Vintage Factor Analysis with Varimax Performs Statistical Inference

Karl Rohe and Muzhe Zeng

March 11, 2022

To be read before The Royal Statistical Society at an online meeting organized by the Emerging Applications Section and Discussion Meetings Committee on Wednesday, 11 May 2022, Dr. C. Grazian in the Chair

Abstract

Psychologists developed Multiple Factor Analysis to decompose multivariate data into a small number of interpretable factors without any *a priori* knowledge about those factors [Thurstone, 1935]. In this form of factor analysis, the Varimax *factor rotation* redraws the axes through the multidimensional factors to make them sparse and thus make them more interpretable [Kaiser, 1958]. Charles Spearman and many others objected to factor rotations because the factors seem to be rotationally invariant [Thurstone, 1947, Anderson and Rubin, 1956]. These objections are still reported in all contemporary multivariate statistics textbooks. However, this vintage form of factor analysis has survived and is widely popular because, empirically, the factor rotation often makes the factors easier to interpret. We argue that the rotation makes the factors easier to interpret because, in fact, the Varimax factor rotation performs statistical inference. We show that Principal Components Analysis (PCA) with the Varimax axes provides a unified spectral estimation strategy for a broad class of semi-parametric factor models, including the Stochastic Blockmodel and a natural variation of Latent Dirichlet Allocation (i.e., "topic modeling"). In addition, we show that Thurstone's widely employed sparsity diagnostics implicitly assess a key *leptokurtic* condition that makes the axes statistically identifiable in these models. Taken together, this shows that the know-how of Vintage Factor Analysis performs statistical inference, reversing nearly a century of statistical thinking on the topic. We illustrate these techniques use on two large bibliometric examples (a citation network and a text corpus). With a sparse eigensolver, PCA with Varimax is both fast and stable. Combined with Thurstone's straightforward diagnostics, this vintage approach is suitable for a wide array of modern applications.

Keywords: Factor analysis, Independent Component Analysis, Spectral Clustering, Little Jiffy, orthoblique

Outside the language of mathematical statistics, Louis Leon Thurstone, Henry Kaiser, and other psychologists developed the first forms of Multiple Factor Analysis, or what is referred to herein as Vintage Factor Analysis [Thurstone, 1935, 1947, Kaiser, 1958]. There are two simultaneous aims of Vintage Factor Analysis. The first aim is to provide a low dimensional approximation of the observed data; in this sense, it is like Principal Components Analysis (PCA).[1] The second aim is to ensure that each factor (i.e. each axis in the lower dimensional representation) corresponds to a "scientifically meaningful category" [Thurstone, 1935]. A Varimax rotation of the principal components is a simple and popular way to find such meaningful dimensions [Kaiser, 1958, Jolliffe, 2002].

For example, suppose n students take an exam with d questions, producing a d dimensional vector of data for each individual. Principal components analysis with k=2 dimensions will roughly approximate the students' d dimensional data; this is the first aim of factor analysis. In order to make those two dimensions more interpretable, Varimax draws different axes through the two dimensional space; a fancier way to say this is that it rotates the points. Selecting the axes does not change the quality of the lower dimensional approximation. After inspecting how each question embeds in the k=2 Varimax coordinates, an analyst might find the Varimax axes to be meaningful; *linguistic* questions fall onto one axis and *mathematical* questions onto the other. This form of data analysis is often called "exploratory" because the factor dimensions are computed from the data without requiring an hypothesis to specify them.

The key source of the controversy is the second aim, producing axes that correspond to what Thurstone called scientifically meaningful categories. Anderson and Rubin [1956] showed that under the Gaussian factor model, the factors are *rotationally invariant*; there is nothing in the data to suggest where the axes should be drawn. Contemporary multivariate analysis textbooks all discuss the result from Anderson and Rubin [1956], but then go on to report the empirical benefits of the factor rotation (e.g. Ramsay and Silverman [2007], Johnson and Wichern [2007], Bartholomew et al. [2011]). For example, after discussing rotational invariance, Jolliffe [2002] says "The simplification achieved by rotation can help in interpreting the factors or rotated PCs."

Maxwell's Theorem starts to resolve this enigma [Maxwell [1860] and Feller [1971] Chapter 3, Section 4]. It characterizes the multivariate Gaussian distribution as the only distribution of independent random variables that is rotationally invariant. So, if the factors are independent random variables and come from any non-Gaussian distribution, then the axes are partially identifiable with the potential to identify scientifically meaningful categories. See Figure 1 for an example in $k = 2$ dimensions.[2]

Maxwell's theorem and some of the core factor analysis methodologies have been rediscovered and further developed in the literature on Independent Components Analysis (ICA) [Hyvärinen et al., 2004]. More recently, Anandkumar et al. [2014] showed how a

[1] PCA is not the preferred approach in Vintage Factor Analysis. See Remark 5.3 for a further discussion.

[2] A common point of confusion is to presume that the factors must be Gaussian if we are using PCA; see Section 5 and Remark 5.1 to see how PCA performs with non-Gaussian factors.

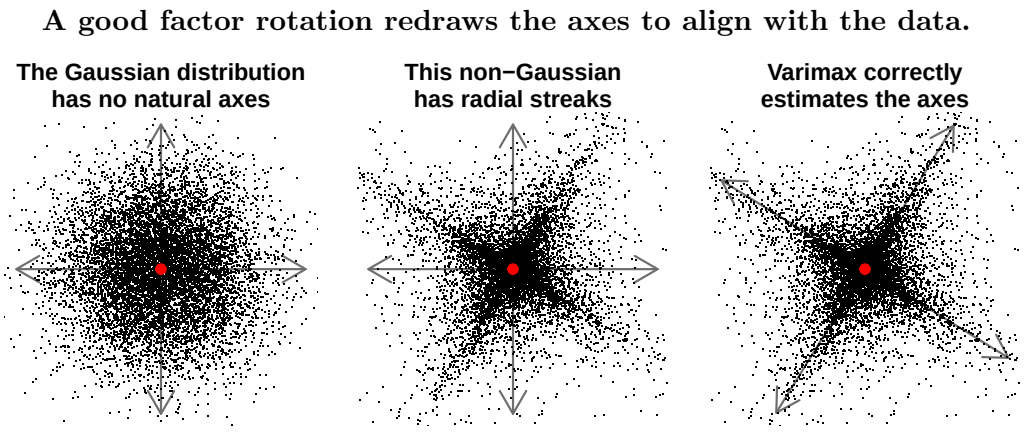


Figure 1: Maxwell's Theorem characterizes the multivariate Gaussian distribution (left panel) as the only rotationally invariant distribution of independent variables. The center panel and the right panel give the same data; the only difference is that the right panel gives the axes that Varimax estimates.

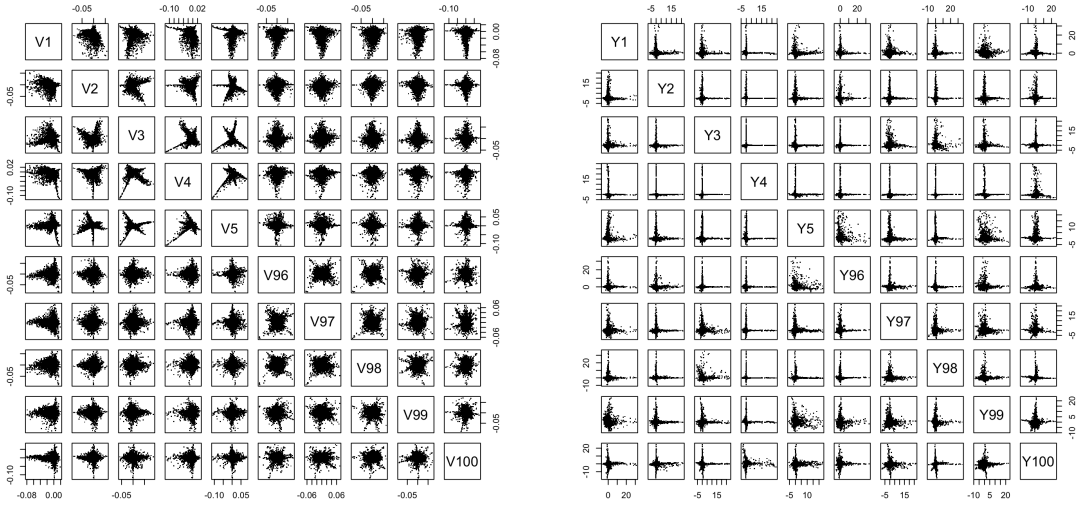
tensor decomposition can estimate a broad class of factor models that is closely related to the class studied herein. The current paper demonstrates that tensor methods are not required; an old approach with historical precedence to ICA is sufficient. This old approach comes with a suite of know-how and diagnostic practices that are described in Section 4. This old approach provides a unified spectral estimation strategy and diagnostic practices that can be applied to many different problems in multivariate statistics. It relates Projection Pursuit, Independent Components Analysis, Non-Negative Matrix Decompositions, Latent Dirichlet Allocation, and Stochastic Blockmodeling.

Figure 2 shows a motivating data example with a $22,688 \times 22,688$ matrix of citations among 22,688 academic journals, where $A_{ij} \in \{0, 1\}$ indicates if the papers in journal i cite the papers in journal j. Each panel in Figure 2(a) plots a pair of principal components against one another. Each panel in Figure 2(b) plots these components after the Varimax rotation (i.e. with the Varimax axes). Section 2 describes this procedure in more detail. See Section 3.1 for further details on the data and the data analysis in Figure 2.

All of the panels in Figure 2 display *radial streaks*, a phrase used in Thurstone [1947] to identify the axes. In Figure 2(b), the streaks are aligned with the coordinate axes. This is precisely the desired outcome of a factor rotation because *when the axes are aligned with the streaks, the resulting components are approximately sparse.* For this reason, this paper refers to Varimax rotated PCA as Vintage Sparse PCA (`vsp`). Vu and Lei [2013] referred to the vintage notion of subspace sparsity as *column-wise sparsity.* See Chen and Rohe [2020] for further discussion.

Theorem 6.1, the main result of this paper, shows that `vsp` can estimate the following semi-parametric factor model.

In this data example, the principal components (left) have radial streaks. Varimax draws new axes that align with the streaks (right). Varimax rotated PCA is Vintage Sparse PCA, `vsp`.



(a) Principal Components

(b) After Varimax rotation

Figure 2: In this example, the data is a $22,688 \times 22,688$ matrix of citations among 22,688 academic journals. Each small panel on the left is a scatter plot of two principal components. Each small panel on the right is a scatter plot of two Varimax rotated components. See Section 3.1 for more details.

Definition 1. *Let $Z \in \mathbb{R}^{n\times k}$ and $Y \in \mathbb{R}^{d\times k}$ be latent factor matrices. Under the* ***semi-parametric factor model****, we observe $A \in \mathbb{R}^{n\times d}$ which has independent elements and has expectation*

$$\mathbb{E}(A|Z,Y) = ZBY^T, \quad \textit{where } B \in \mathbb{R}^{k\times k} \textit{ is not necessarily diagonal.} \tag{1}$$

This model is semi-parametric because it does not make parametric assumptions on the distribution of Z, Y or $A|Z,Y$. Section C in the Appendix describes how this model includes the Stochastic Blockmodel, several of its generalizations, and Latent Dirichlet Allocation.

Importantly, in the semi-parametric factor model, the columns of Z are not the principal components of $\mathbb{E}(A|Z,Y)$. However, if the elements of Z are independently generated from a *leptokurtic* distribution, then a Varimax rotation of the principal components estimates Z. This means that the Varimax axes for the principal components will align with the axes (i.e. columns) of Z; they will have the same set of coordinates (up to statistical errors). The leptokurtic condition on the elements of Z is the key identifying assumption

for Varimax and `vsp`.

Definition 2. *For a random variable $X \in \mathbb{R}$ with four finite moments, let $\eta = \mathbb{E}(X)$ and define the jth centered moment as $\eta_j = \mathbb{E}(X - \eta)^j$ for $j = 2, 4$. The kurtosis of X is $\kappa = \eta_4/\eta_2^2$. The random variable X and its distribution are* ***leptokurtic*** *if $\kappa > 3$.*

For any Gaussian random variable, $\kappa = 3$. As such, $\kappa \neq 3$ indicates a non-Gaussian distribution. Roughly speaking, when $\kappa > 3$, the distribution has a heavier tail than Gaussian. Kurtosis κ was originally named and used by Pearson around 1900 to measure whether a symmetric distribution was Gaussian [Fiori and Zenga, 2009]. See Section 4 for further discussion of leptokurtosis.

After reading Section 1 and the algorithm in Section 2, one can read Sections 3, 4, 5 and 6 in any order. Section 7 should be read after Sections 5.1 and 5.2.

Section 1 introduces Varimax and gives both algebraic and geometric intuition for why it prefers "sparse axes". Section 2 describes the `vsp` algorithm and some variations on the algorithm. Section 3 illustrates how to interpret the results of `vsp` by applying it to a large citation network and a large text corpus. Section 4 provides intuition for the sparsity diagnostics developed in Thurstone [1935, 1947] to show that they implicitly assess the leptokurtic assumption. Section 5 gives the population results for PCA with latent variable models and population results for Varimax applied to these population principal components. Section 6 gives the main theoretical result, Theorem 6.1. Section 7 discusses what happens when the latent variables are not independent.

