# A Indefinite orthogonal transformations in practice

In the next two sections, we respond to the possible argument that concerns about distortion by an indefinite orthogonal $\mathbf{Q}$ arise only as an artifact of the GRDPG formalism; or, at least, that such concerns are of little relevance if sole interest in spectral embedding is to allow inference for stochastic block models.

## A.1 Distortion under the stochastic block model

A practical experiment that reveals the presence of indefinite orthogonal transformation implicit in spectral embedding is to simulate graphs from two stochastic block models that differ only in their community proportions. As in Section 4.1 we will use $K = 2$ and the probability matrix $\mathbf{B}^{(1)}$, with community proportions now set to $(0.5, 0.5)$ and $(0.05, 0.95)$ respectively, as opposed to $(0.2, 0.8)$. Resulting spectral embeddings on $n = 4000$ nodes are shown in the left-hand panel of Figure 11, in orange and purple respectively. While each exhibits two clusters, there is very little overlap between the orange and purple point clouds. This discrepancy is almost entirely due to distortion by indefinite orthogonal transformation. Since in simulation $\mathbf{Q}$ can be identified, by inversion the indefinite orthogonal transformation that takes us from the purple to the orange point cloud can be found, and those conforming point clouds are shown in the right-hand panel. The centres of corresponding clusters are now in much closer agreement, and remaining discrepancy is down to statistical error. One cannot make sense of the geometric relationship between the two point clouds without the notion of an indefinite orthogonal transformation.

If an indefinite orthogonal transformation is applied to $\hat{\mathbf{X}}$ before clustering using $K$-means (with Euclidean distance), a different partition of the points is obtained. One might have hoped that the spectral decomposition would produce an embedding somehow optimally configured for clustering using Euclidean $K$-means, but there is no obvious statistical argument to prefer $\hat{\mathbf{X}}$. For example, within-class variance was previously used to compare spectral embeddings under the stochastic block model (Sarkar and Bickel, 2015), but both the empirical (cluster assignment estimated) and oracle (cluster assignment known) within-class variances are minimised for a different, non-degenerate configuration. There are no such concerns with Gaussian mixture modelling, which is invariant. The arguments here support, but go beyond, previous analyses showing the suboptimality of $K$-means clustering versus Gaussian mixture modelling under the non-negative-definite stochastic block model (Tang and Priebe, 2018), since in that special case $K$-means clustering is at least invariant.

## A.2 Distortion under the degree-corrected stochastic block model

Two graphs on $n = 5000$ nodes are now simulated from a two-community degree-corrected stochastic block model, with block matrix $\mathbf{B}^{(1)}$, and community proportions $(0.5, 0.5)$, changing only the degree distributions. In the first case we set $w_i \overset{i.i.d}{\sim}$ uniform$[0, 1]$, whereas in the second $w_i \overset{ind}{\sim}$ Beta$(1, 5)$ when $Z_i = 1$ and $w_i \overset{ind}{\sim}$ Beta$(5, 1)$ when $Z_i = 2$. Resulting spectral embeddings into $\mathbb{R}^2$ are shown in the left-hand panel of Figure 12, in orange and purple respectively. Because $\mathbf{B}^{(1)}$ has one positive and one negative eigenvalue, our theory predicts that each point cloud should live close to the union of two rays through the origin whose joint configuration is predicted *only* up to indefinite orthogonal transformation in $\mathbb{O}(1, 1)$. In particular the hyperbolic angle between the two rays

$$\operatorname{arcosh}(\mathrm{v}_1^\top \mathbf{I}_{1,1} \mathrm{v}_2),$$

is a population quantity that subject to regularity conditions (e.g. as given in Theorem 3) can be estimated consistently as $n \to \infty$ and is not dependent on the node weights. This would make a natural measure of distance between the communities and, in general, when $p = 1$ and $q \geq 1$ the point cloud $\hat{X}_i / (\hat{X}_i^\top \mathbf{I}_{1,q} \hat{X}_i)$ equipped with distance $d(x, y) = \operatorname{arcosh}(x^\top \mathbf{I}_{1,q} y)$ is an embedding into hyperbolic space that accounts for degree heterogeneity. Accordingly, the hyperbolic angle between the two orange rays and that between the two purple rays are equal in the right panel of Figure 12. On the other hand the Euclidean (or ordinary) angle $\arccos(u_1^\top u_2)$, where $u_1$, $u_2$ are unit-norm vectors on each ray, is visibly different between the two point clouds.

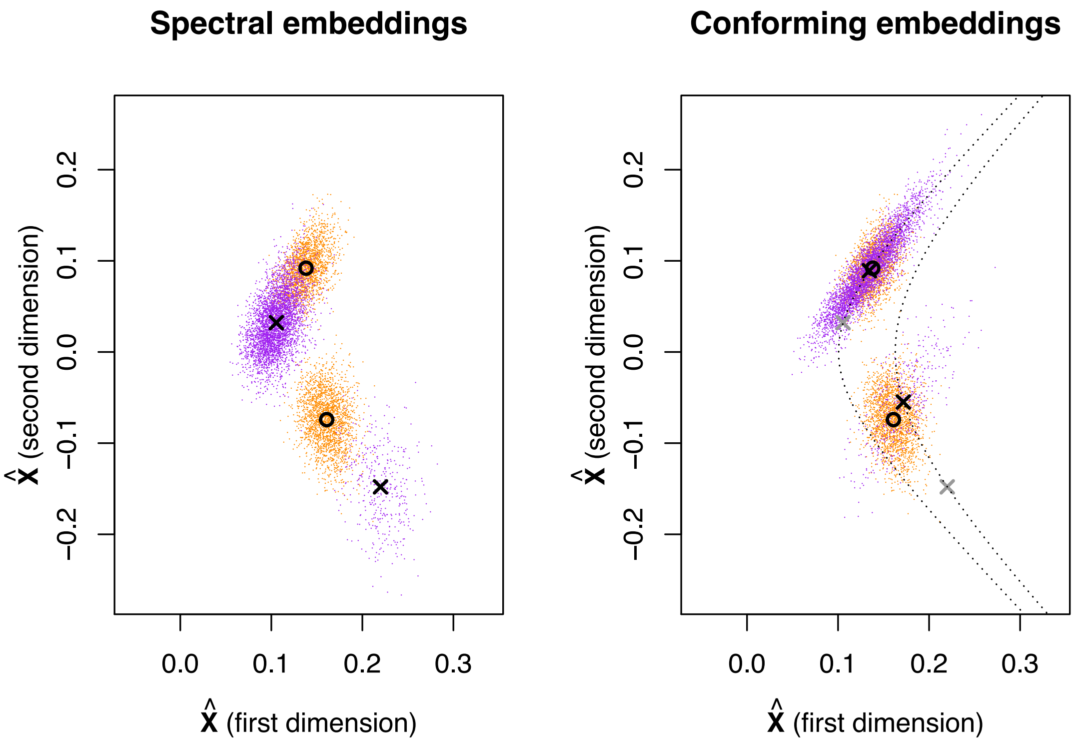


Figure 11: A practical manifestation of $\mathbf{Q}$. Two graphs are generated from two-community stochastic block models with the same block matrix, but differing community proportions, $(0.5, 0.5)$ (orange) versus $(0.05, 0.95)$ (purple). Left: adjacency spectral embedding into $\mathbb{R}^2$, with circles (respectively crosses) indicating the cluster centres of the orange (respectively purple) point cloud. Right: the purple point cloud is re-configured to align with the orange point cloud. The dotted lines show the orbits along which the cluster centres were moved, and the grey crosses their original position. The cluster centres of both point clouds (black circles and crosses) are now close.

