$\mathrm{tr}\left(\Phi(U_n^\top S_{m,n}^\top S_{m,n}U_n - I_p)\right) = \mathrm{tr}\left(\Phi\sum_{i=1}^n u_{n,i}u_{n,i}^\top\left(\frac{n}{m}B_i - 1\right)\right) = \sum_{i=1}^n u_{n,i}^\top\Phi u_{n,i}\left(\frac{n}{m}B_i - 1\right).$$

Denote $J_i = (1/\gamma_n - 1)^{-1/2}(\frac{n}{m}B_i - 1)$ for all $i \in [n]$, so that $\mathbb{E}J_i = 0, \mathbb{E}J_i^2 = 1$. Then from Lemma 3.12, if U_n satisfies $\frac{\max_{i\in[n]}(u_{n,i}^\top\Phi u_{n,i})^2}{\sum_{i=1}^n(u_{n,i}^\top\Phi u_{n,i})^2} \to 0$ in Condition $\mathfrak{P}'$ from (28), the above term is asymptotically normal.

G.8 Proofs of Results in Section D

G.8.1 Proof of Theorem D.1

By Lemma D.4, since $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, Condition 2.8 holds with $\tau_{m,n} = 1$ and $G_n = I_{p^2} + P_p + \Gamma_n$. Moreover, for all $i \in [p]$, $(G_n)_{(ii),(ii)} = 2 + (\Gamma_n)_{(ii),(ii)}$, and thus from (10),

$$\Delta_i = \Delta_i(\Xi, G_n) = \sum_{k\neq i, l\neq i} \frac{\lambda_i(\Xi)\sqrt{\lambda_k(\Xi)\lambda_l(\Xi)}}{(\lambda_i(\Xi) - \lambda_k(\Xi))\,(\lambda_i(\Xi) - \lambda_l(\Xi))} \left(\delta_{kl} + (\Gamma_n)_{(ij),(kl)}\right) v_k(\Xi) v_l(\Xi)^\top.$$

Since $(X_n)_{n\geqslant 1}$ satisfies Condition 2.3 B, the conclusion holds due to Theorem 2.9.

G.8.2 Proof of Theorem D.2

We first show the following result, which establishes the asymptotic normality of sketched least squares estimators. The result can be regarded as a special case of Theorem 3.2 in Zhang et al. (2023) for a fixed dimension p, but with the weaker condition that the moments of the sketching matrices have a $(4+\delta)$-th moment, as opposed to having all moments finite. Also, it turns out that for i.i.d. sketching, using our general framework would require assuming that $(U_n)_{n\geqslant 1}$ satisfies the conditions of Lemma D.4. However, by a more direct analysis, we can avoid this condition; and so the proof we present deviates from our general framework.

In the remainder, we denote $X_n = (x_{n,1}^\top, \ldots, x_{n,p}^\top)^\top \in \mathbb{R}^{n\times p}$, $y_n = (y_{n,1}, \ldots, y_{n,p})^\top \in \mathbb{R}^p$.

Lemma G.6 (Asymptotic Normality in OLS with i.i.d. Sketching)**.** *Under the conditions of Theorem D.2, with*

$$\Sigma_n := (X_n^\top X_n)^{-1} \left\{ \sum_{i=1}^n \varepsilon_{n,i}^2 \left[X_n^\top X_n + (\kappa_{n,4} - 3) x_{n,i} x_{n,i}^\top \right] \right\} (X_n^\top X_n)^{-1}, \tag{88}$$

we have $m^{1/2}\Sigma_n^{-1/2}(\hat\beta_{m,n} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$. *Similarly, defining*

$$\begin{aligned}\Sigma_n^{(\mathrm{p})} := {} & (X_n^\top X_n)^{-1} \|y_n - \varepsilon_n\|^2 \\ & + (X_n^\top X_n)^{-1} \left[(\kappa_{n,4} - 3) \sum_{i=1}^n (y_{n,i} - \varepsilon_{n,i})^2 x_{n,i} x_{n,i}^\top \right] (X_n^\top X_n)^{-1} + \beta_n \beta_n^\top,\end{aligned} \tag{89}$$

we have $m^{1/2}(\Sigma_n^{(\mathrm{p})})^{-1/2}(\hat\beta_{m,n}^{(\mathrm{p})} - \beta_n) \Rightarrow \mathcal{N}(0, I_p)$.

Proof. Consider a sequence of vectors $(a_n, \tilde a_n)_{n\geqslant 1}$ such that $a_n, \tilde a_n \in \mathbb{S}^{n-1}$ for all n. From Lemma D.3, we have Condition 2.2.

For the first claim, (98) in the proof of Lemma D.3 in Appendix G.8.3 shows that the sequence of variances $\sigma_n^2(a_n, \tilde a_n) = 1 + (a_n^\top \tilde a_n)^2 + (\kappa_{n,4} - 3)\sum_{i=1}^n (a_{n,i}\tilde a_{n,i})^2$ is uniformly bounded away from zero and infinity. Therefore, Condition 2.4 holds for the sequence $(U_n b_n, \bar\varepsilon_n)_{n\geqslant 1}$, and

we have $M_n = I_p + (\kappa_{n,4} - 3)U_n^\top \operatorname{diag}(\bar{\varepsilon}_n \odot \bar{\varepsilon}_n)U_n$. Thus,

$$\begin{aligned} V_n L_n^{-1} M_n L_n^{-1} V_n^\top &= V_n L_n^{-2} V_n^\top + (\kappa_{n,4} - 3) V_n L_n^{-1} U_n^\top \operatorname{diag}(\bar{\varepsilon}_n \odot \bar{\varepsilon}_n) U_n L_n^{-1} V_n^\top \\ &= (X_n^\top X_n)^{-1} + (\kappa_{n,4} - 3)(X_n^\top X_n)^{-1} X_n^\top \operatorname{diag}(\varepsilon_n \odot \varepsilon_n) X_n (X_n^\top X_n)^{-1} / \|\varepsilon_n\|^2 \\ &= (X_n^\top X_n)^{-1} + (\kappa_{n,4} - 3)(X_n^\top X_n)^{-1} \left(\sum_{i=1}^n \varepsilon_{n,i}^2 x_{n,i} x_{n,i}^\top \right) (X_n^\top X_n)^{-1} / \|\varepsilon_n\|^2. \end{aligned}$$

Therefore, the claim for $\hat{\beta}_{m,n}$ follows from (5) in Theorem 2.5; where we remark that all necessary conditions hold due to the stated assumptions.

For the second claim, we define $\tau_{m,n} = \kappa_{n,4}$, so that by the Hölder inequality, $\tau_{m,n} = o(m)$. Also, $\sigma_n^2(a_n, \tilde{a}_n) = (1 + (a_n^\top \tilde{a}_n)^2 + (\kappa_{n,4} - 3)\sum_{i=1}^n (a_{n,i}\tilde{a}_{n,i})^2)/\kappa_{n,4} = \sum_{i=1}^n (a_{n,i}\tilde{a}_{n,i})^2 + o(1)$. From the proof of Lemma D.3, $\sigma_n^2(a_n, \tilde{a}_n)$, are uniformly bounded away from zero and infinity for $n \geqslant 1$. Similarly, Condition 2.4 holds for $(U_n b_n, \overline{X_n \beta_n})_{n \geqslant 1}$, and using Lemma D.3, we have $M_n^{(\mathrm{p})} = I_p + \|X_n\beta_n\|^{-2} L_n V_n \beta_n \beta_n^\top V_n^\top L_n + \|X_n\beta_n\|^{-2}(\kappa_{n,4} - 3) U_n^\top \operatorname{diag}((X_n\beta_n) \odot (X_n\beta_n)) U_n \in \mathbb{S}_+^{p \times p}$. Therefore, we can calculate

$$\begin{aligned} \Sigma_n^{(\mathrm{p})} &= \|X_n\beta_n\|^2 V_n L_n^{-1} M_n^{(\mathrm{p})} L_n^{-1} V_n^\top \\ &= \|X_n\beta_n\|^2 (X_n^\top X_n)^{-1} + \beta_n \beta_n^\top + (\kappa_{n,4} - 3) V_n L_n^{-1} U_n^\top \operatorname{diag}((X_n\beta_n) \odot (X_n\beta_n)) U_n L_n^{-1} V_n^\top \\ &= (X_n^\top X_n)^{-1} \|y_n - \varepsilon_n\|^2 \\ &\qquad + (X_n^\top X_n)^{-1} \left[(\kappa_{n,4} - 3) \sum_{i=1}^n (y_{n,i} - \varepsilon_{n,i})^2 x_{n,i} x_{n,i}^\top \right] (X_n^\top X_n)^{-1} + \beta_n \beta_n^\top, \end{aligned}$$

the claim for $\hat{\beta}_{m,n}^{(\mathrm{p})}$ follows from Theorem B.2. □

We will also use the following proposition, which shows that $\hat{\Sigma}_{m,n}$ and $\hat{\Sigma}_{m,n}^{(\mathrm{p})}$ are ratio-consistent estimators of Σ_n and $\Sigma_n^{(\mathrm{p})}$, respectively.

