THE TWO-TO-INFINITY NORM AND SINGULAR SUBSPACE GEOMETRY WITH APPLICATIONS TO HIGH-DIMENSIONAL STATISTICS[1]

BY JOSHUA CAPE, MINH TANG AND CAREY E. PRIEBE

Johns Hopkins University

The singular value matrix decomposition plays a ubiquitous role throughout statistics and related fields. Myriad applications including clustering, classification, and dimensionality reduction involve studying and exploiting the geometric structure of singular values and singular vectors.

This paper provides a novel collection of technical and theoretical tools for studying the geometry of singular subspaces using the two-to-infinity norm. Motivated by preliminary deterministic Procrustes analysis, we consider a general matrix perturbation setting in which we derive a new Procrustean matrix decomposition. Together with flexible machinery developed for the two-to-infinity norm, this allows us to conduct a refined analysis of the induced perturbation geometry with respect to the underlying singular vectors even in the presence of singular value multiplicity. Our analysis yields singular vector entrywise perturbation bounds for a range of popular matrix noise models, each of which has a meaningful associated statistical inference task. In addition, we demonstrate how the two-to-infinity norm is the preferred norm in certain statistical settings. Specific applications discussed in this paper include covariance estimation, singular subspace recovery, and multiple graph inference.

Both our Procrustean matrix decomposition and the technical machinery developed for the two-to-infinity norm may be of independent interest.

1. Introduction.

1.1. *Background.* The geometry of singular subspaces is of fundamental importance throughout a wide range of fields including statistics, machine learning, computer science, applied mathematics, and network science. Singular vectors (resp., eigenvectors) together with their corresponding subspaces and singular values (resp., eigenvalues) appear throughout various statistical applications including

---

principal component analysis [7, 27, 29, 39], covariance matrix estimation [19–21, 26], spectral clustering [31, 45, 47, 56], and graph inference [51, 52, 54].

Singular subspaces are also studied in random matrix theory, a discipline which has come to have a profound influence on the development of high-dimensional statistical theory [2, 59]. In random matrix theory, topics of interest include the phenomenon of eigenvector delocalization [46] as well as the spectral behavior of (often low rank) matrices undergoing random perturbations [3, 41]. For an overview of recent work on the properties of eigenvectors of random matrices, see [40]. Further discussion of how random matrix theory has come to have an impact on statistics can be found in [42].

From a computational perspective, optimization algorithms in signal processing and compressed sensing are often concerned with the behavior of singular vectors and subspaces [17]. The study of algorithmic performance on manifolds and manifold learning, especially involving the Grassmann and Stiefel manifolds, motivates related interest in a collection of Procrustes-type problems [4, 6, 16]. Procrustes analysis occupies an established area within the theoretical study of statistics on manifolds [12] and arises in applications including diffusion tensor imaging [14] and shape analysis [15]. See [23] for an extended treatment of both theoretical and numerical aspects concerning Procrustes-type problems.

Foundational results from matrix theory concerning perturbations of singular values and vectors date back to the original work of Weyl [58], Davis and Kahan [13], and Wedin [57], among others. Indeed, these results form the backbone of much of the linear algebraic machinery that has since been developed for use in statistics. The references [5, 25, 48] provide an overview of classical perturbation results and historical developments.

1.2. *Overview.* This paper provides a collection of technical and theoretical tools for studying the perturbations of singular vectors and subspaces with respect to the two-to-infinity norm (defined below). Our main theoretical results are first presented quite generally and then followed by concrete consequences thereof to facilitate direct statistical applications. In this work, we establish perturbation bounds for both low and high rank matrices. Among the advantages of our methods is that we allow singular value multiplicity and merely leverage a population singular value gap assumption in the spirit of [61].

As a special case of our general framework and methods, we improve upon results in [21] wherein the authors obtain an ℓ_∞ norm perturbation bound for singular vectors of low rank matrices exhibiting specific coherence structure. In this way, beyond the stated theorems in this paper, our results can be applied analogously to robust covariance estimation involving heavy-tailed random variables.

Our Procrustes analysis complements the study of perturbation bounds for singular subspaces in [8]. When considered in tandem, we demonstrate a Procrustean setting in which one recovers nearly rate-matching upper and lower bounds with respect to the two-to-infinity norm.

Yet another contribution of this work is that we extend and complement current spectral methodology for graph inference and embedding [32, 37, 51]. To the best of our knowledge, we obtain among the first-ever estimation bounds for multiple graph inference in the presence of edge correlation.

1.3. *Setting.*   This paper formulates and analyzes a general matrix decomposition for the aligned difference between real matrices U and $\hat{U}$, each consisting of r orthonormal columns (i.e., partial isometries; Stiefel matrices; orthogonal r-frames), given by

$$(1.1) \qquad\qquad \hat{U} - UW,$$

where W denotes an $r \times r$ orthogonal matrix. We focus on (but are strictly speaking not limited to) a certain "nice" choice of W which corresponds to an "optimal" Procrustes transformation in a sense that will be made precise.

Along with the matrix decomposition considerations presented in this paper, we provide technical machinery for the two-to-infinity subordinate vector norm on matrices, which for $A \in \mathbb{R}^{p_1 \times p_2}$ is given by

$$(1.2) \qquad\qquad \|A\|_{2\to\infty} := \sup_{\|x\|_2=1} \|Ax\|_\infty.$$

Together, these allow us to obtain a suite of perturbation bounds within an additive perturbation framework of the singular value decomposition.

The two-to-infinity norm yields finer uniform control on the entries of a matrix than the more common spectral and Frobenius norms. We shall demonstrate that, in certain settings, the two-to-infinity norm is preferable to these and to other norms. In particular, matrices exhibiting bounded coherence in the sense of [9] form a popular and widely-encountered class of matrices for which the two-to-infinity norm is demonstrably an excellent choice.

The two-to-infinity norm has previously appeared in the statistics literature, including in [37] wherein it is leveraged to prove that adjacency spectral embedding achieves perfect clustering for certain stochastic block model graphs. More recently, it has also appeared in the study of random matrices when a fraction of the matrix entries are modified [44]. In general, however, the two-to-infinity norm has received far less attention than other norms. Among the aims of this paper is to advocate for the more widespread consideration of the two-to-infinity norm.

1.4. *Sample application*: *covariance estimation.*   We pause here to present an application of our work and methods to estimating the top singular vectors of a structured covariance matrix.

Denote a random vector Y by $Y := (Y^{(1)}, Y^{(2)}, \ldots, Y^{(d)})^\top \in \mathbb{R}^d$, and let $Y, Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed (i.i.d.) mean zero multivariate Gaussian random (column) vectors with common covariance matrix $\Gamma \in \mathbb{R}^{d \times d}$. Denote the singular value decomposition of Γ by $\Gamma \equiv U\Sigma U^\top + U_\perp \Sigma_\perp U_\perp^\top$,

where $[U|U_\perp] \equiv [u_1|u_2|\ldots|u_d] \in \mathbb{R}^{d \times d}$ is an orthogonal matrix. The singular values of Γ are indexed in nonincreasing order, $\sigma_1(\Gamma) \geq \sigma_2(\Gamma) \geq \cdots \geq \sigma_d(\Gamma)$, with $\Sigma := \operatorname{diag}(\sigma_1(\Gamma), \sigma_2(\Gamma), \ldots, \sigma_r(\Gamma)) \in \mathbb{R}^{r \times r}$, $\Sigma_\perp := \operatorname{diag}(\sigma_{r+1}(\Gamma), \sigma_{r+2}(\Gamma), \ldots, \sigma_d(\Gamma)) \in \mathbb{R}^{(d-r) \times (d-r)}$, $\delta_r(\Gamma) := \sigma_r(\Gamma) - \sigma_{r+1}(\Gamma) > 0$, and $r \ll d$. Here, Σ may be viewed as containing the "signal" (i.e., spike) singular values of interest, while $\Sigma_\perp$ contains the remaining "noise" (i.e., bulk) singular values. The singular values of Γ are not assumed to be distinct; rather, the assumption $\delta_r(\Gamma) > 0$ simply specifies a singular value population gap between Σ and $\Sigma_\perp$.

Let $\hat{\Gamma}_n$ denote the empirical covariance matrix $\hat{\Gamma}_n := \frac{1}{n} \sum_{k=1}^{n} Y_k Y_k^\top$ with decomposition $\hat{\Gamma}_n \equiv \hat{U} \hat{\Sigma} \hat{U}^\top + \hat{U}_\perp \hat{\Sigma}_\perp \hat{U}_\perp^\top$. Let $E_n := \hat{\Gamma}_n - \Gamma$ denote the difference between the empirical and theoretical covariance matrices. Below, $C, c, c_1, c_2, \ldots$ are positive constants (possibly related) of minimal interest. We use $\operatorname{Var}(Y^{(i)})$ to denote the variance of $Y^{(i)}$ and $o(\cdot)$ to denote little-o notation.

We are interested in the regime where the sample size n and the covariance matrix dimension d are simultaneously allowed to grow. In this regime, an important measure of complexity is given by the effective rank of Γ, defined as $\mathfrak{r}(\Gamma) := \operatorname{trace}(\Gamma)/\sigma_1(\Gamma)$ [29].

THEOREM 1.1 (Application: covariance estimation). *In Section* 1.4, *assume that* $\max\{\mathfrak{r}(\Gamma), \log d\} = o(n)$, $\sigma_1(\Gamma)/\sigma_r(\Gamma) \leq c_1$, $\delta_r(\Gamma) \geq c_2 \sigma_r(\Gamma) > 0$, *and* $\|U\|_{2\to\infty} \leq c_3 \sqrt{r/d}$. *Let* $\nu(Y) := \max_{1 \leq i \leq d} \sqrt{\operatorname{Var}(Y^{(i)})}$. *Then, there exists an* $r \times r$ *orthogonal matrix* W_U *and a constant* $C > 0$ *such that with probability at least* $1 - d^{-2}$,

$$\|\hat{U} - U W_U\|_{2\to\infty} \leq C \sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}} \left(\frac{\nu(Y)r}{\sqrt{\sigma_r(\Gamma)}} + \frac{\sigma_{r+1}(\Gamma)}{\sigma_r(\Gamma)} \right)$$
$$+ C \left(\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n} \right) \left(\sqrt{\frac{\sigma_{r+1}(\Gamma)}{\sigma_r(\Gamma)}} + \sqrt{\frac{r}{d}} \right).$$

REMARK 1.2. In the setting of Theorem 1.1, spectral norm probabilistic concentration [28, 29] can be applied to yield a naïve two-to-infinity norm bound of the form

$$(1.3) \qquad \|\hat{U} - U W_U\|_{2\to\infty} \leq C \sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}}.$$

When Γ exhibits the additional spike structure $\Gamma \equiv U(\Lambda + c^2 I)U^\top + c^2 U_\perp U_\perp^\top$ with $\sigma_1(\Gamma) \geq c_4(d/r)$, then $\sqrt{\sigma_{r+1}(\Gamma)}, \nu(Y) \leq c_5 \sqrt{\sigma_1(\Gamma)} \sqrt{r/d}$, and so the bound in Theorem 1.1 simplifies to the form

$$(1.4) \qquad \|\hat{U} - U W_U\|_{2\to\infty} \leq C \sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}} \sqrt{\frac{r^3}{d}}.$$

The bound in equation (1.4) manifestly improves upon equation (1.3) since here $r \ll d$ and d is taken to be large.

1.5. *Organization.*   The rest of this paper is organized as follows. Section 2 establishes notation, motivates the use of the two-to-infinity norm in a Procrustean context, and presents the additive matrix perturbation model considered in this paper. Section 3 collects our main results which fall into two categories: matrix decompositions and matrix perturbation theorems. Section 4 demonstrates how this paper improves upon and complements existing work in the literature by way of considering three statistical applications involving covariance estimation (see also Section 1.4), singular subspace recovery, and multiple graph inference. Section 5 offers some final remarks. Section 6 contains the technical machinery developed for this paper as well as proofs of our main theorems.

2. Preliminaries.

2.1. *Notation.*   All vectors and matrices in this paper are taken to be real-valued. The symbols $:=$ and $\equiv$ are used to assign definitions and to denote formal equivalence. For any positive integer n, let $[n] := \{1, 2, \ldots, n\}$. We use C_α to denote a general constant that may change from line to line unless otherwise specified and that possibly depends only on α (either a parameter or an indexing value). Let $\mathcal{O}(\cdot)$ denote standard big-O notation, possibly with an underlying probabilistic qualifying statement. These conventions are simultaneously upheld when we write $\mathcal{O}_\alpha(\cdot)$. We let $\mathbb{O}_{p,r}$ denote the set of all $p \times r$ real matrices with orthonormal columns so that $\mathbb{O}_p \equiv \mathbb{O}_{p,p}$ denotes the set of orthogonal matrices in $\mathbb{R}^{p \times p}$.

For (column) vectors $x, y \in \mathbb{R}^{p_1}$ where $x \equiv (x_1, \ldots, x_{p_1})^\top$, the standard Euclidean inner product between x and y is denoted by $\langle x, y \rangle$. The classical ℓ_p vector norms are given by $\|x\|_p := (\sum_{i=1}^{p} |x_i|^p)^{1/p}$ for $1 \le p < \infty$, and $\|x\|_\infty := \max_i |x_i|$. This paper also makes use of several standard matrix norms. Letting $\sigma_i(A)$ denote the ith largest singular value of A, then $\|A\|_2 := \sigma_1(A)$ denotes the spectral norm of A, $\|A\|_F := \sqrt{\sum_i \sigma_i^2(A)}$ denotes the Frobenius norm of A, $\|A\|_1 := \max_j \sum_i |a_{ij}|$ denotes the maximum absolute column sum of A, and $\|A\|_\infty := \max_i \sum_j |a_{ij}|$ denotes the maximum absolute row sum of A. Additionally, we consider $\|A\|_{\max} := \max_{i,j} |a_{ij}|$.

2.2. *Norm relations.*   A central focus of this paper is on the two-to-infinity norm, which for the matrix A is given by $\|A\|_{2\to\infty} := \sup_{\|x\|_2=1} \|Ax\|_\infty$. Proposition 6.1 establishes the elementary fact that this norm corresponds to the maximum Euclidean row norm of A. As such, the two-to-infinity norm is easy to interpret and straightforward to compute. In certain settings, $\|\cdot\|_{2\to\infty}$ will be shown to serve as an attractive surrogate for $\|\cdot\|_{\max}$ in light of additional algebraic properties that $\|\cdot\|_{2\to\infty}$ enjoys.