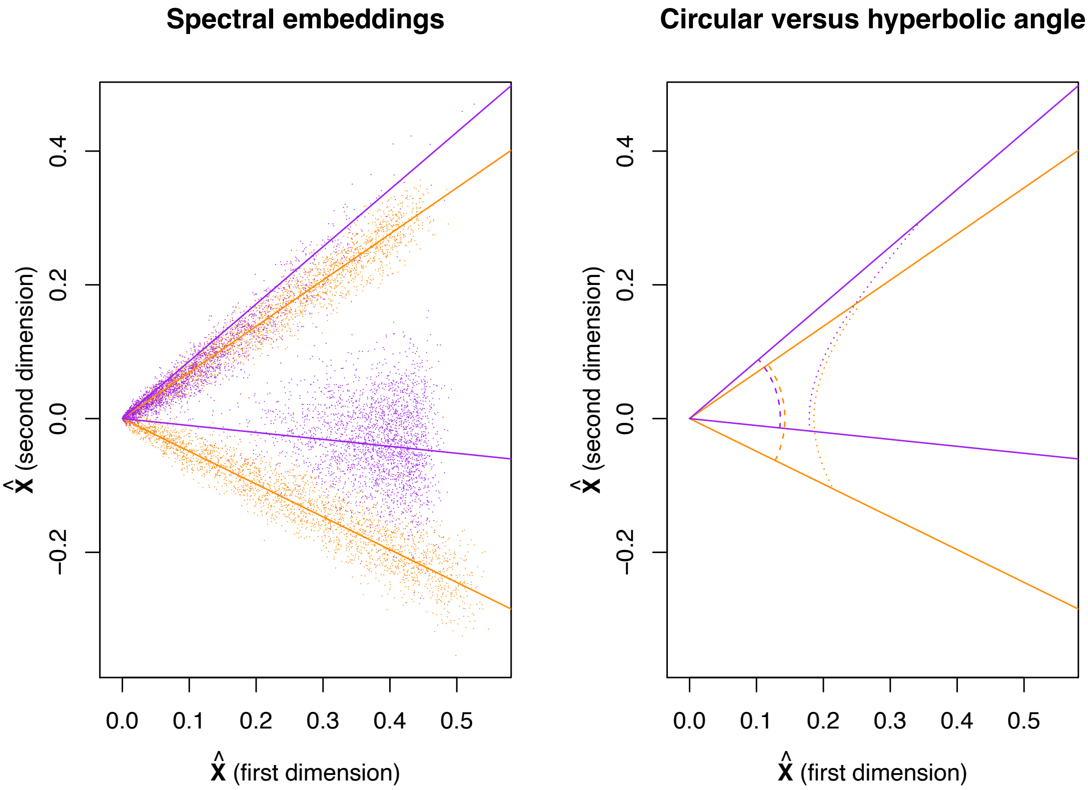


Figure 12: Distortion of angle under the degree-corrected stochastic block model. Two graphs are generated from two-community degree-corrected stochastic block models with the same probability matrix and community proportions, but differing weight distributions, uniform (orange) versus Beta-distributed (purple). Left: adjacency spectral embedding into $\mathbb{R}^2$. In each case the point cloud is a noisy observation of two rays through the origin, shown with orange and purple lines respectively. Right: Euclidean (dashed) and hyperbolic (dotted) angles between the rays. The two hyperbolic angles are identical but the two Euclidean angles differ.

# B Uniqueness

There are a number of reasonable alternative latent position models which, broadly described, assign the nodes to elements $X_1, \ldots, X_n$ of a set $\mathcal{X}$ and, with this assignment held fixed, set

$$A_{ij} \overset{ind}{\sim} \text{Bernoulli}\left\{f(X_i, X_j)\right\},$$

for $i < j$, where $f : \mathcal{X}^2 \to [0,1]$ is some symmetric function. For example, Hoff et al. (2002) considered the choice $f(x,y) = \text{logistic}\,(\alpha - \|x - y\|)$. What is special about the GRDPG?

One argument for considering the GRDPG is that it provides essentially the only way of faithfully reproducing mixtures of connectivity probability profiles as convex combinations in latent space. This idea is now made formal.

**Property 8** (Reproducing mixtures of connectivity probability profiles)**.** Suppose that $\mathcal{X}$ is a convex subset of a real vector space, and that $S$ is a subset of $\mathcal{X}$ whose convex hull is $\mathcal{X}$. We say that a symmetric function $f : \mathcal{X}^2 \to [0,1]$ reproduces mixtures of connectivity probability profiles from $S$ if, whenever $x = \sum_r \alpha_r u_r$, where $u_r \in S$, $0 \le \alpha_r \le 1$ and $\sum \alpha_r = 1$, we have

$$f(x,y) = \sum_r \alpha_r f(u_r, y),$$

for any $y$ in $\mathcal{X}$.

This property helps interpretation of latent space. For example, suppose $X_1, \ldots, X_4 \in S$, and $X_1 = 1/2 X_2 + 1/2 X_3$. In a latent position model where $f$ satisfies the above, we can either think of $\mathbf{A}_{14}$ as being directly generated through $\mathbf{A}_{14} \overset{ind}{\sim} \text{Bernoulli}\{f(X_1, X_4)\}$, or by first flipping a coin, and generating an edge with probability $f(X_2, X_4)$ if it comes up heads, or with probability $f(X_3, X_4)$ otherwise.

In choosing a latent position model to represent the mixed membership stochastic block model, it would be natural to restrict attention to kernels satisfying Property 8, since they allow the simplex representation illustrated in Figure 3, with vertices $S = \{\mathrm{v}_1, \ldots, \mathrm{v}_K\}$ representing communities, and latent positions within it reflecting the nodes' community membership preferences.

We now find that in finite dimension, any such choice amounts to a GRDPG model in at most one extra dimension:

**Theorem 9.** *Suppose $\mathcal{X}$ is a subset of $\mathbb{R}^l$, for some $l \in \mathbb{N}$. The function $f$ reproduces mixtures of connectivity probability profiles if and only if there exist integers $p \ge 1$, $q \ge 0$, $d = p + q \le l + 1$, a matrix $\boldsymbol{T} \in \mathbb{R}^{d \times l}$, and a vector $\nu \in \mathbb{R}^d$ so that $f(x,y) = (\boldsymbol{T}x + \nu)^\top \boldsymbol{I}_{p,q}(\boldsymbol{T}y + \nu)$, for all $x, y \in \mathcal{X}$.*

The mixed membership stochastic block model is an example where this additional dimension is required: in Figure 3 the model is represented as a GRDPG model in $d = 3$ dimensions, but the latent positions live on a 2-dimensional subset.

## B.1 Proof of Theorem 9

Let $\text{aff}(C)$ denote the *affine hull* of a set $C \subseteq \mathbb{R}^d$,

$$\text{aff}(C) = \left\{ \sum_{i=1}^n \alpha_i u_i ; n \in \mathbb{N}, u_i \in C, \alpha_i \in \mathbb{R}, \sum_{i=1}^n \alpha_i = 1 \right\}.$$

We say that a function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a bi-affine form if it is an affine function when either argument is fixed, i.e., $g\{\lambda x_1 + (1-\lambda)x_2, y\} = \lambda g(x_1, y) + (1-\lambda) g(x_2, y)$ and $g\{x, \lambda y_1 + (1-\lambda) y_2\} = \lambda g(x, y_1) + (1-\lambda) g(x, y_2)$, for any $x, y, x_1, x_2, y_1, y_2 \in \mathbb{R}^d$, $\lambda \in \mathbb{R}$. We say that a function $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a bilinear form if it is bi-affine and $h(x,y) = 0$ if either argument is zero.

The proof of Theorem 9 is a direct consequence of the following two lemmas.