Proposition G.7 (Consistency via i.i.d. Sketching for Covariance Matrices)**.** *Under the conditions of Theorem D.2, we have*

$$\|\Sigma_n^{-1}\hat{\Sigma}_{m,n} - I_p\|_{\mathrm{Fr}} = o_P(1) \text{ and } \|(\Sigma_n^{(\mathrm{p})})^{-1}\hat{\Sigma}_{m,n}^{(\mathrm{p})} - I_p\|_{\mathrm{Fr}} = o_P(1).$$

Proof. We consider $y_n \in \mathbb{S}^{n-1}$ and X_n of a unit spectral radius first. For each $i \in [m]$, we denote the i-th row of $\sqrt{m} S_{m,n}$ by $s_i^\top$. To prove the first claim, due to (51) and (88), it suffices to show that

$$\|\widetilde{X}_{m,n}^\top \widetilde{X}_{m,n} - X_n^\top X_n\|_{\mathrm{Fr}} = o_P(1) \tag{90}$$

and

$$\left\| m\widetilde{X}_{m,n}^\top \operatorname{diag}(\tilde{\varepsilon}_{m,n} \odot \tilde{\varepsilon}_{m,n}) \widetilde{X}_{m,n} - \sum_{i=1}^n \varepsilon_{n,i}^2 \left[X_n^\top X_n + (\kappa_{n,4} - 3) x_{n,i} x_{n,i}^\top \right] \right\|_{\mathrm{Fr}} = o_P(1). \tag{91}$$

We can conclude (90) from (52). Also, since $\sqrt{m}\tilde{\varepsilon}_{n,i} = s_i^\top (y_n - X_n \hat{\beta}_{m,n})$, to establish (91), we just need to show that for any fixed $c \in \mathbb{R}^p$ and $c \neq 0$, for $\sigma_n^2 := c^\top \left\{ \sum_{i=1}^n \varepsilon_{n,i}^2 \left[X_n^\top X_n + (\kappa_{n,4} - 3) x_{n,i} x_{n,i}^\top \right] \right\} c$,

we have

$$\sigma_n^2 - \frac{1}{m}\sum_{i=1}^m \left(c^\top X_n^\top s_i s_i^\top (y_n - X_n\hat{\beta}_{m,n})\right)^2 = o_P(1). \tag{92}$$

From (95) with $a_n = X_n c$ and $\tilde{a}_n = \varepsilon_n = y_n - X_n\beta_n$, and using the law of large numbers for triangular arrays (see, for instance, Theorem 2.2.6 in Durrett (2019)), we deduce that

$$\frac{1}{m}\sum_{i=1}^m \sigma_n^{-2}\left(c^\top X_n^\top s_i s_i^\top (y_n - X_n\beta_n)\right)^2 \to_P 1. \tag{93}$$

We claim that

$$\frac{1}{m}\sum_{i=1}^m \sigma_n^{-2}\left(c^\top X_n^\top s_i s_i^\top X_n(\beta_n - \hat{\beta}_{m,n})\right)^2 \to_P 0. \tag{94}$$

This claim follows from the following three facts:

$$\|\beta_n - \hat{\beta}_{m,n}\| = O_P(m^{-1/2}\sigma_n),\ \max_{i\in[m]}\|s_i^\top X_n\| = o_P(m^{1/2}), \quad \text{and } c^\top X_n^\top S_{m,n}^\top S_{m,n} X_n c = O_P(1).$$

The first claim follows from Lemma G.6; the second is a consequence of Markov's inequality and a union bound, and the third one follows from $\|a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top \tilde{a}_n\| = O_P(m^{-1/2})$.

Then, from (93), (94), we can conclude (92). Thus, the first claim of Proposition G.7 holds when $\|y_n\| = 1$ and $\|X_n\| = 1$.

For general y_n and X_n, we let $y_n' = y_n/\|y_n\|$ and $X_n' = X_n/\|X_n\|$, and we consider the observed data $S_{m,n}(X_n', y_n')$. By applying the above argument with $\sigma_n^2 = c^\top\{\sum_{i=1}^n(\varepsilon_{n,i}/\|y_n\|)^2 [X_n^\top X_n + (\kappa_{n,4} - 3)x_{n,i}x_{n,i}^\top]/\|X_n\|^2\}c$ rescaled by $\|X_n\|$ and $\|y_n\|$, and corresponding rescaled estimators, as well as all other appropriate quantities rescaled, the $\|X_n\|$ and $\|y_n\|$ terms cancel out, leading to the desired result for general X_n and y_n.

Next, we consider the second conclusion from Proposition G.7. Similarly to the analysis of the first conclusion, it suffices to consider $y_n \in \mathbb{S}^{n-1}$. By choosing $a_n = X_n c$ and $\tilde{a}_n = X_n\beta_n$ in (95), we can apply the law of large numbers for triangular arrays (see, for instance, Theorem 2.2.6 in Durrett (2019)) to obtain that with

$$\begin{aligned}(\sigma_n^{(\mathrm{p})})^2 &:= (c^\top X_n^\top X_n c)(\beta_n^\top X_n^\top X_n\beta) + (c^\top X_n^\top X_n\beta_n)^2 + (\kappa_{n,4} - 3)c^\top X_n^\top \operatorname{diag}(X_n\beta_n \odot X_n\beta_n)X_n c \\ &= c^\top\left(\|y_n - \varepsilon_n\|^2\left[X_n^\top X_n + (\kappa_{n,4} - 3)x_{n,i}x_{n,i}^\top\right] + X_n^\top y_n y_n^\top X_n\right)c,\end{aligned}$$

we have $\frac{1}{m}\sum_{i=1}^m(\sigma_n^{(\mathrm{p})})^{-2}\left(c^\top X_n^\top s_i s_i^\top X_n\beta_n\right)^2 \to_P 1$. The conclusion follows similarly as for sketch-and-solve above, using that $\|\beta_n - \hat{\beta}_{m,n}^{(\mathrm{p})}\| = O_P(m^{-1/2}\sigma_n^{(\mathrm{p})})$. □

Combining the above two results, the conclusion follows.