For $A \in \mathbb{R}^{p_1 \times p_2}$, the standard relations between the ℓ_p norms for $p \in \{1, 2, \infty\}$ permit a quantitative comparison of $\|\cdot\|_{2\to\infty}$ to $\|\cdot\|_{\max}$ and $\|\cdot\|_\infty$. In particular,

these matrix norms are related through matrix column dimension via

$$\frac{1}{\sqrt{p_2}}\|A\|_{2\to\infty} \le \|A\|_{\max} \le \|A\|_{2\to\infty} \le \|A\|_\infty \le \sqrt{p_2}\|A\|_{2\to\infty}.$$

In contrast, the relationship between $\|\cdot\|_{2\to\infty}$ and $\|\cdot\|_2$ depends on the matrix row dimension (see Proposition 6.3) via

$$\|A\|_{2\to\infty} \le \|A\|_2 \le \sqrt{p_1}\|A\|_{2\to\infty}.$$

As an example, consider the rectangular matrix $A := \{1/\sqrt{p_2}\}^{p_1 \times p_2}$, for which $\|A\|_{2\to\infty} = 1$ while $\|A\|_2 = \|A\|_F = \sqrt{p_1}$. This example, together with the above norm relations, demonstrates that possibly $\|A\|_{2\to\infty} \ll \|A\|_2$ when the row dimension of A is large relative to the column dimension, that is, $p_1 \gg p_2$. Bounding $\|A\|_{2\to\infty}$ would then be preferred to bounding $\|A\|_2$ when seeking more refined (e.g., entrywise) control of A. The same observation holds with respect to the Frobenius norm which satisfies the well-known, rank-based relation with the spectral norm given by

$$\|A\|_2 \le \|A\|_F \le \sqrt{\mathrm{rank}(A)}\|A\|_2.$$

We pause to point out that the two-to-infinity norm is not in general submultiplicative for matrices. Moreover, the "constrained" submultiplicative behavior of $\|\cdot\|_{2\to\infty}$ (see Proposition 6.5), when taken together with the noncommutativity of matrix multiplication and standard properties of more common matrix norms, yields substantial flexibility when bounding matrix products and passing between norms. For this reason, a host of bounds beyond those presented in this paper follow naturally from the matrix decomposition results in Section 3.1. The relative strength of derived bounds will depend upon underlying, application-specific properties and assumptions.

2.3. *Singular subspaces and Procrustes analysis.* Let $\mathcal{U}$ and $\hat{\mathcal{U}}$ denote the subspaces for which the columns of $U, \hat{U} \in \mathbb{O}_{p,r}$ form orthonormal bases, respectively. From the classical CS matrix decomposition, a natural measure of distance between these subspaces (corresponding matrices) is given via the canonical (i.e., principal) angles between $\mathcal{U}$ and $\hat{\mathcal{U}}$. More specifically, for the singular values of $U^\top \hat{U}$, denoted by $\{\sigma_i(U^\top \hat{U})\}_{i=1}^r$ and indexed in nonincreasing order, the canonical angles are the main diagonal elements of the $r \times r$ diagonal matrix

$$\Theta(\hat{U}, U) := \mathrm{diag}\big(\cos^{-1}\big(\sigma_1(U^\top \hat{U})\big), \cos^{-1}\big(\sigma_2(U^\top \hat{U})\big), \ldots, \cos^{-1}\big(\sigma_r(U^\top \hat{U})\big)\big).$$

For an in-depth review of the CS decomposition and canonical angles see, for example, [5, 48]. An extensive summary of the relationships between $\sin\Theta$ distances, specifically $\|\sin\Theta(\hat{U}, U)\|_2$ and $\|\sin\Theta(\hat{U}, U)\|_F$, as well as various other distance measures, is provided in [8]. This paper focuses on $\sin\Theta$ distance in relation to Procrustes analysis.

Given two matrices A and B together with a set of matrices $\mathbb{S}$ and a norm $\|\cdot\|$, a general version of the Procrustes problem is to investigate

$$\inf_{S \in \mathbb{S}} \|A - BS\|.$$

For $U, \hat{U} \in \mathbb{O}_{p,r}$ and $\eta \in \{\max, 2 \to \infty, 2, F\}$, this paper specifically considers

$$(2.1) \qquad \inf_{W \in \mathbb{O}_r} \|\hat{U} - UW\|_{\eta}.$$

For each choice of η, the corresponding infimum in equation (2.1) is provably achieved by the compactness of $\mathbb{O}_r$ together with properties of norms in finite-dimensional vector spaces. As such, let $W_{\eta}^{\star} \in \mathbb{O}_r$ denote a corresponding Procrustes solution under η (where dependence upon the underlying matrices U and $\hat{U}$ is implicit from context). Unfortunately, these solutions are not analytically tractable in general, save under the Frobenius norm, in which case $W_U \equiv W_F^{\star}(U, \hat{U})$ corresponds to the the classical orthogonal Procrustes problem solution [23] given explicitly by $W_U \equiv W_1 W_2^{\top}$ when the singular value decomposition of $U^{\top}\hat{U} \in \mathbb{R}^{r \times r}$ is written as $U^{\top}\hat{U} \equiv W_1 \Sigma_U W_2^{\top}$.

For each η, it is therefore natural to study the behavior of

$$(2.2) \qquad \|\hat{U} - UW_U\|_{\eta}.$$

To this end, $\sin \Theta$ distances and the above Procrustes problems are related in the sense that (e.g., [8])

$$\big\|\sin \Theta(\hat{U}, U)\big\|_F \leq \|\hat{U} - UW_U\|_F \leq \sqrt{2}\big\|\sin \Theta(\hat{U}, U)\big\|_F$$

and

$$\big\|\sin \Theta(\hat{U}, U)\big\|_2 \leq \big\|\hat{U} - UW_2^{\star}\big\|_2 \leq \|\hat{U} - UW_U\|_2 \leq \sqrt{2}\big\|\sin \Theta(\hat{U}, U)\big\|_2.$$

By Lemma 6.8, $\|\hat{U} - UW_U\|_2$ can be bounded differently in a manner suggesting that the performance of W_U is "close" to the performance of $W_2^{\star}$ under $\|\cdot\|_2$, namely

$$\|\hat{U} - UW_U\|_2 \leq \big\|\sin \Theta(\hat{U}, U)\big\|_2 + \big\|\sin \Theta(\hat{U}, U)\big\|_2^2.$$

Loosely speaking, it follows that the discrepancy between W_U and $W_2^{\star}$ in the spectral norm Procrustes problem behaves as $\mathcal{O}(\|\sin \Theta(\hat{U}, U)\|_2^2)$.

As for the two-to-infinity norm, simply considering the naïve relationship between $\|\cdot\|_{2 \to \infty}$ and $\|\cdot\|_2$ yields

$$\frac{1}{\sqrt{p}}\big\|\sin \Theta(\hat{U}, U)\big\|_2 \leq \big\|\hat{U} - UW_{2 \to \infty}^{\star}\big\|_{2 \to \infty} \leq \|\hat{U} - UW_U\|_{2 \to \infty}.$$

These observations collectively suggest that direct analysis of $\hat{U} - UW_U$ may yield meaningfully tighter bounds on $\|\hat{U} - UW_U\|_{2 \to \infty}$ in settings wherein

$\|\hat{U} - U W_U\|_{2\to\infty} \ll \|\hat{U} - U W_U\|_2$ when $p \gg r$. In such a regime, $\|\cdot\|_{2\to\infty}$ and $\|\cdot\|_{\max}$ differ by at most a (relatively small) r-dependent factor, so it is conceivable that $\|\cdot\|_{2\to\infty}$ may serve as a decent proxy for $\|\cdot\|_{\max}$.

We now proceed to introduce a matrix perturbation framework in which $\hat{U}$ represents a perturbation (i.e., estimate) of U. We then formulate a Procrustean matrix decomposition in Section 3.1 by further decomposing the underlying matrices whose spectral norm bounds give rise to the above quantities $\|\sin\Theta(\hat{U}, U)\|_2$ and $\|\sin\Theta(\hat{U}, U)\|_2^2$. Together with two-to-infinity norm machinery and model-based analysis, we subsequently derive a collection of perturbation bounds and demonstrate their utility in problems of statistical estimation.

2.4. *Perturbation framework for the singular value decomposition.* For rectangular matrices $X, E \in \mathbb{R}^{p_1 \times p_2}$, let X denote an unobserved matrix, let E denote an unobserved perturbation (i.e., error) matrix, and let $\hat{X} := X + E$ denote an observed matrix that amounts to an additive perturbation of X by E. For X and $\hat{X}$, their respective partitioned singular value decompositions are given in block matrix form by

$$X = \begin{bmatrix} U & U_\perp \end{bmatrix} \cdot \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma_\perp \end{bmatrix} \cdot \begin{bmatrix} V^\top \\ V_\perp^\top \end{bmatrix} = U\Sigma V^\top + U_\perp \Sigma_\perp V_\perp^\top$$

and

$$\hat{X} := X + E = \begin{bmatrix} \hat{U} & \hat{U}_\perp \end{bmatrix} \cdot \begin{bmatrix} \hat{\Sigma} & 0 \\ 0 & \hat{\Sigma}_\perp \end{bmatrix} \cdot \begin{bmatrix} \hat{V}^\top \\ \hat{V}_\perp^\top \end{bmatrix}$$

$$= \hat{U}\hat{\Sigma}\hat{V}^\top + \hat{U}_\perp \hat{\Sigma}_\perp \hat{V}_\perp^\top.$$

Above, $U \in \mathbb{O}_{p_1,r}$, $V \in \mathbb{O}_{p_2,r}$, $[U|U_\perp] \in \mathbb{O}_{p_1}$, and $[V|V_\perp] \in \mathbb{O}_{p_2}$. The matrices $\Sigma \in \mathbb{R}^{r \times r}$ and $\Sigma_\perp \in \mathbb{R}^{(p_1-r)\times(p_2-r)}$ contain the singular values of X, where $\Sigma = \mathrm{diag}(\sigma_1(X), \ldots, \sigma_r(X))$ and $\Sigma_\perp$ contains the remaining singular values $\sigma_{r+1}(X), \ldots$ on its main diagonal, possibly padded with additional zeros, such that $\sigma_1(X) \geq \cdots \geq \sigma_r(X) > \sigma_{r+1}(X) \geq \cdots \geq 0$. The quantities $\hat{U}, \hat{U}_\perp, \hat{V}, \hat{V}_\perp, \hat{\Sigma}$, and $\hat{\Sigma}_\perp$ are defined analogously.

This paper is primarily interested in the situation when $\sigma_r(X) \gg \sigma_{r+1}(X)$, although our results and framework hold more generally when Σ is redefined to contain a collection of sequential singular values that are separated from the remaining singular values in $\Sigma_\perp$. In such a modified setting, one would have $\Sigma = \mathrm{diag}(\sigma_s(X), \ldots, \sigma_{s+r}(X))$ for some positive integers s and r, where subsequent bounds and necessary bookkeeping would depend both on the two-sided gap $\min\{\sigma_{s-1}(X) - \sigma_s(X), \sigma_{s+r}(X) - \sigma_{s+r+1}(X)\}$ and on the magnitude of the perturbation E, as in [61].

3. Main results.

3.1. *A Procrustean matrix decomposition and its variants.* Below, Theorem 3.1 states our main matrix decomposition in general form. Remark 3.2 subsequently provides accompanying discussion and is designed to offer a more intuitive, high-level explanation of the decomposition considerations presented here. The formal procedure for deriving Theorem 3.1 is based on geometric considerations presented in Section 6.3.

THEOREM 3.1 (Procrustean matrix decomposition). *In the setting of Sections* 2.3 *and* 2.4, *if $\hat{X}$ has rank at least r, then $\hat{U} - UW_U \in \mathbb{R}^{p_1 \times r}$ admits the decomposition*

$$\hat{U} - UW_U = (I - UU^\top)EVW_V\hat{\Sigma}^{-1} \tag{3.1}$$

$$+ (I - UU^\top)E(\hat{V} - VW_V)\hat{\Sigma}^{-1} \tag{3.2}$$

$$+ (I - UU^\top)X(\hat{V} - VV^\top\hat{V})\hat{\Sigma}^{-1} \tag{3.3}$$

$$+ U(U^\top\hat{U} - W_U). \tag{3.4}$$

This decomposition still holds when replacing the $r \times r$ orthogonal matrices W_U and W_V with any real $r \times r$ matrices T_1 and T_2, respectively. The analogous decomposition for $\hat{V} - VW_V$ is given by replacing $U, \hat{U}, V, \hat{V}, E, X, W_U$, and W_V above with $V, \hat{V}, U, \hat{U}, E^\top, X^\top, W_V$, and W_U, respectively.

REMARK 3.2 (Intuition for Theorem 3.1). The decomposition presented in Theorem 3.1 can be loosely motivated in the following way. When X and $\hat{X}$ have rank at least r, then by Section 2.4, $U \equiv XV\Sigma^{-1}$ and $\hat{U} \equiv \hat{X}\hat{V}\hat{\Sigma}^{-1} = X\hat{V}\hat{\Sigma}^{-1} + E\hat{V}\hat{\Sigma}^{-1}$. It is thus conceivable that the difference between U and $\hat{U}$ behaves to leading order as $EV\Sigma^{-1}$ (modulo proper orthogonal transformation) under suitable perturbation and structural assumptions. Indeed, we repeatedly observe such first-order behavior via the matrix term $(I - UU^\top)EVW_V\hat{\Sigma}^{-1}$ when $\|U\|_{2\to\infty} \ll 1$.

For the purpose of obtaining upper bounds, passing from $\hat{\Sigma}^{-1}$ to Σ^{-1} amounts to transitioning from $\sigma_r(\hat{X})$ to $\sigma_r(X)$; this can be achieved via Weyl's inequality [5] provided the perturbation E is suitably small in norm relative to $\sigma_r(X)$.