Key Notation: Let $\mathscr{O}(k) = \{R \in \mathbb{R}^{k\times k} : R^T R = RR^T = I_k\}$ denote the set of $k \times k$ orthonormal matrices. Let $\mathbf{1}_a \in \mathbb{R}^a$ be a column vector of ones. Let I_d denote the $d \times d$ identity matrix. For $x \in \mathbb{R}^d$, let $diag(x) \in \mathbb{R}^{d\times d}$ be a diagonal matrix with $diag(x)_{ii} = x_i$. For $M \in \mathbb{R}^{a\times b}$, define $M_i \in \mathbb{R}^b$ as the ith row of M and $\|M\|_{p\to\infty} = \max_i \|M_i\|_p$, for $p \geq 1$ and ℓ_p norm for vectors $\|\cdot\|_p$. Let $\|M\|_F$ be the Frobenius norm, $\|M\|$ be the spectral norm, $\|M\|_\infty$ be the maximum absolute row sum of M, and $\|M\|_{\max}$ be the maximum element of M in absolute value. For sequences $x_n, y_n \in \mathbb{R}$, define $x_n \asymp y_n$ to mean that $x_n \to \infty$ and $y_n \to \infty$ and there exists an N, ϵ, and c all in $(0, \infty)$ such that $x_n/y_n \in (\epsilon, c)$ for all $n > N$. Define $x_n \succeq y_n$ to mean that for any $\epsilon \in (0, \infty)$, there exists an $N < \infty$ such that for all $n > N$, $x_n/y_n > \epsilon > 0$. Define $[k] = \{1, \ldots, k\}$.

1 Varimax

Varimax is the most popular way of computing a factor rotation [Kaiser, 1958]. It is contained in the base `R` packages and, akin to `kmeans`, is so popular that it is often not properly cited. Ramsay and Silverman [2005] describes Kaiser's Varimax as an "invaluable tool in multivariate analysis".

Given an $n \times k$ matrix U, with columns that form an orthonormal basis (e.g. as in PCA), the Varimax rotation is the $k \times k$ orthogonal matrix that maximizes the following

function

$$v(R,U) = \sum_{\ell=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \left([UR]_{i\ell}^4 - \left(\frac{1}{n} \sum_{q=1}^{n} [UR]_{q\ell}^2 \right)^2 \right). \tag{2}$$

Kaiser [1958] suggests preprocessing U by normalizing each row to have sum of squares equal to one. We do not use this normalization herein.[3] In later work, Kaiser suggested removing this normalization [Kaiser, 1970, Kaiser and Rice, 1974].

Varimax is not convex; each solution has $k!\,2^k$ optima, all corresponding to the identical set of axes, but simply reorder the coordinates ($k!$) and changing their sign (2^k), neither of which changes the value of (2). In `R`, `varimax` is optimized via projected gradient ascent.

1.1 Varimax and sparsity

To see why the Varimax axes prefer sparsity, imagine a single point $(x_1, x_2) \in \mathbb{R}^2$ on the unit circle, $x_1^2 + x_2^2 = 1$. In this case, optimizing the axes is equivalent to deciding where to put this point on the circle. The Varimax objective is $x_1^4 + x_2^4 - 1$. To maximize $x_1^4 + x_2^4$, notice that

$$x_1^4 + x_2^4 = (x_1^2 + x_2^2)^2 - 2x_1^2 x_2^2 = 1 - 2x_1^2 x_2^2.$$

This is maximized at any "sparse point," where either $x_1 = 0$ or $x_2 = 0$. This argument extends to a single point on the unit sphere in higher dimensions, $x \in \mathbb{R}^d$,

$$\sum_{i=1}^{d} x_i^4 = (\sum_{i=1}^{d} x_i^2)^2 - 2(\sum_{i,j} x_i^2 x_j^2) = 1 - 2(\sum_{i,j} x_i^2 x_j^2).$$

This is maximized whenever all but one of the components is equal to zero.

Of course, we are not typically interested in sparsely representing a single point, but multiple points. To reach towards this, define $R(\theta)$ as a rotation matrix in $\mathbb{R}^2$,

$$R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

The left panel in Figure 3 gives a single data point x. The thicker blue line is the curve $(\theta, v(\theta, x))$ in polar coordinates, where the radius $v(\theta, x)$ is the Varimax objective after rotating x by $R(\theta)$. An angle θ^* that maximizes $v(\theta, x)$ (i.e. the radius of the blue line) is an angle that gives the optimal Varimax rotation, $R(\theta^*)$; there are four optimal values, all of which give the same axes. The optimal axes are displayed in thinner blue lines. They sparsely represent the single data point.

[3] In `R`, the function `varimax` has a default argument `normalize = TRUE`. Note that when U has orthogonal columns (as is the case for PCA) and normalization is not used, then the second term in Varimax is a constant function of the matrix R. In such cases, this term can be ignored without changing the optimum.

In the right panel of Figure 3, there are 5000 points $x_1, \ldots, x_{5000}$ distributed with radial streaks. Each data point creates $v(\theta, x_i)$, a "four petal flower," as in the left panel. Then, the Varimax objective function is the sum of these flowers, $\sum_{i=1}^{5000} v(\theta, x_i)$. The sum of the flowers is displayed as the thicker blue line in the right panel. The thinner blue lines gives the optimal axes, which align with the radial streaks in this data.

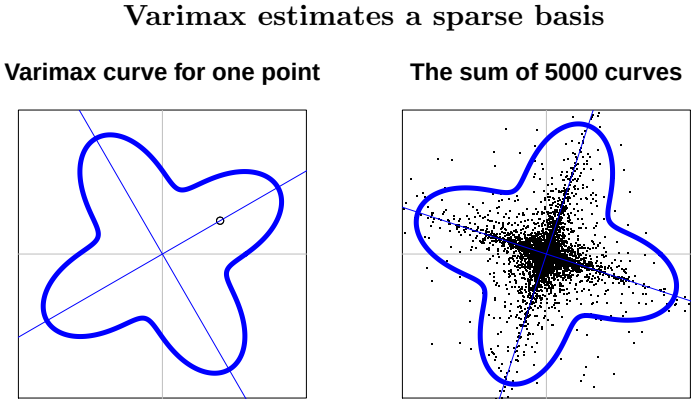


Figure 3: These curves are in polar coordinates, where the radius of the curve is the Varimax objective value for that angle. The optimal axes are displayed in blue. These axes provide an approximately sparse representation for the points because most points are close to the axes.

2 `vsp`: Vintage Sparse PCA

This section describes the methodological details of Vintage Sparse PCA (`vsp`). First, the algorithm is stated. Then, Remarks 2.1 and 2.2 describe ways in which `vsp` can be modified for certain settings; Table 1 summarizes these settings.

Algorithm: `vsp`

- Input $A \in \mathbb{R}^{n \times d}$ and desired number of dimensions k.

1. **Centering** (optional). Define row, column, and grand means,

$$\widehat{\mu}_r = A\mathbf{1}_d/d \in \mathbb{R}^n, \quad \widehat{\mu}_c = \mathbf{1}_n^T A/n \in \mathbb{R}^d, \quad \widehat{\mu}_\cdot = \mathbf{1}_n^T A\mathbf{1}_d/(nd) \in \mathbb{R}.$$

Here $\widehat{\mu}_r$ is a column vector and $\widehat{\mu}_c$ is a row vector. Define

$$\widetilde{A} = A - \widehat{\mu}_r\mathbf{1}_d^T - \mathbf{1}_n\widehat{\mu}_c + \widehat{\mu}_\cdot\mathbf{1}_n\mathbf{1}_d^T \in \mathbb{R}^{n \times d}. \tag{3}$$

2. **SVD**. If centering is being used, then compute the top k left and right singular vectors of $\widetilde{A}$, $\widehat{U} \in \mathbb{R}^{n\times k}$ and $\widehat{V} \in \mathbb{R}^{d\times k}$. These are the principal components and their loadings. Let $\widehat{D} \in \mathbb{R}^{k\times k}$ be a diagonal matrix containing the corresponding singular values. So, $\widetilde{A} \approx \widehat{U}\widehat{D}\widehat{V}^T$. If centering is not being used, then use the original input matrix A instead of $\widetilde{A}$. If A is large and sparse, steps 1 and 2 can be accelerated. See Remark B.

3. **Varimax**. Compute the orthogonal matrices that maximize Varimax, $v(R, \widehat{U})$ and $v(R, \widehat{V})$. Define them as $R_{\widehat{U}}, R_{\widehat{V}} \in \mathscr{O}(k)$ respectively.

- Output:

$$\widehat{Z} = \sqrt{n}\widehat{U}R_{\widehat{U}}, \quad \widehat{Y} = \sqrt{d}\widehat{V}R_{\widehat{V}}, \quad \text{and} \quad \widehat{B} = R_{\widehat{U}}^T\widehat{D}R_{\widehat{V}}/\sqrt{nd} \tag{4}$$

In modern applications where the row sums (or column sums) of A are highly heterogeneous, the degree-normalized version of A can be input into `vsp`.

Remark 2.1. *[Optional degree-normalization step] Define the row "degree", the row regularization parameter, and the diagonal degree matrix as*

$$deg_r = |A|\mathbf{1}_d \in \mathbb{R}^n, \quad \tau_r = \mathbf{1}_n^T deg_r/n \in \mathbb{R}, \quad D_r = diag(deg_r + \tau_r \mathbf{1}_n) \in \mathbb{R}^{n\times n},$$

where $|A|_{ij} = |A_{ij}|$. Similarly, define the column quantities deg_c, τ_c, D_c with $deg_c = \mathbf{1}_n^T|A| \in \mathbb{R}^d$ and $\tau_c = deg_c\mathbf{1}_d/d$. Define the normalized matrix as $L = D_r^{-1/2}AD_c^{-1/2}$. Then, input L to `vsp` *(instead of A). When using L,* `vsp` *estimates a normalized version of Z and Y. To undo this, the output of* `vsp` *could be renormalized as $D_r^{1/2}\widehat{Z}$ and $D_c^{1/2}\widehat{Y}$.*

Normalizing the matrix with the regularizer τ improves the statistical performance of spectral estimators derived from a sparse random matrix [Chaudhuri et al., 2012, Amini et al., 2013, Le et al., 2017]. In many empirical examples, the τ_r and τ_c prevent large outliers in the elements of the singular vectors that are created as an artifact of noise in sparse matrices [Zhang and Rohe, 2018]. In this paper, the degree-normalization step is used for the analyses in Section 3, but it is not studied in the main theorem.

The optional centering step (step 1 of `vsp`) plays a surprising role. In particular, Proposition 5.1 in Section 5 shows that if A is centered in step 1, then `vsp` estimates the centered factors in the semi-parametric factor model (i.e., $Z - \mathbb{E}(Z)$). See Remark 5.2 and Section 7.1.2 for more discussion. To estimate Z, instead of its column centered version, the output of `vsp` can be recentered as follows.

Remark 2.2. *[Optional recentering step] After running* `vsp` *with the centering step, it is possible to use the quantities already computed to recenter the estimated factors $\widehat{Z}$ and $\widehat{Y}$ as a post-processing step. This enables* `vsp` *to estimate Z instead of $Z - \mathbb{E}(Z)$. Define*

$$\widehat{\mu}_Z = \sqrt{n}\widehat{\mu}_c\widehat{V}\widehat{D}^{-1}R_{\widehat{U}}, \quad \text{and} \quad \widehat{\mu}_Y = \sqrt{d}\widehat{\mu}_r^T\widehat{U}\widehat{D}^{-1}R_{\widehat{V}} \tag{5}$$

and recenter the estimated factors as follows: $\widehat{Z} + \mathbf{1}_n \widehat{\mu}_Z$ *and* $\widehat{Y} + \mathbf{1}_d \widehat{\mu}_Y$*. If the renormalization step in Remark 2.1 is also used, then recenter before renormalizing. Section 5 and Appendix F.1 justify the estimator* $\widehat{\mu}_Z$*.*

Table 1 below lists the variations of `vsp` that are defined above.

Option	**Motivated when ...**
Centering	factor modeling, topic modeling, soft-clustering. See Remarks 2.2 and 5.2, Sections 7.1.2 and C.4
Recentering	the factor means are desired. See Theorem 6.1, Remark 5.2, Section F.1.
Avoid centering	hard-clustering, Stochastic Blockmodeling. See Sections 7.1.2 and C.3.
Degree-normalization	heterogeneous column sums or row sums in A. Used in the data example.
Renormalization	we want to estimate the distribution of the factors Z. See Remark 2.1.

Table 1: The motivation for each of the optional steps in `vsp`.

3 An example with Academic Bibliometrics

This section uses `vsp` to study academic citation patterns and abstracts from a corpus of over 200 million academic publications that are curated and provided by the Semantic Scholar project [Ammar et al., 2018].[4] In order to (1) identify academic areas or disciplines and (2) identify the large journals within these disciplines, Section 3.1 applies `vsp` to the citation patterns among academic journals. Then, in order to understand where and how "factor analysis" is used, Section 3.2 applies `vsp` to all abstracts that contain the phrase "factor analysis."

3.1 `vsp` on journal citations

We apply `vsp` to the citation patterns among academic journals and find that the columns of $\widehat{Y}$ identify academic disciplines or areas. For a small value of k, `vsp` factorizes journals into high level groupings (e.g. medicine, biology, physical sciences, mathematics, etc). For a large value of k, the academic areas are more resolved (e.g. pure mathematics vs. applied mathematics). This section uses degree-normalization, renormalization, centering, and recentering.