G.8.3 Proof of Lemma D.3

We start by observing that $a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n - a_n^\top\tilde{a}_n = \frac{1}{m}\sum_{i=1}^m(s_i^\top a_n\tilde{a}_n^\top s_i - a_n^\top\tilde{a}_n)$. By applying equation (9.8.6) in Bai and Silverstein (2010), each of the above i.i.d. centered random variables

has variance

$$\sigma_n^2 := \mathbb{E}(s_i^\top a_n \tilde{a}_n^\top s_i - a_n^\top \tilde{a}_n)^2 = (a_n^\top a_n)(\tilde{a}_n^\top \tilde{a}_n) + (a_n^\top \tilde{a}_n)^2 + (\kappa_{n,4} - 3)\sum_{i=1}^n (a_{n,i}\tilde{a}_{n,i})^2. \tag{95}$$

To establish the asymptotic normality, we use the Lindeberg-Feller central limit theorem for triangular arrays for $\xi_{n,i} := (s_i^\top a_n \tilde{a}_n^\top s_i - a_n^\top \tilde{a}_n)/\sigma_n$, so that $\xi_{n,1}, \ldots, \xi_{n,m}$ are i.i.d. random variables with unit variance, whose distribution may depend on n. To check the Lindeberg condition, we denote $r = 1 + \delta/4 > 1$ and use the Hölder inequality to obtain

$$\mathbb{E}\left[\xi_{n,1}^2 I\{|\xi_{n,1}| > \varepsilon\sqrt{m}\}\right] \leqslant \mathbb{E}[\xi_{n,1}^{2r}]^{1/r} \cdot \mathbb{P}\left[|\xi_{n,1}| > \varepsilon\sqrt{m}\right]^{1-1/r}. \tag{96}$$

Applying Chebyshev's inequality, we have

$$\mathbb{P}\left[|\xi_{n,1}| > \varepsilon\sqrt{m}\right] \leqslant \frac{\mathbb{E}[\xi_{n,1}^2]}{\varepsilon^2 m} = \frac{1}{\varepsilon^2 m}. \tag{97}$$

Furthermore, by Lemma B.26 in Bai and Silverstein (2010) for the centered quadratic form $\xi_{n,1}$, we find $\mathbb{E}[\xi_{n,1}^{2r}] \leqslant C(\kappa_{n,4}^r + \kappa_{n,4r})/\sigma_n^{2r}$, where C is an absolute constant. Therefore, $\mathbb{E}\left[\xi_{n,1}^2 I\{|\xi_{n,1}| > \varepsilon\sqrt{m}\}\right] \leqslant [C(\kappa_{n,4}^r + \kappa_{n,4r})]^{1/r}/\sigma_n^2 \cdot (\varepsilon^2 m)^{1/r-1}$.

Now consider the first condition from Condition $\mathfrak{P}$ defined in (47). We observe that if $\kappa_{n,4} \geqslant 3$, then $1 \leqslant \sigma_n^2 \leqslant 2 + \kappa_{n,4} - 3 \leqslant \kappa_{n,4+\delta}^{4/(4+\delta)} - 1 < +\infty$. Otherwise, if $1 + \delta' \leqslant \kappa_{n,4} < 3$, we have

$$\begin{aligned}\sigma_n^2 &= \frac{\kappa_{n,4} - 1}{2} + \frac{\kappa_{n,4} - 1}{2}\left(\sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}\right)^2 + \frac{3 - \kappa_{n,4}}{2}\left[\sum_{i\neq j} a_{n,i}^2 \tilde{a}_{n,j}^2 + \sum_{i\neq j} a_{n,i}\tilde{a}_{n,i} a_{n,j}\tilde{a}_{n,j}\right] \\ &= \frac{\kappa_{n,4} - 1}{2} + \frac{\kappa_{n,4} - 1}{2}\left(\sum_{i=1}^n a_{n,i}\tilde{a}_{n,i}\right)^2 + \frac{3 - \kappa_{n,4}}{4}\sum_{i\neq j}(a_{n,i}\tilde{a}_{n,j} + a_{n,j}\tilde{a}_{n,i})^2 \geqslant \frac{\delta'}{2}.\end{aligned} \tag{98}$$

Thus σ_n^2 are bounded uniformly away from zero. Hence, $\mathbb{E}\left[\xi_{n,1}^2 I\{|\xi_{n,1}| > \varepsilon\sqrt{m}\}\right] \to 0$.

Next, consider the second condition from Condition $\mathfrak{P}$ defined in (47), In this case, since $\kappa_{n,4} \to \infty$, $\sigma_n^2/\kappa_{n,4} = \sum_{i=1}^n (a_{n,i}\tilde{a}_{n,i})^2 + o(1)$ from the expression in (95), so $\sigma_n^2/\kappa_{n,4}$ are uniformly bounded away from zero and infinity. Therefore, we can choose a $\delta'' > 0$ in this case such that $\sigma_n^2/\kappa_{n,4} \geqslant \delta''$. Consequently, using Hölder's inequality again, and as $\kappa_{n,4+\delta}^{4/(4+\delta)}/\kappa_{n,4} = o(m^{\delta/(4+\delta)})$, $\mathbb{E}\left[\xi_{n,1}^2 I\{|\xi_{n,1}| > \varepsilon\sqrt{m}\}\right]$ is bounded above by

$$\frac{C(\kappa_{n,4}^r + \kappa_{n,4r})^{1/r}}{(\varepsilon^2 m)^{1-1/r}\kappa_{n,4}(\delta' \wedge \delta'')^2} \leqslant \frac{C'\kappa_{n,4r}^{1/r}}{(\varepsilon^2 m)^{1-1/r}\kappa_{n,4}(\delta' \wedge \delta'')^2} = o\left(\frac{C'(m^{1-1/r})}{(\varepsilon^2 m)^{1-1/r}(\delta' \wedge \delta'')^2}\right) \to 0.$$

By the Lindeberg-Feller central limit theorem, the conclusion follows.

G.8.4 Proof of Lemma D.4

By the subsequence argument as in the proof of (5) from Theorem 2.5, it suffices to show that for any $\Phi \in \mathbb{R}^{p\times p}$,

$$\sqrt{m}\left(\mathrm{vec}(\Phi)^\top D_p^\top G_n D_p \mathrm{vec}(\Phi)\right)^{-1/2} \mathrm{vec}(\Phi)^\top D_p^\top \mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) \Rightarrow \mathcal{N}(0,1)\,. \tag{99}$$

Since $D_p\mathrm{vec}(\Phi) = \mathrm{vec}\left(\tilde{\Phi}\right)$, where $\tilde{\Phi}$ is an upper triangular matrix, we only need to show that for any upper triangular $\Phi \in \mathbb{R}^{p\times p}$,

$$\sqrt{m}\left(\mathrm{vec}(\Phi)^\top G_n \mathrm{vec}(\Phi)\right)^{-1/2} \mathrm{vec}(\Phi)^\top \mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) \Rightarrow \mathcal{N}(0,1)\,.$$

For each $i \in [m]$, we denote by $s_i^\top$ the i-th row of $\sqrt{m}S_{m,n}$, where $S_{m,n}^\top = (s_1, \ldots, s_m)$. Since $S_{m,n}^\top S_{m,n} = \frac{1}{m}\sum_{i=1}^m s_i s_i^\top$, we can write

$$\begin{aligned}\sqrt{m}\mathrm{vec}(\Phi)^\top \mathrm{vec}(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p) &= \sqrt{m}\,\mathrm{tr}\left(\Phi\left(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p\right)\right)\\ &= \frac{1}{\sqrt{m}}\sum_{i=1}^m \left(s_i^\top U_n \Phi U_n^\top s_i - \mathrm{tr}(\Phi)\right).\end{aligned}$$

Denoting $\xi_{n,i} = s_i^\top U_n \Phi U_n^\top s_i - \mathrm{tr}(\Phi)$ for all $i \in [n]$, $\xi_{n,1}, \ldots, \xi_{n,m}$ are i.i.d. random variables. According to (9.8.6) in Bai and Silverstein (2010),

$$\mathrm{Var}\left[\xi_{n,1}\right] = (\kappa_{n,4} - 3)\sum_{i=1}^n (U_n\Phi U_n^\top)_{ii}^2 + 2\,\mathrm{tr}\left(U_n\Phi U_n^\top (U_n \Phi U_n^\top)^\top\right).$$

Now

$$\begin{aligned}\mathrm{tr}\left(U_n\Phi U_n^\top (U_n\Phi U_n^\top)^\top\right) &= \mathrm{tr}(\Phi U_n^\top U_n \Phi^\top U_n^\top U_n)\\ &= \mathrm{tr}(\Phi\Phi^\top) = \mathrm{vec}(\Phi)^\top \mathrm{vec}(\Phi) = \mathrm{vec}(\Phi)^\top P_p \mathrm{vec}(\Phi),\end{aligned}$$

and thus $2\,\mathrm{tr}(U_n\Phi U_n^\top (U_n\Phi U_n^\top)^\top) = \mathrm{vec}(\Phi)^\top (I_{p^2} + P_p)\mathrm{vec}(\Phi)$.