Subsequent results in Section 3.2 will demonstrate that lines (3.2)–(3.4) amount to situational residual and approximation error terms. Namely, with respect to the two-to-infinity norm:

- Line (3.2) can be much smaller than $\|\sin\Theta(\hat{V}, V)\|_2$ as a function of the relative magnitudes of E and $\hat{\Sigma}^{-1}$.
- Line (3.3) can be much smaller than $\|\sin\Theta(\hat{V}, V)\|_2$ as a function of the multiplicative singular value gap $\sigma_{r+1}(X)/\sigma_r(\hat{X})$.

- Line (3.4) can be much smaller than $\| \sin \Theta(\hat{U}, U) \|_2^2$ as a function of $\|U\|_{2\to\infty}$, specifically when $\|U\|_{2\to\infty} \ll 1$.

Theorem 3.1 can be rewritten in terms of the spectral matrix decomposition when X and E are both symmetric matrices. For ease of reference, we state this special case in the form of a corollary.

COROLLARY 3.3. *Let $X, E \in \mathbb{R}^{p\times p}$ be symmetric matrices. Rephrase Section 2.4 to hold for the spectral matrix decomposition in terms of the eigenvalues and eigenvectors of X and $\hat{X}$. Provided $\hat{X}$ has rank at least r, then $\hat{U} - UW_U \in \mathbb{R}^{p\times r}$ admits the decomposition*

$$\hat{U} - UW_U = (I - UU^\top)EUW_U\hat{\Sigma}^{-1}$$
$$+ (I - UU^\top)E(\hat{U} - UW_U)\hat{\Sigma}^{-1}$$
$$+ (I - UU^\top)X(\hat{U} - UU^\top\hat{U})\hat{\Sigma}^{-1}$$
$$(3.5) \qquad + U(U^\top\hat{U} - W_U).$$

REMARK 3.4 (The orthogonal matrix W_U). This paper does not assume that the leading r singular values of X or $\hat{X}$ are distinct. As such, in general $\hat{U}$ alone cannot hope to recover U in the presence of singular value multiplicity. Indeed, $\hat{U}$ can only be viewed as an estimate of U up to an orthogonal transformation, and our specific choice of W_U is based upon the aforementioned Procrustes-based considerations.

To reiterate, W_U depends upon the perturbed quantity $\hat{U}$, which in turn depends upon the perturbation E. Consequently, W_U is unknown (resp., random) when E is assumed unknown (resp., random). We note that statistical inference methodologies and applications are often either invariant under or equivalent modulo orthogonal transformations as a source of nonidentifiability. For example, K-means clustering applied to the rows of U in Euclidean space is equivalent to clustering the rows of UW_U.

It will subsequently prove convenient to work with the following modified versions of Theorem 3.1 which are stated below as corollaries.

COROLLARY 3.5. *The decomposition in Theorem 3.1 can be rewritten as*

$$\hat{U} - UW_U = (I - UU^\top)E(VV^\top)VW_V\hat{\Sigma}^{-1}$$
$$+ (I - UU^\top)(E + X)(\hat{V} - VW_V)\hat{\Sigma}^{-1}$$
$$(3.6) \qquad + U(U^\top\hat{U} - W_U).$$

COROLLARY 3.6.   *Corollary 3.5 can be equivalently written as*

$$\hat{U} - UW_U = (U_\perp U_\perp^\top)E(VV^\top)VW_V\hat{\Sigma}^{-1}$$
$$+ (U_\perp U_\perp^\top)E(VV^\top)V(V^\top\hat{V} - W_V)\hat{\Sigma}^{-1}$$
$$+ (U_\perp U_\perp^\top)E(V_\perp V_\perp^\top)(\hat{V} - VV^\top\hat{V})\hat{\Sigma}^{-1}$$
$$+ (U_\perp U_\perp^\top)X(V_\perp V_\perp^\top)(\hat{V} - VV^\top\hat{V})\hat{\Sigma}^{-1}$$
$$(3.7) \qquad + U(U^\top\hat{U} - W_U).$$

For Corollaries 3.5 and 3.6, the first term following the equality sign in each display equation is shown in practice to be the leading order term of interest. This point shall be made more precise and quantitative below.

3.2. *General perturbation theorems.*   This section presents a collection of perturbation theorems derived via a unified methodology that combines Theorem 3.1, its variants, the two-to-infinity norm machinery in Section 6.1, and the geometric observations in Section 6.2. We bound $\hat{U} - UW_U$, while similar bounds hold for $\hat{V} - VW_V$ under the appropriate modifications detailed in Theorem 3.1.

THEOREM 3.7 (Baseline two-to-infinity norm bound).   *Provided $\sigma_r(X) > \sigma_{r+1}(X) \geq 0$ and $\sigma_r(X) \geq 2\|E\|_2$, then*

$$\|\hat{U} - UW_U\|_{2\to\infty} \leq 2\left(\frac{\|(U_\perp U_\perp^\top)E(VV^\top)\|_{2\to\infty}}{\sigma_r(X)}\right)$$
$$+ 2\left(\frac{\|(U_\perp U_\perp^\top)E(V_\perp V_\perp^\top)\|_{2\to\infty}}{\sigma_r(X)}\right)\|\sin\Theta(\hat{V}, V)\|_2$$
$$+ 2\left(\frac{\|(U_\perp U_\perp^\top)X(V_\perp V_\perp^\top)\|_{2\to\infty}}{\sigma_r(X)}\right)\|\sin\Theta(\hat{V}, V)\|_2$$
$$(3.8) \qquad + \|\sin\Theta(\hat{U}, U)\|_2^2\|U\|_{2\to\infty}.$$

Let $C_{X,U}$ and $C_{X,V}$ denote upper bounds on the quantities $\|(U_\perp U_\perp^\top)X\|_\infty$ and $\|(V_\perp V_\perp^\top)X^\top\|_\infty$, respectively, and define $C_{E,U}$ and $C_{E,V}$ analogously. Theorem 3.8 provides a uniform perturbation bound for $\|\hat{U} - UW_U\|_{2\to\infty}$ and $\|\hat{V} - VW_V\|_{2\to\infty}$. When $\text{rank}(X) = r$, Corollary 3.9 presents a weaker but simpler version of the bound in Theorem 3.8.

THEOREM 3.8 (Uniform perturbation bound for rectangular matrices).   *Suppose $\sigma_r(X) > \sigma_{r+1}(X) > 0$ and that*

$$\sigma_r(X) \geq \max\{2\|E\|_2, (2/\alpha)C_{E,U}, (2/\alpha')C_{E,V}, (2/\beta)C_{X,U}, (2/\beta')C_{X,V}\}$$

for constants $0 < \alpha, \alpha', \beta, \beta' < 1$ such that $\delta := (\alpha + \beta)(\alpha' + \beta') < 1$. Then

$$(1 - \delta)\|\hat{U} - U W_U\|_{2\to\infty} \le 2\left(\frac{\|(U_\perp U_\perp^\top)E(VV^\top)\|_{2\to\infty}}{\sigma_r(X)}\right)$$

$$+ 2\left(\frac{\|(V_\perp V_\perp^\top)E^\top(UU^\top)\|_{2\to\infty}}{\sigma_r(X)}\right)$$

$$+ \|\sin\Theta(\hat{U}, U)\|_2^2 \|U\|_{2\to\infty}$$

(3.9)
$$+ \|\sin\Theta(\hat{V}, V)\|_2^2 \|V\|_{2\to\infty}.$$

If $\mathrm{rank}(X) = r$ *so that* $\sigma_{r+1}(X) = 0$, *then the above bound holds for* $\delta := \alpha \times \alpha' < 1$ *under the weaker assumption that*

$$\sigma_r(X) \ge \max\{2\|E\|_2, (2/\alpha)C_{E,U}, (2/\alpha')C_{E,V}\}.$$

COROLLARY 3.9 (Uniform perturbation bound for low rank matrices). *Suppose* $\sigma_r(X) > \sigma_{r+1}(X) = 0$ *and that*

$$\sigma_r(X) \ge \max\{2\|E\|_2, (2/\alpha)C_{E,U}, (2/\alpha')C_{E,V}\}$$

for some constants $0 < \alpha, \alpha' < 1$ *such that* $\delta := \alpha \times \alpha' < 1$. *Then*

(3.10) $$(1 - \delta)\|\hat{U} - U W_U\|_{2\to\infty} \le 12 \times \max_{\eta \in \{1,\infty\}}\left\{\frac{\|E\|_\eta}{\sigma_r(X)}\right\} \times \max_{Z \in \{U,V\}}\{\|Z\|_{2\to\infty}\}.$$

4. Applications. This section applies our perturbation theorems and two-to-infinity norm machinery to three statistical settings corresponding to, among others, the results in [8, 21, 37], thereby yielding Theorem 4.2, Theorem 4.3, and Theorem 4.7. Each of these theorems (including Theorem 1.1 presented earlier) is obtained by combining general considerations with application-specific analysis.

Moving forward, consider the following structural matrix property which arises within the context of low rank matrix recovery.

DEFINITION 4.1 (Coherence [9]). Let $\mathcal{U}$ be a subspace of dimension r in $\mathbb{R}^p$, and let $P_{\mathcal{U}}$ be the orthogonal projection onto $\mathcal{U}$. Then the *coherence* of $\mathcal{U}$ (vis-à-vis the standard basis $\{e_i\}$) is defined to be

(4.1) $$\mu(\mathcal{U}) := \left(\frac{p}{r}\right)\max_{i\in[p]}\|P_{\mathcal{U}}e_i\|_2^2.$$

For $U \in \mathbb{O}_{p,r}$, the (orthonormal) columns of U span a subspace of dimension r in $\mathbb{R}^p$, so it is natural to abuse notation and to interchange U with its underlying subspace $\mathcal{U}$. In this case, $P_{\mathcal{U}} \equiv UU^\top$, and so Propositions 6.1 and 6.6 lead to the equivalent formulation

$$\mu(U) := \left(\frac{p}{r}\right)\|U\|_{2\to\infty}^2.$$

Observe that $1 \leq \mu(U) \leq p/r$, where the upper and lower bounds are achieved for U consisting of all standard basis vectors and of vectors whose entries each have magnitude $1/\sqrt{p}$, respectively. Since the columns of U are mutually orthogonal with unit Euclidean norm, the magnitude of $\mu(U)$ can be viewed as quantifying the row-wise accumulation of "mass" in U.

The bounded coherence property [9] corresponds to the existence of a positive constant $C_\mu \geq 1$ such that

$$(4.2) \qquad \|U\|_{2 \to \infty} \leq C_\mu \sqrt{\frac{r}{p}}.$$

Bounded coherence arises naturally in the random orthogonal (matrix) model and influences the recoverability of low rank matrices via nuclear norm minimization when sampling only a subset of the matrix entries [9]. Bounded coherence is also closely related to the notion of eigenvector delocalization in random matrix theory [46]. Examples of matrices whose row and column space factors exhibit bounded coherence can be found, for example, in the study of networks. Specifically, it is not difficult to check that bounded coherence holds for the top eigenvectors of the (non-random) low rank edge probability matrices corresponding to the Erdős–Rényi random graph model and the balanced K-block stochastic block model.

4.1. *Singular vector* (*entrywise*) *perturbation bound.* In [21], the authors consider low rank matrices exhibiting bounded coherence. For such matrices, the results therein provide singular vector perturbation bounds under the ℓ_∞ vector norm, which are then applied to robust covariance estimation.

In this paper, Corollary 3.9 states a perturbation bound that is similar in kind to those in [21]. It is worth emphasizing that upper bounding $\|\hat{U} - U W_U\|_{2 \to \infty}$ immediately bounds both $\|\hat{U} - U W_U\|_{\max}$ and $\inf_{W \in \mathbb{O}_r} \|\hat{U} - U W\|_{\max}$, thereby providing ℓ_∞-type bounds for the perturbed singular vectors up to orthogonal transformation, the analogue of sign flips for well-separated, distinct singular values (similarly for V, $\hat{V}$ and W_V). The joint, symmetric nature of the singular value gap assumption controls the dependence of $\|\hat{U} - U W_U\|_{2 \to \infty}$ and $\|\hat{V} - V W_V\|_{2 \to \infty}$ on one another and takes into account the underlying matrix dimensions.

For symmetric matrices, Theorem 4.2 improves upon [21] and implicitly applies to the applications discussed therein.

THEOREM 4.2 (Application: eigenvector (entrywise) perturbation bound). *Let $X, E \in \mathbb{R}^{p \times p}$ be symmetric matrices where* $\mathrm{rank}(X) = r$ *and X has spectral decomposition* $X = U \Lambda U^\top + U_\perp \Lambda_\perp U_\perp^\top \equiv U \Lambda U^\top$ *with leading eigenvalues* $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_r| > 0$. *If* $|\lambda_r| \geq 4\|E\|_\infty$, *then there exists an orthogonal matrix* $W_U \in \mathbb{O}_r$ *such that*

$$(4.3) \qquad \|\hat{U} - U W_U\|_{2 \to \infty} \leq 14 \left(\frac{\|E\|_\infty}{|\lambda_r|} \right) \|U\|_{2 \to \infty}.$$

Theorem 4.2 provides a user-friendly, deterministic perturbation bound that permits repeated eigenvalues in X and makes no assumption on the behavior of $\|U\|_{2\to\infty}$. When bounded coherence does hold, combining equation (4.2) with Theorem 4.2 immediately yields the bound

$$\|\hat{U} - UW_U\|_{2\to\infty} \le 14C_\mu\left(\frac{\sqrt{r}\|E\|_\infty}{\sqrt{p}|\lambda_r|}\right).$$

It is worth emphasizing that stronger (albeit more complicated) bounds are obtained in the proof leading up to the statement of Theorem 4.2.

4.2. *Singular subspace perturbation and random matrices.* This section interfaces the results in this paper with the spectral and Frobenius-based rate-optimal singular subspace perturbation bounds obtained in [8].