[4] `http://s2-public-api.prod.s2.allenai.org/corpus/`

In Figure 2 and in this subsection, the data matrix A is a $22,688 \times 22,688$ matrix. For each $i \in 1, \ldots, 22,688$, the ith row and column of A corresponds to a unique journal name in the Semantic Scholar database (after putting all letters in lower case and removing all punctuation). For computational ease, we took a simple random sample of 5% of the paper citations.[5] If there were more than five citations from the papers in journal i to the papers in journal j in this 5% sample, then $A_{ij} = 1$, otherwise $A_{ij} = 0$. There were roughly 100,000 journals that appeared in the database, but only 22,688 remain after the sampling and thresholding described above. While A is a square matrix, it is not symmetric because a citation is directed from one paper to another.

This matrix is sparse with heterogeneous row and column sums. There are 474,841 non-zero elements in A, roughly 1/1000 of the elements, making the average row and column sum roughly 20. The median row sum is four. The median column sum is two. PLOS ONE has the largest row sum, 5,556. Nature has the largest column sum, 4,413. The next table gives the column and row sums for Journal of the Royal Statistical Society-Series B (JRSS-B) , Annals of Statistics (AOS), Journal of the American Statistical Association (JASA), Annals of Probability (AOP), Nature, PLOS ONE, Proceedings of the National Academy of Sciences (PNAS), and The New England Journal of Medicine (NEJM).

	JRSS-B	AOS	JASA	AOP	Nature	PLOS ONE	PNAS	NEJM
column sum	178	146	462	59	4413	3176	3928	3209
row sum	16	45	51	28	522	5556	1283	284

Because the column and row sums of A have a heavy tail, we used the degree-normalization described in Remark 2.1. The sparsity in the data matrix makes `vsp` quick to compute. In `R`, on a 2.3 GHz Macbook Pro, it takes 1.3 seconds for $k = 10$, 13 seconds for $k = 50$, and 23 seconds for $k = 50$.

Notice that the columns of A measure how widely a journal is cited. For this reason, the $\widehat{Y}$ matrix in `vsp`, which embeds the *columns* of A, reveals how widely a journal *receives* citations. We will refer to each column of $\widehat{Y}$ as a factor. So, if element $\widehat{Y}_{ij}$ is large, it suggests that journal i is a more central or prestigious journal in factor j. Because the rows of A measure how a journal cites other journals, the elements in $\widehat{Z}$ reveal how widely the journal sends citations [Rohe et al., 2016a]. Here, we will focus on $\widehat{Y}$.

Figure 4 plots the largest 300 squared singular values of L. Inspecting this scree plot, it seems that the typical analyst would hesitate to make k larger than 50. However, with $k = 100$ there continues to be radial streaks in $\widehat{V}$ that Varimax aligns with the axes in $\widehat{Y}$; Figure 2 shows columns $1, 2, 3, 4, 5, 96, 97, 98, 99$, and 100 of $\widehat{V}$ (on the left) and $\widehat{Y}$ (on the right). The leading columns of $\widehat{V}$ have a few radial streaks when they are plotted against one another. The trailing columns of $\widehat{V}$ show multiple streaks within each plot. The leading columns of $\widehat{Y}$ have streaks that are tightly aligned with the axes; the trailing columns, even

[5] Specifically, the population of this sample is the edges (u, v) between *papers*.

with $k = 100$ are axis aligned. These later factors are more diffuse, suggesting that they contain more noise.[6]

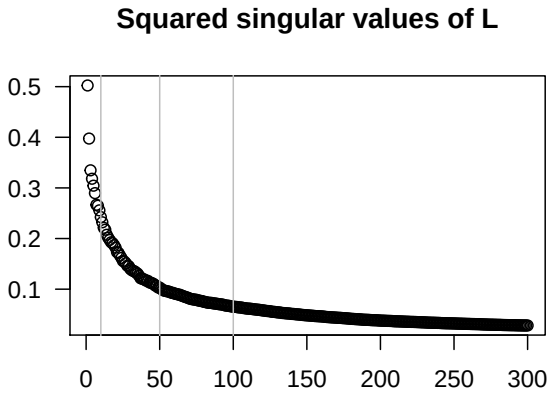


Figure 4: The first 300 squared singular values of L are plotted, along with lines at $k = 10, 50$, and 100.

3.1.1 Journal factors with $k = 10$

We interpret the meaning of a factor in $\widehat{Y}$ by (1) finding external features that correlate with that column and then (2) examining the journals that have the largest values in that column. For external features, we construct the document-term matrix from the journal titles. Define $X \in \{0,1\}^{22{,}688 \times 2397}$ where $X_{i\ell} \in \{0,1\}$ indicates whether the title for journal i contains word ℓ. Due to the sparse and heterogeneous nature of $\widehat{Y}$ and X, simple correlations are unstable. We have found better results with the following "best feature function" `bff` [Wang and Rohe, 2016]. For each factor j, define the sets $in(j) = \{i : \widehat{Y}_{ij} \geq 0\}$ and $out(j) = \{i : \widehat{Y}_{ij} < 0\}$. Define the importance of word ℓ in factor j as

$$\texttt{bff}(j,\ell) = \sqrt{\frac{\sum_{i \in in(j)} \widehat{Y}_{ij} X_{i\ell}}{\sum_{i \in in(j)} \widehat{Y}_{ij}}} - \sqrt{\frac{\sum_{i \in out(j)} X_{i\ell}}{|out(j)|}}.$$

Using $k = 10$, `vsp` finds a high level grouping of disciplines. For each factor $j = 1, \ldots, 10$, the largest seven elements of `bff` are given below.

1. medicine, surgery, clinical, american, cancer, official, oncology

[6] In later work, Chen et al. [2021] developed a resampling procedure to examine whether a column of $\widehat{V}$ is statistically significant. Figure 3 in that paper shows that the first 150 eigenvectors on a symmetrized version of the journal citation graph are all highly statistically significant.

2. molecular, cell, biology, immunology, microbiology, genetics, nature
3. psychology, psychiatry, neuroscience, brain, neurology, behavior, psychological
4. materials, chemistry, physics, chemical, physical, energy, polymer, engineering
5. ecology, plant, biology, evolution, microbiology, marine, environmental
6. geology, earth, geological, geophysical, planetary, atmospheric, geophysics
7. ieee, on, conference, transactions, computer, pattern, vision
8. mathematical, mathematics, arxiv, physics, geometry, analysis, differential
9. economics, economic, review, management, finance, statistics, financial
10. oral, dentistry, dental, surgery, orthodontics, maxillofacial, periodontology

Figure 5 plots factor 1 "medicine" against all of the others; each dot is a journal. "Medicine" has a mixing pattern with factor 2 "small-scale biology", because multiple journals rank highly in both. With factor 3 "psych/neuro", there is less mixing, but still some. For the other factors, there is nearly zero mixing, making the radial streaks increasingly pronounced.

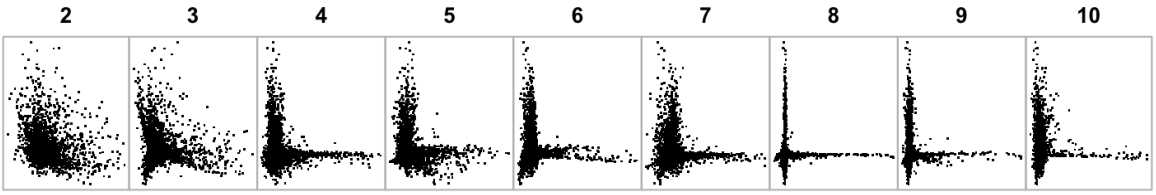


Figure 5: Each dot is a journal. The vertical axis gives factor 1, the "medicine" factor. The horizontal axis gives the other factors, $2, \ldots, 10$. If you squint, some panels have multiple horizontal streaks (e.g. factor 6); with a larger choice of k, these streaks reveal themselves to be buds that unfurl and branch into their own axes, in a hierarchical tree fashion.

The factors in $\widehat{Z}$ identify the same academic areas as $\widehat{Y}$. The leading `bff` terms for $\widehat{Z}$ are given in Section D.1 and the top eleven journals for both $\widehat{Z}$ and $\widehat{Y}$ are given in Section D.2. The difference between $\widehat{Z}$ and $\widehat{Y}$ is that the largest elements in $\widehat{Y}$ tend to identify the more prestigious journals in that academic area, whereas the largest elements in $\widehat{Z}$ tend to identify the journals that publish the most papers (and thus send the most citations) in that academic area. For example, the largest five elements in the first factor of $\widehat{Y}$ are highly prestigious journals: JAMA, The New England Journal of Medicine, The Lancet, Annals of Internal Medicine, and BMJ, in descending order. In the first column of $\widehat{Z}$, *none* of these journals are among the largest 20 elements. Instead, the leading journal in the first column of $\widehat{Z}$ is Medicine, an open access journal akin to PLOS ONE.

3.1.2 The middle B matrix

Interpreting $\widehat{B}$ can be challenging. Vintage factor analysis does not typically include this matrix.[7] In PCA, there is a diagonal matrix of eigenvalues that is mildly analogous to B; typically these eigenvalues are absorbed into the components. When $\widehat{B}$ is not strictly diagonal, it describes how the factors in $\widehat{Z}$ relate to the factors in $\widehat{Y}$. The Stochastic Blockmodel, further discussed in Section C.3, provides the most expedient interpretation for the B matrix. It is the first parametric model to include the "middle B matrix" [Holland et al., 1983]. Under the Stochastic Blockmodel, the matrices Z (and Y) have a single one in each row and the rest of the elements are zero. If $Z_{ij} = 1$, then we say "person i is in block j." In that model, B_{uv} gives the probability that a person in block u is friends with a person in block v. Zhang et al. [2014], Jin and Ke [2017] generalized this model to allow people to have non-negative, weighted memberships in each block. In this generalization, the middle B matrix has an analogous interpretation. In order to adopt that interpretation here, the elements of $\widehat{Z}, \widehat{Y}$, and $\widehat{B}$ must be non-negative.

Figure 6, in the left panel, gives the matrix $\widehat{B}$. Indeed, it is hard to interpret. The middle panel gives the *non-negative interpretation*, defined as follows. For any matrix M, define M_+ to be equal to M, except setting the negative elements to zero. Define non-negative interpretation (nni) for $\widehat{B}$ as

$$\widehat{B}^{nni} = \left[(\widehat{Z}_+^T \widehat{Z}_+)^{-1} \widehat{Z}_+^T A \widehat{Y}_+ (\widehat{Y}_+^T \widehat{Y}_+)^{-1} \right]_+ . \tag{6}$$

In Figure 6, the non-negative interpretation of $\widehat{B}$ has a clear diagonal structure, which is consistent with our understanding that journals in the same area are more likely to cite one another than journals from separate areas.

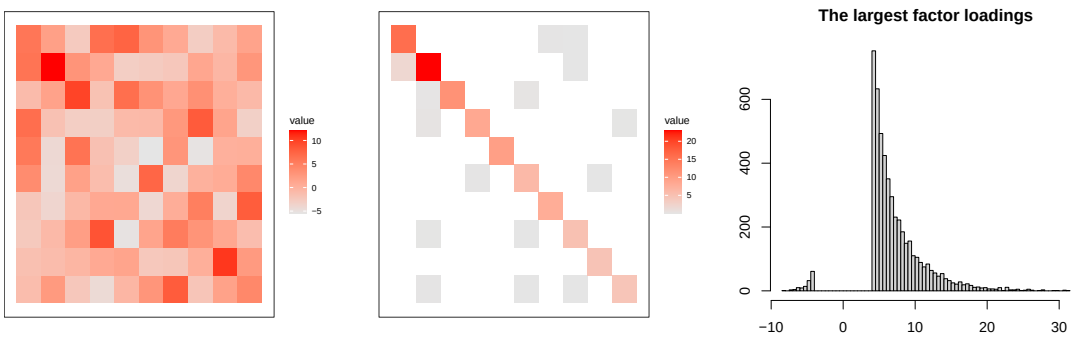
While it seems strange to threshold away all of the negative values, this step is not as severe as it first sounds. The right panel in Figure 6 gives a histogram of the elements in $\widehat{Z}$ and $\widehat{Y}$ that are larger than 4 in absolute value. The largest values are all positive. This is because, empirically, the factors estimated by Varimax tend to be "one-sided," with large skewness.[8] Following Kaiser and Rice [1974], we change the signs of all factors to ensure the skewness is positive. With $k = 10$, the median skewness of the 20 factors in $\widehat{Z}$ and $\widehat{Y}$ is 8.1 and all but one of the factors has skewness greater than 2. Because of this, thresholding away the negative values enables a clearer interpretation of $\widehat{B}$.

3.1.3 The factors become more refined as k increases

As k increases, the factors provide a more refined specification of academic areas; this refinement is roughly hierarchical, e.g. "medicine" splits into different areas. However, it is

[7] It does appear in Harris and Kaiser [1964]; see Section 7.2 for more.

[8] The skewness for a random variable is $\eta_3/\eta_2^{3/2}$ where the η's are the centered moments defined in Definition 2. Symmetric random variables have zero skewness and the exponential distribution, which seems quite skewed, has skewness of two. In this section, we are discussing empirical moments.