On the other hand, $(\kappa_{n,4}-3)\sum_{i=1}^n (U_n\Phi U_n^\top)_{ii}^2 = \mathrm{vec}(\Phi)^\top \Gamma_n \mathrm{vec}(\Phi)$, where for every $k_1, k_2, k_3, k_4 \in [p]$,

$$(\Gamma_n)_{(k_1k_2),(k_3k_4)} = (\kappa_{n,4} - 3)\sum_{h=1}^n (U_n)_{hk_1}(U_n)_{hk_2}(U_n)_{hk_3}(U_n)_{hk_4}.$$

Therefore, $\mathrm{Var}\left[\xi_{n,1}\right] = \mathrm{vec}(\Phi)^\top G_n \mathrm{vec}(\Phi)$. In order to apply the central limit theorem for triangular arrays $(1/\sqrt{m})\cdot\sum_{i=1}^m \left(\xi_{n,i}/\sqrt{\mathrm{Var}\left[\xi_{n,i}\right]}\right)$, we need to verify the Lindeberg-Feller condition that for every fixed $\varepsilon > 0$, $\mathbb{E}\left[\frac{\xi_{n,1}^2}{\mathrm{Var}[\xi_{n,1}]} I\left\{|\xi_{n,1}| > \varepsilon\sqrt{m\,\mathrm{Var}\left[\xi_{n,1}\right]}\right\}\right] \to 0$. With $r = 1 + \delta/4 > 1$, by an argument identical to the one in the proof of Lemma D.3, we can conclude (96) and the first inequality in (97). Then $\mathbb{E}[\xi_{n,1}^2]$ and $\mathbb{E}[\xi_{n,1}^{2r}]$ can be bounded as in that argument via Lemma B.26 in Bai and Silverstein (2010), and using that Φ does not depend on m and n.

Now consider the first condition from Condition $\mathfrak{P}$ defined in (47). One can check as in the proof of Lemma D.3, calculating similarly as in (98) that $(\delta'/2)\cdot\|\Phi\|_{\mathrm{Fr}}^2 \leqslant \mathrm{vec}(\Phi)^\top G_n \mathrm{vec}(\Phi) \leqslant (\kappa_{n,4+\delta}^{4/(4+\delta)} - 1)\|\Phi\|_{\mathrm{Fr}}^2$.

Next, consider the second condition from Condition $\mathfrak{P}$ defined in (47), so that $\kappa_{n,4} \to \infty$,

and $\kappa_{n,4+\delta}^{4/(4+\delta)}/\kappa_{n,4} = o(m^{\delta/(4+\delta)})$. Moreover, $\mathrm{vec}(\Phi)^\top G_n \mathrm{vec}(\Phi)/\kappa_{n,4} = \sum_{i=1}^n (U_n \Phi U_n^\top)_{ii}^2 + o(1)$, so $\mathrm{vec}(\Phi)^\top G_n \mathrm{vec}(\Phi)/\kappa_{n,4}$ are uniformly bounded away from zero and infinity, as in the previous proof for Lemma D.3. Therefore,

$$\mathbb{E}\left[\frac{\xi_{n,1}^2}{\mathrm{Var}\,[\xi_{n,1}]} I\left\{|\xi_{n,1}| > \varepsilon\sqrt{m\,\mathrm{Var}\,[\xi_{n,1}]}\right\}\right] \leqslant \frac{C'\kappa_{n,4r}^{1/r}}{(\varepsilon^2 m)^{1-1/r}\delta'^2} \to 0,$$

verifying the Lindeberg-Feller condition.

G.8.5 Proof of Theorem D.5

By Lemma D.8, if $(U_n)_{n\geqslant 1}$ satisfies Condition $\mathfrak{P}'$, Condition 2.8 `Sym-1` holds with $G = I_{p^2} + P_p$ and $\tau_{m,n} = 1 - \gamma_n$. Therefore, Condition 2.8 `Sym-1'` holds with $\psi = 0$. Then under Condition 2.3 `B`, we finish the proof based on Theorem 2.9.

G.8.6 Proof of Theorem D.6

From Lemma D.7, we know that Condition 2.4 `1-dim'` holds for $(U_n b_n, \bar{\varepsilon}_n)_{n\geqslant 1}$ with $\psi = 0$. Then by Theorem 2.5, we conclude the first result, and the second one follows from Theorem B.2 by the same argument.

G.8.7 Proof of Lemma D.7

If $a_n \neq \pm\tilde{a}_n$, define $a_n' = (I - a_n a_n^\top)\tilde{a}_n/\|(I - a_n a_n^\top)\tilde{a}_n\|$, otherwise, define a_n' as an arbitrary unit length vector orthogonal to a_n, so that we have (75). Thus, we can write

$$a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n = (a_n^\top \tilde{a}_n) a_n^\top S_{m,n}^\top S_{m,n} a_n + (1 - (a_n^\top \tilde{a}_n)^2)^{1/2} a_n^\top S_{m,n}^\top S_{m,n} a_n'. \tag{100}$$

Now, let $Z = (Z_1, \ldots, Z_n)^\top, \widetilde{Z} = (\widetilde{Z}_1, \cdots, \widetilde{Z}_n)^\top$ be two independent vectors, each containing independent standard normal entries. For a vector $v \in \mathbb{R}^n$ and some integers m, m' with $1 \leqslant m \leqslant m' \leqslant n$, we denote $v_{m:m'} = (v_m, \ldots, v_m')^\top$. We then claim that we have the following representation:

$$S_{m,n} a_n =_d \gamma_n^{-1/2} \frac{Z_{1:m}}{\|Z\|}, \quad S_{m,n} a_n' =_d \gamma_n^{-1/2} \frac{\widetilde{Z}_{1:m} - \frac{Z^\top \widetilde{Z}}{\|Z\|^2} Z_{1:m}}{\sqrt{\|\widetilde{Z}\|^2 - \frac{(Z^\top \widetilde{Z})^2}{\|Z\|^2}}}. \tag{101}$$

This follows because $\gamma_n^{1/2} S_{m,n}(a_n, a_n')$ has the same distribution as an $m \times 2$ submatrix of a Haar distributed matrix. By generating the Haar matrix using the Gram-Schmidt process starting from $(Z, \widetilde{Z})$, we obtain the representation from (101). Specifically, the first column is given by $w_1 = Z/\|Z\|$, and the second is obtained as $[\widetilde{Z} - (w_1^\top \widetilde{Z}) w_1]/\|\widetilde{Z} - (w_1^\top \widetilde{Z}) w_1\|$. Taking the first m coordinates of these columns gives us the representation in (101).

We define the following quantities:

$$\mathcal{X}_{n,1} = m^{-1}\|Z_{1:m}\|^2,\ \mathcal{X}_{n,2} = (n-m)^{-1}\|Z_{(m+1):n}\|^2,$$
$$\mathcal{C}_{n,1} = m^{-1}\sum_{i=1}^{m} Z_i\widetilde{Z}_i,\ \mathcal{C}_{n,2} = (n-m)^{-1}\sum_{i=m+1}^{n} Z_i\widetilde{Z}_i,$$

and $\tilde{\mathcal{X}}_{n,1}$, $\tilde{\mathcal{X}}_{n,2}$ are defined by replacing the coordinates of Z in the definitions of $\mathcal{X}_{n,1}, \mathcal{X}_{n,2}$ with the corresponding ones of $\widetilde{Z}$. According to the central limit theorem, the asymptotic distribution as $m, n-m \to \infty$ of the vector

$$\Big(\sqrt{m}(\mathcal{X}_{n,1}-1), \sqrt{n-m}(\mathcal{X}_{n,2}-1), \sqrt{m}\mathcal{C}_{n,1}, \sqrt{n-m}\mathcal{C}_{n,2}, \sqrt{m}(\tilde{\mathcal{X}}_{n,1}-1), \sqrt{n-m}(\tilde{\mathcal{X}}_{n,2}-1)\Big) \tag{102}$$