Consider the setting wherein $X \in \mathbb{R}^{p_1 \times p_2}$ is a fixed rank r matrix with $r \ll p_1 \ll p_2$ and $\sigma_r(X) \ge C(p_2/\sqrt{p_1})$. Let $E \in \mathbb{R}^{p_1 \times p_2}$ be a random matrix with independent standard normal entries. Then by [8], the following bounds hold for the left and right singular subspaces with high probability:

$$\left\|\sin\Theta(\hat{U}, U)\right\|_2 \le C\left(\frac{\sqrt{p_1}}{\sigma_r(X)}\right) \quad \text{and} \quad \left\|\sin\Theta(\hat{V}, V)\right\|_2 \le C\left(\frac{\sqrt{p_2}}{\sigma_r(X)}\right).$$

Here, working with V and $\hat{V}$ is desirable though comparatively more difficult. Theorem 4.3 demonstrates how (even relatively coarse) two-to-infinity norm analysis allows one to recover upper and lower bounds for $\|\hat{V} - VW_V\|_{2\to\infty}$ that at times nearly match. For ease of presentation, Theorem 4.3 is stated simply as holding with high probability.

THEOREM 4.3 (Application: singular subspace recovery). *Let $X, E \in \mathbb{R}^{p_1 \times p_2}$ be as in Section* 4.2. *There exists an orthogonal matrix $W_V \in \mathbb{O}_r$ and $C_r > 0$ such that, with high probability,*

$$\|\hat{V} - VW_V\|_{2\to\infty}$$

$$(4.4) \qquad \le C_r\left(\frac{\log(p_2)}{\sigma_r(X)}\right)\left(1 + \left(\frac{p_1}{\sigma_r(X)}\right) + \left(\frac{\sqrt{p_1}}{\log(p_2)}\right)\|V\|_{2\to\infty}\right).$$

If in addition $\sigma_r(X) \ge cp_1$ and $\|V\|_{2\to\infty} \le c_r/\sqrt{p_2}$ for some $c, c_r > 0$, then with high probability

$$(4.5) \qquad \|\hat{V} - VW_V\|_{2\to\infty} \le C_r\left(\frac{\log(p_2)}{\sigma_r(X)}\right).$$

The lower bound $\frac{1}{\sqrt{p_2}}\|\sin\Theta(\hat{V}, V)\|_2 \le \|\hat{V} - VW_V\|_{2\to\infty}$ always holds by Proposition 6.3 *and Lemma* 6.7.

4.3. *Statistical inference for random graphs.* In the study of networks, community detection and clustering are tasks of central interest. A network (i.e., a graph $\mathcal{G} \equiv (\mathcal{V}, \mathcal{E})$ consisting of a vertex set $\mathcal{V}$ and edge set $\mathcal{E}$) may be represented by its adjacency matrix, $A \equiv A_{\mathcal{G}}$, which captures the edge connectivity of the nodes in the network. For inhomogeneous independent edge random graph models, the adjacency matrix can be viewed as a random perturbation of an underlying (often low rank) edge probability matrix P, where in expectation $P \equiv \mathbb{E}[A]$. In the notation of Section 2.4, the matrix P corresponds to X, the matrix $A - P$ corresponds to E, and the matrix A corresponds to $\hat{X}$. By viewing $\hat{U}$ (here the matrix of leading eigenvectors of A) as an estimate of U (here the matrix of leading eigenvectors of P), the results in Section 3 immediately apply.

Spectral methods and related optimization problems for random graphs employ the spectral decomposition of the adjacency matrix (or matrix-valued functions thereof, such as the Laplacian matrix and its variants) [45, 47, 49, 54]. For example, [30] presents a general spectral-based, dimension-reduction community detection framework which incorporates the spectral norm distance between the leading eigenvectors of A and P. Taken in the context of [30] and indeed the wider statistical network analysis literature, this paper complements existing work and paves the way for expanding the toolkit of network analysts to include more Procrustean considerations and two-to-infinity norm machinery.

Much of the existing literature for graphs and network models concerns the popular stochastic block model (SBM) [24] and its variants. The related random dot product graph model (RDPG model) [60] has recently been developed in a series of papers as both a tractable and flexible random graph model amenable to spectral methods [22, 37, 49–52, 54]. In the RDPG model, the graph's eigenvalues and eigenvectors are closely related to the model's generative latent positions. In particular, the leading eigenvectors of the adjacency matrix can be used to estimate the latent positions when properly scaled by the leading eigenvalues.

In the context of the wider RDPG literature, this paper extends both the treatment of the two-to-infinity norm in [37] and Procrustes matching for graphs in [51]. Our bounds in Section 3 imply an eigenvector version of Lemma 5 in [37] that does not require the matrix-valued model parameter P to have distinct eigenvalues. Our Procrustes analysis also suggests a refinement of the test statistic formulation in the two-sample graph inference hypothesis testing framework of [51].

Our level of generality permits the consideration of random graph models that allow edge dependence structure, such as the (C, c, γ) property [41] (see below). Indeed, moving beyond independent edge models represents an important direction for future work in the field of statistical network analysis.

DEFINITION 4.4 ((C, c, γ) concentration [41]). A $p_1 \times p_2$ random matrix E is said to be (C, c, γ)-*concentrated* if, given a trio of positive constants (C, c, γ), for all unit vectors $u \in \mathbb{R}^{p_1}$, $v \in \mathbb{R}^{p_2}$, and for every $t > 0$,

$$(4.6) \qquad \mathbb{P}\big[|\langle Ev, u\rangle| > t\big] \leq C\exp(-ct^{\gamma}).$$

REMARK 4.5 (Probabilistic concentration and the perturbation E). The proofs of our main results demonstrate the importance of bounding $\|EV\|_{2\to\infty}$ and $\|U^\top EV\|_2$ in the perturbation framework of Section 2.4. When E satisfies the (C, c, γ) concentration property in Definition 4.4, these quantities can be easily controlled via standard Bernstein and Hoeffding-type probabilistic bounds encountered throughout statistics.

In the statistical network analysis literature, current active research directions include the development of random graph models exhibiting edge correlation and the development of inference methodology for multiple graphs. Here, we briefly consider the ρ-correlated stochastic block model [36] and the omnibus embedding matrix for multiple graphs [43] employed in [11, 32, 35]. The ρ-correlated stochastic block model provides a simple yet easily interpretable and tractable model for dependent random graphs [34], while the omnibus embedding matrix provides a framework for performing spectral analysis on multiple graphs by leveraging graph (dis)similarities [11, 35, 43].

DEFINITION 4.6 (ρ-correlated SBM graphs [34]). Let $\mathcal{G}^n$ denote the set of labeled, n-vertex, simple, undirected graphs. Two n-vertex random graphs $(G^1, G^2) \in \mathcal{G}^1 \times \mathcal{G}^2$ are said to be ρ-correlated SBM$(\kappa, \vec{n}, b, \Lambda)$ graphs (abbreviated ρ-SBM) if:

1. $G^1 := (\mathcal{V}, \mathcal{E}(G^1))$ and $G^2 := (\mathcal{V}, \mathcal{E}(G^2))$ are marginally SBM$(\kappa, \vec{n}, b, \Lambda)$ random graphs; that is, for each $i = 1, 2$,
   (a) The vertex set $\mathcal{V}$ is the union of κ blocks $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_\kappa$, which are disjoint sets with respective cardinalities $n_1, n_2, \ldots, n_\kappa$;
   (b) The block membership function $b : \mathcal{V} \mapsto [\kappa]$ is such that for each $v \in \mathcal{V}$, $b(v)$ denotes the block of v; that is, $v \in \mathcal{V}_{b(v)}$;
   (c) The block adjacency probabilities are given by the symmetric matrix $\Lambda \in [0, 1]^{\kappa \times \kappa}$; that is, for each pair of vertices $\{j, l\} \in \binom{\mathcal{V}}{2}$, the adjacency of j and l is an independent Bernoulli trial with probability of success $\Lambda_{b(j),b(l)}$.
2. The random variables

$$\{\mathbb{I}[\{j, k\} \in \mathcal{E}(G^i)]\}_{i=1,2;\{j,k\}\in\binom{\mathcal{V}}{2}}$$

are collectively independent except that for each $\{j, k\} \in \binom{\mathcal{V}}{2}$, the correlation between $\mathbb{I}[\{j, k\} \in \mathcal{E}(G^1)]$ and $\mathbb{I}[\{j, k\} \in \mathcal{E}(G^2)]$ is $\rho \geq 0$.

The following theorem provides a guarantee for estimating the leading eigenvectors of a multiple graph omnibus matrix when the graphs are not independent. Theorem 4.7 is among the first of its kind and complements the recent, concurrent work on joint graph embedding in [32].

THEOREM 4.7 (Application: multiple graph inference).   *Let (G^1, G^2) be a pair of ρ-correlated SBM$(\kappa, \overrightarrow{n}, b, \Lambda)$ graphs as in Definition 4.6 with $n \times n$ (symmetric, binary) adjacency matrices (A^1, A^2). Let the model omnibus matrix $\mathfrak{O}$ and adjacency omnibus matrix $\hat{\mathfrak{O}}$ be given by*

$$\mathfrak{O} := \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \otimes \mathcal{Z}\Lambda\mathcal{Z}^\top \quad \text{and} \quad \hat{\mathfrak{O}} := \begin{bmatrix} A^1 & \dfrac{A^1 + A^2}{2} \\ \dfrac{A^1 + A^2}{2} & A^2 \end{bmatrix},$$

where $\otimes$ denotes the matrix Kronecker product and $\mathcal{Z}$ is the $n \times \kappa$ matrix of vertex-to-block assignments such that $P := \mathcal{Z}\Lambda\mathcal{Z}^\top \in [0, 1]^{n \times n}$ denotes the edge probability matrix. Let $\mathrm{rank}(\Lambda) = r$, and hence $\mathrm{rank}(\mathfrak{O}) = r$. For $i = 1, 2$, suppose that the maximum expected degree of G^i, Δ, satisfies $\Delta \gg \log^4(n)$, along with $\sigma_r(\mathfrak{O}) \geq c\Delta$ for some $c > 0$. Let $U, \hat{U} \in \mathbb{O}_{2n,r}$ denote the matrices whose columns are the normalized eigenvectors corresponding to the largest eigenvalues of $\mathfrak{O}$ and $\hat{\mathfrak{O}}$ given by the diagonal matrices Σ and $\hat{\Sigma}$, respectively. For $W_U \in \mathbb{O}_r$ as in Section 2.3, with probability $1 - o(1)$ as $n \to \infty$,

$$\|\hat{U} - UW_U\|_{2\to\infty} = \mathcal{O}_r\left(\frac{\log n}{\Delta}\right).$$

In contrast, spectral norm analysis implies the weaker two-to-infinity norm bound $\|\hat{U} - UW_U\|_{2\to\infty} = \mathcal{O}_r(\frac{1}{\sqrt{\Delta}})$.

REMARK 4.8 (Edge correlation).   The implicit dependence upon the correlation factor ρ in Theorem 4.7 can be made explicit by a more careful analysis of constant factors and the probability statement. This is not our present concern.

5. Discussion and conclusion.   This paper develops a flexible Procrustean matrix decomposition and its variants together with machinery for the two-to-infinity norm in order to study the perturbation of singular vectors and subspaces. We have demonstrated both implicitly and explicitly the widespread applicability of our framework and results to a host of popular matrix noise models, namely matrices that have:

- independent and identically distributed entries (Section 4.2);
- independent and identically distributed rows (Section 1.4);
- independent but not identically distributed entries (Section 4.3);
- neither independent nor identically distributed entries (Section 4.1).

Each application presented in this paper requires model-specific analysis. Namely, one must determine which formulation of the Procrustean matrix decomposition to use, how to effectively transition between norms, and how to analyze the resulting quantities. For example, in Section 4.1 the product term

$\|E\|_\infty \|U\|_{2\to\infty}$ is meaningful when coupled with the bounded coherence assumption, whereas the term $\|EU\|_{2\to\infty}$ is analyzed directly in order to prove Theorem 4.7. Similarly, with respect to covariance estimation (Theorems 1.1 and 4.2), context-specific differences motivate idiosyncratic approaches when deriving the stated bounds.

This paper focuses on decomposing the matrix $\hat{U} - UW_U$ and on establishing the two-to-infinity norm as a useful tool for matrix perturbation analysis. In the time since this work was first made publicly available, there has been a flurry of activity within the statistics, computer science, and mathematics communities devoted to obtaining precise perturbation bounds for singular vectors and eigenvectors [1, 10, 18, 38, 53]. Among the observations made earlier in this paper, it is useful to keep in mind that

$$\inf_{W\in\mathbb{O}_r} \|\hat{U} - UW\|_{\max} \le \|\hat{U} - UW_U\|_{\max} \le \|\hat{U} - UW_U\|_{2\to\infty}.$$

Ample open problems and applications exist for which it is and will be productive to utilize the two-to-infinity norm and matrix decompositions in the future. It is our hope that the level of generality and flexibility presented in this paper will facilitate the more widespread use of the two-to-infinity norm in statistics.

6. Proofs.

6.1. *Technical tools for the two-to-infinity norm.* For $A \in \mathbb{R}^{p_1 \times p_2}$, consider the vector norm on matrices $\|\cdot\|_{2\to\infty}$ defined by

$$\|A\|_{2\to\infty} := \sup_{\|x\|_2=1} \|Ax\|_\infty.$$

Let e_i denote the ith standard basis vector, and let $A_i \in \mathbb{R}^{p_2}$ denote the ith row of A. The following proposition says that $\|A\|_{2\to\infty}$ corresponds to the maximum row-wise Euclidean norm of A.

PROPOSITION 6.1. *For $A \in \mathbb{R}^{p_1 \times p_2}$, then $\|A\|_{2\to\infty} = \max_{i\in[p_1]} \|A_i\|_2$.*

PROOF. The definition of $\|\cdot\|_{2\to\infty}$ and the Cauchy–Schwarz inequality together yield that $\|A\|_{2\to\infty} \le \max_{i\in[p_1]} \|A_i\|_2$, since

$$\|A\|_{2\to\infty} := \sup_{\|x\|_2=1} \|Ax\|_\infty = \sup_{\|x\|_2=1} \max_{i\in[p_1]} |\langle Ax, e_i\rangle|$$

$$\le \max_{i\in[p_1]} \|A_i\|_2.$$

Barring the trivial case $A \equiv 0$, let $e_\star$ denote the standard basis vector in $\mathbb{R}^{p_1}$ with index given by $\operatorname{argmax}_{i\in[p_1]} \|A_i\|_2 > 0$, noting that for each $i \in [p_1]$, $A_i = A^\top e_i$.