(a) $\widehat{B}$ (b) $\widehat{B}^{nni}$ (c) Largest elements of $\widehat{Z}$ and $\widehat{Y}$

Figure 6: The $\widehat{B}$ matrix can be hard to interpret. $\widehat{B}^{nni}$ provides a clearer picture; it is constructed by thresholding away the negative values in $\widehat{Z}, \widehat{Y}$ as in Equation (6). The right panel justifies this thresholding, by showing that the largest values in $\widehat{Z}, \widehat{Y}$ are positive.

not perfectly hierarchical. This can be seen in the loadings for AOP (Annals of Probability) for k increasing from 10, to 50, to 100. In the factoring with $k = 10$, AOS, JASA, and JRSS-B have their largest loading values in both factors 7 and 9; that is, the rows of $\widehat{Y}$ corresponding to these journals have their largest values in columns 7 and 9. Meanwhile, AOP has its largest loading value in factor 8 (mathematics). None of these journals rank among the highest forty journals in these factors. Increasing to $k = 50$, AOS, JASA, JRSS-B, and AOP combine into a "Probability and Statistics" factor. Despite the fact that there is another factor of prestigious math journals (Inventiones Mathematicae, Annals of Mathematics, etc), AOP has its highest loading in the "Probability and Statistics" factor. The journals with the largest 20 elements in the "Probability and Statistics" factor are given in a text box below, in decreasing order. AOS, JASA, JRSS-B, and AOP all rank highly in this factor. This merging pattern is completely sensible and yet not strictly hierarchical.

The top 20 journals in "Probability and Statistics" in $k = 50$ factoring. annals of statistics, annals of mathematical statistics, journal of statistical planning and inference, journal of multivariate analysis, biometrika, statistics probability letters, journal of the royal statistical society series b statistical methodology, statistical science, scandinavian journal of statistics, annals of probability, technometrics, journal of computational and graphical statistics, comput stat data anal, journal of the american statistical association, bernoulli, journal of applied probability, stochastic processes and their applications, annals of the institute of statistical mathematics, biometrics, probability theory and related fields.

Increasing to $k = 100$, the "Probability and Statistics" factor from $k = 50$ splits in a hierarchical fashion into two separate factors, one for "Statistics" and one for "Probability." Table 2 gives the first `bff` term for each of these 100 factors in $\widehat{Y}$. The leading five journals in each of these factors is given in Appendix D.3.

gastroenterology	microbiology	infectious	marketing	alcohol	urology	comb
cardiovascular	microbiology	management	materials	control	animal	food
communications	neuroscience	nephrology	mechanics	ecology	cancer	ieee
pharmaceutical	parasitology	obstetrics	neurology	ecology	comput	ieee
otolaryngology	pharmacology	psychiatry	nutrition	geology	energy	ieee
rehabilitation	rheumatology	psychology	numerical	nursing	health	oper
transportation	atmospheric	psychology	political	optical	marine	oral
communication	dermatology	quaternary	radiology	physics	nature	soil
endocrinology	probability	statistics	sociology	physics	sports	inf
environmental	accounting	toxicology	circuits	polymer	speech	de
ophthalmology	anesthesia	veterinary	genetics	sensing	vision	
astrophysics	analytical	chemistry	language	surgery	aging	
geotechnical	entomology	economics	language	surgery	child	
mathematical	immunology	education	robotics	surgery	fuzzy	
mathematical	immunology	geography	software	tourism	plant	

Table 2: For each of the $k = 100$ factors, this table gives the top `bff` term. While there are 9 `bff` terms that repeat for more than one factor (e.g. mathematical), these repeated factors identify subdisciplines within these areas (e.g. one of the math factors finds "applied math" journals and the other appears to find "pure math" journals). See Appendix D.3 for the leading five journals in all 100 factors.

3.1.4 How should we choose k?

In this example, the screeplot in Figure 4 suggests a value of k much smaller than 100. However, there is no evidence of over-factoring in the $k = 100$ factoring above. First, in Figure 2, there are still radial streaks in the principal components all the way up to the 96th, 97th, 98th, 99th, and 100th components and these streaks rotate to become axis aligned. Second, in Appendix D.3, the journals with the largest loadings in each of the 100 factors neatly identify academic areas.

While there is certainly an upper bound for k, beyond which the factors behave like noise and fail to provide meaningful insights, in this example with academic journals, the screeplot is not helpful in detecting this upper bound. When inspecting the screeplot to select k, do not mind the eigen-gap too much.

In addition to the meaningful factoring at $k = 100$, the factors at $k = 10$ are also meaningful. This is a common empirical phenomenon; many times the factors have some-

thing resembling a hierarchical structure. Perhaps the $k = 10$ results are more suited for a certain task at hand. The *Cheshire Cat Rule* says that there is not a single correct answer for the choice of k, that the answer depends upon where you want to go.

> Alice: Would you tell me, please, which way I ought to go from here?
> The Cheshire Cat: That depends a good deal on where you want to get to.
> Alice: I don't much care where.
> The Cheshire Cat: Then it doesn't much matter which way you go.
>
> — Lewis Carroll, Alice in Wonderland

3.2 How and where is "factor analysis" used?

This section describes where and how "factor analysis" is used. We study the 144,136 papers in the Semantic Scholar database that contain "factor analysis" in the title or abstract (case insensitive) and for which the abstract is classified as English by Compact Language Detector 3 [Salcianu and et al, 2020].

3.2.1 Where is factor analysis used?

In order to find *where* factor analysis is used, we examine where these papers appear in the $\widehat{Y}$ journal embedding above. Of the 144,136 papers, 64,873 were published in a journal that was included in that analysis. For each of these 64,873 papers, take its journal's row of $\widehat{Y}$ from the $k = 100$ analysis and place it into the rows of a new $64,873 \times 100$ matrix. Columns of this matrix with a large sum correspond to places in the academic literature where factor analysis appears.

In descending order, the largest 16 of these 100 journal factors are child–psychology, psychiatry–psychiatric, psychology–social, nursing–nurse, health–care, rehabilitation–occupational, environmental–water, aging–gerontology, nutrition–obesity, alcohol–health, education–educational, analytical–chromatography, tourism–hospitality, toxicology–environmental, management–business, and statistics–statistical. The column with the smallest sum is probability–annals. Each of these factor names is constructed from the first two `bff` terms for that factor (Table 2 only gives the first).

This exploratory analysis has numerous lurking variables, such as the number of papers published within each factor and the likelihood that a paper using factor analysis includes "factor analysis" in the title or abstract. That said, it is not surprising that psychology, psychiatry, and statistics rank high, while probability ranks low.

3.2.2 How is factor analysis used?

In order to explore how "factor analysis" is used, we study the document-term matrix A constructed via the tidytext package [Silge and Robinson, 2016] constructed with the 144,136 abstracts. There are 240,331 unique words in the corpus. So, $A \in \{0, 1\}^{144,136 \times 240,331}$

with A_{ij} indicating whether abstract i contains word j. Just as in Section 3.1, A is sparse with highly heterogeneous column sums. It contains 16.8 million non-zero elements, which averages to 117 terms per document. The column sums of A are highly skewed; the median is 1, the average is 70, and 12 terms appear in over 100,000 documents. Stop words (e.g. the, of, and, to) have not been removed.

With $k = 50$, `vsp` takes 72 seconds.[9] For comparison, constructing A from the $144,136$ abstracts represented as character strings requires 23 seconds in tidytext.

Seven of the $k = 50$ factors appear to focus on words that are more "methodological." These seven are listed below; in bold font is a name we assigned based upon our interpretation, following that are the ten words with the largest loading values in $\widehat{Y} \in \mathbb{R}^{240,331 \times 50}$. In addition, we find 32 "subject area" factors. These subject area factors use discipline specific words. These 32 factors echo the journal factors listed in Section 3.2.1 (e.g. Environment, Nutrition, Psychology, etc). See Section E.1 for these factors. The final eleven factors are artifacts and anomolies that are further discussed in Section E.2.

The seven methodological factors.
item–response–theory: consistency, cronbach, internal, validity, reliability, retest, version, alpha, psychometric, properties
modern–factor–models: algorithm, bayesian, estimation, carlo, monte, simulation, algorithms, likelihood, inference, markov
confirmatory–factor–analysis(cfa): invariance, across, fit, confirmatory, measurement, scalar, configural, multigroup, cfa, metric
structural–equation–modeling(sem): equation, structural, modeling, sem, confirmatory, model, mediating, intention, modelling, amos
cfa–sem–summaries: rmsea, cfi, gfi, df, agfi, nfi, tli, srmr, root, approximation
qualitative–research: literature, review, development, develop, management, implementation, process, experts, qualitative, interviews
vintage–factor–analysis: olkin, kaiser, meyer, bartlett, sphericity, kmo, rotation, varimax, principal, adequacy

[9] In `R` on a 2020 MacBook Pro with 2.3 GHz Intel i7.

4 Thurstone's diagnostics assess whether Varimax can identify the axes

> Factors are *rotational invariant* because redrawing the factor axes does not change the fit to the data. In linear regression with more features than samples ($p > n$), there is also an invariance. However, we now know that if there is a sparse solution, it can be unique. Decades before sparsity became popular for removing the invariance in $p > n$ regression, Thurstone proposed using sparsity to remove rotational invariance in factor analysis. His sparsity diagnostics are still used routinely in practice. Theorem 4.1 shows that sparsity implies the key leptokurtic condition that is sufficient for Varimax to identify the axes. In this way, Vintage Factor Analysis performs statistical inference.

Step 2 of `vsp` approximates $\widetilde{A}$ with the leading k singular vectors, $\widetilde{A} \approx \widehat{U}\widehat{D}\widehat{V}^T$. Step 3 computes the Varimax rotations of $\widehat{U}$ and $\widehat{V}$. However, for any rotation matrices $R_1, R_2 \in \mathscr{O}(k)$, rotating $\widehat{U}$ and $\widehat{V}$ does not change the approximation to $\widetilde{A}$,

$$\widehat{U}\widehat{D}\widehat{V}^T = (\widehat{U}R_1)(R_1^T\widehat{D}R_2)(\widehat{V}R_2)^T,$$

where the rotated factor matrices $\widehat{U}R_1$ and $\widehat{V}R_2$ still have orthonormal columns. As such, no rotation can improve the approximation to $\widetilde{A}$. Many have interpreted this to imply that we can never estimate factor rotations from data. This is the misunderstanding of rotational invariance.

In an attempt to resolve the rotational invariance, Thurstone developed a new type of data analysis to find rotations $R_{\widehat{U}} \in \mathscr{O}(k)$ such that $\widehat{U}R_{\widehat{U}}$ is sparse [Thurstone, 1935, 1947]. He developed a suite of tools and diagnostics to assess this sparsity and many of these remain in use today. They are described in modern textbooks, built into the base `R` packages for factor analysis, and used routinely in practice. Section 4.1 describes these diagnostic practices. Section 4.2 and Theorem 4.1 show how these diagnostics can be reinterpreted as assessing whether the factors come from a leptokurtic distribution which is a key condition for Varimax to be able to identify the correct factor rotation in Theorems 5.1 and 6.1.

4.1 Thurstone's simple structure and diagnostics

Thurstone [1935] and Thurstone [1947] propose using sparsity to remove the rotational invariance. "In numerical terms this is a demand for the smallest number of non-vanishing entries in each row of the ... factor matrix. It seems strange indeed, and it was entirely unexpected, that so simple and plausible an idea should meet with a storm of protest from the statisticians" [p333 Thurstone [1947]]. Thurstone refers to this sparsity as *simple structure*. Thurstone's use of sparsity is analogous to the modern use of sparsity in high dimensional regression and underdetermined systems of linear equations. In these more

modern problems, without any sparsity constraint, there is a large space of plausible solutions. However, under certain conditions, the sparse solution is unique. This intuition is analogous to Thurstone's intuition for resolving rotational invariance.

Thurstone implemented techniques to find rotations which produce sparse solutions, but he struggled to find any assurance that the computed solution is the sparsest solution (i.e. unique). "When [a solutions has] been found which produces a simple structure, it is of considerable scientific interest to know whether the simple structure is unique... The necessary and sufficient conditions for uniqueness of a simple structure need to be investigated. In the absence of a complete solution to this problem, five criteria will here be listed which probably constitute sufficient conditions for the uniqueness of a simple structure" [p334 Thurstone [1947]]. Thurstone's five conditions are quoted in Appendix A. They motivate his *radial streaks* diagnostic, illustrated in Figure 2.

If the diagnostic plots do not show radial streaks, Thurstone suggests that one should proceed more cautiously. A few pages after giving the five criteria for simple structure, Thurstone gives a diagnostic plot with points evenly spaced inside a circle (i.e., like the Gaussian in Figure 1) and explains what happens when you have loadings that appear to come from a rotationally invariant distribution. "A figure such as [this] leaves one unconvinced, no matter where the axes are drawn, unless an interpretation can be found that seems right. Random configurations like this seldom yield clear interpretations, but they are not, of course, physically impossible."

The current paper creates a statistical theory around Thurstone's key ideas by presuming that the factors are generated as random variables from a statistical model and using the Varimax estimator. Thurstone does not presume the latent factors are generated from a probability distribution, and as such, he lacks any statistical notion of the true axes to be inferred. His notion of uniqueness is more akin to the uniqueness of an optimization solution.