follows a multivariate normal distribution with mean zero and a diagonal covariance matrix with the vector $(2,2,1,1,2,2)$ on the diagonal. Finally, we have

$$\|S_{m,n}a_n\|^2 = \frac{n}{m}\frac{\|Z_{1:m}\|^2}{\|Z_{1:m}\|^2 + \|Z_{(m+1):n}\|^2} = \frac{\mathcal{X}_{n,1}}{\gamma_n\mathcal{X}_{n,1} + (1-\gamma_n)\mathcal{X}_{n,2}},$$

and

$$a_n^\top S_{m,n}^\top S_{m,n}a_n' = \frac{\mathcal{C}_{n,1} - \frac{\mathcal{X}_{n,1}}{\gamma_n\mathcal{X}_{n,1}+(1-\gamma_n)\mathcal{X}_{n,2}}(\gamma_n\mathcal{C}_{n,1} + (1-\gamma_n)\mathcal{C}_{n,2})}{\sqrt{\gamma_n\mathcal{X}_{n,1} + (1-\gamma_n)\mathcal{X}_{n,2}}\left[\gamma_n\mathcal{X}_{n,1} + (1-\gamma_n)\mathcal{X}_{n,2} - \frac{(\gamma_n\mathcal{C}_{n,1}+(1-\gamma_n)\mathcal{C}_{n,2})^2}{\gamma_n\mathcal{X}_{n,1}+(1-\gamma_n)\mathcal{X}_{n,2}}\right]}.$$

Combining (100) with the above two equations, we define the function $g : (0,\infty)^6 \to \mathbb{R}$ such that for all values of $\mathcal{X}_{n,p,i}$, $\mathcal{C}_{n,p,i}$, $\tilde{\mathcal{X}}_{n,p,i}$, for $i = 1,2$, we have

$$a_n^\top S_{m,n}^\top S_{m,n}\tilde{a}_n = g(\mathcal{X}_{n,1}, \mathcal{X}_{n,2}, \mathcal{C}_{n,1}, \mathcal{C}_{n,2}, \tilde{\mathcal{X}}_{n,1}, \tilde{\mathcal{X}}_{n,2}).$$

By direct calculation, we find that

$$\begin{aligned}&\nabla g(1,1,0,0,1,1)\\ &= (a_n^\top\tilde{a}_n(1-\gamma_n), -a_n^\top\tilde{a}_n(1-\gamma_n), \sqrt{1-(a_n^\top\tilde{a}_n)^2}(1-\gamma_n), -\sqrt{1-(a_n^\top\tilde{a}_n)^2}(1-\gamma_n), 0, 0).\end{aligned}$$

Using the delta method, we have

$$\begin{aligned}\sqrt{\frac{m}{1-\gamma_n}}&\Big(g(\mathcal{X}_{n,1}, \mathcal{X}_{n,2}, \mathcal{C}_{n,1}, \mathcal{C}_{n,2}, \tilde{\mathcal{X}}_{n,1}, \tilde{\mathcal{X}}_{n,2}) - g(1,1,0,0,1,1)\Big)\\ &= a_n^\top\tilde{a}_n\sqrt{1-\gamma_n}\sqrt{m}(\mathcal{X}_{n,1}-1) - a_n^\top\tilde{a}_n\sqrt{\gamma_n}\sqrt{n-m}(\mathcal{X}_{n,2}-1)\\ &+ \sqrt{1-(a_n^\top\tilde{a}_n)^2}\sqrt{1-\gamma_n}\sqrt{m}\mathcal{C}_{n,1} - \sqrt{1-(a_n^\top\tilde{a}_n)^2}\sqrt{\gamma_n}\sqrt{n-m}\mathcal{C}_{n,2} + o_P(1).\end{aligned}$$

Dividing both sides by $\sqrt{1+(a_n^\top\tilde{a}_n)^2}$ and using the limiting distribution in (102), we can conclude the desired result.

G.8.8 Proof of Lemma D.8

Let $\tilde{U}_n = (U_n, U_{n,\perp})$, so $S_{m,n}U_n = S_{m,n}\tilde{U}_n(I_p, 0)^\top$. Since the distribution of $S_{m,n}$ is rotationally invariant, i.e., $S_{m,n} =_d S_{m,n}\tilde{U}_n$, we have

$$U_n^\top S_{m,n}^\top S_{m,n} U_n = (I_p, 0)\tilde{U}_n^\top S_{m,n}^\top S_{m,n}\tilde{U}_n(I_p, 0)^\top =_d (I_p, 0)S_{m,n}^\top S_{m,n}(I_p, 0)^\top.$$

We can represent $S_0 = (ZZ^\top)^{-1/2}Z$, where Z is an $m \times n$ matrix with i.i.d. $\mathcal{N}(0, 1)$ entries, see for example, Theorem 8.2.5 in Gupta and Nagar (1999). Thus, we can write $S_0 = (Z_1Z_1^\top + Z_2Z_2^\top)^{-1/2}(Z_1, Z_2)$, where $Z_1 \in \mathbb{R}^{m\times p}$ and $Z_2 \in \mathbb{R}^{m\times(n-p)}$. Since $S_{m,n} = \gamma_n^{-1/2}S_0$, we have

$$(I_p, 0)S_{m,n}^\top S_{m,n}(I_p, 0)^\top = \gamma_n^{-1}Z_1^\top(Z_1Z_1^\top + Z_2Z_2^\top)^{-1}Z_1.$$

Applying the Sherman-Morrison-Woodbury formula, we obtain

$$P_1 := Z_1^\top(Z_1Z_1^\top + Z_2Z_2^\top)^{-1}Z_1 = Z_1^\top(Z_2Z_2^\top)^{-1}Z_1\left(I + Z_1^\top(Z_2Z_2^\top)^{-1}Z_1\right)^{-1}.$$

Now, to show that $Z_1^\top(Z_2Z_2^\top)^{-1}Z_1$ is asymptotically normal, we have that if the following almost sure limits exist:

$$\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}\left(\left(Z_2Z_2^\top/(n-p)\right)^{-1}_{ii}\right)^2 = \omega\ ;\ \lim_{m\to\infty}\frac{1}{m}\operatorname{tr}\left(\left(Z_2Z_2^\top/(n-p)\right)^{-2}\right) = \theta\ , \tag{103}$$

then by Theorem 11.8 in Yao et al. (2015) with $A_n = \left(Z_2Z_2^\top/(n-p)\right)^{-1}$ and $W = Z_1^\top$,

$$R_m := \frac{1}{\sqrt{m}}\left(Z_1^\top\left(Z_2Z_2^\top/(n-p)\right)^{-1}Z_1 - \operatorname{tr}\left(\left(Z_2Z_2^\top/(n-p)\right)^{-1}\right)I_p\right) \Rightarrow R, \tag{104}$$

where R follows a Gaussian distribution described below. By the Skorokhod representation theorem, instead of the almost sure limits in (103), it is enough to assume convergence in probability, i.e.,

$$\frac{1}{m}\sum_{i=1}^{m}\left(\left(Z_2Z_2^\top/(n-p)\right)^{-1}_{ii}\right)^2 \to_P \omega; \qquad \frac{1}{m}\operatorname{tr}\left(\left(Z_2Z_2^\top/(n-p)\right)^{-2}\right) \to_P \theta. \tag{105}$$

Let x_i, $i \in [p]$, represent i.i.d. random variables with a standard normal distribution. Then, recalling Wick's formula $\mathbb{E}(x_ix_jx_{i'}x_{j'}) = \delta_{ij}\delta_{i'j'} + \delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{ji'}$, according to Lemma 11.7 of Yao et al. (2015), the covariance matrix of R is given by

$$\operatorname{Cov}[R_{ij}, R_{i'j'}] = \omega\left\{\mathbb{E}(x_ix_jx_{i'}x_{j'}) - \delta_{ij}\delta_{i'j'}\right\} + (\theta - \omega)\left\{\delta_{ij}\delta_{i'j} + \delta_{ii'}\delta_{jj'}\right\} = \theta\left\{\delta_{ij'}\delta_{i'j} + \delta_{ii'}\delta_{jj'}\right\}.$$