For the unit Euclidean norm vector $x_\star := \|A^\top e_\star\|_2^{-1}(A^\top e_\star)$,

$$\|A\|_{2\to\infty} = \sup_{\|x\|_2=1} \max_{i\in[p_1]} |\langle Ax, e_i\rangle| \geq |\langle Ax_\star, e_\star\rangle| = \|A^\top e_\star\|_2$$

$$= \max_{i\in[p_1]} \|A_i\|_2.$$

This establishes the stated equivalence. $\square$

REMARK 6.2. The two-to-infinity norm is subordinate with respect to the ℓ_2 and ℓ_∞ vector norms in the sense that for any $x \in \mathbb{R}^{p_2}$, $\|Ax\|_\infty \leq \|A\|_{2\to\infty}\|x\|_2$. However, $\|\cdot\|_{2\to\infty}$ is not submultiplicative for matrices in general. For example, $\|AB\|_{2\to\infty} = \sqrt{5} > \sqrt{4} = \|A\|_{2\to\infty}\|B\|_{2\to\infty}$ when

$$A \equiv B := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad AB = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}.$$

PROPOSITION 6.3. *For $A \in \mathbb{R}^{p_1\times p_2}$, then*

$$(6.1) \qquad \|A\|_{2\to\infty} \leq \|A\|_2 \leq \min\{\sqrt{p_1}\|A\|_{2\to\infty}, \sqrt{p_2}\|A^\top\|_{2\to\infty}\}.$$

PROOF. The first inequality is obvious since

$$\|A\|_{2\to\infty} = \sup_{\|x\|_2=1} \max_{i\in[p_1]} |\langle Ax, e_i\rangle| \leq \sup_{\|x\|_2=1} \sup_{\|y\|_2=1} |\langle Ax, y\rangle|$$

$$= \|A\|_2.$$

The second inequality holds by an application of the Cauchy–Schwarz inequality together with the vector norm relationship $\|Ax\|_2 \leq \sqrt{p_1}\|Ax\|_\infty$ for $Ax \in \mathbb{R}^{p_1}$. In particular,

$$\sup_{\|x\|_2=1} \sup_{\|y\|_2=1} |\langle Ax, y\rangle| \leq \sup_{\|x\|_2=1} \|Ax\|_2 \leq \sqrt{p_1} \sup_{\|x\|_2=1} \|Ax\|_\infty$$

$$= \sqrt{p_1}\|A\|_{2\to\infty}.$$

By the transpose-invariance of the spectral norm, similarly

$$\|A\|_2 = \|A^\top\|_2 \leq \sqrt{p_2}\|A^\top\|_{2\to\infty}. \qquad \square$$

REMARK 6.4. Proposition 6.3 is sharp. Indeed, for the second inequality, take $A := \{1/\sqrt{p_2}\}^{p_1\times p_2}$. Then $\|A\|_{2\to\infty} = 1$ and $\|A^\top\|_{2\to\infty} = \sqrt{p_1/p_2}$ while $\|A\|_2 = \sqrt{p_1}$. For "tall, skinny" rectangular matrices, the two-to-infinity norm can be much smaller than the spectral norm.

PROPOSITION 6.5. *For $A \in \mathbb{R}^{p_1\times p_2}$, $B \in \mathbb{R}^{p_2\times p_3}$ and $C \in \mathbb{R}^{p_4\times p_1}$, then*

$$(6.2) \qquad\qquad\qquad \|AB\|_{2\to\infty} \leq \|A\|_{2\to\infty}\|B\|_2;$$

$$(6.3) \qquad\qquad\qquad \|CA\|_{2\to\infty} \leq \|C\|_\infty\|A\|_{2\to\infty}.$$

PROOF.   The subordinate property of $\|\cdot\|_{2\to\infty}$ yields that for all $x \in \mathbb{R}^{p_3}$, $\|ABx\|_\infty \le \|A\|_{2\to\infty}\|Bx\|_2$, hence maximizing over all unit vectors x yields equation (6.2). Equation (6.3) follows from Hölder's inequality coupled with the fact that the vector norms ℓ_1 and ℓ_∞ are dual to one another. Explicitly,

$$
\begin{aligned}
\|CA\|_{2\to\infty} &= \sup_{\|x\|_2=1} \max_{i\in[p_1]} |\langle CAx, e_i \rangle| \\
&\le \sup_{\|x\|_2=1} \max_{i\in[p_1]} \|C^\top e_i\|_1 \|Ax\|_\infty \\
&\le \left(\sup_{\|y\|_1=1} \|C^\top y\|_1 \right)\left(\sup_{\|x\|_2=1} \|Ax\|_\infty \right) = \|C^\top\|_1 \|A\|_{2\to\infty} \\
&= \|C\|_\infty \|A\|_{2\to\infty}. \qquad \square
\end{aligned}
$$

PROPOSITION 6.6.   *For $A \in \mathbb{R}^{r\times s}$, $U \in \mathbb{O}_{p_1,r}$ and $V \in \mathbb{O}_{p_2,s}$,*

$$
(6.4) \qquad \|A\|_2 = \|UA\|_2 = \|AV^\top\|_2 = \|UAV^\top\|_2;
$$

$$
(6.5) \qquad \|A\|_{2\to\infty} = \|AV^\top\|_{2\to\infty}.
$$

However, $\|UA\|_{2\to\infty}$ need not equal $\|A\|_{2\to\infty}$.

PROOF.   The first statement follows from Proposition 6.5, the fact that the spectral norm is submultiplicative, and since $U^\top U$ and $V^\top V$ are both the identity matrix. As for the final claim, consider the matrices

$$
(6.6) \quad U := \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}, \qquad A := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad UA = \begin{bmatrix} 1/\sqrt{2} & \sqrt{2} \\ 1/\sqrt{2} & 0 \end{bmatrix},
$$

for which $\|UA\|_{2\to\infty} = \sqrt{5/2} > \sqrt{2} = \|A\|_{2\to\infty}$. $\square$

6.2. Singular subspace geometric bounds.
Let $U, \hat{U} \in \mathbb{O}_{p\times r}$ and $W_U \in \mathbb{O}_r$ denote the Frobenius-optimal Procrustes transformation. We shall use the fact that $\|\sin\Theta(\hat{U},U)\|_2 = \|U_\perp^\top \hat{U}\|_2 = \|(I - UU^\top)\hat{U}\hat{U}^\top\|_2$ ([5], Chapter 7).

LEMMA 6.7.   *Let $T \in \mathbb{R}^{r\times r}$ be arbitrary. Then*

$$
(6.7) \qquad \|\sin\Theta(\hat{U},U)\|_2 = \|\hat{U} - UU^\top \hat{U}\|_2 \le \|\hat{U} - UT\|_2,
$$

$$
(6.8) \qquad \frac{1}{2}\|\sin\Theta(\hat{U},U)\|_2^2 \le \|U^\top \hat{U} - W_U\|_2 \le \|\sin\Theta(\hat{U},U)\|_2^2.
$$

PROOF.   The matrix difference $(\hat{U} - UU^\top \hat{U}) \in \mathbb{R}^{p\times r}$ represents the residual of $\hat{U}$ after orthogonally projecting onto the subspace spanned by the columns of U.

Note that $\|A\|_2^2 = \|A^\top A\|_2 = \sup_{\|x\|_2=1} |\langle A^\top Ax, x\rangle|$, and so several intermediate steps of computation yield that, for any $T \in \mathbb{R}^{r \times r}$,

$$
\begin{aligned}
\|\hat{U} - UU^\top \hat{U}\|_2^2 &= \sup_{\|x\|_2=1} |\langle (\hat{U} - UU^\top \hat{U})^\top (\hat{U} - UU^\top \hat{U})x, x\rangle| \\
&= \sup_{\|x\|_2=1} |\langle (I - \hat{U}^\top UU^\top \hat{U})x, x\rangle| \\
&\leq \sup_{\|x\|_2=1} \left(|\langle (I - \hat{U}^\top UU^\top \hat{U})x, x\rangle| + \|(T - U^\top \hat{U})x\|_2^2 \right) \\
&= \sup_{\|x\|_2=1} |\langle (\hat{U} - UT)^\top (\hat{U} - UT)x, x\rangle| \\
&= \|\hat{U} - UT\|_2^2.
\end{aligned}
$$

On the other hand, by Proposition 6.6 and the above observation,

$$
\begin{aligned}
\|\hat{U} - UU^\top \hat{U}\|_2 &= \|\hat{U}\hat{U}^\top - UU^\top \hat{U}\hat{U}^\top\|_2 = \|(I - UU^\top)\hat{U}\hat{U}^\top\|_2 \\
&= \|\sin \Theta(\hat{U}, U)\|_2.
\end{aligned}
$$

The matrix difference $(U^\top \hat{U} - W_U) \in \mathbb{R}^{r \times r}$ represents the extent to which $U^\top \hat{U}$ with singular value decomposition $W_1 \Sigma_U W_2^\top$ is "almost" the Frobenius-optimal Procrustes transformation $W_U \equiv W_1 W_2^\top$. The orthogonal invariance of the spectral norm together with the interpretation of canonical angles between $\hat{U}$ and U, denoted by $\{\theta_i\}$ with $\cos(\theta_i) = \sigma_i(U^\top \hat{U}) \in [0, 1]$, yields

$$
\|U^\top \hat{U} - W_U\|_2 = \|W_1 \Sigma_U W_2^\top - W_1 W_2^\top\|_2 = \|\Sigma_U - I_r\|_2 = 1 - \min_i \cos(\theta_i).
$$

Thus, both

$$
\|U^\top \hat{U} - W_U\|_2 \leq 1 - \min_i \cos^2(\theta_i) = \max_i \sin^2(\theta_i) = \|\sin \Theta(\hat{U}, U)\|_2^2
$$

and

$$
\begin{aligned}
\|U^\top \hat{U} - W_U\|_2 &\geq \frac{1}{2}\left(1 - \min_i \cos^2(\theta_i)\right) = \frac{1}{2}\max_i \sin^2(\theta_i) \\
&= \frac{1}{2}\|\sin \Theta(\hat{U}, U)\|_2^2.
\end{aligned} \qquad \square
$$

LEMMA 6.8. *The quantity $\|\hat{U} - UW_U\|_2$ satisfies the lower bound*

$$
(6.9) \qquad \|\sin \Theta(\hat{U}, U)\|_2 \leq \|\hat{U} - UW_2^\star\|_2 \leq \|\hat{U} - UW_U\|_2
$$

and satisfies the upper bound

$$
(6.10) \qquad \|\hat{U} - UW_U\|_2 \leq \|\sin \Theta(\hat{U}, U)\|_2 + \|\sin \Theta(\hat{U}, U)\|_2^2.
$$

Taken together with Lemma 1 in [8], an improved upper bound is given by

$$
(6.11) \qquad \|\hat{U} - UW_U\|_2 \leq \min\{1 + \|\sin \Theta(\hat{U}, U)\|_2, \sqrt{2}\}\|\sin \Theta(\hat{U}, U)\|_2.
$$

PROOF.   The lower bound follows from setting $T = W_2^\star$ in Lemma 6.7 together with the definition of $W_2^\star$. Again by Lemma 6.7 and, together with the triangle inequality,

$$\|\hat{U} - U W_U\|_2 \leq \|\hat{U} - U U^\top \hat{U}\|_2 + \|U(U^\top \hat{U} - W_U)\|_2$$
$$\leq \|\sin \Theta(\hat{U}, U)\|_2 + \|\sin \Theta(\hat{U}, U)\|_2^2.$$

The proof of Lemma 1 in [8] establishes that

$$\inf_{W \in \mathbb{O}_r} \|\hat{U} - U W\|_2 \leq \|\hat{U} - U W_U\|_2 \leq \sqrt{2}\|\sin \Theta(\hat{U}, U)\|_2,$$

which completes the proof. □

For ease of reference and notation, Theorem 6.9 below states a version of the Davis–Kahan $\sin \Theta$ theorem [13] in the language of [61]. This amounts to a recasting of Theorem VII.3.2 in [5], and so we omit the proof.

THEOREM 6.9.   *Let $X, \hat{X} \in \mathbb{R}^{p \times p}$ be symmetric matrices with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$, respectively. Write $E := \hat{X} - X$ and fix $1 \leq r \leq s \leq p$. Assume that $\delta_{\text{gap}} := \min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$ where $\lambda_0 := \infty$ and $\lambda_{p+1} := -\infty$. Let $d = s - r + 1$ and let $V := [v_r | v_{r+1} | \ldots | v_s] \in \mathbb{R}^{p \times d}$ and $\hat{V} := [\hat{v}_r | \hat{v}_{r+1} | \ldots | \hat{v}_s] \in \mathbb{R}^{p \times d}$ have orthonormal columns satisfying $X v_j = \lambda_j v_j$ and $\hat{X} \hat{v}_j = \hat{\lambda}_j \hat{v}_j$ for $j = r, r+1, \ldots, s$. Then*

$$(6.12) \qquad \|\sin \Theta(\hat{V}, V)\|_2 \leq \left(\frac{2\|E\|_2}{\delta_{\text{gap}}} \right).$$

6.3. *Theorem* 3.1.

PROOF OF THEOREM 3.1.   The matrices X and $\hat{X}$ have rank at least r, so $\hat{U} \equiv \hat{X} \hat{V} \hat{\Sigma}^{-1}$ and $U W_U \equiv X V \Sigma^{-1} W_U$ by the block matrix formulation in Section 2.4. The explicit correspondence between W_U and $U^\top \hat{U}$ along with subsequent left-multiplication by the matrix U motivates the introduction of the projected quantity $\pm U U^\top \hat{U}$ and leads to

$$\hat{U} - U W_U = (\hat{U} - U U^\top \hat{U}) + (U U^\top \hat{U} - U W_U)$$
$$= (I - U U^\top) \hat{X} \hat{V} \hat{\Sigma}^{-1} + U(U^\top \hat{U} - W_U).$$

The matrix $U(U^\top \hat{U} - W_U)$ can be meaningfully bounded in both spectral and two-to-infinity norm by Lemma 6.8 and Proposition 6.5. Ignoring U for the moment, the difference $U^\top \hat{U} - W_U$ represents the geometric approximation error between $U^\top \hat{U}$ and the orthogonal matrix W_U.