Thurstone computed rotations by hand and human judgement. Only after Thurstone's death in 1955 did it become popular to compute factor rotations such as Varimax on "electronic computers" with numerical optimization techniques. In $k = 2$ dimensions, Kaiser [1958] gives a a unique closed form solution to Varimax. In $k > 2$, if one assumes the models in this paper, then the maximizer to Varimax is unique (up to permutations and sign flips). However, under lesser assumptions, uniqueness remains an open problem.

4.1.1 Simple structure in contemporary multivariate statistics

Contemporary textbooks on multivariate statistics still suggest that the rotated factors or the rotated principal components should be inspected to see if they are sparse [Mardia et al., 1979, Jolliffe, 2002, Johnson and Wichern, 2007, Bartholomew et al., 2011]. These textbooks all share the empirical observation that it is often easier to interpret factors which have been rotated for sparsity. The given reason is that sparse factors are "simpler." While this appears to use Thurstone's word, these texts do not discuss whether or not

this simple structure might resolve the problem of rotational invariance. Rather, it is an empirical observation that sparse and simple solutions are easier to interpret. For example, Ramsay and Silverman [2007] says "It is well known in classical multivariate analysis that an appropriate rotation of the principal components can, on occasion, give components ... more informative than the original components themselves." Johnson and Wichern [2007] says "A rotation of the factors often reveals a simple structure and aids interpretation." Bartholomew et al. [2011] says "Rotation assumes a very important role when we come to the interpretation of latent variables."

The notion that the data analyst should inspect the factors for sparsity is built into the `print` function for factor loadings in the base `R` packages; if a loading is less than the `print` argument `cutoff` then instead of printing a number, it appears as a whitespace.

This paper shows that sparsity does not merely make the factors simpler; sparsity enables statistical identification and inference. *Sparsity* and *radial streaking* are two distinctively non-Gaussian patterns. As such, Thurstone's visualizations and diagnostics can be reinterpreted as assessing whether the factors are generated from a non-Gaussian distribution and thus, by Maxwell's theorem, whether the rotation is statistically identifiable. Moreover, Theorem 4.1 in the next subsection shows that sparsity implies leptokurticity, the key identifying assumption for Varimax.

4.2 Kurtosis and sparsity

The next theorem shows that Thurstone's sparsity diagnostics can be reinterpreted as assessing an identifying assumption for Varimax.

Theorem 4.1. *Any random variable X that satisfies $\frac{5}{6} < \mathbb{P}(X = 0) < 1$ and has four finite moments is leptokurtic.*

For example, suppose $X \sim Bernoulli(p)$ with $q = P(X = 0)$. Theorem 4.1 implies that if $q > 5/6 \approx .83$, then X is leptokurtic. For comparison, in this specific case of the Bernoulli distribution, X is leptokurtic if q (or p) is greater than $(\sqrt{3}+1)(2\sqrt{3})^{-1} \approx 0.79$. While Theorem 4.1 does not provide a sharp results for the Bernoulli distribution, .83 is close to .79. Moreover, Theorem 4.1 applies to any random variable, it does not make any parametric assumptions, and the moment assumptions are only so that kurtosis is defined. See Section G.1 in the Appendix for a proof. This theorem assumes hard sparsity (i.e., $\mathbb{P}(X = 0) > 0$) for technical convenience. See Appendix G.2 for a discussion about softer forms of sparsity.

5 Mathematical intuition for `vsp` with population results

This section studies each of the three steps in `vsp` by studying their population behavior. Statistical convergence around the population quantities is rigorously treated in Theorem 6.1 in Section 6.

5.1 Population results for two layers of randomness

The semi-parametric factor model is a latent variable model with two sequential layers of randomness. In the first layer of randomness, the latent variables Z and Y are generated. In the second layer, the observed matrix A is generated, conditionally on the latent variables. To parallel these two layers, there are two types of population results given in the next two subsections.

Propositions 5.1 and 5.2 study the first two steps of `vsp` applied to the population matrix

$$\mathscr{A} = \mathbb{E}(A|Z, Y) = ZBY^T \tag{7}$$

instead of A. These propositions imply that the population principal components can be expressed as $\widetilde{Z}R$, where $\widetilde{Z} \in \mathbb{R}^{n\times k}$ is Z after column centering and $R \in \mathbb{R}^{k\times k}$ is defined below. If the nk many random variables in $Z \in \mathbb{R}^{n\times k}$ are mutually independent, then R converges to a rotation matrix. These results in Section 5.2 allows for the randomness in Z and Y, but they remove the second layer of randomness by using $\mathscr{A}$ instead of A. Then, Section 5.3 studies the population version of the Varimax step. To do this, take the expectation of the Varimax objective function, evaluated at the population principal components (i.e., $\widetilde{Z}R$), where the expectation is over the distribution of Z. This expectation removes the randomness in Z. Under the identification assumptions for Varimax defined below, Theorem 5.1 shows that the rotation that maximizes this function is $R^T \in \mathscr{O}(k)$. So, rotating the population principal components with the population Varimax rotation yields the original factors, $(\widetilde{Z}R)R^T = \widetilde{Z}$.

5.2 PCA for latent variable models; population results

Define $\bar{Z} \in \mathbb{R}^{n\times k}$ such that $\bar{Z}_{ij}$ equals the sample mean of the jth column of Z. Similarly for $\bar{Y} \in \mathbb{R}^{d\times k}$. Define

$$\widetilde{Z} = Z - \bar{Z} \quad \text{and} \quad \widetilde{Y} = Y - \bar{Y}. \tag{8}$$

Proposition 5.1. *[Step 1 of `vsp`] Centering $\mathscr{A}$ to get $\widetilde{\mathscr{A}}$ as in Equation* (3)*, has the effect of centering Z and Y.*

$$\widetilde{\mathscr{A}} = \widetilde{Z}B\widetilde{Y}^T$$

This does not require any distributional assumptions on Z or Y.

A proof is given in Appendix F. The next proposition gives the SVD of $\widetilde{\mathscr{A}} = \widetilde{Z}B\widetilde{Y}^T$. Define

$$\widehat{\Sigma}_Z = \widetilde{Z}^T\widetilde{Z}/n, \quad \widehat{\Sigma}_Y = \widetilde{Y}^T\widetilde{Y}/d,$$

and define $\widetilde{R}_U, \widetilde{R}_V \in \mathscr{O}(k)$, and diagonal matrix $\widetilde{D}$ to be the SVD of $\widehat{\Sigma}_Z^{1/2}B\widehat{\Sigma}_Y^{1/2} \in \mathbb{R}^{k\times k}$,

$$\widehat{\Sigma}_Z^{1/2}B\widehat{\Sigma}_Y^{1/2} = \widetilde{R}_U^T\widetilde{D}\widetilde{R}_V. \tag{9}$$

The next proposition shows that the rotation matrices $\widetilde{R}_U$ and $\widetilde{R}_V$ convert the factor matrices $\widetilde{Z}$ and $\widetilde{Y}$ into the principal components and loadings U and V.

Proposition 5.2. *[Step 2 of* `vsp`*] Define the following matrices,*

$$U = n^{-1/2}\widetilde{Z}\ \widehat{\Sigma}_Z^{-1/2}\widetilde{R}_U^T, \quad D = \sqrt{nd}\widetilde{D}, \quad V = d^{-1/2}\widetilde{Y}\widehat{\Sigma}_Y^{-1/2}\widetilde{R}_V^T. \tag{10}$$

Then, $\widetilde{\mathscr{A}} = UDV^T$, where U and V contain the left and right singular vectors of $\widetilde{\mathscr{A}}$ and D contains the singular values of $\widetilde{\mathscr{A}}$. This does not require any distributional assumptions on Z or Y.

The proof requires demonstrating the equality $\widetilde{\mathscr{A}} = UDV^T$ and showing that U and V have orthonormal columns. Substituting in the definitions reveals this result. Taken together, Propositions 5.1 and 5.2 show that the first two steps of `vsp` on $\mathscr{A}$ compute $U \propto \widetilde{Z}\ \widehat{\Sigma}_Z^{-1/2}\widetilde{R}_U^T$; these are the principal components of $\mathscr{A}$.

Remark 5.1. *[Relationship between PCA and the factors] Proposition 5.2 relates PCA on the population matrix $\mathscr{A}$ to the factors Z. This is because the population principal components are the columns of the matrix*

$$U = n^{-1/2}\widetilde{Z}\ \widehat{\Sigma}_Z^{-1/2}\widetilde{R}_U^T. \tag{11}$$

So, the principal components are the centered latent factors $\widetilde{Z}$, orthogonalized with $\widehat{\Sigma}_Z^{-1/2}$, and rotated by a $k \times k$ nuisance matrix $\widetilde{R}_U^T$. Despite the fact that PCA is typically considered a second order technique, this result implies that the principal components themselves do not retain any first or second order information about the latent factors, but retain all other distributional information. With Maxwell's Theorem, this suggests that higher order techniques such as Varimax hold the possibility of identifying the nuisance matrix. In fact, Theorem 5.1 below shows that Varimax can identify the nuisance matrix.

5.3 Population results for Varimax

The Varimax problem applied to the population principal components U in Equation (11) is

$$\arg\max_{R \in \mathscr{O}(k)} v(R, \widetilde{Z}\ \widehat{\Sigma}_Z^{-1/2}\widetilde{R}_U^T). \tag{12}$$

Despite the fact that these are the population principal components, this is still a sample quantity because Z is random. This randomness is from the first stage of randomness in the semi-parametric factor model. The next theorem gives a population result for the M-estimator in (12) by studying the expected value of v over Z, to show that it can identify $\widetilde{R}_U$. Assumption 1 gives the identification assumptions on the distribution of Z that will be used in both the population result for Varimax (Theorem 5.1) and the main theorem (Theorem 6.1).

Assumption 1. *[The identification assumptions for Varimax] The matrix $Z \in \mathbb{R}^{n\times k}$ satisfies the identification assumptions for Varimax if all of the following conditions hold on the rows $Z_i \in \mathbb{R}^k$ for $i = 1, \ldots, n$:*

i) the vectors $Z_1, Z_2, \ldots, Z_n$ are iid,

ii) each vector $Z_i \in \mathbb{R}^k$ is composed of k independent random variables (not necessarily identically distributed),

iii) $Var(Z_{ij}) = 1$ for all j,[10] *and*

iv) the elements of Z_i are leptokurtic.

Let $\widetilde{Z}_1$ be the first row of $\widetilde{Z}$. Define $Z^o = Z_1 - \mathbb{E}(Z_1) \in \mathbb{R}^k$. Theorem 5.1 shows that the rotation matrix R that maximizes the expected Varimax objective function, $\mathbb{E}(v(R, Z^o\widetilde{R}_U^T))$, is $\widetilde{R}_U$. In this formulation, several quantities from the sample maximization problem (12) have been replaced. First, the sample objective function v in Equation (2) has been replaced with its expectation over the distribution of Z. Then, $\bar{Z}$ has been replaced by $\mathbb{E}(Z_1)$ and $\Sigma_Z^{-1/2}$ has been replaced with its limiting quantity under Assumption 1 (i.e., the identity matrix).

Because the Varimax objective function does not change if the estimated factors are reordered or if some of the estimated factors have a sign change, the maximizer of Varimax is actually an equivalence class that allows for these operations. Define the set

$$\mathcal{P}(k) = \{P \in \mathscr{O}(k) : P_{ij} \in \{-1, 0, 1\}\}. \tag{13}$$

It is the full set of matrices that allow for column reordering and sign changes.

Theorem 5.1. *[step 3] Suppose that $Z \in \mathbb{R}^{n\times k}$ satisfies the identification assumptions for Varimax (Assumption 1). Let $Z_1 \in \mathbb{R}^k$ be the first row of Z. Define $Z^o = Z_1 - \mathbb{E}(Z_1)$. For any nuisance rotation matrix $\widetilde{R} \in \mathscr{O}(k)$,*

$$\arg\max_{R\in\mathscr{O}(k)} \mathbb{E}(v(R, Z^o\widetilde{R}^T)) = \{\widetilde{R}P : P \in \mathcal{P}(k)\} \tag{14}$$

The output step of `vsp` right multiplies the principal components $\sqrt{n}U \approx \widetilde{Z}\widetilde{R}_U^T$ with a matrix which maximizes Varimax. In the population results, this matrix is $\widetilde{R}_U$. Thus, the Varimax rotation reveals the unrotated factors, $(\widetilde{Z}\widetilde{R}_U^T)\widetilde{R}_U = \widetilde{Z}$.

Remark 2.2 describes a method to recenter the factors $\widetilde{Z}$ to get Z. Section F.1 in the Appendix gives a population justification for this recentering step.

[10] The third assumption in Varimax is not restrictive because the matrix B can absorb a rescaling of the variables. That is, let $Z^{rescaled} \in \mathbb{R}^{n\times k}$ satisfy the first two conditions and presume that $\mathscr{A} = Z^{rescaled}B^{rescaled}Y^T$. Define $\Sigma_Z = Cov(Z_i^{rescaled})$, $Z = Z^{rescaled}\Sigma^{-1/2}$, and $B = \Sigma^{1/2}B^{rescaled}$. Because $Z^{rescaled}$ satisfies the second condition, Σ_Z is diagonal. So, $Z = Z^{rescaled}\Sigma^{-1/2}$ retains independent components and now satisfies the third condition. Moreover, $\mathscr{A} = ZBY^T$.