**Lemma 10.** *Suppose $\mathcal{X}$ is a convex subset of $\mathbb{R}^l$, for some $l \in \mathbb{N}$. Then $f$ reproduces mixtures of connectivity probability profiles on $S$ if and only if it can be extended to a symmetric bi-affine form $g : \mathit{aff}(\mathcal{X}) \times \mathit{aff}(\mathcal{X}) \to \mathbb{R}$.*

**Lemma 11.** *Suppose $g : \mathit{aff}(\mathcal{X}) \times \mathit{aff}(\mathcal{X}) \to \mathbb{R}$ is a bi-affine form. Let $\ell = \dim\{\mathrm{aff}(\mathcal{X})\} \leq l$. Then there exist a matrix $\boldsymbol{R} \in \mathbb{R}^{(\ell+1)\times l}$, a vector $\mu \in \mathbb{R}^{\ell+1}$, and a bilinear form $h : \mathbb{R}^{(\ell+1)} \times \mathbb{R}^{(\ell+1)} \to \mathbb{R}$ such that $g(x, y) = h(\boldsymbol{R}x + \mu, \boldsymbol{R}y + \mu)$, for all $x, y \in \mathrm{aff}(\mathcal{X})$.*

As is well-known, because $h$ is a symmetric bilinear form on a finite-dimensional real vector space, it can be written $h(x, y) = x^\top \mathbf{J} y$ where $\mathbf{J} \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$ is a symmetric matrix. Write $\mathbf{J} = \mathbf{V}_d \mathbf{S}_d \mathbf{V}_d^\top$ where $\mathbf{V}_d \in \mathbb{R}^{(\ell+1)\times d}$ has orthonormal columns, $\mathbf{S}_d \in \mathbb{R}^{d\times d}$ is diagonal and has $p \geq 0$ positive followed by $q \geq 0$ negative eigenvalues on its diagonal, and $d = p + q = \mathrm{rank}(\mathbf{J})$. Next, define $\mathbf{M} = \mathbf{V}_d |\mathbf{S}_d|^{1/2}$. Then,

$$\begin{aligned} f(x, y) &= g(x, y) = h(\mathbf{R}x + \mu, \mathbf{R}y + \mu) \\ &= \{\mathbf{M}(\mathbf{R}x + \mu)\}^\top \mathbf{I}_{p,q} \{\mathbf{M}(\mathbf{R}y + \mu)\} = (\mathbf{T}x + \nu)^\top \mathbf{I}_{p,q} (\mathbf{T}y + \nu), \end{aligned}$$

where $\mathbf{T} = \mathbf{MR}$ and $\nu = \mathbf{M}\mu$. Since $f(x, x) \geq 0$ on $\mathcal{X} \times \mathcal{X}$, we must have $p > 0$ unless $f$ is uniformly zero over $\mathcal{X} \times \mathcal{X}$.

*Proof of Lemma 10.* The "if" part of the proof is straightforward. Here, we prove the "only if". By definition, any $x, y \in \mathrm{aff}(\mathcal{X}) = \mathrm{aff}(S)$ can be written $x = \sum \alpha_r u_r$, $y = \sum \beta_r v_r$ where $u_r, v_r \in S$, $\alpha_r, \beta_r \in \mathbb{R}$, and $\sum \alpha_r = \sum \beta_r = 1$. For any such $x, y$, we define $g(x, y) = \sum_{r,s} \alpha_r \beta_s f(u_r, v_s)$.

Suppose that $\sum \alpha_r u_r = \sum \gamma_r t_r, \sum \beta_r v_r = \sum \delta_r w_r$ where $t_r, w_r \in S$, $\gamma_r, \delta_r, \in \mathbb{R}$, and $\sum \gamma_r = \sum \delta_r = 1$. Rearrange the first equality to $\sum \alpha'_r u'_r = \sum \gamma'_r t'_r$ by moving any $\alpha_r u_r$ term where $\alpha_r < 0$ to the right — so that the corresponding new coefficient is $\alpha'_s = -\alpha_r$, for some $s$ — and any $\gamma_r t_r$ term where $\gamma_r < 0$ to the left, so that the corresponding new coefficient is $\gamma'_s = -\gamma_r$, for some $s$. Both linear combinations now involve only non-negative scalars. Furthermore, $\sum \alpha_r = \sum \gamma_r$ $(= 1)$ implies $\sum \alpha'_r = \sum \gamma'_r = c$, for some $c \geq 0$.

Then, $\sum (\alpha'_r/c) u'_r = \sum (\gamma'_r/c) t'_r$ are two convex combinations, therefore,

$$\sum (\alpha'_r/c) f(u'_r, v) = f\left\{\sum (\alpha'_r/c) u'_r, v\right\} = f\left\{\sum (\gamma'_r/c) t'_r, v\right\} = \sum (\gamma'_r/c) f(t'_r, v),$$

for any $v \in S$, so that $\sum \alpha_r f(u_r, v) = \sum \gamma_r f(t_r, v)$. Therefore,

$$\begin{aligned} \sum_{r,s} \alpha_r \beta_s f(u_r, v_s) &= \sum_s \beta_s \left\{\sum_r \gamma_r f(t_r, v_s)\right\} \\ &= \sum_r \gamma_r \left\{\sum_s \beta_s f(v_s, t_r)\right\} = \sum_{r,s} \gamma_r \delta_s f(t_r, w_s), \end{aligned}$$

so that $g$ is well-defined. The function $g$ is symmetric and it is also clear that $g\{\lambda x_1 + (1-\lambda) x_2, y\} = \lambda g(x_1, y) + (1 - \lambda) g(x_2, y)$ for any $\lambda \in \mathbb{R}$, making it bi-affine by symmetry. □

The proof technique now used is known as the homogenisation trick in geometry (Gallier, 2000).

*Proof of Lemma 11.* Let $x_0, x_1, \ldots, x_\ell \in \mathbb{R}^l$ be an affine basis of $\mathrm{aff}(\mathcal{X})$. Then there exists an affine transformation $x \to \mathbf{R}x + \mu$, mapping $x_0$ to $z_0 = (0, \ldots, 0, 1) \in \mathbb{R}^{\ell+1}$, $x_1$ to $z_1 = (1, 0, \ldots, 0, 1)$, and so forth, finally mapping $x_\ell$ to $z_\ell = (0, \ldots, 0, 1, 1)$, where $\mathbf{R} \in \mathbb{R}^{(\ell+1)\times l}$ and $\mu \in \mathbb{R}^{\ell+1}$. The vectors $z_0, \ldots, z_\ell$ form a basis of $\mathbb{R}^{\ell+1}$, so that if we set $h(z_i, z_j) = g(x_i, x_j)$ for $0 \leq i, j \leq \ell$, then the value $h$ is well-defined over $\mathbb{R}^{(\ell+1)} \times \mathbb{R}^{(\ell+1)}$ by bilinearity and basis expansion. Since any $x, y \in \mathrm{aff}(\mathcal{X})$ can be written $x = \sum_{r=0}^{\ell} \alpha_r x_r$, $y = \sum_{r=0}^{\ell} \beta_r x_r$ where $\alpha_r, \beta_r \in \mathbb{R}$, and $\sum \alpha_r = \sum \beta_r = 1$, we have

$$\begin{aligned} g(x, y) &= \sum_{r,s} \alpha_r \beta_s g(x_r, x_s) = \sum_{r,s} \alpha_r \beta_s h(z_r, z_s) \\ &= h\left(\sum \alpha_r z_r, \sum \beta_r z_r\right) = h(\mathbf{R}x + \mu, \mathbf{R}y + \mu). \end{aligned}$$

□

# C Proof of Theorems 3 and 4

Broadly speaking, extending prior results on adjacency spectral embedding from the random dot product graph to the GRDPG requires new methods of analysis, that together represent the main technical contribution of this paper (mainly Theorems 3 and 4). Further extending results to the case of Laplacian spectral embedding, while mathematically involved, follows *mutatis mutandis* the machinery developed in Tang and Priebe (2018). Analogous Laplacian-based results (Theorems 6 and 7) are therefore stated without proof.