It is not immediately clear how to control $(I - UU^\top)\hat{X}\hat{V}\hat{\Sigma}^{-1}$ given the dependence on the perturbed quantity $\hat{X}$. If instead we replace $\hat{X}$ with X and consider the matrix $(I - UU^\top)X\hat{V}\hat{\Sigma}^{-1}$, then by the block matrix form in Section 2.4 one can check that $(I - UU^\top)X = X(I - VV^\top)$. Since $(I - UU^\top)$ is an orthogonal projection, and hence is idempotent, then $(I - UU^\top)X\hat{V}\hat{\Sigma}^{-1} = (I - UU^\top)X(\hat{V} - VV^\top\hat{V})\hat{\Sigma}^{-1}$. Therefore,

$$(I - UU^\top)\hat{X}\hat{V}\hat{\Sigma}^{-1} = (I - UU^\top)E\hat{V}\hat{\Sigma}^{-1} + (I - UU^\top)X(\hat{V} - VV^\top\hat{V})\hat{\Sigma}^{-1}.$$

By Lemma 6.7 and Proposition 6.5, the terms comprising the matrix product $(I - UU^\top)X(\hat{V} - VV^\top\hat{V})\hat{\Sigma}^{-1}$ can be suitably controlled in norm. At times, it shall be useful to further decompose $(I - UU^\top)X(\hat{V} - VV^\top\hat{V})\hat{\Sigma}^{-1}$ as

$$\left((I - UU^\top)X(\hat{V} - VW_V)\hat{\Sigma}^{-1}\right) + \left((I - UU^\top)XV(W_V - V^\top\hat{V})\hat{\Sigma}^{-1}\right),$$

where the second term vanishes since $(I - UU^\top)XV$ vanishes.

As for the matrix $(I - UU^\top)E\hat{V}\hat{\Sigma}^{-1}$, we do not assume explicit control of $\hat{V}$, so we rewrite the above matrix product in terms of V and a corresponding residual quantity. A natural choice is to incorporate the orthogonal factor W_V. Specifically, introducing $\pm(I - UU^\top)EVW_V\hat{\Sigma}^{-1}$ yields

$$(I - UU^\top)E\hat{V}\hat{\Sigma}^{-1} = (I - UU^\top)E(\hat{V} - VW_V)\hat{\Sigma}^{-1} + (I - UU^\top)EVW_V\hat{\Sigma}^{-1}.$$

Gathering right-hand sides of the above equations yields Theorem 3.1. Corollaries 3.3 and 3.5 are evident given that $U^\top U$ and $V^\top V$ are both simply the identity matrix. Corollary 3.6 is obtained from Corollary 3.5 by additional straightforward algebraic manipulations. In applications, $(I - UU^\top)EVW_V\hat{\Sigma}^{-1} \approx EVW_V\hat{\Sigma}^{-1}$ can be shown to function as the leading order term. □

6.4. *Theorem 3.7.*

PROOF OF THEOREM 3.7.   The assumption $\sigma_r(X) \geq 2\|E\|_2$ implies that $\sigma_r(\hat{X}) \geq \frac{1}{2}\sigma_r(X)$ since by Weyl's inequality for singular values, $\sigma_r(\hat{X}) \geq \sigma_r(X) - \|E\|_2 \geq \frac{1}{2}\sigma_r(X)$. The result then follows from Corollary 3.6 together with Proposition 6.5 and Lemma 6.7. □

6.5. *Theorem 3.8.*

PROOF OF THEOREM 3.8.   By Corollary 3.5, consider the decomposition

$$\begin{aligned}
\hat{U} - UW_U &= (I - UU^\top)E(VV^\top)VW_V\hat{\Sigma}^{-1} \\
&\quad + (I - UU^\top)(E + X)(\hat{V} - VW_V)\hat{\Sigma}^{-1} \\
&\quad + U(U^\top\hat{U} - W_U).
\end{aligned}$$

Applying Proposition 6.5 and Lemma 6.7 yields

$$
\begin{aligned}
\|\hat{U} - U W_U\|_{2\to\infty} \leq{} & \|(U_\perp U_\perp^\top) E(VV^\top)\|_{2\to\infty}/\sigma_r(\hat{X}) \\
& + (C_{E,U} + C_{X,U})\|\hat{V} - V W_V\|_{2\to\infty}/\sigma_r(\hat{X}) \\
& + \|\sin\Theta(\hat{U}, U)\|_2^2 \|U\|_{2\to\infty}
\end{aligned}
$$

and similarly

$$
\begin{aligned}
\|\hat{V} - V W_V\|_{2\to\infty} \leq{} & \|(V_\perp V_\perp^\top) E^\top (UU^\top)\|_{2\to\infty}/\sigma_r(\hat{X}) \\
& + (C_{E,V} + C_{X,V})\|\hat{U} - U W_U\|_{2\to\infty}/\sigma_r(\hat{X}) \\
& + \|\sin\Theta(\hat{V}, V)\|_2^2 \|V\|_{2\to\infty}.
\end{aligned}
$$

By assumption,

$$
\sigma_r(X) \geq \max\{2\|E\|_2, (2/\alpha)C_{E,U}, (2/\alpha')C_{E,V}, (2/\beta)C_{X,U}, (2/\beta')C_{X,V}\}
$$

for some constants $0 < \alpha, \alpha', \beta, \beta' < 1$ such that $\delta := (\alpha + \beta)(\alpha' + \beta') < 1$. The assumption $\sigma_r(X) \geq 2\|E\|_2$ implies that $\sigma_r(\hat{X}) \geq \sigma_r(X) - \|E\|_2 \geq \frac{1}{2}\sigma_r(X)$ by Weyl's inequality for singular values. Combining the above observations and rearranging terms yields

$$
\begin{aligned}
(1-\delta)\|\hat{U} - U W_U\|_{2\to\infty} \leq{} & 2\|(U_\perp U_\perp^\top) E(VV^\top)\|_{2\to\infty}/\sigma_r(X) \\
& + 2(\alpha + \beta)\|(V_\perp V_\perp^\top) E^\top (UU^\top)\|_{2\to\infty}/\sigma_r(X) \\
& + \|\sin\Theta(\hat{U}, U)\|_2^2 \|U\|_{2\to\infty} \\
& + (\alpha + \beta)\|\sin\Theta(\hat{V}, V)\|_2^2 \|V\|_{2\to\infty}.
\end{aligned}
$$

The first claim follows since $(\alpha + \beta) < 1$. When $\mathrm{rank}(X) = r$, then $U_\perp U_\perp^\top X$ vanishes. Corollary 3.3 then yields the simpler form

$$
\begin{aligned}
\hat{U} - U W_U ={} & (I - UU^\top) E(VV^\top) V W_V \hat{\Sigma}^{-1} \\
& + (I - UU^\top) E(\hat{V} - V W_V)\hat{\Sigma}^{-1} \\
& + U(U^\top \hat{U} - W_U),
\end{aligned}
$$

and similarly for $\hat{V} - V W_V$. In this case, the bound holds without needing to consider either $C_{X,U}$ or $C_{X,V}$. $\square$

6.6. *Corollary 3.9.*

PROOF OF COROLLARY 3.9.   By Theorem 3.8,

$$
(1-\delta)\|\hat{U} - U W_U\|_{2\to\infty} \leq 2\|(U_\perp U_\perp^\top) E(VV^\top)\|_{2\to\infty}/\sigma_r(X)
$$

$$+ 2\|(V_\perp V_\perp^\top)E^\top(UU^\top)\|_{2\to\infty}/\sigma_r(X)$$

$$+ \|\sin\Theta(\hat{U}, U)\|_2^2 \|U\|_{2\to\infty}$$

$$+ \|\sin\Theta(\hat{V}, V)\|_2^2 \|V\|_{2\to\infty}.$$

Applying Wedin's $\sin\Theta$ theorem together with the general matrix fact that $\|E\|_2 \leq \max\{\|E\|_\infty, \|E\|_1\}$ and the assumption $\sigma_r(X) \geq 2\|E\|_2$ yields

$$\max_{Z\in\{U,V\}}\{\|\sin\Theta(\hat{Z}, Z)\|_2\} \leq \min\left\{\left(\frac{2\times\max\{\|E\|_\infty, \|E\|_1\}}{\sigma_r(X)}\right), 1\right\}.$$

By invoking properties of the two-to-infinity norm, it follows that

$$\|(U_\perp U_\perp^\top)E(VV^\top)\|_{2\to\infty} \leq \|E(VV^\top)\|_{2\to\infty} + \|(UU^\top)E(VV^\top)\|_{2\to\infty}$$

$$\leq \|EV\|_{2\to\infty} + \|U\|_{2\to\infty}\|U^\top EV\|_2$$

$$\leq \|E\|_\infty\|V\|_{2\to\infty} + \|U\|_{2\to\infty}\max\{\|E\|_\infty, \|E\|_1\}$$

$$\leq 2\times\max_{\eta\in\{1,\infty\}}\{\|E\|_\eta\}\times\max_{Z\in\{U,V\}}\{\|Z\|_{2\to\infty}\}.$$

Similarly,

$$\|(V_\perp V_\perp^\top)E^\top(UU^\top)\|_{2\to\infty}$$

$$\leq \|E\|_1\|U\|_{2\to\infty} + \|V\|_{2\to\infty}\max\{\|E\|_\infty, \|E\|_1\}$$

$$\leq 2\times\max_{\eta\in\{1,\infty\}}\{\|E\|_\eta\}\times\max_{Z\in\{U,V\}}\{\|Z\|_{2\to\infty}\}.$$

Hence

$$(1-\delta)\|\hat{U} - UW_U\|_{2\to\infty} \leq 12\times\max_{\eta\in\{1,\infty\}}\left\{\frac{\|E\|_\eta}{\sigma_r(X)}\right\}\times\max_{Z\in\{U,V\}}\{\|Z\|_{2\to\infty}\},$$

as claimed. $\square$

6.7. *Theorem* 1.1.

PROOF OF THEOREM 1.1.   For ease of presentation, we use $C > 0$ to denote various constants that are allowed to depend on one another. Both n and d are taken to be large.

By hypothesis $\max\{\mathfrak{r}(\Gamma), \log d\} = o(n)$, where $\mathfrak{r}(\Gamma) := \operatorname{trace}(\Gamma)/\sigma_1(\Gamma)$ denotes the effective rank of Γ. In the present multivariate Gaussian covariance matrix setting, it follows from [28, 29] that there exists some constant $C > 0$ such that with probability at least $1 - \frac{1}{3}d^{-2}$,

$$\|E_n\|_2 \leq C\sigma_1(\Gamma)\sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}}.$$

By hypothesis $\sigma_1(\Gamma)/\sigma_r(\Gamma) \leq C$, and so together with the above observations, then $\sigma_r(\Gamma) \geq 2\|E_n\|_2$ with high probability. Theorem 3.7 thus yields

$$\|\hat{U} - UW_U\|_{2\to\infty} \leq C\|(U_\perp U_\perp^\top)E_n(UU^\top)\|_{2\to\infty}/\sigma_r(\Gamma)$$
$$+ C\|(U_\perp U_\perp^\top)E_n(U_\perp U_\perp^\top)\|_{2\to\infty}\|\sin\Theta(\hat{U}, U)\|_2/\sigma_r(\Gamma)$$
$$+ C\|(U_\perp U_\perp^\top)\Gamma(U_\perp U_\perp^\top)\|_{2\to\infty}\|\sin\Theta(\hat{U}, U)\|_2/\sigma_r(\Gamma)$$
$$+ \|\sin\Theta(\hat{U}, U)\|_2^2\|U\|_{2\to\infty}.$$

Moving forward, we record several important observations:

- By Proposition 6.5, $\|(U_\perp U_\perp^\top)E_n(UU^\top)\|_{2\to\infty} \leq \|U_\perp U_\perp^\top\|_\infty \|E_n U\|_{2\to\infty}$.
- By the (bounded coherence) assumption that $\|U\|_{2\to\infty} \leq C\sqrt{r/d}$, then

$$\|U_\perp U_\perp^\top\|_\infty = \|I - UU^\top\|_\infty \leq 1 + \sqrt{d}\|UU^\top\|_{2\to\infty} \leq (1+C)\sqrt{r}.$$

- The random (Gaussian) vector $U_\perp^\top Y$ has covariance matrix $U_\perp \Sigma_\perp U_\perp^\top$, so by [28, 29] there exists some constant $C > 0$ such that with probability at least $1 - \frac{1}{3}d^{-2}$,

$$\|(U_\perp U_\perp^\top)E_n(U_\perp U_\perp^\top)\|_2 \leq C\sigma_{r+1}(\Gamma)\sqrt{\frac{\max\{\mathfrak{r}(\Sigma_\perp), \log d\}}{n}}$$
$$\leq C\sqrt{\sigma_{r+1}(\Gamma)}\sqrt{\sigma_1(\Gamma)}\sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}},$$

where the final inequality holds since

$$\mathfrak{r}(\Sigma_\perp) = \left(\frac{\sigma_1(\Gamma)}{\sigma_{r+1}(\Gamma)}\right)\left(\mathfrak{r}(\Gamma) - \frac{\text{tr}(\Sigma)}{\sigma_1(\Gamma)}\right) \leq \left(\frac{\sigma_1(\Gamma)}{\sigma_{r+1}(\Gamma)}\right)\mathfrak{r}(\Gamma).$$

- Note that $\|(U_\perp U_\perp^\top)\Gamma(U_\perp U_\perp^\top)\|_{2\to\infty} = \|U_\perp \Sigma_\perp U_\perp^\top\|_{2\to\infty} \leq \sigma_{r+1}(\Gamma)$.
- Theorem 6.9 yields the bound $\|\sin\Theta(\hat{U}, U)\|_2 \leq C\|E_n\|_2/\delta_r(\Gamma)$ with population gap given by $\delta_r(\Gamma) := \sigma_r(\Gamma) - \sigma_{r+1}(\Gamma) \geq c_2\sigma_r(\Gamma) > 0$.