Remark 5.2. *[The role of centering] A version of Proposition 5.2 still holds for the SVD of $\mathscr{A}$ (without centering) by replacing $\widehat{\Sigma}_Z$ with Z^TZ/n and replacing $\widehat{\Sigma}_Y$ with Y^TY/d in Equation* (10)*. Even if the elements of the matrix Z are independent and have unit variance, then the columns of Z will not be asymptotically orthogonal (unless $\mathbb{E}(Z) = 0$). As such, right multiplying $U = Z(Z^TZ/n)^{-1/2}\widetilde{R}_U^T$ with an orthogonal rotation (i.e., the one estimated by Varimax) cannot reveal Z. This highlights the role of centering in* `vsp`*; centering $\mathscr{A}$ has the effect of centering the latent variables, which in turn makes the latent factors asymptotically orthogonal under the assumption of independence. This allows Varimax to unmix them with an orthogonal matrix. This point is further discussed in Section 7.*

Remark 5.3. *The first step in Vintage Factor Analysis is to extract the factors. In this paper, we extract the factors with PCA. However, this is not the preferred technique in the classical approach to factor analysis. To see why, define $\mathscr{A} = \mathbb{E}(A|Z,Y) = ZBY^T$ and notice that the diagonal elements of $n^{-1}\mathscr{A}\mathscr{A}^T$ are less than or equal to the diagonal elements of the expected sample covariance matrix $n^{-1}\mathbb{E}(A^TA|Z,Y)$. PCA does not adjust for this excess along the diagonal of the sample covariance matrix and this makes PCA biased. Traditional approaches in Vintage Factor Analysis attempt to estimate the diagonal elements of $n^{-1}\mathscr{A}\mathscr{A}^T$ and replace those estimates down the diagonal of $n^{-1}AA^T$. One of the more common approaches begins with the observation that the diagonal elements of $\mathscr{A}\mathscr{A}^T$ are the diagonal elements of UD^2U^T. So, compute a low rank eigendecomposition of $AA^T \approx \widehat{U}\widehat{D}^2\widehat{U}^T$, replace the diagonal elements of AA^T with the diagonal elements of $\widehat{U}\widehat{D}^2\widehat{U}^T$, then iteratively recompute the eigendecomposition and replace the diagonal elements, until convergence. This problem is still an area of research (e.g. Bertsimas et al. [2017], Zhang et al. [2018]). Alternatively, Bartholomew et al. [2011] suggests specifying a parametric model for both the latent variables Z and the manifest variables A, then using Bayesian and/or likelihood based approaches.*

6 The main theorem

Theorem 6.1 is the main result for this paper. This theorem does not presume a parametric form for the random variables in Z or A. Instead, it uses the identifying assumptions for Varimax (Assumption 1) and two further assumptions on the tails of the distributions for Z and A.

Recall that $\widehat{\mu}_Z$ estimates the column means of Z defined in Remark 2.2. Let $\widehat{Z}_i$ be the ith row of $\widehat{Z}$. Theorem 6.1 shows that for *every* $i \in 1, \ldots, n$, $\widehat{Z}_i + \widehat{\mu}_Z$ converges to Z_i (after allowing for a permutation and sign flip).

Assumption 2. *Each column of Z and Y is generated from a distribution that does not change asymptotically and has a moment generating function in some fixed $\epsilon > 0$ neighborhood around zero.*

Let $\mathscr{A}$ be defined in Equation (7). Define the mean and maximum of $\mathscr{A}$ as

$$\rho_n = \frac{1}{nd}\sum_{i,j}\mathscr{A}_{ij} \quad \text{and} \quad \bar{\rho}_n = \max_{i,j}|\mathscr{A}_{ij}|. \tag{15}$$

Theorem 6.1 allows for A to contain mostly zeros by assuming that as n and d grow, $B_n = \rho_n B$ for some fixed matrix $B \in \mathbb{R}^{k\times k}$. If $\rho_n \to 0$, then A is sparse. This is analogous to the asymptotics in Bickel and Chen [2009] for the Stochastic Blockmodel.

Assumption 3. *For any valid subscripts i and j, eventually in n,*

$$\mathbb{E}[(A_{ij} - \mathscr{A}_{ij})^m] \leq (m-1)!\max\{\bar{\rho}_n^{m/2}, \bar{\rho}_n\}, \quad \textit{for all } m \geq 2,$$

where this expectation is conditional on Z, Y.

Assumption 3 controls the tail behavior of the random variables in the elements of A. This assumption is more inclusive than sub-Gaussian. For example, this assumption is satisfied when A contains Poisson random variables, as happens in Latent Dirichlet Allocation in Section C.4. This assumption is also satisfied if A contains Bernoulli random variables, as happens in Stochastic Blockmodeling. See Sections J.1.5 and J.2.1 in the Appendix for further discussion.

The quantity

$$\Delta_n = n\rho_n$$

controls the asymptotic rate in Theorem 6.1. So, it is helpful to have some sense for it. For example, suppose that (i) A contains Bernoulli elements, (ii) each row and column sum of $\mathscr{A}$ grows at a similar rate, (iii) $n \asymp d$, and (iv) $\rho_n \to 0$, then Δ_n is roughly the expected number of ones in each row and column of A.

Theorem 6.1. *Suppose that $A \in \mathbb{R}^{n\times d}$ is generated from a semi-parametric factor model that satisfies Assumptions 1, 2, and 3. Presume that asymptotically, $\mathscr{A} = \rho_n ZBY^T$ for some fixed and full rank matrix B. In the asymptotic regime where $n \asymp d$ and $\Delta_n \succeq \log^{11.1} n$,*

$$||(\widehat{Z} + \mathbf{1}_n\widehat{\mu}_Z) - ZP_n||_{2\to\infty} = O_p(\Delta_n^{-.24}\log^{2.75} n), \tag{16}$$

where $\widehat{Z}$ is the estimate produced by `vsp` *(with step 1) applied to A and $\widehat{\mu}_Z$ is the estimate defined in Equation* (5)*.*

Theorem 6.1 shows convergence in $2 \to \infty$ norm. This means that every row of $\widehat{Z}+\mathbf{1}_n\widehat{\mu}_Z$ converges to the corresponding row of Z in ℓ_2. The P_n matrix accounts for the fact that we do not attempt to identify the order of the columns in Z, or their sign. If $\widehat{Z}$ is used without recentering by $\mathbf{1}_n\widehat{\mu}_Z$, then a similar result holds for estimating $\widetilde{Z}$. By symmetry, if Y satisfies the identification assumptions for Varimax, then `vsp` can also estimate Y. If both Z and Y satisfy the identification assumptions for Varimax, then B can also be recovered, even when it is not diagonal. The proof for Theorem 6.1 begins in Appendix G.3. Corollaries C.1 and C.2 in the Appendix extend Theorem 6.1 to the Stochastic Blockmodel and Latent Dirichlet Allocation.

7 Correlated factors or "Why should the radial streaks be orthogonal?"

Because Varimax provides an orthogonal rotation, it constructs orthogonal axes. One common concern in the factor analysis literature is that orthogonal axes cannot detect latent factors that are correlated. For example, in Figure 5, the panel with the title "3" has radial streaks that are slightly wider than the vertical and horizontal axes; we will call this phenomenon *the appearance of non-orthogonal factors*. This non-orthogonality can be far more severe than what appears in Figure 5. Despite the fact that correlated factors are an often discussed problem, this section shows how severe cases can be an artifact of a common data processing step that is not included in `vsp`.

`vsp` easily handles correlated factors; Section 7.1 gives more intuition for how and why. Then, Sections 7.1.1 and 7.1.2 describe how two data analytic choices can create the appearance of non-orthogonal factors (even when the factors are independent). Section 7.2 shows how "the middle B matrix" in the semi-parametric factor model provides a path towards deeper understanding of correlated factors, a path that we reserve for future research. If the slight misalignment of streaks, such as in panel "3" discussed above, needs a solution, then the `vsp` solution could be refined via an iterative approach that involves soft thresholding (e.g. Chen and Rohe [2020]).

7.1 `vsp` can handle correlated factors

Proposition 5.1 and Proposition 5.2 do not make any probabilistic assumptions; both are simply results of linear algebra. Together, these propositions show that if the data matrix is not centered, then the principal components are

$$U = Z(Z^T Z)^{-1/2} R_U^T$$

for some rotation matrix R_U. Alternatively, if the data matrix is centered, then the principal components are a function of the centered latent factors $\widetilde{Z}$,

$$U = \widetilde{Z}(\widetilde{Z}^T \widetilde{Z})^{-1/2} \widetilde{R}_U^T$$

for some other rotation matrix $\widetilde{R}_U$.

In the principal components U, the latent factors Z have been orthogonalized via $(Z^T Z)^{-1/2}$. As such, if the original latent factors are correlated, they become orthogonal in the principal components U. So, a set of orthogonal Varimax axes could potentially uncover the orthogonalized factors $Z(Z^T Z)^{-1/2}$. This is good news. If underlying correlated factors had radial streaks, those radial streaks will be preserved in $Z(Z^T Z)^{-1/2}$. Those streaks will not necessarily be perfectly orthogonal. However, they are often close, as in panel "3" of Figure 5.

This assessment aligns with Kaiser's. In "A Second Generation Little Jiffy", Kaiser discusses *orthoblique*, which rotates the unit length eigenvectors[11] via Varimax without row normalization [Harris and Kaiser, 1964, Kaiser, 1970]. This differs from `vsp` only in some preprocessing steps. Kaiser says, orthoblique has "the tremendous advantage of being 99% of the way" to the solution for recovering correlated factors. He develops a much more complicated winsorizing technique and makes the following remark.

> One final comment about this Kaiser-Tukey winsorizing business: above when I said that we were 99% of the way with orthoblique, I was not using a figure of speech. In some 40 or 50 studies involving hundreds of factors the average correlation between an original Harris orthoblique factor [i.e. `vsp`] and its winsorized counterpart was .99. It is clear that we have gone to a lot of trouble to apply a very mild polish. [Kaiser, 1970]

The fact that `vsp` easily handles correlated factors is an empirical phenomenon that does not contradict any of the technical results in this paper. In the technical results, the independence of elements in Z is a *sufficient* condition, not a necessary condition.

7.1.1 Scaling eigenvectors creates the appearance of non-orthogonal factors

A key difference between common factor analysis practice and `vsp`/orthoblique, is that `vsp`/orthoblique use unit length eigenvectors, whereas common practice scales each eigenvector by the square root of its eigenvalue. For example, the popular `psych` package in `R` does this scaling [Revelle, 2017].

This subsection describes how the common practice of scaling the eigenvectors creates the appearance of non-orthogonal factors, *even if the factors are independent*. Then, Subsection 7.1.2 explores one (necessarily unexciting) place where the remaining 1% from Kaiser's calculation might come from.

For simplicity, suppose that we are not centering and that Z^TZ is the identity matrix. So, $U = ZR_U^T$. We hope that Varimax provides $R^* = R_U$. If it does, then `vsp` rotates and recovers,

$$UR^* = Z(R_U^T R_U) = Z.$$

This is, essentially, why `vsp` works. However, suppose D is a diagonal matrix containing the singular values of $\mathscr{A}$ (i.e. the square root of the eigenvalues of $\mathscr{A}\mathscr{A}^T$). If we scale U by D before rotation, then the two rotation matrices cannot cancel out like they do above,

$$UDR^* = Z(R_U^T DR^*).$$

[11] In this section, we refer to the columns of U as eigenvectors, not principal components or singular vectors, because they are also eigenvectors of AA^T. In the historical literature cited, "the eigenvectors" are typically coming from matrices that have been preprocessed in ways discussed in 5.3.

By scaling with D, the appearance of non-orthogonal factors could become much more severe than in Figure 5.

For example, if Z contains independent, mean zero, and unit variance factors, but Y contains correlated factors, then the singular values of $\mathscr{A} = ZY$ in the diagonal matrix D will be proportional to the eigenvalues of $(Y^TY)^{1/2}$ (see Proposition 5.2). In general, Varimax will not be able to recover Z from UD. Moreover, it will appear as though the factors in Z are not orthogonal; in fact, the factors in Z are orthogonal, but they are not if you rotate them with R_U and then scale by a diagonal matrix that is determined by Y, not Z.

Given the numerous data analytic choices that must be made in the course of performing factor analysis, Henry Kaiser proposed a sequence of default procedures "Little Jiffy," "A second generation Little Jiffy," and finally "Little Jiffy, Mark IV" [Kaiser, 1970, Kaiser and Rice, 1974]. All of these default procedures apply Varimax (without row normalization) to the unit length eigenvectors (of variously transformed matrices); this is the procedure used in this paper too.

Kaiser uses unit length vectors because of an observation in Harris and Kaiser [1964] that it solves the rotation problem when each row of Z has exactly one non-zero element; they call this "Independent Cluster" structure. This structure in Z is analogous to the Degree Corrected Stochastic Blockmodel discussed in Section C.3. However, if the structure in Z is not this nice, Harris and Kaiser [1964] says this: "If the 'ideal' common part of any one or more variables is of complexity greater than one [i.e. more than one non-zero element in that row of Z], then rotating ... will not yield [the] 'ideal' solution." In general, this observation is true. However, if the latent factors are independent and leptokurtic random variables, then the rows of Z can have multiple non-zero elements and Theorem 6.1 shows that rotating the principal components with Varimax *can* reveal these structures.