## C.1 Preliminaries

This proof synthesizes and adapts both the proof architecture and machinery developed in the papers Tang et al. (2017); Cape et al. (2019b,a). We invoke the probabilistic concentration phenomena for GRDPGs presented in Lemma 7 of Tang and Priebe (2018) as well as an eigenvector matrix series decomposition concisely analyzed in Cape et al. (2019a). This proof is not, however, a trivial corollary of earlier results, for it requires additional technical considerations and insight.

We recall the setting of Theorems 3 and 4 wherein the rows of $\mathbf{X}$ are independent replicates of the random vector $\rho_n^{1/2}\xi$, $\xi \sim F$, and for $i < j$ the $ij$-th entries of $\mathbf{A}$ are independent Bernoulli random variables with mean $X_i^\top \mathbf{I}_{p,q} X_j$. Here we shall allow self-loops for mathematical convenience since (dis)allowing self-loops is immaterial with respect to the asymptotic theory we pursue.

Now for $\mathbf{P} = \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$, denote the low-rank spectral decomposition of $\mathbf{P}$ by $\mathbf{P} = \mathbf{U}\mathbf{S}\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{O}_{n,d}$ and $\mathbf{S} \in \mathbb{R}^{d\times d}$. Write $\mathbf{U} \equiv [\mathbf{U}_{(+)}|\mathbf{U}_{(-)}]$ with $\mathbf{U}_{(+)} \in \mathbb{O}_{n,p}$ and $\mathbf{U}_{(-)} \in \mathbb{O}_{n,q}$ to indicate the orthonormal eigenvectors corresponding to the $p$ positive and $q$ negative non-zero eigenvalues of $\mathbf{P}$, written in block-diagonal matrix form as $\mathbf{S} = \mathbf{S}_{(+)} \bigoplus \mathbf{S}_{(-)} \in \mathbb{R}^{(p+q)\times(p+q)} \equiv \mathbb{R}^{d\times d}$. Denote the full spectral decomposition of $\mathbf{A}$ by $\mathbf{A} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{U}}^\top + \hat{\mathbf{U}}_\perp\hat{\mathbf{S}}_\perp\hat{\mathbf{U}}_\perp^\top$, where $\hat{\mathbf{U}} \in \mathbb{O}_{n,d}$ denotes the matrix of leading (orthonormal) eigenvectors of $\mathbf{A}$ and $\hat{\mathbf{S}} \in \mathbb{R}^{d\times d}$ denotes the diagonal matrix containing the $d$ largest-in-magnitude eigenvalues of $\mathbf{A}$ arranged in decreasing order. Here, the matrix $\hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{U}}^\top$ corresponds to the best canonical rank $d$ representation of $\mathbf{A}$. Also above, write $\hat{\mathbf{U}} \equiv [\hat{\mathbf{U}}_{(+)}|\hat{\mathbf{U}}_{(-)}]$ such that the columns of $\hat{\mathbf{U}}_{(+)}$ and $\hat{\mathbf{U}}_{(-)}$ consist of orthonormal eigenvectors corresponding to the largest $p$ positive and $q$ negative non-zero eigenvalues of $\mathbf{A}$, respectively.

We remark at the onset that for the GRDPG model, asymptotically almost surely $\|\mathbf{U}\|_{2\to\infty} = O(n^{-1/2})$ and $|\mathbf{S}_{ii}|, |\hat{\mathbf{S}}_{ii}| = \Theta((n\rho_n))$ for each $i = 1, \ldots, d$. Simultaneously, $\|\mathbf{A}-\mathbf{P}\| = O_{\mathbb{P}}((n\rho_n)^{1/2})$ (regarding the latter, see for example Lu and Peng (2013); Lei and Rinaldo (2015)).

Before going into the details of the proof, we first show that $\mathbf{U}^\top\hat{\mathbf{U}}$ is sufficiently close to an orthogonal matrix $\mathbf{W}_*$ with block diagonal structure that is simultaneously an element of $\mathbb{O}(p,q)$. To this end, the matrix $\mathbf{U}^\top\hat{\mathbf{U}}$ can be written in block form as

$$\mathbf{U}^\top\hat{\mathbf{U}} = \left[\begin{array}{c|c} \mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(+)} & \mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(-)} \\ \hline \mathbf{U}_{(-)}^\top\hat{\mathbf{U}}_{(+)} & \mathbf{U}_{(-)}^\top\hat{\mathbf{U}}_{(-)} \end{array}\right] \in \mathbb{R}^{d\times d}, \tag{2}$$

where $\mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(+)} \in \mathbb{R}^{p\times p}$, $\mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(-)} \in \mathbb{R}^{p\times q}$, $\mathbf{U}_{(-)}^\top\hat{\mathbf{U}}_{(+)} \in \mathbb{R}^{q\times p}$, and $\mathbf{U}_{(-)}^\top\hat{\mathbf{U}}_{(-)} \in \mathbb{R}^{q\times q}$.

Write the singular value decomposition of $\mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(+)} \in \mathbb{R}^{p\times p}$ as $\mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(+)} \equiv \mathbf{W}_{(+),1}\mathbf{\Sigma}_{(+)}\mathbf{W}_{(+),2}^\top$, and define the orthogonal matrix $\mathbf{W}_{(+)}^\star := \mathbf{W}_{(+),1}\mathbf{W}_{(+),2}^\top \in \mathbb{O}_p$. Similarly, let $\mathbf{W}_{(-)}^\star \in \mathbb{O}_q$ denote the orthogonal (product) matrix corresponding to $\mathbf{U}_{(-)}^\top\hat{\mathbf{U}}_{(-)}$. Now let $\mathbf{W}_\star$ denote the structured orthogonal matrix

$$\mathbf{W}_\star = \left[\begin{array}{c|c} \mathbf{W}_{(+)}^\star & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{W}_{(-)}^\star \end{array}\right] \in \mathbb{O}_d. \tag{3}$$

Observe that $\mathbf{W}_\star\mathbf{I}_{p,q}\mathbf{W}_\star^\top = \mathbf{I}_{p,q}$, hence simultaneously $\mathbf{W}_\star \in \mathbb{O}(p,q)$. Via the triangle inequality, the spectral norm quantity $\|\mathbf{U}^\top\hat{\mathbf{U}} - \mathbf{W}_\star\|$ is bounded above by four times the largest spectral norm of its blocks. The main diagonal blocks can be analyzed in a straightforward manner via canonical angles and satisfy

$$\|\mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(+)} - \mathbf{W}_{(+)}^\star\|, \|\mathbf{U}_{(-)}^\top\hat{\mathbf{U}}_{(-)} - \mathbf{W}_{(-)}^\star\| = O_{\mathbb{P}}((n\rho_n)^{-1}). \tag{4}$$

More specifically, let $\sigma_1, \sigma_2, \ldots, \sigma_p$ be the singular values of $\mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(+)}$. Then $\sigma_i = \cos(\theta_i)$ where $\theta_i$ are the principal angles between the subspaces spanned by $\mathbf{U}_{(+)}$ and $\hat{\mathbf{U}}_{(+)}$. The definition of

$\mathbf{W}_{(+)}$ implies

$$\|\mathbf{U}_{(+)}^{\top}\hat{\mathbf{U}}_{(+)}-\mathbf{W}_{(+)}^{\star}\|_F = \|\boldsymbol{\Sigma}_{(+)}-\mathbf{I}\|_F = \Big(\sum_{i=1}^{p}(1-\sigma_i)^2\Big)^{1/2} \leq \sum_{i=1}^{p}(1-\sigma_i^2) = \|\mathbf{U}_{(+)}\mathbf{U}_{(+)}^{\top}-\hat{\mathbf{U}}_{(+)}\hat{\mathbf{U}}_{(+)}^{\top}\|_F^2.$$

By the Davis-Kahan $\sin\Theta$ theorem (see e.g. Section VII.3 of Bhatia (1997) or Yu et al. (2015)), we have

$$\|\mathbf{U}_{(+)}^{\top}\hat{\mathbf{U}}_{(+)} - \mathbf{W}_{(+)}^{\star}\|_F \leq \|\mathbf{U}_{(+)}\mathbf{U}_{(+)}^{\top} - \hat{\mathbf{U}}_{(+)}\hat{\mathbf{U}}_{(+)}^{\top}\|_F^2 \leq \frac{C\|\mathbf{A}-\mathbf{P}\|^2}{\lambda_p(\mathbf{P})^2} = O_{\mathbb{P}}((n\rho_n)^{-1}),$$

where $\lambda_p(\mathbf{P})$ is the smallest positive eigenvalue of $\mathbf{P}$. The bound $\|\mathbf{U}_{(-)}^{\top}\hat{\mathbf{U}}_{(-)}-\mathbf{W}_{(-)}^{\star}\| = O_{\mathbb{P}}((n\rho_n)^{-1})$ is derived similarly.