Now, $Z_2Z_2^\top \sim W_m(n-p, I_m)$ follows a Wishart distribution. Hence, $(Z_2Z_2^\top)^{-1} \sim IW_m(n-p+m+1, I_m)$ follows an inverted Wishart distribution, see e.g., Theorem 3.4.1 in Gupta and Nagar (1999). Furthermore, Corollary 3.4.2.1 in Gupta and Nagar (1999) allows us to deduce that each diagonal element of $(Z_2Z_2^\top)^{-1}$ follows an inverted Gamma distribution $Q_i := (Z_2Z_2^\top)^{-1}_{ii} \sim$

InvGamma($\frac{n-p-m+1}{2}, \frac{1}{2}$). Using the properties of this distribution,

$$\mathbb{E}(n-p)^2 Q_i^2 = \frac{(n-p)^2}{(n-p-m-1)(n-p-m-3)} \to \frac{1}{(1-c)^2},$$

as well as

$$\begin{aligned}\operatorname{Var}\left[Q_i^2\right] &= \mathbb{E}Q_i^4 - (\mathbb{E}Q_i^2)^2 \\ &= \frac{8(n-p-m-4)}{(n-p-m-1)^2(n-p-m-3)^2(n-p-m-5)(n-p-m-7)};\end{aligned}$$

Applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[(n-p)^2\left(\frac{1}{m}\sum_{i=1}^m (Q_i^2 - \mathbb{E}Q_i^2)\right)\right]^2 &\leqslant \frac{(n-p)^4}{m}\mathbb{E}\sum_{i=1}^m (Q_i^2 - \mathbb{E}Q_i^2)^2 = (n-p)^4 \operatorname{Var}\left[Q_1^2\right] \\ &= \frac{8(n-p-m-4)(n-p)^4}{(n-p-m-1)^2(n-p-m-3)^2(n-p-m-5)(n-p-m-7)},\end{aligned}$$

which tends to zero as $n, m \to \infty$ with $m/n \to c > 0$. Therefore, by using Chebyshev's inequality, we conclude that

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^m \left((Z_2 Z_2^\top/(n-p))_{ii}^{-1}\right)^2 - \mathbb{E}(n-p)^2 Q_1^2 \to_P 0.$$

Consequently, we have $\omega = 1/(1-c)^2$.

Furthermore, $Z_2 Z_2^\top/(n-p) \sim W_m(n-p, \frac{1}{n-p} I_m)$, and by equation (9.3.9.g) in Section 9.3.9 of Holgersson and Pielaszkiewicz (2020) it follows that $\theta = 1/(1-c)^3$, and the proof of (105) is complete. As a result, (104) follows.

Moreover, based on Theorem 4.2 in Pielaszkiewicz and Holgersson (2020), we conclude that

$$\frac{n-p}{\sqrt{m}}\left(\operatorname{tr}\left((Z_2 Z_2^\top)^{-1}\right) - \frac{m}{n-m-p-1}\right) \to_P 0.$$

Thus $\sqrt{m}\left(Z_1^\top (Z_2 Z_2^\top)^{-1} Z_1 - \frac{m}{n-m-p-1} I_p\right) \Rightarrow cR$. By the delta method, we can deduce that

$$\sqrt{m}\left(Z_1^\top (Z_2 Z_2^\top)^{-1} Z_1 (I + Z_1^\top (Z_2 Z_2^\top)^{-1} Z_1)^{-1} - \frac{m}{n-p-1} I_p\right) \Rightarrow (1-c)^2 cR.$$

Hence, $\sqrt{m}\left(P_1 - \frac{m}{n-p-1} I_p\right) \Rightarrow (1-c)^2 cR$, and thus

$$\sqrt{m}\left(\frac{m}{n} U_n^\top S_{m,n}^\top S_{m,n} U_n - \frac{m}{n-p-1} I_p\right) \Rightarrow (1-c)^2 cR.$$

Equivalently, $\sqrt{m}\left(U_n^\top S_{m,n}^\top S_{m,n} U_n - I_p\right) \Rightarrow (1-c)^2 R$. Since the covariance matrix of $\operatorname{vec}(R)$ is $\theta(I_{p^2} + P_p)$, that of $\operatorname{vec}\left((1-c)^2 R\right)$ is $(1-c)^4\theta(I_{p^2} + P_p) = (1-c)(I_{p^2} + P_p)$. Therefore, by Slutsky's theorem, Condition 2.8 holds for $G = I_{p^2} + P_p$.

G.9 Other proofs

G.9.1 Proof of Lemma 3.14

Proof. For brevity, we denote $S_{1,\tilde{m},n}$ and $S_{2,m,\tilde{m}}$ by S_1 and S_2, respectively. By Lemma D.3, conditionally on $S_1 a_n$ and $S_1 \tilde{a}_n$, we have

$$\sqrt{\frac{m}{\|S_1 a_n\|^2 \|S_1 \tilde{a}_n\|^2 + (a_n^\top S_1^\top S_1 \tilde{a}_n)^2}} \left(a_n^\top S_1^\top S_2^\top S_2 S_1 \tilde{a}_n - a_n^\top S_1^\top S_1 \tilde{a}_n \right) \Rightarrow \mathcal{N}(0,1).$$

The result is still true unconditionally over the randomness in $S_1 a_n$ and $S_1 \tilde{a}_n$. Using (25) we find

$$\sqrt{\tilde{m}}(a_n^\top S_1^\top S_1 \tilde{a}_n - a_n^\top \tilde{a}_n) = O_P(1). \tag{106}$$

From this it can be readily verified that

$$\sqrt{\tilde{m}} \left(\|S_1 a_n\|^2 \|S_1 \tilde{a}_n\|^2 + (a_n^\top S_1^\top S_1 \tilde{a}_n)^2 - (1 + (a_n^\top \tilde{a}_n)^2) \right) = O_P(1).$$

Therefore we have

$$\sqrt{\frac{m}{1 + (a_n^\top \tilde{a}_n)^2}} \left(a_n^\top S_1^\top S_2^\top S_2 S_1 \tilde{a}_n - a_n^\top S_1^\top S_1 \tilde{a}_n \right) \Rightarrow \mathcal{N}(0,1).$$

This combined with (106) and the condition $\tilde{m}/m \to 0$ completes the proof. □

G.9.2 Proof of Theorem 4.2

Proof. Using the multivariate version of Taylor's theorem, we have

$$\begin{aligned}
\Delta_{t+1} &= \Delta_t - \tilde{H}_t^{-1}(\nabla f(\theta_t) - \nabla f(\theta_*)) \\
&= \Delta_t - \tilde{H}_t^{-1} \cdot \nabla^2 f(\theta_*)\Delta_t - \tilde{H}_t^{-1} \left[\int_0^1 (1-u) \nabla^3 f(\theta_* + u(\theta_t - \theta_*)) du \right] (\theta_t - \theta_*)^2 \\
&= (I - \tilde{H}_t^{-1} H_*)\Delta_t + \|\tilde{H}_t^{-1}\| O(p^{3/2} \|\Delta_t\|^2).
\end{aligned} \tag{107}$$

Here $\nabla^3 f$ denotes the third order derivative tensor of f and the notation $\nabla^3 f(\theta)(v)^2$ for some vector v refers to a vector whose i-th entry is given by $\sum_{j,k} \frac{\partial^3 f(\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} v_j v_k$. In the last step we use $\max_{\sum \alpha_i = 3} \frac{\partial^3 f}{\partial\theta_1^{\alpha_1} \cdots \partial\theta_p^{\alpha_p}}$ is bounded and Big-O notation here is interpreted as $\|\Delta_{t+1} - (I - \tilde{H}_t^{-1} H_*)\Delta_t)\| = \|\tilde{H}_t^{-1}\| O(p^{3/2} \|\Delta_t\|^2)$.