7.1.2 The role of centering

The appearance of non-orthogonal factors can also happen as a result of improper centering. The last section has a simple suggestion for data analysis: use the unit length eigenvectors, do not scale them by their eigenvalues. Unfortunately, for the problem of centering, there is not a simple suggestion. The good news is that this is likely a small problem; akin to the 1% in Kaiser's calculation.

In order for Varimax to be asymptotically unbiased in recovering Z (or $\widetilde{Z}$) from U, the orthogonalizing matrix $(Z^TZ)^{-1/2}$ or $(\widetilde{Z}^T\widetilde{Z})^{-1/2}$ should converge to a diagonal matrix; diagonal matrices are acceptable because if Z has radial streaks aligned with the axes, then scaling it by a diagonal matrix would keeps the streaks aligned with the axes. However, in certain settings described below where Z has orthogonal radial streaks, $(Z^TZ)^{-1/2}$ is not diagonal. In this situation, the orthogonalizing matrix $(Z^TZ)^{-1/2}$ will skew the orthogonal streaks and thus give the factors the appearance of non-orthogonality. A similar phenomenon can hold for $\widetilde{Z}$.

Case I (independent and non-zero mean): Suppose the entries of Z are independent *with non-zero mean*, then $E(Z^TZ) = \Sigma + n\mu_z\mu_z^T$, where Σ is a diagonal matrix and μ_z is the expectation of one row of Z. This means that $(Z^TZ)^{-1/2}$ does not converge to a diagonal matrix. However, if the data matrix is centered, then U is determined by $\widetilde{Z}$ which has asymptotically orthogonal columns. Thus, $(\widetilde{Z}^T\widetilde{Z})^{-1/2}$ will converge to a diagonal matrix. In this case, if we center the data matrix, compute the principal components, and rotate with Varimax, then we can hope to recover $\widetilde{Z}$, then uncenter to recover Z.
Case II (*Independent clusters*): Suppose that the latent factor matrix Z has exactly one non-zero element in each row, as in the Stochastic Blockmodel or what Harris and Kaiser [1964] and others call *Independent Clusters*. In this setting, Z does not have independent entries, but it does have orthogonal columns. So, $(Z^TZ)^{-1/2}$ is diagonal. In this case, centering *removes* the orthogonality; $(\widetilde{Z}^T\widetilde{Z})^{-1/2}$ is *not* diagonal. This is the opposite of Case I. In Case II, if we compute the principal components (without centering), and rotate with Varimax, then we can hope to recover Z.
Case III (Both independent clusters and factors): Suppose there are $k = k_1 + k_2$ columns in Z and the first k_1 columns correspond to k_1 independent clusters and the last k_2 columns correspond to independent factors. In this setting, neither $(Z^TZ)^{-1/2}$ nor $(\widetilde{Z}^T\widetilde{Z})^{-1/2}$ will be diagonal. This is a troubling scenario that centering alone cannot fix.
Case IV (Mean zero factors): If the latent factors already have mean zero, centering will not change anything.

To summarize, centering ensures that independent factors are orthogonal (Case I). However, factors that are already orthogonal, can become non-orthogonal after centering (Case II). In these cases above, the appearance of non-orthogonal factors is not an artifact of latent factors being correlated (in any interesting fashion). In our experience, centering or not centering has a minimal, yet non-zero, effect on the non-orthogonality of the factors.

7.2 The middle B matrix contains information about factor correlations

One way of understanding the "middle B matrix" in the semi-parametric factor model is that it describes the correlation among the factors. The Stochastic Blockmodel is the only previous statistical model (that we are aware of) that parameterizes such a matrix. In that model, the Z matrix records block memberships and B_{uv} gives the probability of a connection between a node in block u and a node in block v (see Sections 3.1.2 and C.3). This B matrix is not typically imaged as describing the correlation among some latent factors (i.e. "blocks"), but it certainly could be (e.g. "highly correlated blocks form more connections").

Outside of the Stochastic Blockmodel, suppose that the Z factors are correlated; the Y factors are centered, independent, and leptokurtic; and B is proportional to the identity matrix. Moreover, suppose that $\widehat{Z}$ converges to the orthogonalized factors $Z(Z^TZ)^{-1/2}$, then $\widehat{B}$ estimates $(Z^TZ)^{1/2}$ (e.g. see Equation (9)). So, if the data generating model does not have a B matrix (set to identity), then the *estimated* B matrix records information

about the correlation among factors. In fact, Harris and Kaiser [1964] and Kaiser and Rice [1974] discuss a quantity that they call L^* (or `L`, or `LSTAR`) that is analogous to $\widehat{B}$. Harris and Kaiser [1964] says "The matrix L designates the intercorrelation of the factors."

Perhaps more directly, hierarchical clustering is one way of imagining how clusters/factors could be correlated; more correlated factors are closer in the hierarchy. In some parameterizations of the hierarchical Stochastic Blockmodel, the hierarchical structure is not parameterized in the Z matrix, but rather in the B matrix [Lei et al., 2020]. This is consistent with the idea that B records information about factor correlations.

Taken together, this all suggests that the B matrix provides a path to understanding "correlation among the factors." Understanding this phenomenon is an active area of research in our lab.

8 Discussion

PCA with Varimax is a vintage data analysis technique. Theorem 6.1 shows that it provides a unified spectral estimation strategy for a broad class of semi-parametric factor models. The reason is that (1) the principal component subspace is the same subspace as the latent factor subspace and (2) under the leptokurtic assumption, Varimax draws a set of axes through this space such that each axis aligns with one of the latent factors; this is the intuition gained in Section 5. The leptokurtosis condition is satisfied if the factors are sparse and this condition can be examined in the data. In fact, Section 4 reinterprets the diagnostics practices developed in Thurstone [1935, 1947] as examining that leptokurtic condition. Taken together, the results in this paper show that the Vintage Factor Analysis know-how developed by Thurstone and Kaiser performs statistical inference. This know-how has survived for nearly a century, despite the conventional wisdom that the factor rotation cannot perform statistical inference.

Acknowledgements: Thank you to De Huang for valuable discussions. Thank you to Joshua Cape for helpful comments on an early draft on this paper. Thank you to Alex Hayes for help creating an `R` package of the code. Thank you to E Auden Krauska, Dan Bolt, Alex Hayes, Fan Chen, Stephen Stigler, Anru Zhang, Miaoyan Wang, Shuqi Yu, Sijia Fang, Sebastien Roch, and Gemma Moran for helpful discussions during the course of this research. Thank you to the referees and the editors for valuable feedback that improved this paper. This research is supported in part by NSF Grants DMS-1612456 and DMS-1916378 and ARO Grants W911NF-15-1-0423 and W911NF-20-1-0051.

References

Louis Leon Thurstone. *The vectors of mind: Multiple-factor analysis for the isolation of primary traits.* University of Chicago Press, 1935.

Henry F Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

Louis Leon Thurstone. *Multiple factor analysis.* University of Chicago Press: Chicago, 1947.

Theodore W Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150, 1956.

I.T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer, 2002. ISBN 9780387954424.

J O Ramsay and B W Silverman. *Applied functional data analysis: methods and case studies.* Springer, 2007.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis.* Pearson Education International. Pearson Prentice Hall, 2007. ISBN 9780135143506.

D.J. Bartholomew, M. Knott, and I. Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach.* Wiley Series in Probability and Statistics. Wiley, 2011. ISBN 9780470971925.

James Clerk Maxwell. V. illustrations of the dynamical theory of gases. part i. on the motions and collisions of perfectly elastic spheres. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 19(124):19–32, 1860.

W. Feller. *An Introduction to Probability Theory and its Applications, Volume 2.* John Wiley and Sons, Inc., 1971.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Vincent Q Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

Fan Chen and Karl Rohe. A new basis for sparse pca. *arXiv preprint arXiv:2007.00596*, 2020.

Anna M Fiori and Michele Zenga. Karl pearson and the origin of kurtosis. *International Statistical Review*, 77(1):40–50, 2009.

J. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005. ISBN 9780387400808.

Henry F Kaiser. A second generation little jiffy. *Psychometrika*, 35(4):401–415, 1970.

Henry F Kaiser and John Rice. Little jiffy, mark iv. *Educational and psychological measurement*, 34(1):111–117, 1974.

Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, pages 35–1, 2012.

Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41 (4):2097–2122, 2013.

Can M Le, Elizaveta Levina, and Roman Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

Yilin Zhang and Karl Rohe. Understanding regularized spectral clustering via graph conductance. *arXiv preprint arXiv:1806.01468*, 2018.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*, 2018.

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45): 12679–12684, 2016a.

Fan Chen, Sebastien Roch, Karl Rohe, and Shuqi Yu. Estimating graph dimension with cross-validated eigenvalues. *arXiv preprint arXiv:2108.03336*, 2021.

S Wang and Karl Rohe. Coauthorship and citation networks for statisticians: Comment. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

Chester W Harris and Henry F Kaiser. Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29(4):347–362, 1964.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting overlapping communities in networks using spectral methods. *arXiv preprint arXiv:1412.3432*, 2014.

Jiashun Jin and Zheng Tracy Ke. A sharp lower bound for mixed-membership estimation. *arXiv preprint arXiv:1709.05603*, 2017.

Alex Salcianu and et al. Google compact language detector v3 (cld3). 2020. doi: 10.21105/joss.00037. URL `https://github.com/google/cld3#readme`.

Julia Silge and David Robinson. tidytext: Text mining and analysis using tidy data principles in r. *JOSS*, 1(3), 2016. doi: 10.21105/joss.00037. URL `http://dx.doi.org/10.21105/joss.00037`.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Probability and mathematical statistics. 10th printing in 1995. Academic Press, 1979. ISBN 9780124712508.

Dimitris Bertsimas, Martin S Copenhaver, and Rahul Mazumder. Certifiably optimal low rank factor analysis. *The Journal of Machine Learning Research*, 18(1):907–959, 2017.

Anru Zhang, T Tony Cai, and Yihong Wu. Heteroskedastic pca: Algorithm, optimality, and applications. *Annals of Statistics (to appear); arXiv preprint arXiv:1810.08316*, 2018.

Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106 (50):21068–21073, 2009.

William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2017. URL `https://CRAN.R-project.org/package=psych`. R package version 1.7.8.

Lihua Lei, Xiaodong Li, and Xingmei Lou. Consistency of spectral clustering on hierarchical stochastic block models. *arXiv preprint arXiv:2004.14531*, 2020.

Karl Rohe, Zeng Muzhe, and Alex Hayes. Vintage sparse pca for non-parametric factor analysis. `https://github.com/karlrohe/vsp`, 2020.

Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2017. URL `https://CRAN.R-project.org/package=Matrix`. R package version 1.2-12.

Yixuan Qiu, Jiali Mei, and authors of the ARPACK library. See file AUTHORS for details. *rARPACK: Solvers for Large Scale Eigenvalue and SVD Problems*, 2016. URL `https://CRAN.R-project.org/package=rARPACK`. R package version 0.11-0.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

Aiyou Chen and Peter J Bickel. Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, 53(10):3625–3632, 2005.

Aiyou Chen and Peter J Bickel. Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825–2855, 2006.

Tianwen Wei. A convergence and asymptotic analysis of the generalized symmetric fastica algorithm. *IEEE transactions on signal processing*, 63(24):6445–6458, 2015.

Jari Miettinen, Sara Taskinen, Klaus Nordhausen, and Hannu Oja. Fourth moments and independent component analysis. *Statistical science*, 30(3):372–390, 2015.

Richard J Samworth and Ming Yuan. Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973–3002, 2012.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

Majid Janzamin, Rong Ge, Jean Kossaifi, Anima Anandkumar, et al. Spectral learning on matrices and tensors. *Foundations and Trends® in Machine Learning*, 12(5-6):393–536, 2019.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, pages 309–336, 2011.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45): 12679–12684, 2016b.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Joshua Cape, Minh Tang, and Carey E Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250, 2019a.

Joshua Cape, Minh Tang, Carey E Priebe, et al. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439, 2019b.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Spectral statistics of erdős–rényi graphs i: local semicircle law. *The Annals of Probability*, 41(3B):2279–2375, 2013.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating mixed memberships with sharp eigenvector deviations. *arXiv preprint arXiv:1709.00407*, 2017.

Sara A Van de Geer. *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge, 2000.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

Moody T Chu and Nickolay T Trendafilov. Orthomax rotation problem. a differential equation approach. *Behaviormetrika*, 25(1):13–23, 1998.

Richard J Sherin. A matrix formulation of kaiser's varimax criterion. *Psychometrika*, 31 (4):535–538, 1966.

H Neudecker. On the matrix formulation of kaiser's varimax criterion. *Psychometrika*, 46 (3):343–345, 1981.

Jos MF ten Berge. A joint treatment of varimax rotation and the problem of diagonalizing symmetric matrices simultaneously in the least-squares sense. *Psychometrika*, 49(3): 347–358, 1984.

John Riordan. Moment recurrence relations for binomial, poisson and hypergeometric frequency distributions. *The Annals of Mathematical Statistics*, 8(2):103–111, 1937.

A Thurstone's five criteria for simple structure

In the quote below, Thurstone's original mathematical notation has been replaced with the notation in this paper.