We now bound the quantities $\|\mathbf{U}_{(+)}^{\top}\hat{\mathbf{U}}_{(-)}\|$. Let $\boldsymbol{u}_{i,(+)}$ and $\hat{\boldsymbol{u}}_{j,(-)}$ be arbitrary columns of $\mathbf{U}_{(+)}$ and $\hat{\mathbf{U}}_{(-)}$, respectively. Note that the $ij$-th entry of $\mathbf{U}_{(+)}^{\top}\hat{\mathbf{U}}_{(-)}$ is $(\boldsymbol{u}_{i,(+)})^{\top}\hat{\boldsymbol{u}}_{j,(-)}$ and that $\lambda_{i,(+)}(\boldsymbol{u}_{i,(+)})^{\top}\hat{\boldsymbol{u}}_{j,(-)} = (\boldsymbol{u}_{i,(+)})^{\top}\mathbf{P}\hat{\boldsymbol{u}}_{j,(-)}$, $\hat{\lambda}_{j,(-)}(\boldsymbol{u}_{i,(+)})^{\top}\hat{\boldsymbol{u}}_{j,(-)} = (\boldsymbol{u}_{i,(+)})^{\top}\mathbf{A}\hat{\boldsymbol{u}}_{j,(-)}$ where $\lambda_{i,(+)}$ (resp. $\hat{\lambda}_{j,(-)}$) is the $i$-th (resp. $j$-th) largest in modulus positive eigenvalue (resp. negative eigenvalue) of $\mathbf{P}$ (resp. $\mathbf{A}$). We therefore have

$$\begin{aligned}(\boldsymbol{u}_i^{(+)})^{\top}\hat{\boldsymbol{u}}_j^{(-)} &= (\hat{\lambda}_{j,(-)} - \lambda_{i,(+)})^{-1}(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})\hat{\boldsymbol{u}}_{j,(-)} \\ &= (\hat{\lambda}_{j,(-)} - \lambda_{i,(+)})^{-1}(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})\mathbf{U}_{(-)}\mathbf{U}_{(-)}^{\top}\hat{\boldsymbol{u}}_{j,(-)} \\ &+ (\hat{\lambda}_{j,(-)} - \lambda_{i,(+)})^{-1}(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})(\mathbf{I}-\mathbf{U}_{(-)}\mathbf{U}_{(-)}^{\top})\hat{\boldsymbol{u}}_{j,(-)}.\end{aligned}$$

The term $(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})\mathbf{U}_{(-)}$ is a vector in $\mathbb{R}^q$, and conditional on $\mathbf{P}$, each element of $(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})\mathbf{U}_{(-)}$ can be written as a sum of independent random variables. Hence, by Hoeffding's inequality, $\|(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})\mathbf{U}_{(-)}\| = O_{\mathbb{P}}(\log n)$. Furthermore, by the Davis-Kahan theorem, $\|(\mathbf{I}-\mathbf{U}_{(-)}\mathbf{U}_{(-)}^{\top})\hat{\boldsymbol{u}}_{j,(-)}\| = O_{\mathbb{P}}((n\rho_n)^{-1/2})$. We therefore have

$$\|(\hat{\lambda}_{j,(-)} - \lambda_{i,(+)})^{-1}(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})\mathbf{U}_{(-)}\mathbf{U}_{(-)}^{\top}\hat{\boldsymbol{u}}_{j,(-)}\| = O_{\mathbb{P}}((n\rho_n^{-1})\log n); \tag{5}$$

$$\|(\hat{\lambda}_{j,(-)} - \lambda_{i,(+)})^{-1}(\boldsymbol{u}_{i,(+)})^{\top}(\mathbf{A}-\mathbf{P})(\mathbf{I}-\mathbf{U}_{(-)}\mathbf{U}_{(-)}^{\top})\hat{\boldsymbol{u}}_{j,(-)}\| = O_{\mathbb{P}}((n\rho_n)^{-1}). \tag{6}$$

Equations (5) and (6) together imply

$$\|\mathbf{U}_{(+)}^{\top}\hat{\mathbf{U}}_{(-)}\| = O_{\mathbb{P}}((n\rho_n)^{-1}(\log n)), \tag{7}$$

thus $\|\mathbf{U}^{\top}\hat{\mathbf{U}} - \mathbf{W}_{\star}\| = O_{\mathbb{P}}((n\rho_n)^{-1}(\log n))$.

## C.2 Proof details

We now proceed with the proof of Theorem 3 and Theorem 4. The matrix relation $\hat{\mathbf{U}}\hat{\mathbf{S}} = \mathbf{A}\hat{\mathbf{U}} = (\mathbf{P}+(\mathbf{A}-\mathbf{P}))\hat{\mathbf{U}}$ yields the matrix equation $\hat{\mathbf{U}}\hat{\mathbf{S}} - (\mathbf{A}-\mathbf{P})\hat{\mathbf{U}} = \mathbf{P}\hat{\mathbf{U}}$. The spectra of $\hat{\mathbf{S}}$ and $\mathbf{A}-\mathbf{P}$ are disjoint asymptotically almost surely, so $\hat{\mathbf{U}}$ can be written as a matrix series of the form (see e.g. Theorem VII.2.1 and Theorem VII.2.2 of Bhatia (1997))

$$\hat{\mathbf{U}} = \sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{P}\hat{\mathbf{U}}\hat{\mathbf{S}}^{-(k+1)} = \sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}\mathbf{S}\mathbf{U}^{\top}\hat{\mathbf{U}}\hat{\mathbf{S}}^{-(k+1)}. \tag{8}$$

By scaling the matrix $\hat{\mathbf{U}}$ by $|\hat{\mathbf{S}}|^{1/2}$, observing that $\hat{\mathbf{S}} = \mathbf{I}_{p,q}|\hat{\mathbf{S}}|$, and applying a well-thought-out "plus zero" trick, we arrive at the decomposition

$$\begin{aligned}
\hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2} &= \sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}\mathbf{S}\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}|\hat{\mathbf{S}}|^{-k-1/2} \\
&= \sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}\mathbf{I}_{p,q}|\mathbf{S}|^{-k+1/2}\mathbf{W}_\star\mathbf{I}_{p,q}^{k+1} \\
&\quad+\sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}\mathbf{I}_{p,q}|\mathbf{S}|^{-k+1/2}(\mathbf{U}^\top\hat{\mathbf{U}}-\mathbf{W}_\star)\mathbf{I}_{p,q}^{k+1} \\
&\quad+\sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}\mathbf{S}(\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}|\hat{\mathbf{S}}|^{-k-1/2}-|\mathbf{S}|^{-k-1/2}\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}) \\
&:= \mathbf{V}_1+\mathbf{V}_2+\mathbf{V}_3.
\end{aligned}$$

### C.2.1 The matrix $\mathbf{V}_1$

Diagonal matrices commute, as do the matrices $\mathbf{I}_{p,q}$ and $\mathbf{W}_\star$, so $\mathbf{V}_1$ can be written as

$$\mathbf{V}_1 \equiv \sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}|\mathbf{S}|^{-k+1/2}\mathbf{W}_\star\mathbf{I}_{p,q}^{k+2} = \mathbf{U}|\mathbf{S}|^{1/2}\mathbf{W}_\star + (\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{S}|^{-1/2}\mathbf{W}_\star\mathbf{I}_{p,q} + \mathbf{R}_{\mathbf{V}_1}, \tag{9}$$

where $\mathbf{R}_{\mathbf{V}_1} = \sum_{k=2}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}|\mathbf{S}|^{-k+1/2}\mathbf{W}_\star\mathbf{I}_{p,q}^{k+2}$. We now use the following slight restatement of Lemma 7.10 from Erdős et al. (2013). This result was also noted in Mao et al. (2021).