Let $\tilde{H}_{t*} = H_{*,l} S_t^\top S_t H_{*,r}$. We can bound the difference between $\tilde{H}_t = H_{t,l} S_t^\top S_t H_{t,r}$ and $\tilde{H}_{t*}$ locally near θ_* first. If follows from the triangle inequality that

$$\|\tilde{H}_t - \tilde{H}_{t*}\| \leqslant \|(H_{t,l} - H_{*,l}) S_t^\top S_t H_{t,r}\| + H_{*,l} S_t^\top S_t (H_{t,r} - H_{*,r})\|.$$

By the Lipschitz continuity of H_l and H_r, the Bounded Quadratic Form condition, and due to the independence between $H_{t,l}$ and S_t, it follows that with respect to the randomness in S_t, we have $\|\tilde{H}_t - \tilde{H}_{t*}\| = O_P(\|\Delta_t\|)$. Then

$$\tilde{H}_t^{-1} H_* - \tilde{H}_{t*}^{-1} H_* = O_P(\|\Delta_t\|). \tag{108}$$

Let $H_{*,r} = U_* L_* V_*^\top$. We write

$$\tilde{H}_{t*}^{-1} H_* = (V_* L_* U_*^\top S_t^\top S_t U_* L_* V_*)^{-1} V_* L_*^2 V_*^\top = V_* L_*^{-1} (U_*^\top S_t^\top S_t U_*)^{-1} L_* V_*^\top,$$

and thus

$$I_p - \tilde{H}_{t*}^{-1} H_* = V_* L_*^{-1} \left[I - (U_*^\top S_t^\top S_t U_*)^{-1} \right] L_* V_*^\top.$$

For U satisfying $U^\top U = I_p$, a Taylor series expansion of $U^\top S_{m,n}^\top S_{m,n} U \mapsto (U^\top S_{m,n}^\top S_{m,n} U)^{-1}$ at I_p shows

$$\begin{aligned}
&I_p - (U^\top S_{m,n}^\top S_{m,n} U)^{-1} \\
&= (U^\top S_{m,n}^\top S_{m,n} U - I_p) - (I_p - U^\top S_{m,n}^\top S_{m,n} U)^2 (U^\top S_{m,n}^\top S_{m,n} U)^{-1} \\
&= U^\top (S_{m,n}^\top S_{m,n} - I_p) U + O_P\left(\frac{1}{m}\right).
\end{aligned}$$

Therefore, we have

$$I_p - \tilde{H}_{t*}^{-1} H_* = V_* L_*^{-1} \left[U_*^\top (S_{t,m,n}^\top S_{t,m,n} - I_n) U_* \right] L_* V_*^\top + O_P\left(\frac{1}{m}\right).$$

Combining this with (108), and applying (107), we obtain

$$m^{1/2} \Delta_{t+1} = m^{1/2} V_* L_*^{-1} \left[U_*^\top (S_{t,m,n}^\top S_{t,m,n} - I_n) U_* \right] L_* V_*^\top \Delta_t + O(m^{-1/2} \|\Delta_t\|) + O(m^{1/2} \|\Delta_t\|^2).$$

This iteration together with $\|\Delta_{t_0}\| = o(m^{-1/2})$ implies that $\|m^{1/2} \Delta_{t+1}\| = O_P(\|\Delta_t\|)$. Then we can check by induction that for any $t > t_0$ such that $t - t_0$ is bounded,

$$m^{(t-t_0)/2} \Delta_t = m^{(t-t_0)/2} V_* L_*^{-1} \Omega_{t-1} \cdots \Omega_{t_0} L_* V_*^\top \Delta_{t_0} + m^{1/2} O(\|\Delta_{t_0}\|^2). \tag{109}$$

Dividing both sides by $\|\Delta_{t_0}\|$, we conclude (35).

Multiplying both sides of (109) on the left by $L_* V_*^\top$, dividing by $\|L_* V_*^\top \Delta_{t_0}\|$, and then considering the square of L_2-norm of both sides, we obtain

$$\frac{m^{(t-t_0)} \|\Delta_t\|_{H_*}^2}{\|\Delta_{t_0}\|_{H_*}^2} \overset{d}{\approx} w^\top \left\{ \prod_{i=t_0}^{t-1} \Omega_i \right\} \left\{ \prod_{i=t-1}^{t_0} \Omega_i \right\} w$$

where $w = L_* V_*^\top \Delta_0 / \|L_* V_*^\top \Delta_0\|$. Note that if U_* satisfies the conditions $\mathfrak{P}'$ for either Hadamard or i.i.d. sketching, then $U_* O$ for any orthogonal matrix O still satisfies the corresponding assumption. Moreover, replacing U_* with $U_* O$ will not affect the asymptotic distributions. Therefore we can replace w with the canonical basis vector e_1. This concludes the proof.

□

H PCA after Sketching with i.i.d. Gaussian Matrices

Here we present how our result from Section D.1 simplifies for a Gaussian sketch, recovering results of Anderson (1963) for individual eigenvalues.

Corollary H.1 (Inference for Eigenvalues/vectors with i.i.d. Gaussian Sketching)**.** *Let* $(X_n)_{n\geqslant 1}$ *satisfy Condition 2.3* B*, and the sketching matrices* $m^{1/2}S_{m,n}$ *have i.i.d. standard normal entries. Then,*

$$\sqrt{\frac{m}{2}}\hat{\Lambda}_{m,i}^{-1}\left(\hat{\Lambda}_{m,i}-\Lambda_{n,i}\right)\Rightarrow\mathcal{N}(0,1), \tag{110}$$

and for any vector $c\in\mathbb{S}^{p-1}$ *satisfying* $\limsup_{n\to\infty}(c^\top v_{n,i})^2<1$*,*

$$\sqrt{m}\left(\sum_{k\neq i}\frac{\hat{\Lambda}_{m,i}\hat{\Lambda}_{m,n,k}}{(\hat{\Lambda}_{m,i}-\hat{\Lambda}_{m,n,k})^2}(c^\top\hat{v}_{m,n,k})^2\right)^{-1/2}c^\top(\hat{v}_{m,n,i}-v_{n,i})\Rightarrow\mathcal{N}(0,1). \tag{111}$$

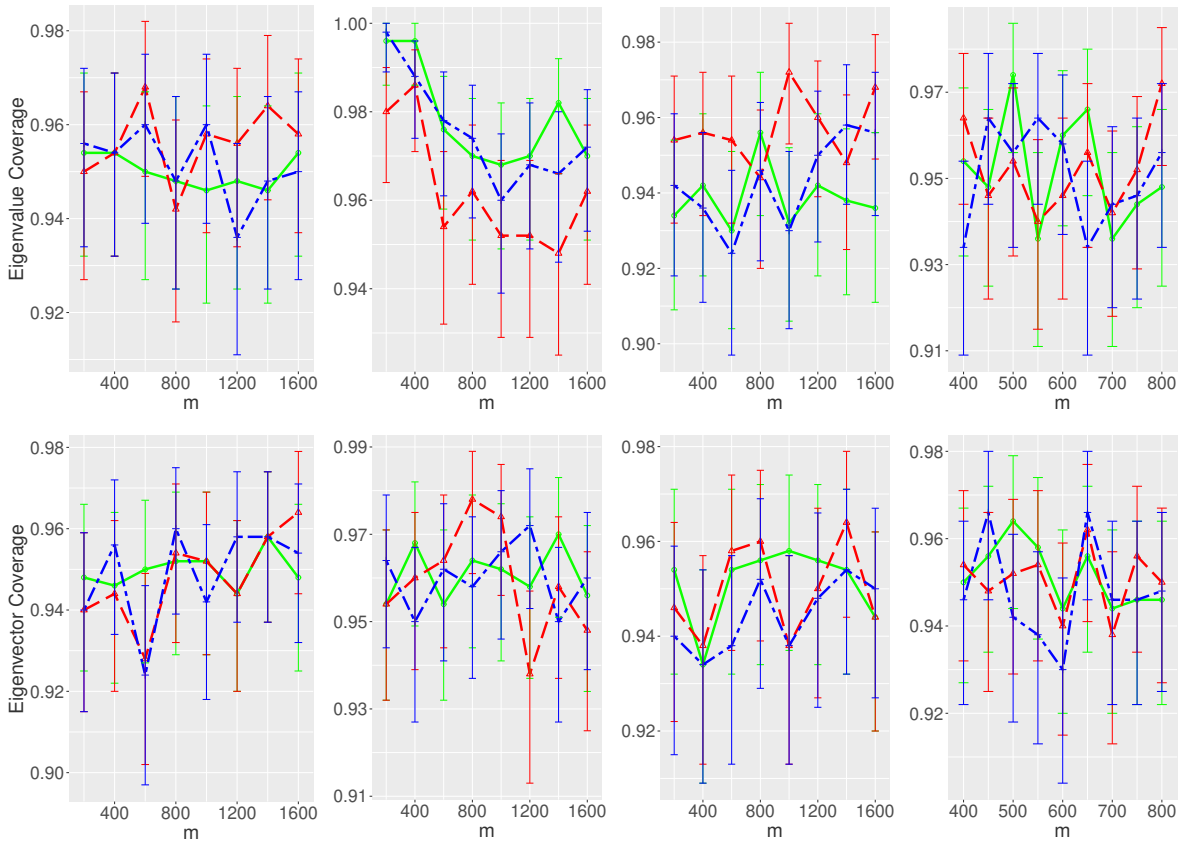
This result is implied by Theorem D.1.