Together these observations yield

$$\|\hat{U} - UW_U\|_{2\to\infty} \leq C\sqrt{r}\|E_n U\|_{2\to\infty}/\sigma_r(\Gamma)$$
$$+ C\left(\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}\right)\sqrt{\sigma_{r+1}(\Gamma)/\sigma_r(\Gamma)}$$
$$+ C\sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}}(\sigma_{r+1}(\Gamma)/\sigma_r(\Gamma))$$
$$+ C\left(\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}\right)\sqrt{r/d}.$$

Now let e_i denote the ith standard basis vector in $\mathbb{R}^d$ and u_j denote the jth column of U. The matrix E_n is symmetric, and $E_n U \in \mathbb{R}^{d \times r}$ can be bounded in two-to-infinity norm in the manner

$$\|E_n U\|_{2\to\infty} \leq \sqrt{r}\|E_n U\|_{\max} = \sqrt{r}\max_{i\in[d],\, j\in[r]}|\langle E_n e_i, u_j\rangle|.$$

For each $(i, j) \in [d] \times [r]$, the scalar $\langle E_n e_i, u_j\rangle$ can be expanded as

$$\langle E_n e_i, u_j\rangle = \frac{1}{n}\sum_{k=1}^{n}\left((u_j^\top Y_k)(Y_k^\top e_i) - u_j^\top \Gamma e_i\right) = \frac{1}{n}\sum_{k=1}^{n}\left(\langle Y_k, u_j\rangle Y_k^{(i)} - \langle \Gamma e_i, u_j\rangle\right).$$

The product of (sub-)Gaussian random variables is subexponential, so for i and j fixed, $(\langle Y_k, u_j\rangle Y_k^{(i)} - \langle \Gamma e_i, u_j\rangle)$ with $1 \leq k \leq n$ is a collection of independent and identically distributed, centered subexponential random variables [55]. To this end, the (univariate) subexponential norm, the (univariate) sub-Gaussian norm, and the vector sub-Gaussian norm are, respectively,

$$\left\|(Y^{(i)})^2\right\|_{\psi_1} := \sup_{p\geq 1} p^{-1}\left(\mathbb{E}[|(Y^{(i)})^2|^p]\right)^{1/p};$$

$$\left\|Y^{(i)}\right\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2}\left(\mathbb{E}[|Y^{(i)}|^p]\right)^{1/p};$$

$$\|Y\|_{\psi_2} := \sup_{\|x\|_2=1}\left\|\langle Y, x\rangle\right\|_{\psi_2}.$$

By properties of these (Orlicz) norms [55], there exists some constant $C > 0$ such that the above subexponential random variables satisfy the bound

$$\left\|\langle Y_k, u_j\rangle Y_k^{(i)} - \langle \Gamma e_i, u_j\rangle\right\|_{\psi_1} \leq 2\left\|\langle Y_k, u_j\rangle Y_k^{(i)}\right\|_{\psi_1} \leq C\left\|\langle Y, u_j\rangle\right\|_{\psi_2}\left\|Y^{(i)}\right\|_{\psi_2}.$$

The random vector Y is mean zero multivariate Gaussian, hence for each $1 \leq i \leq d$, the norm of the ith component satisfies the variance-based bound

$$\left\|Y^{(i)}\right\|_{\psi_2} \leq C\max_{1\leq i\leq d}\sqrt{\mathrm{Var}(Y^{(i)})} \equiv C\nu(Y).$$

For each $j \in [r]$, $\mathrm{Var}(\langle Y, u_j\rangle) = u_j^\top \Gamma u_j = \sigma_j(\Gamma)$, where $\langle Y, u_j\rangle$ is univariate Gaussian, so we have the spectral-based bound $\|\langle Y, u_j\rangle\|_{\psi_2} \leq C\sqrt{\sigma_1(\Gamma)}$. Taken together, these observations establish a uniform bound over all i, j, k of the form

$$\left\|\langle Y_k, u_j\rangle Y_k^{(i)} - \langle \Gamma e_i, u_j\rangle\right\|_{\psi_1} \leq C\nu(Y)\sqrt{\sigma_1(\Gamma)}.$$

By combining a union bound with Bernstein's inequality for subexponential random variables [55], it follows that there exists a constant $C > 0$ such that with probability at least $1 - \frac{1}{3}d^{-2}$,

$$\|E_n U\|_{2\to\infty} \leq C\nu(Y)\sqrt{\sigma_1(\Gamma)}\sqrt{r}\sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}}.$$

The r largest singular values of Γ bound each other up to absolute multiplicative constants for all values of d by assumption. Moreover, $\delta_r(\Gamma) \geq c_2 \sigma_r(\Gamma)$ by assumption. Aggregating the above observations yields that with probability at least $1 - d^{-2}$,

$$\|\hat{U} - U W_U\|_{2\to\infty} \leq C \sqrt{\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n}} \left(\frac{\nu(Y) r}{\sqrt{\sigma_r(\Gamma)}} + \frac{\sigma_{r+1}(\Gamma)}{\sigma_r(\Gamma)} \right)$$
$$+ C \left(\frac{\max\{\mathfrak{r}(\Gamma), \log d\}}{n} \right) \left(\sqrt{\frac{\sigma_{r+1}(\Gamma)}{\sigma_r(\Gamma)}} + \sqrt{\frac{r}{d}} \right),$$

which completes the proof. $\quad\square$

6.8. *Theorem* 4.2.

PROOF OF THEOREM 4.2.   Specializing Corollary 3.5 to the case when X is symmetric with $\operatorname{rank}(X) = r$ yields the decomposition

$$\hat{U} - U W_U = E(\hat{U} - U W_U)\hat{\Lambda}^{-1} - (U U^\top) E(\hat{U} - U W_U)\hat{\Lambda}^{-1} + E U W_U \hat{\Lambda}^{-1}$$
$$- (U U^\top) E U W_U \hat{\Lambda}^{-1} + U(U^\top \hat{U} - W_U).$$

Applying the technical results in Sections 6.1 and 6.2 yields the termwise bounds

$$\|E(\hat{U} - U W_U)\hat{\Lambda}^{-1}\|_{2\to\infty} \leq \|E\|_\infty \|\hat{U} - U W_U\|_{2\to\infty} |\hat{\lambda}_r|^{-1},$$
$$\|(U U^\top) E(\hat{U} - U W_U)\hat{\Lambda}^{-1}\|_{2\to\infty} \leq \|U\|_{2\to\infty} \|E\|_2 \|\hat{U} - U W_U\|_2 |\hat{\lambda}_r|^{-1},$$
$$\|E U W_U \hat{\Lambda}^{-1}\|_{2\to\infty} \leq \|E\|_\infty \|U\|_{2\to\infty} |\hat{\lambda}_r|^{-1},$$
$$\|(U U^\top) E U W_U \hat{\Lambda}^{-1}\|_{2\to\infty} \leq \|U\|_{2\to\infty} \|E\|_2 |\hat{\lambda}_r|^{-1},$$
$$\|U(U^\top \hat{U} - W_U)\|_{2\to\infty} \leq \|U\|_{2\to\infty} \|U^\top \hat{U} - W_U\|_2.$$

The matrix E is symmetric by assumption, therefore, $\|E\|_2 \leq \|E\|_\infty$. Furthermore, $\|\hat{U} - U W_U\|_2 \leq \sqrt{2}\|\sin\Theta(\hat{U}, U)\|_2$ by Lemma 6.8, and Theorem 6.9 guarantees that $\|\sin\Theta(\hat{U}, U)\|_2 \leq 2\|E\|_2 |\lambda_r|^{-1}$. Therefore,

$$\|E(\hat{U} - U W_U)\hat{\Lambda}^{-1}\|_{2\to\infty} \leq \|E\|_\infty \|\hat{U} - U W_U\|_{2\to\infty} |\hat{\lambda}_r|^{-1},$$
$$\|(U U^\top) E(\hat{U} - U W_U)\hat{\Lambda}^{-1}\|_{2\to\infty} \leq 4\|E\|_\infty^2 \|U\|_{2\to\infty} |\hat{\lambda}_r|^{-1} |\lambda_r|^{-1},$$
$$\|E U W_U \hat{\Lambda}^{-1}\|_{2\to\infty} \leq \|E\|_\infty \|U\|_{2\to\infty} |\hat{\lambda}_r|^{-1},$$
$$\|(U U^\top) E U W_U \hat{\Lambda}^{-1}\|_{2\to\infty} \leq \|E\|_\infty \|U\|_{2\to\infty} |\hat{\lambda}_r|^{-1},$$
$$\|U(U^\top \hat{U} - W_U)\|_{2\to\infty} \leq 4\|E\|_\infty^2 \|U\|_{2\to\infty} |\lambda_r|^{-2}.$$

By assumption $|\lambda_r| \geq 4\|E\|_\infty$, so $|\hat{\lambda}_r| \geq \frac{1}{2}|\lambda_r|$. As a consequence, it follows that $\|E\|_\infty |\hat{\lambda}_r|^{-1} \leq 2\|E\|_\infty |\lambda_r|^{-1} \leq \frac{1}{2}$. Thus,

$$\|E(\hat{U} - UW_U)\hat{\Lambda}^{-1}\|_{2\to\infty} \leq \frac{1}{2}\|\hat{U} - UW_U\|_{2\to\infty},$$

$$\|(UU^\top)E(\hat{U} - UW_U)\hat{\Lambda}^{-1}\|_{2\to\infty} \leq 2\|E\|_\infty\|U\|_{2\to\infty}|\lambda_r|^{-1},$$

$$\|EUW_U\hat{\Lambda}^{-1}\|_{2\to\infty} \leq 2\|E\|_\infty\|U\|_{2\to\infty}|\lambda_r|^{-1},$$

$$\|(UU^\top)EUW_U\hat{\Lambda}^{-1}\|_{2\to\infty} \leq 2\|E\|_\infty\|U\|_{2\to\infty}|\lambda_r|^{-1},$$

$$\|U(U^\top\hat{U} - W_U)\|_{2\to\infty} \leq \|E\|_\infty\|U\|_{2\to\infty}|\lambda_r|^{-1}.$$

Hence, $\|\hat{U} - UW_U\|_{\max} \leq \|\hat{U} - UW_U\|_{2\to\infty} \leq 14\left(\frac{\|E\|_\infty}{|\lambda_r|}\right)\|U\|_{2\to\infty}$.   □

6.9. *Theorem* 4.3.

PROOF OF THEOREM 4.3.   Rewriting Corollary 3.5 in terms of the matrix $\hat{V} - VW_V$ as described in Theorem 3.1 yields the decomposition

$$\hat{V} - VW_V = (V_\perp V_\perp^\top)E^\top(UU^\top)UW_U\hat{\Sigma}^{-1}$$
$$+ (V_\perp V_\perp^\top)(E^\top + X^\top)(\hat{U} - UW_U)\hat{\Sigma}^{-1}$$
$$+ V(V^\top\hat{V} - W_V).$$

Observe that $(UU^\top)U \equiv U$, while the assumption $\text{rank}(X) = r$ implies that $(V_\perp V_\perp^\top)X^\top$ vanishes. Applying Proposition 6.5, Lemma 6.7, and Lemma 6.8 to the remaining terms therefore yields

$$\|(V_\perp V_\perp^\top)E^\top UW_U\hat{\Sigma}^{-1}\|_{2\to\infty} \leq \|(V_\perp V_\perp^\top)E^\top U\|_{2\to\infty}/\sigma_r(\hat{X}),$$

$$\|(V_\perp V_\perp^\top)E^\top(\hat{U} - UW_U)\hat{\Sigma}^{-1}\|_{2\to\infty}$$
$$\leq C\|(V_\perp V_\perp^\top)E^\top\|_{2\to\infty}\|\sin\Theta(\hat{U}, U)\|_2/\sigma_r(\hat{X}),$$

$$\|V(V^\top\hat{V} - W_V)\|_{2\to\infty} \leq \|\sin\Theta(\hat{V}, V)\|_2^2\|V\|_{2\to\infty}.$$

The columns of $(V_\perp V_\perp^\top)E^\top \in \mathbb{R}^{p_2 \times p_1}$ are centered independent multivariate normal random vectors with covariance matrix $(V_\perp V_\perp^\top)$, so row i of $(V_\perp V_\perp^\top)E^\top$ is a centered multivariate normal random vector with covariance matrix $\sigma_i^2 I$, where $\sigma_i^2 := (V_\perp V_\perp^\top)_{i,i} \leq 1$ and $I \in \mathbb{R}^{p_1 \times p_1}$ denotes the identity matrix. By Gaussian concentration and applying a union bound with $p_2 \gg p_1$, then $\|(V_\perp V_\perp^\top)E^\top\|_{2\to\infty} \leq C_r\sqrt{p_1}\log(p_2)$ with high probability.

As for $(V_\perp V_\perp^\top)E^\top U \in \mathbb{R}^{p_2 \times r}$, the above argument implies that entry (i, j) is $\mathcal{N}(0, \sigma_i^2)$. Hence by the same approach, $\|(V_\perp V_\perp^\top)E^\top U\|_{2\to\infty} \leq C_r\log(p_2)$ with high probability.

By hypothesis $r \ll p_1 \ll p_2$ and $\sigma_r(X) \geq Cp_2/\sqrt{p_1}$, where $\|E\|_2 \leq C\sqrt{p_2}$ with high probability. In this setting, via [8],

$$\|\sin\Theta(\hat{U}, U)\|_2 \leq C\left(\frac{\sqrt{p_1}}{\sigma_r(X)}\right) \quad \text{and} \quad \|\sin\Theta(\hat{V}, V)\|_2 \leq C\left(\frac{\sqrt{p_2}}{\sigma_r(X)}\right).$$

Combining these observations yield

$$\left(\frac{\|(V_\perp V_\perp^\top)E^\top U\|_{2\to\infty}}{\sigma_r(\hat{X})}\right) \leq C_r\left(\frac{\log(p_2)}{\sigma_r(X)}\right);$$

$$\left(\frac{\|(V_\perp V_\perp^\top)E^\top\|_{2\to\infty}}{\sigma_r(\hat{X})}\right)\|\sin\Theta(\hat{U}, U)\|_2 \leq C_r\left(\frac{\log(p_2)}{\sigma_r(X)}\right)\left(\frac{p_1}{\sigma_r(X)}\right);$$

$$\|\sin\Theta(\hat{V}, V)\|_2^2\|V\|_{2\to\infty} \leq C_r\left(\frac{\log(p_2)}{\sigma_r(X)}\right)\left(\frac{\sqrt{p_1}}{\log(p_2)}\right)\|V\|_{2\to\infty}.$$

Hence, with high probability

$$\|\hat{V} - VW_V\|_{2\to\infty} \leq C_r\left(\frac{\log(p_2)}{\sigma_r(X)}\right)\left(1 + \left(\frac{p_1}{\sigma_r(X)}\right) + \left(\frac{\sqrt{p_1}}{\log(p_2)}\right)\|V\|_{2\to\infty}\right).$$

If in addition $\sigma_r(X) \geq cp_1$ and $\|V\|_{2\to\infty} \leq c_r/\sqrt{p_2}$ for some $c, c_r > 0$, then the above bound simplifies to the form

$$\|\hat{V} - VW_V\|_{2\to\infty} \leq C_r\left(\frac{\log(p_2)}{\sigma_r(X)}\right),$$

which completes the proof. □

6.10. *Theorem* 4.7.