**Lemma 12.** *Assume the setting and notations in Theorem 3. Let $\boldsymbol{u}_j$ be the $j$-th column of $\mathbf{U}$ for $j = 1, 2, \ldots, d$. Then there exists a (universal) constant $c > 1$ such that for all $k \le \log n$*

$$\|(\mathbf{A}-\mathbf{P})^k\mathbf{U}\|_{2\to\infty} \le d^{1/2}\max_{j\in[d]}\|(\mathbf{A}-\mathbf{P})^k\boldsymbol{u}_j\|_\infty = O_{\mathbb{P}}\Big(\frac{d^{1/2}(n\rho_n)^{k/2}\log^{kc}(n)}{n^{1/2}}\Big).$$

Thus, for $c > 0$ as above,

$$\begin{aligned}
\|\mathbf{R}_{\mathbf{V}_1}\|_{2\to\infty} &\le \sum_{k=2}^{\log n}\|(\mathbf{A}-\mathbf{P})^k\mathbf{U}\|_{2\to\infty}\||\mathbf{S}|^{-1}\|^{k-1/2} + \sum_{k>\log n}\|\mathbf{A}-\mathbf{P}\|^k\||\mathbf{S}|^{-1}\|^{k-1/2} \\
&= \sum_{k=2}^{\log n} O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{kc}}{n^{1/2}(n\rho_n)^{k/2-1/2}}\right) + \sum_{k>\log n} O_{\mathbb{P}}\left((n\rho_n)^{-k/2+1/2}\right) \\
&= O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right) + O_{\mathbb{P}}\left((n\rho_n)^{-(\log n)/2}\right) \\
&= O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right) + O_{\mathbb{P}}\left(\tfrac{1}{n^{1/2}(n\rho_n)^{1/2}}\right).
\end{aligned}$$

Moving forward, we set forth to make precise the sense in which

$$\begin{aligned}
\hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2} &= \mathbf{U}|\mathbf{S}|^{1/2}\mathbf{W}_\star + (\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{S}|^{-1/2}\mathbf{W}_\star\mathbf{I}_{p,q} + \mathbf{R}_{\mathbf{V}_1} + \mathbf{V}_2 + \mathbf{V}_3 \\
&\approx \mathbf{U}|\mathbf{S}|^{1/2}\mathbf{W}_\star + (\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{S}|^{-1/2}\mathbf{W}_\star\mathbf{I}_{p,q}.
\end{aligned}$$

### C.2.2 The matrix $\mathbf{V}_2$

For the matrix $\mathbf{V}_2 := \sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{U}\mathbf{I}_{p,q}|\mathbf{S}|^{-k+1/2}(\mathbf{U}^\top\hat{\mathbf{U}}-\mathbf{W}_\star)\mathbf{I}_{p,q}^{k+1}$, it is sufficient to observe that by properties of two-to-infinity norm and the bounds established above,

$$\begin{aligned}
\|\mathbf{V}_2\|_{2\to\infty} &\le \|\mathbf{U}\|_{2\to\infty}\||\mathbf{S}|^{1/2}\|\|\mathbf{U}^\top\hat{\mathbf{U}}-\mathbf{W}_\star\| + \|(\mathbf{A}-\mathbf{P})\mathbf{U}\|_{2\to\infty}\||\mathbf{S}|^{-1}\|^{1/2}\|\mathbf{U}^\top\hat{\mathbf{U}}-\mathbf{W}_\star\| + \|\mathbf{R}_{\mathbf{V}_2}\|_{2\to\infty} \\
&= O_{\mathbb{P}}\left(\tfrac{\log n}{n^{1/2}(n\rho_n)^{1/2}}\right) + O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{c+1}}{n^{1/2}(n\rho_n)}\right) + o_{\mathbb{P}}(\|\mathbf{R}_{\mathbf{V}_1}\|_{2\to\infty}) \\
&= O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right).
\end{aligned}$$

In the above, we write that the random variable $Y \in \mathbb{R}$ is $o_{\mathbb{P}}(f(n))$ if for any positive constant $c > 0$ and any $\epsilon > 0$ there exists an $n_0$ such that for all $n \geq n_0$, $|Y| \leq \epsilon f(n)$ with probability at least $1 - n^{-c}$.

### C.2.3 The matrix $\mathbf{V}_3$

The matrix $\mathbf{V}_3$ is given by $\mathbf{V}_3 = \sum_{k=0}^{\infty}(\mathbf{A}-\mathbf{P})^k\mathbf{US}(\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}|\hat{\mathbf{S}}|^{-k-1/2} - |\mathbf{S}|^{-k-1/2}\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1})$. For each $k = 0,1,2,\ldots$, define the matrix $\mathbf{M}_k := (\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}|\hat{\mathbf{S}}|^{-k-1/2} - |\mathbf{S}|^{-k-1/2}\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1})$. Entry $ij$ of the matrix $\mathbf{M}_k$ can be written as

$$\begin{aligned}
(\mathbf{M}_k)_{ij} &= \langle u_i, \hat{u}_j\rangle(\mathbf{I}_{p,q}^{k+1})_{jj}\left[|\hat{\mathbf{S}}_{jj}|^{-k-1/2} - |\mathbf{S}_{ii}|^{-k-1/2}\right] \\
&= \langle u_i, \hat{u}_j\rangle(\mathbf{I}_{p,q}^{k+1})_{jj}\left[|\hat{\mathbf{S}}_{jj}|^{-2k-1} - |\mathbf{S}_{ii}|^{-2k-1}\right]\left[|\hat{\mathbf{S}}_{jj}|^{-k-1/2} + |\mathbf{S}_{ii}|^{-k-1/2}\right]^{-1} \\
&= -\langle u_i, \hat{u}_j\rangle(\mathbf{I}_{p,q}^{k+1})_{jj}\left[|\hat{\mathbf{S}}_{jj}|^{2k+1} - |\mathbf{S}_{ii}|^{2k+1}\right]\left[|\hat{\mathbf{S}}_{jj}|^{-k-1/2} + |\mathbf{S}_{ii}|^{-k-1/2}\right]^{-1}|\hat{\mathbf{S}}_{jj}|^{-2k-1}|\mathbf{S}_{ii}|^{-2k-1} \\
&= -\langle u_i, \hat{u}_j\rangle(\mathbf{I}_{p,q}^{k+1})_{jj}\left[|\hat{\mathbf{S}}_{jj}| - |\mathbf{S}_{ii}|\right]\left[\sum_{l=0}^{2k}|\hat{\mathbf{S}}_{jj}|^l|\mathbf{S}_{ii}|^{2k-l}\right]\left[|\hat{\mathbf{S}}_{jj}|^{-k-1/2} + |\mathbf{S}_{ii}|^{-k-1/2}\right]^{-1}|\hat{\mathbf{S}}_{jj}|^{-2k-1}|\mathbf{S}_{ii}|^{-2k-1}.
\end{aligned}$$