I Additional Experimental Results

In this section, we present some additional experimental results. For each setting, we use 500 Monte Carlo trials. For the simulated datasets given in Section 5, we use $p=15, n=2048, c=(1,0,\ldots,0)^\top$ as in the figures in the main text, and vary m from 200 to 1600.

In Figures 8 and 9, we show coverage results of 95% intervals using sketched PCA and LS, for various sketching methods and datasets; along with 95% Clopper-Pearson intervals for the coverage. We provide details in the figure caption.

In Figure 10, we show the empirical variances of sketched PCA solutions over 500 Monte Carlo trials using various sketching methods and datasets.



(a) $k = 2$, Case 1 (b) $k = 1$, Case 2 (c) $k = 1$, Case 3 (d) $k = 1$, HGDP Dataset

Figure 8: Coverage rates of λ_k and $c^\top v_k$ in a variety of cases, using the same protocol as in Figure 3: (a) $k = 2$, Case 1. (b) $k = 1$, Case 2. (c) $k = 1$, Case 3. (d) $k = 1$, HGDP dataset.

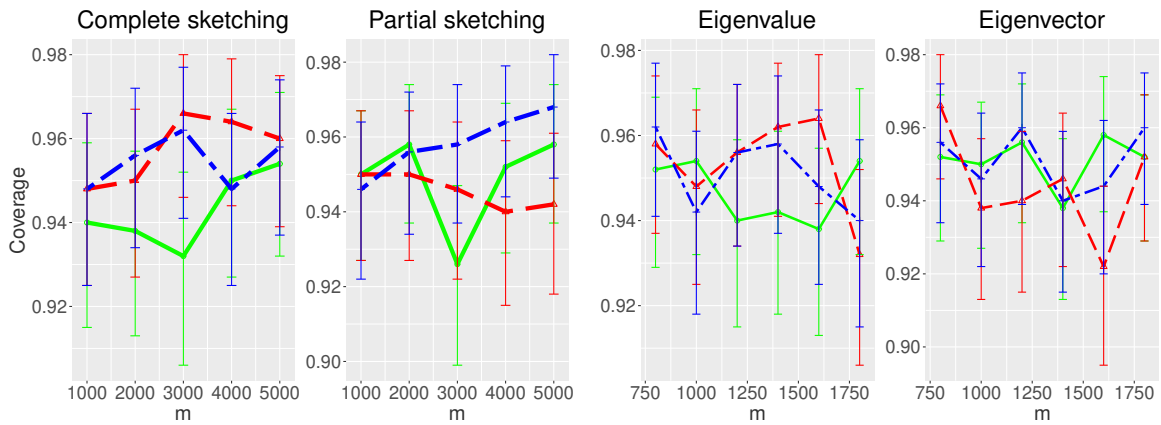
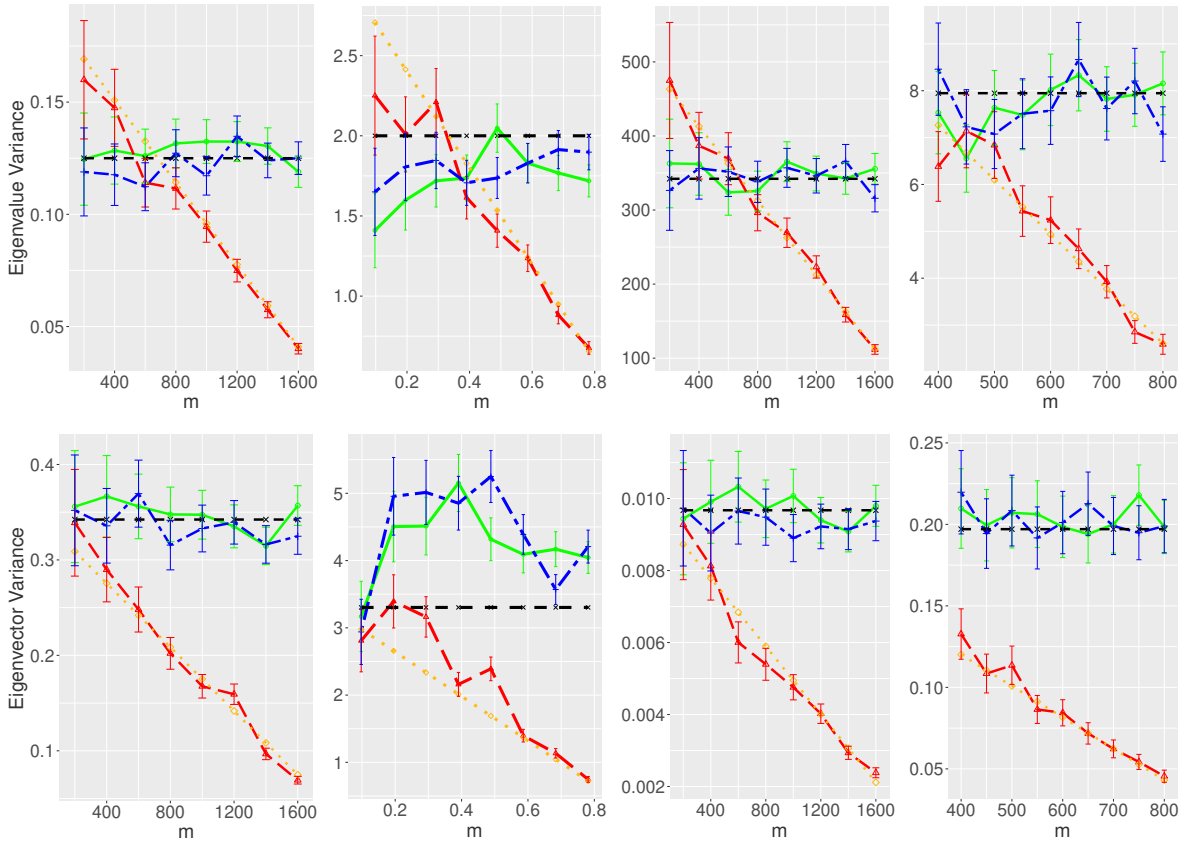


Figure 9: Coverage of 95% intervals for the first coordinate of the least squares solution using the New York flights dataset; as well as the largest eigenvalue and associated eigenvector from the Million Songs dataset. Also shown are 95% Clopper-Pearson intervals for the coverage, where we use the same protocol as in Figure 3.



(a) $k = 2$, Case 1 (b) $k = 1$, Case 2 (c) $k = 1$, Case 3 (d) $k = 1$, HGDP Dataset

Figure 10: The variance of $\sqrt{m}\hat{\Lambda}_{m,n,k}$ and $\sqrt{m}c^\top \hat{v}_{m,n,k}$ for various cases, using the same protocol as in Figure 6: (a) $k = 2$, Case 1. (b) $k = 1$, Case 2. (c) $k = 1$, Case 3. (d) $k = 1$, HGDP dataset.