PROOF OF THEOREM 4.7.   We seek to bound $\|\hat{U} - UW_U\|_{2\to\infty}$ and allow the constant $C > 0$ to change from line to line. Our analysis considers appropriate groupings of matrix elements in order to handle the graph correlation structure.

By assumption $\mathrm{rank}(\mathfrak{D}) = r$ which implies that the matrix $(I - UU^\top)\mathfrak{D}$ vanishes. Via Corollary 3.3, this yields the bound

$$\|\hat{U} - UW_U\|_{2\to\infty} \leq \|(I - UU^\top)(\hat{\mathfrak{D}} - \mathfrak{D})UW_U\hat{\Sigma}^{-1}\|_{2\to\infty}$$
$$+ \|(I - UU^\top)(\hat{\mathfrak{D}} - \mathfrak{D})(\hat{U} - UW_U)\hat{\Sigma}^{-1}\|_{2\to\infty}$$
$$+ \|U\|_{2\to\infty}\|U^\top\hat{U} - W_U\|_2,$$

which can be further weakened to the form

$$\|\hat{U} - UW_U\|_{2\to\infty} \leq \|(\hat{\mathfrak{D}} - \mathfrak{D})U\|_{2\to\infty}\|\hat{\Sigma}^{-1}\|_2$$
$$+ \|U\|_{2\to\infty}\|U^\top(\hat{\mathfrak{D}} - \mathfrak{D})U\|_2\|\hat{\Sigma}^{-1}\|_2$$
$$+ \|\hat{\mathfrak{D}} - \mathfrak{D}\|_2\|\hat{U} - UW_U\|_2\|\hat{\Sigma}^{-1}\|_2$$
$$+ \|U\|_{2\to\infty}\|U^\top\hat{U} - W_U\|_2.$$

Applying the triangle inequality to the block matrix $\hat{\mathfrak{D}} - \mathfrak{D}$ yields a spectral norm bound of the form

$$\|\hat{\mathfrak{D}} - \mathfrak{D}\|_2 \leq C \times \max\{\|A^1 - P\|_2, \|A^2 - P\|_2\}.$$

By assumption, for $i = 1, 2$, the maximum expected degree of G^i, Δ, satisfies $\Delta \gg \log^4(n)$, hence $\|A^i - P\|_2 = \mathcal{O}(\sqrt{\Delta})$ with probability $1 - o(1)$ by [33]. The assumption $\sigma_r(\mathfrak{D}) \geq c\Delta$ implies that $\sigma_r(\hat{\mathfrak{D}}) \geq C\Delta$ with probability $1 - o(1)$ by Weyl's inequality, so $\|\hat{\Sigma}^{-1}\|_2 = \mathcal{O}(1/\Delta)$. Combining these observations with Theorem 6.9 and the proof of Lemma 6.8 produces

$$\|\hat{U} - U W_U\|_2 \leq C\|\sin \Theta(\hat{U}, U)\|_2 \leq C\|\hat{\mathfrak{D}} - \mathfrak{D}\|_2/\sigma_r(\mathfrak{D}) = \mathcal{O}(1/\sqrt{\Delta}),$$

which we note simultaneously provides a naïve bound for $\|\hat{U} - U W_U\|_{2\to\infty}$. As for the matrix, $(\hat{\mathfrak{D}} - \mathfrak{D})U \in \mathbb{R}^{2n \times r}$,

$$\|(\hat{\mathfrak{D}} - \mathfrak{D})U\|_{2\to\infty} \leq \sqrt{r} \max_{i \in [2n], j \in [r]} |\langle (\hat{\mathfrak{D}} - \mathfrak{D})u_j, e_i \rangle|.$$

For all $1 \leq k \leq n$, $U_{(k+n),j} = U_{k,j}$, and for each $1 \leq i \leq n$ and $1 \leq j \leq r$,

$$\langle (\hat{\mathfrak{D}} - \mathfrak{D})u_j, e_i \rangle = e_i^\top (\hat{\mathfrak{D}} - \mathfrak{D})u_j = \sum_{k=1}^n \left(\frac{3}{2} A_{i,k}^1 + \frac{1}{2} A_{i,k}^2 - 2P_{i,k} \right) U_{k,j}.$$

Above, the roles of A^1 and A^2 are interchanged for $n + 1 \leq i \leq 2n$.

For any $1 \leq i \leq n$, the above expansion is a sum of independent (in k), bounded, mean zero random variables taking values in $[-2U_{k,j}, 2U_{k,j}]$.

Hence by Hoeffding's inequality, with probability $1 - o(1)$ in n,

$$\|(\hat{\mathfrak{D}} - \mathfrak{D})U\|_{2\to\infty} = \mathcal{O}_r(\log n).$$

Similarly, for the matrix $U^\top (\hat{\mathfrak{D}} - \mathfrak{D})U \in \mathbb{R}^{r \times r}$,

$$\|U^\top (\hat{\mathfrak{D}} - \mathfrak{D})U\|_2 \leq r \max_{i \in [r], j \in [r]} |\langle (\hat{\mathfrak{D}} - \mathfrak{D})u_j, u_i \rangle|,$$

so for $1 \leq i, j \leq r$, then

$$\langle (\hat{\mathfrak{D}} - \mathfrak{D})u_j, u_i \rangle = u_i^\top (\hat{\mathfrak{D}} - \mathfrak{D})u_j = \sum_{1 \leq l < k \leq n} 4(A_{l,k}^1 + A_{l,k}^2 - 2P_{l,k})U_{k,j}U_{l,i}.$$

This sum decomposes as a sum of independent, mean zero, bounded random variables taking values in $[-8U_{k,j}U_{l,i}, 8U_{k,j}U_{l,i}]$. By another application of Hoeffding's inequality, with probability $1 - o(1)$,

$$\|U^\top (\hat{\mathfrak{D}} - \mathfrak{D})U\|_2 = \mathcal{O}_r(\log n).$$

Lemma 6.7 bounds $\|U^\top \hat{U} - W_U\|_2$ by $\|\sin \Theta(\hat{U}, U)\|_2^2$ which is $\mathcal{O}(1/\Delta)$. Cumulatively, this demonstrates that $\|\hat{U} - U W_U\|_{2\to\infty} = \mathcal{O}_r((\log n)/\Delta)$ with probability $1 - o(1)$ as $n \to \infty$. $\quad\square$

REFERENCES

[1] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. Preprint. Available at arXiv:1709.09565.

[2] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2567175

[3] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201

[4] BENIDIS, K., SUN, Y., BABU, P. and PALOMAR, D. P. (2016). Orthogonal sparse eigenvectors: A Procrustes problem. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4683–4686.

[5] BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[6] BOJANCZYK, A. W. and LUTOBORSKI, A. (1999). The Procrustes problem for orthogonal Stiefel matrices. *SIAM J. Sci. Comput.* **21** 1291–1304. MR1740396

[7] CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458

[8] CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46** 60–89. MR3766946

[9] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

[10] CAPE, J., TANG, M. and PRIEBE, C. E. (2018). Signal-plus-noise matrix models: Eigenvector deviations and fluctuations. *Biometrika*. To appear. Available at arXiv:1802.00381.

[11] CHEN, L., VOGELSTEIN, J. T., LYZINSKI, V. and PRIEBE, C. E. (2016). A joint graph inference case study: The C. elegans chemical and electrical connectome. *Worm* **5** 1–8.

[12] CHIKUSE, Y. (2003). *Statistics on Special Manifolds*. *Lecture Notes in Statistics* **174**. Springer, New York. MR1960435

[13] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450

[14] DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3** 1102–1123. MR2750388

[15] DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis with Applications in R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3559734

[16] EDELMAN, A., ARIAS, T. A. and SMITH, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. MR1646856

[17] ELDAR, Y. C. and KUTYNIOK, G. (2012). *Compressed Sensing*: *Theory and Applications*. Cambridge Univ. Press, Cambridge. MR2961961

[18] ELDRIDGE, J., BELKIN, M. and WANG, Y. (2018). Unperturbed: Spectral analysis beyond Davis-Kahan. In: *Proceedings of Algorithmic Learning Theory. Proceedings of Machine Learning Research* 83, PMLR 321–358.

[19] FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653

[20] FAN, J., RIGOLLET, P. and WANG, W. (2015). Estimation of functionals of sparse covariance matrices. *Ann. Statist.* **43** 2706–2737. MR3405609

[21] FAN, J., WANG, W. and ZHONG, Y. (2017). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** 1–42. MR3827095

[22] FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990

[23] GOWER, J. C. and DIJKSTERHUIS, G. B. (2004). *Procrustes Problems. Oxford Statistical Science Series* **30**. Oxford Univ. Press, Oxford. MR2051013

[24] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

[25] HORN, R. A. and JOHNSON, C. R. (2012). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR2978290

[26] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[27] JOLLIFFE, I. T. (1986). *Principal Component Analysis. Springer Series in Statistics*. Springer, New York. MR0841268

[28] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768

[29] KOLTCHINSKII, V. and LOUNICI, K. (2017). New asymptotic results in principal component analysis. *Sankhyā A* **79** 254–297. MR3707422

[30] LE, C. M., LEVINA, E. and VERSHYNIN, R. (2016). Optimization via low-rank approximation for community detection in networks. *Ann. Statist.* **44** 373–400. MR3449772

[31] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605

[32] LEVIN, K., ATHREYA, A., TANG, M., LYZINSKI, V. and PRIEBE, C. E. (2017). A central limit theorem for an omnibus embedding of random dot product graphs. Preprint. Available at arXiv:1705.09355.

[33] LU, L. and PENG, X. (2013). Spectra of edge-independent random graphs. *Electron. J. Combin.* **20** 1–18. MR3158266

[34] LYZINSKI, V. (2018). Information recovery in shuffled graphs via graph matching. *IEEE Trans. Inform. Theory* **64** 3254–3273. MR3798375

[35] LYZINSKI, V., PARK, Y., PRIEBE, C. E. and TROSSET, M. (2017). Fast embedding for JOFC using the raw stress criterion. *J. Comput. Graph. Statist.* **26** 786–802. MR3765344

[36] LYZINSKI, V., SUSSMAN, D. L., FISHKIND, D. E., PAO, H., CHEN, L., VOGELSTEIN, J. T., PARK, Y. and PRIEBE, C. E. (2015). Spectral clustering for divide-and-conquer graph matching. *Parallel Comput.* **47** 70–87. MR3371642

[37] LYZINSKI, V., SUSSMAN, D. L., TANG, M., ATHREYA, A. and PRIEBE, C. E. (2014). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Stat.* **8** 2905–2922. MR3299126

[38] MAO, X., SARKAR, P. and CHAKRABARTI, D. (2017). Estimating mixed memberships with sharp eigenvector deviations. Preprint. Available at arXiv:1709.00407.

[39] NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013

[40] O'ROURKE, S., VU, V. and WANG, K. (2016). Eigenvectors of random matrices: A survey. *J. Combin. Theory Ser. A* **144** 361–442. MR3534074

[41] O'ROURKE, S., VU, V. and WANG, K. (2018). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra Appl.* **540** 26–59. MR3739989

[42] PAUL, D. and AUE, A. (2014). Random matrix theory in statistics: A review. *J. Statist. Plann. Inference* **150** 1–29. MR3206718

[43] PRIEBE, C. E., MARCHETTE, D. J., MA, Z. and ADALI, S. (2013). Manifold matching: Joint optimization of fidelity and commensurability. *Braz. J. Probab. Stat*. **27** 377–400. MR3064729

[44] REBROVA, E. and VERSHYNIN, R. (2018). Norms of random matrices: Local and global problems. *Adv. Math*. **324** 40–83. MR3733881

[45] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist*. **39** 1878–1915. MR2893856

[46] RUDELSON, M. and VERSHYNIN, R. (2015). Delocalization of eigenvectors of random matrices with independent entries. *Duke Math. J*. **164** 2507–2538. MR3405592

[47] SARKAR, P. and BICKEL, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist*. **43** 962–990. MR3346694

[48] STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory. Computer Science and Scientific Computing*. Academic Press, Boston, MA. MR1061154

[49] SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc*. **107** 1119–1128. MR3010899

[50] SUSSMAN, D. L., TANG, M. and PRIEBE, C. E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Trans. Pattern Anal. Mach. Intell*. **36** 48–57.

[51] TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist*. **26** 344–354. MR3640191

[52] TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V. and PRIEBE, C. E. (2017). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli* **23** 1599–1630. MR3624872

[53] TANG, M., CAPE, J. and PRIEBE, C. E. (2017). Asymptotically efficient estimators for stochastic blockmodels: The naive MLE, the rank-constrained MLE, and the spectral. Preprint. Available at arXiv:1710.10936.

[54] TANG, M. and PRIEBE, C. E. (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Statist*. **46** 2360–2415. MR3845021

[55] VERSHYNIN, R. (2018). High-dimensional probability: An introduction with applications in data science. Online version 2018-02-09.

[56] VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput*. **17** 395–416. MR2409803

[57] WEDIN, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR0309968

[58] WEYL, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann*. **71** 441–479. MR1511670

[59] YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis. Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. MR3468554

[60] YOUNG, S. J. and SCHEINERMAN, E. R. (2007). Random dot product graph models for social networks. In *Algorithms and Models for the Web-Graph. Lecture Notes in Computer Science* **4863** 138–149. Springer, Berlin. MR2504912

[61] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS
JOHNS HOPKINS UNIVERSITY
3400 NORTH CHARLES STREET
BALTIMORE, MARYLAND 21218
USA
E-MAIL: joshua.cape@jhu.edu
        minh@jhu.edu
        cep@jhu.edu