For each $k$, further define a matrix $\mathbf{H}_k \in \mathbb{R}^{d\times d}$ entrywise as

$$(\mathbf{H}_k)_{ij} := \left[\sum_{l=0}^{2k}|\hat{\mathbf{S}}_{jj}|^l|\mathbf{S}_{ii}|^{2k-l}\right]\left[|\hat{\mathbf{S}}_{jj}|^{-k-1/2} + |\mathbf{S}_{ii}|^{-k-1/2}\right]^{-1}|\hat{\mathbf{S}}_{jj}|^{-2k-1}|\mathbf{S}_{ii}|^{-2k-1}, \tag{10}$$

where it follows that

$$(\mathbf{H}_k)_{ij} = O_{\mathbb{P}}((k+1)(n\rho_n)^{-k-3/2}). \tag{11}$$

Letting $\circ$ denote the Hadamard matrix product, we arrive at the decomposition

$$\mathbf{M}_k = -\mathbf{H}_k \circ (\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}|\hat{\mathbf{S}}| - |\mathbf{S}|\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}). \tag{12}$$

The matrices $\mathbf{U}^\top\hat{\mathbf{U}}$ and $\mathbf{I}_{p,q}$ approximately commute. More precisely,

$$\left(\mathbf{I}_{p,q}\mathbf{U}^\top\hat{\mathbf{U}} - \mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}\right) = \left[\begin{array}{c|c}\mathbf{0} & 2\mathbf{U}_{(+)}^\top\hat{\mathbf{U}}_{(-)} \\ \hline -2\mathbf{U}_{(-)}^\top\hat{\mathbf{U}}_{(+)} & \mathbf{0}\end{array}\right] \in \mathbb{R}^{d\times d}, \tag{13}$$

so by Eq. (7), the spectral norm of this matrix difference behaves as $O_{\mathbb{P}}((n\rho_n)^{-1}(\log n))$. This approximate commutativity is important in light of further decomposing the matrix $\mathbf{M}_k$ as

$$\begin{aligned}
\mathbf{M}_k &= -\mathbf{H}_k \circ (\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}|\hat{\mathbf{S}}| - |\mathbf{S}|\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^{k+1}) \\
&= -\mathbf{H}_k \circ \left(\left(\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}|\hat{\mathbf{S}}|\mathbf{I}_{p,q}^k - |\mathbf{S}|\mathbf{I}_{p,q}\mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}^k\right) + |\mathbf{S}|\left(\mathbf{I}_{p,q}\mathbf{U}^\top\hat{\mathbf{U}} - \mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}\right)\mathbf{I}_{p,q}^k\right) \\
&= -\mathbf{H}_k \circ \left(\left(\mathbf{U}^\top\hat{\mathbf{U}}\hat{\mathbf{S}} - \mathbf{S}\mathbf{U}^\top\hat{\mathbf{U}}\right)\mathbf{I}_{p,q}^k + |\mathbf{S}|\left(\mathbf{I}_{p,q}\mathbf{U}^\top\hat{\mathbf{U}} - \mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}\right)\mathbf{I}_{p,q}^k\right).
\end{aligned}$$

We note that $\mathbf{U}^\top\hat{\mathbf{U}}\hat{\mathbf{S}} - \mathbf{S}\mathbf{U}^\top\hat{\mathbf{U}} = \mathbf{U}^\top(\mathbf{A}-\mathbf{P})\hat{\mathbf{U}} = \mathbf{U}^\top(\mathbf{A}-\mathbf{P})\mathbf{U}\mathbf{U}^\top\hat{\mathbf{U}} + \mathbf{U}^\top(\mathbf{A}-\mathbf{P})(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\hat{\mathbf{U}}$ and once again, by Hoeffding's inequality and the Davis-Kahan theorem, we have $\|\mathbf{U}^\top\hat{\mathbf{U}}\hat{\mathbf{S}} - \mathbf{S}\mathbf{U}^\top\hat{\mathbf{U}}\| = O_{\mathbb{P}}(\log n)$, so $\mathbf{M}_k$ can be bounded as

$$\begin{aligned}
\|\mathbf{M}_k\| &\leq \|\mathbf{H}_k\|\|(\mathbf{U}^\top\hat{\mathbf{U}}\hat{\mathbf{S}} - \mathbf{S}\mathbf{U}^\top\hat{\mathbf{U}})\mathbf{I}_{p,q}^k + |\mathbf{S}|(\mathbf{I}_{p,q}\mathbf{U}^\top\hat{\mathbf{U}} - \mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q})\mathbf{I}_{p,q}^k\| \\
&\leq d\|\mathbf{H}_k\|_{\max}\left[\|\mathbf{U}^\top\hat{\mathbf{U}}\hat{\mathbf{S}} - \mathbf{S}\mathbf{U}^\top\hat{\mathbf{U}}\| + \||\mathbf{S}|\|\|\mathbf{I}_{p,q}\mathbf{U}^\top\hat{\mathbf{U}} - \mathbf{U}^\top\hat{\mathbf{U}}\mathbf{I}_{p,q}\|\right] \\
&= O_{\mathbb{P}}(d(k+1)(n\rho_n)^{-k-3/2})\left[O_{\mathbb{P}}(\log n) + O_{\mathbb{P}}((n\rho_n)\times(n\rho_n)^{-1}(\log n))\right] \\
&= O_{\mathbb{P}}(d(k+1)(\log n)(n\rho_n)^{-k-3/2}).
\end{aligned}$$

Hence, for the matrix $\mathbf{V}_3$,

$$\|\mathbf{V}_3\|_{2\to\infty} \leq \|\mathbf{USM}_0\|_{2\to\infty} + \|(\mathbf{A}-\mathbf{P})\mathbf{USM}_1\|_{2\to\infty} + \sum_{k=2}^{\infty}\|(\mathbf{A}-\mathbf{P})^k\mathbf{USM}_k\|_{2\to\infty},$$

where

$$\|\mathbf{USM}_0\|_{2\to\infty} \le \|\mathbf{U}\|_{2\to\infty}\|\mathbf{S}\|\mathbf{M}_0\| = O_{\mathbb{P}}\left(\tfrac{d(\log n)}{n^{1/2}(n\rho_n)^{1/2}}\right),$$
$$\|(\mathbf{A}-\mathbf{P})\mathbf{USM}_1\|_{2\to\infty} \le \|(\mathbf{A}-\mathbf{P})\mathbf{U}\|_{2\to\infty}\|\mathbf{S}\|\|\mathbf{M}_1\| = O_{\mathbb{P}}\left(\tfrac{d^{3/2}(\log n)^{c+1}}{n^{1/2}(n\rho_n)}\right),$$

and

$$\begin{aligned}
\sum_{k=2}^{\infty}\|(\mathbf{A}-\mathbf{P})^k\mathbf{USM}_k\|_{2\to\infty} &\le \sum_{k=2}^{\log n}\|(\mathbf{A}-\mathbf{P})^k\mathbf{USM}_k\|_{2\to\infty} + \sum_{k>\log n}\|(\mathbf{A}-\mathbf{P})^k\mathbf{USM}_k\|_{2\to\infty} \\
&\le \left(\tfrac{d(\log n)^2}{n\rho_n}\right)\sum_{k=2}^{\log n} O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{kc}}{n^{1/2}(n\rho_n)^{k/2-1/2}}\right) + \left(\tfrac{d(\log n)}{n\rho_n}\right)\sum_{k>\log n} O_{\mathbb{P}}(k(n\rho_n)^{-k/2+1/2}) \\
&= \left(\tfrac{d(\log n)^2}{n\rho_n}\right) O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right) + \left(\tfrac{d(\log n)}{n\rho_n}\right) O_{\mathbb{P}}\left((\log n)(n\rho_n)^{-(\log n)/2}\right) \\
&= \left(\tfrac{d(\log n)^2}{n\rho_n}\right) O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right) + \left(\tfrac{d(\log n)^2}{n\rho_n}\right) O_{\mathbb{P}}\left(\tfrac{1}{n^{1/2}(n\rho_n)^{1/2}}\right).
\end{aligned}$$

Since $(n\rho_n) = \omega(d(\log n)^{4c})$ for $c > 1$, we have $(n\rho_n) = \omega(d(\log n)^2)$. It follows that

$$\|\mathbf{V}_3\|_{2\to\infty} = O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right).$$

#### C.2.4 First and second-order characterization

In summary, so far we have shown that

$$\hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2} = \mathbf{U}|\mathbf{S}|^{1/2}\mathbf{W}_\star + (\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{S}|^{-1/2}\mathbf{W}_\star\mathbf{I}_{p,q} + \mathbf{R}, \tag{14}$$

for some (residual) matrix $\mathbf{R} \in \mathbb{R}^{n\times d}$ satisfying $\|\mathbf{R}\|_{2\to\infty} = O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right)$.

Now let $\mathbf{Q_X}$ be such that $\mathbf{X} = \mathbf{U}|\mathbf{S}|^{1/2}\mathbf{Q_X}$. Rearranging the terms in Eq. (14) and multiplying first by $\mathbf{W}_\star^\top$ followed by $\mathbf{Q_X}$ yields

$$\begin{aligned}
\hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2}\mathbf{W}_\star^\top\mathbf{Q_X} - \mathbf{U}|\mathbf{S}|^{1/2}\mathbf{Q_X} &= (\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{S}|^{-1/2}\mathbf{I}_{p,q}\mathbf{Q_X} + \mathbf{RW}_\star^\top\mathbf{Q_X} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{S}|^{1/2}\mathbf{Q_X}\mathbf{Q_X}^{-1}|\mathbf{S}|^{-1}\mathbf{I}_{p,q}\mathbf{Q_X} + \mathbf{RW}_\star^\top\mathbf{Q_X} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{X}\mathbf{Q_X}^{-1}|\mathbf{S}|^{-1}\mathbf{I}_{p,q}\mathbf{Q_X} + \mathbf{RW}_\star^\top\mathbf{Q_X} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{X}\mathbf{Q_X}^{-1}|\mathbf{S}|^{-1}\mathbf{I}_{p,q}\mathbf{Q_X}\mathbf{I}_{p,q}\mathbf{I}_{p,q} + \mathbf{RW}_\star^\top\mathbf{Q_X} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{X}\mathbf{Q_X}^{-1}|\mathbf{S}|^{-1}(\mathbf{Q_X}^{-1})^\top\mathbf{I}_{p,q} + \mathbf{RW}_\star^\top\mathbf{Q_X} \\
&= (\mathbf{A}-\mathbf{P})\mathbf{X}(\mathbf{Q_X}^\top|\mathbf{S}|\mathbf{Q_X})^{-1}\mathbf{I}_{p,q} + \mathbf{RW}_\star^\top\mathbf{Q_X}.
\end{aligned}$$

Both $\hat{\mathbf{X}} = \hat{\mathbf{U}}|\hat{\mathbf{S}}|^{1/2}$ and $\mathbf{Q_X}^\top|\mathbf{S}|\mathbf{Q_X} = \mathbf{X}^\top\mathbf{X}$, yielding the crucial equivalence

$$\hat{\mathbf{X}}\mathbf{W}_\star^\top\mathbf{Q_X} - \mathbf{X} = (\mathbf{A}-\mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}_{p,q} + \mathbf{RW}_\star^\top\mathbf{Q_X}. \tag{15}$$

Theorem 3 holds by observing that

$$\|\mathbf{RW}_\star^\top\mathbf{Q_X}\|_{2\to\infty} = O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^{2c}}{n^{1/2}(n\rho_n)^{1/2}}\right)$$

and

$$\|(\mathbf{A}-\mathbf{P})\mathbf{X}(\mathbf{Q_X}^\top|\mathbf{S}|\mathbf{Q_X})^{-1}\mathbf{I}_{p,q}\|_{2\to\infty} = \|(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{S}|^{-1/2}\mathbf{I}_{p,q}\mathbf{Q_X}\|_{2\to\infty} = O_{\mathbb{P}}\left(\tfrac{d^{1/2}(\log n)^c}{n^{1/2}}\right),$$

where Lemma 5 was implicitly invoked.

For the purpose of establishing Theorem 4, the $i$-th row of Eq. 15, when scaled by $n^{1/2}$, can be written as

$$n^{1/2}(\mathbf{Q_X}^\top\mathbf{W}_\star\hat{X}_i - X_i) = n^{1/2}\mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{X})^{-1}((\mathbf{A}-\mathbf{P})\mathbf{X})_i + n^{1/2}\mathbf{Q_X}^\top\mathbf{W}_\star R_i,$$

where the vector $n^{1/2}\mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{X})^{-1}((\mathbf{A}-\mathbf{P})\mathbf{X})_i$ can be expanded as

$$\mathbf{I}_{p,q}(n^{-1}\mathbf{X}^\top\mathbf{X})^{-1}\left[n^{-1/2}\sum_j(\mathbf{A}_{ij}-\mathbf{P}_{ij})X_j\right]=\mathbf{I}_{p,q}(n^{-1}\rho_n^{-1}\mathbf{X}^\top\mathbf{X})^{-1}\left[(n\rho_n)^{-1/2}\sum_j(\mathbf{A}_{ij}-\mathbf{P}_{ij})\xi_j\right]$$

by recalling that $X_i=\rho_n^{1/2}\xi_i$. The law of large numbers and the continuous mapping theorem together yield $(n^{-1}\rho_n^{-1}\mathbf{X}^\top\mathbf{X})^{-1}\to\mathbf{E}(\xi\xi)^{-1}\equiv\boldsymbol{\Delta}^{-1}$ almost surely. In addition, the classical multivariate central limit theorem gives the (conditional) convergence in distribution

$$\left((n\rho_n)^{-1/2}\sum_j(\mathbf{A}_{ij}-\mathbf{P}_{ij})\xi_j\middle|\xi_i=x_i\right)\to\mathcal{N}_d(\mathbf{0},\boldsymbol{\Gamma}(x_i)),\tag{16}$$

with explicit covariance matrix given by $\boldsymbol{\Gamma}(x_i)=\mathbb{E}\left\{(x_i^\top\mathbf{I}_{p,q}\xi)(1-\rho_n x_i^\top\mathbf{I}_{p,q}\xi)\xi\xi^\top\right\}$. In addition, by combining Lemma 5 with our earlier analysis, it follows that the (transformed) residual matrix satisfies

$$\|n^{1/2}\mathbf{Q}_{\mathbf{X}}^\top\mathbf{W}_\star R_i\|_{2\to\infty}\le n^{1/2}\|\mathbf{Q}_{\mathbf{X}}\|\|\mathbf{W}_\star\|\|\mathbf{R}\|_{2\to\infty}=O_{\mathbb{P}}\left(\frac{d^{1/2}(\log n)^{2c}}{(n\rho_n)^{1/2}}\right)\xrightarrow{\mathbb{P}}0.$$

The above observations together with an application of Slutsky's theorem yield

$$\mathbb{P}\left\{n^{1/2}(\mathbf{Q}_n\hat{X}_i-X_i)\le z\ \mid X_i=\rho_n^{1/2}x\right\}\to\Phi(z,\boldsymbol{\Sigma}(x))\tag{17}$$

for $\mathbf{Q}_n:=\mathbf{Q}_{\mathbf{X}_n}^\top\mathbf{W}_{\star,n}$ and $\boldsymbol{\Sigma}(x)=\mathbf{I}_{p,q}\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}_{\rho_n}(x)\boldsymbol{\Delta}^{-1}\mathbf{I}_{p,q}$. Application of the Cramér-Wold device yields Theorem 4, concluding the proof.