Five rules for simple factor structure; quoted from Thurstone [1947] p335

We shall describe five useful criteria by which the k reference vectors [i.e., the columns of $R_{\widehat{U}}$] can be determined. These are as follows:

1. Each row of the .... matrix $\widehat{U}R_{\widehat{U}}$ should have at least one zero.
2. For each column ℓ of the factor matrix $\widehat{U}R_{\widehat{U}}$ there should be a distinct set of k linearly independent [rows] whose factor loadings $[\widehat{U}R_{\widehat{U}}]_{j\ell}$ are zero. [sic[a]]
3. For every pair of columns of $\widehat{U}R_{\widehat{U}}$ there should be several [rows] whose entries $[\widehat{U}R_{\widehat{U}}]_{jp}$ vanish in one column but not in the other.
4. For every pair of columns of $\widehat{U}R_{\widehat{U}}$, a large proportion of the tests should have zero entries in both columns. This applies to factor problems with four or five or more common factors.
5. For every pair of columns there should preferably be only a small number of [rows] with non-vanishing entries in both columns.

When these [five] conditions are satisfied, the plot of each pair of columns shows (1) a large concentration of points in two radial streaks, (2) a large number of points at or near the origin, and (3) only a small number of points off the two radial streaks. For a configuration of k dimensions there are $\frac{1}{2}k(k-1)$ diagrams. When all of them satisfy the three characteristics, we say that the structure is 'compelling,' and we have good assurance that the simple structure is unique. In the last analysis it is the appearance of the diagrams that determines, more than any other criterion, which of the hyperplanes of the simple structure are convincing and whether the whole configuration is to be accepted as stable and ready for interpretation.*

*Ever since I found the simple-structure solution for the factor problem, I have never attempted interpretation of a factorial result without first inspecting the diagrams. [footnote original to text]

[a] There cannot be k linearly independent vectors in a $k-1$ dimensional hyperplane.

An example of the diagrams (i.e., plots) that Thurstone proposes are given in Figure 2. Each of those plots displays radial streaks. After the Varimax rotation, those radial streaks align with the coordinate axes, making the rotated factors approximately sparse.

B Fast computation for sparse data matrices

In many contemporary applications, A is sparse (i.e., most elements A_{ij} are zero). In this case, the SVD step should be computed with power methods. These methods are faster and require less memory because they only require matrix-vector multiplication (which are fast for matrices that are sparse). Moreover, if step 1 of `vsp` is being used, then the centered matrix $\widetilde{A}$ should not be explicitly computed. Instead, the matrix-vector multiplications can be computed as the right hand side of the following equality,

$$\widetilde{A}x = Ax - \widehat{\mu}_r(\mathbf{1}_d^T x) - \mathbf{1}_n(\widehat{\mu}_c x) + \widehat{\mu}_{\cdot} \mathbf{1}_n(\mathbf{1}_d^T x), \tag{17}$$

and similarly for $y\widetilde{A}$. When computed naively, the left hand side of Equation (17) requires $O(nd)$ operations. However, the right hand size requires $O(\texttt{nnz})$ operations, where `nnz` is the number of non-zero elements in A. In the bibliometrics example displayed in Figure 2, `nnz` is three orders of magnitude smaller than nd. Using Equation (17) also dramatically reduces the amount of memory required to store the matrices. This can be used in conjunction with the degree-normalization step in Remark 2.1. This is implemented in an `R` package available on GitHub [Rohe et al., 2020] using the `R` packages Matrix and rARPACK [Bates and Maechler, 2017, Qiu et al., 2016].

C Modern factor models as semi-parametric factor models

> The semi-parametric factor model is related to ICA, the Stochastic Blockmodel, and Latent Dirichlet Allocation. Corollaries C.1 and C.2 show that with some slight variations on the preprocessing of A, `vsp` can estimate the Stochastic Blockmodel and Latent Dirichlet Allocation.

C.1 Relationship to Independent Components Analysis

Independent Components Analysis (ICA) uses a type of semi-parametric factor model that is motivated by blind-source separation in signal processing. In the typical formulation of ICA, we observe a multivariate time series $\mathscr{A}_t = Z_t M \in \mathbb{R}^k$ for $t = 1, \ldots, n$, where $Z_t \in \mathbb{R}^k$ contains independent and non-Gaussian random variables. The aim is to estimate M^{-1}, to unmix the observed signals in $\mathscr{A}_t$, and reveal the independent components Z_t. There are multiple ICA results that share some similarities to Theorems 5.1 and 6.1 (e.g. Comon [1994], Hyvärinen et al. [2004], Chen and Bickel [2005, 2006], Wei [2015], Miettinen et al. [2015], Samworth and Yuan [2012]). To see the connection to the current paper, let $M \in \mathbb{R}^{k \times d}$ be potentially rectangular and defined as $M = BY^T$. To enable the regime $d = k$, the results for ICA typically presume that $\mathscr{A} = ZM$ is observed with little or no noise. In contrast, Theorem 6.1 covers situations where (i) d grows at the same rate as n, (ii) there is an abundance of noise in A, and (iii) A is mostly zeros (i.e., sparse). This

allows the theorem to cover contemporary factor models such as the Stochastic Blockmodel and Latent Dirichlet Allocation.

C.2 Tensor decompositions

Motivated in part by the issue of rotational invariance of PCA, Kruskal [1977] showed how a tensor decomposition called the CP decomposition is unique; it decomposes a tensor into a set of factors that are not rotationally invariant. In Section 4, Kruskal discusses how this three way decomposition does not suffer from the same problem of rotation that "consumes considerable attention and effort" in factor analysis. In an elegant formulation, Anandkumar et al. [2014] showed how these tensor spectral methods could be applied to estimate the latent factors in a model class similar to the semi-parametric factor model. Where the principal components of A are the eigenvectors of a matrix that contains the second order moments, $n^{-1}\mathbb{E}(\widetilde{A}^T\widetilde{A})_{uv} = \mathbb{E}(\widetilde{A}_{iu}\widetilde{A}_{iv})$, the elements of this higher order tensor contain the third order (or higher) moments; for example, $T \in \mathbb{R}^{d\times d\times d}$ with $T_{u,v,w} = \mathbb{E}(A_{iu}A_{iv}A_{iw})$. Then, for various formulations of T and latent variable models, the CP tensor decomposition of T has components that are equal to the latent factors [Janzamin et al., 2019].

The issue of rotational invariance motivates for the extension from matrices to tensors. For example, in a recent book on using tensors for latent variable modeling, Janzamin et al. [2019] writes in the abstract "PCA and other spectral techniques applied to matrices have several limitations. By limiting to only pairwise moments, they are effectively making a Gaussian approximation on the underlying data." However, despite the fact that PCA is typically imagined as a second order technique, the principal components of A retain the higher-order distributional properties of the latent variables (see Remark 5.1). As such, we need not consider the higher order moments of the manifest variables A in the tensor T. `vsp` uses the higher order moments of the principal components themselves, by applying Varimax directly to the principal components. Given our heuristics around rotational invariance, it is surprising that this can work.

C.3 Stochastic Blockmodels

In social network analysis, $A \in \{0,1\}^{n\times n}$ is the adjacency matrix of a graph on n people.

$$A_{ij} = \begin{cases} 1 & i \text{ friends with } j \\ 0 & o.w. \end{cases}$$

The Stochastic Blockmodel [Holland et al., 1983] is a semi-parametric factor model for generating a random adjacency matrix. Under this model, each individual i is assigned to a single block $z(i) \in \{1,\dots,k\}$ and the probability that i and j are friends is

$$\mathbb{P}(A_{ij} = 1|z(i), z(j)) = B_{z(i),z(j)}, \text{ where } B \in [0,1]^{k\times k}.$$

Define $\mathscr{A} = E(A|Z, B, Y)$. To express $\mathscr{A}$ in the factor model as ZBZ^T, define $Z \in \{0,1\}^{n\times k}$ such that $Z_{ij} = 1$ when $z(i) = j$ and $Z_{ij} = 0$ otherwise. When friendships are symmetric, so is A; in this setting $Y = Z$ and the elements above the diagonal of A are independent. There are four popular generalizations of the Stochastic Blockmodel that have the structure ZBZ^T, and are thus other types of semi-parametric factor models. The Degree-Corrected Stochastic Blockmodel includes an additional degree parameters $\theta_{i,z(i)} > 0$ for each individual i. The probability of friendship becomes $\mathscr{A}_{ij} = \theta_{i,z(i)}\theta_{j,z(j)}B_{z(i),z(j)}$ [Karrer and Newman, 2011]. To express this model as ZBZ^T, define $Z_{ij} = \theta_{i,z(i)}\mathbb{I}\{z(i) = j\}$, where $\mathbb{I} \in \{0,1\}$ is the indicator function. In the Overlapping Stochastic Blockmodel, $Z \in \{0,1\}^{n\times k}$ is sparse [Latouche et al., 2011].[12] In the mixed-membership Stochastic Blockmodel, each row of Z is an independent sample from the Dirichlet distribution [Airoldi et al., 2008]. Later, Zhang et al. [2014] and Jin and Ke [2017] generalized these models to only presume that $Z_i \in \mathbb{R}^k$ is element-wise non-negative. Table 3 summarizes all of these models. While this discussion focuses on unipartite and undirected graphs, graphs that are "two-way," "bipartite," or "directed," can also be modeled in the form $\mathscr{A} = ZBY^T$ [Rohe et al., 2016b].

SBM	the vector $Z_i \in \mathbb{R}^k$ contains	distribution of Z_i
0) Standard SBM	a single one, the rest zeros	multinomial
1) Degree-Corrected	a single positive entry, the rest zeros	not specified
2) Overlapping	a mix of 1s and 0s	independent Bernoulli
3) Mixed Membership	non-negative entries that sum to one	Dirichlet
4) Degree-Corrected, Mixed Membership	non-negative entries	not specified

Table 3: Restrictions on the factor matrix Z create variations on the Stochastic Blockmodel (SBM). There are further differences between these models that are not emphasized by this table.

Estimating the Degree-Corrected Stochastic Blockmodel with `vsp`. Under the Stochastic Blockmodel and the Degree Corrected version, each node i belongs to exactly one cluster. In such "hard clustering" models, the elements in the same row of Z cannot be independent. This implies that Z cannot satisfy Assumption 1 of Theorem 6.1. The next corollary shows that `vsp` *without the centering step* can estimate these models.

Let $\pi \in \mathbb{R}^k$ be a probability distribution on $[k]$. Suppose that $z(1), \ldots, z(n) \sim Multinomial(\pi)$, independently. For each block j, suppose that $\theta_{1,j}, \ldots, \theta_{n,j} \in \mathbb{R}$ are independent random variables generated from a bounded probability distribution f_j. The scale of this distribution is unidentifiable; so for technical convenience, it is presumed that $\mathbb{E}(Z_{ij}^2) = 1$, or equivalently, that $\mathbb{E}(\theta_{i,j}^2) = 1/\pi_j$. This is akin to the third assumption in

[12] The original paper paper on the overlapping Stochastic Blockmodel is not exactly the factor model used here because it includes a logistic link function, $\mathbb{P}(A_{ij} = 1) = logit(Z_i B Z_j^T)$.

the Varimax assumption. This scaling ensures that $\mathbb{E}(Z^T Z)/n$ (i.e. without centering) converges to the identity matrix. If each f_j is a point mass, then this model is equivalent to the SBM.

Corollary C.1. *Suppose that $A_n \in \mathbb{R}^{n \times n}$ is generated from the Degree Corrected Stochastic Blockmodel with $\mathbb{E}(A_n|Z_n) = Z_n B_n Z_n$, where Z_n is generated as described in the proceeding paragraph. Suppose that the probability distributions f_j for $j \in [k]$ are bounded. Define ρ_n as in Equation* (15) *and suppose that there exists a fixed matrix $B \in \mathbb{R}^{k \times k}$ such that $B_n = \rho_n B$.*

Define $\widehat{Z} \in \mathbb{R}^{n \times k}$ as the output of `vsp` *without centering (i.e. skip step 1). In the asymptotic regime where $\Delta_n \succeq \log^{11.1} n$, there exists a sequence permutation and sign-flip matrices $P_n \in \mathcal{P}(k)$ such that*

$$||\widehat{Z} - ZP_n||_{2 \to \infty} = O_p(\Delta_n^{-.24} \log^{2.75} n). \tag{18}$$

A proof is contained in Appendix J.

To see why the centering step creates bias for `vsp` under a hard clustering model, note that `vsp` with the centering step (step 1) estimates $\widetilde{Z}$ (i.e., Z after centering). By construction, $\widehat{Z}$ contains orthogonal columns. However, under the Stochastic Blockmodel, $\widetilde{Z}$ does not. Interestingly, Z *without centering* does contain orthogonal columns and `vsp` without centering can estimate it.

Overlapping and Mixed Membership. Under the Overlapping SBM,

$$Z_{ij} \sim Bernoulli(p_j)$$

independently for all i and j. This will satisfy the identification assumptions for Varimax so long as $p_j \notin [1/2 \pm 1/\sqrt{12}]$ for $j = 1, \ldots, k$. This rather strange condition ensures that Z_{ij} is leptokurtic and thus Varimax can identity the rotation. If Varimax were replaced with an alternative rotation from the ICA literature, then one could remove the awkward condition on the p_j's.

Under the Mixed Membership SBM, Z_i is on the simplex. As such, its elements must sum to one and cannot be statistically independent. This restriction to the simplex also limits the ability of the Mixed Membership model to create a large amount of degree heterogeneity, a common property in empirical networks. As discussed in Section C.4, this problem also arrises for Latent Dirichlet Allocation (LDA). Section C.4 discusses a natural generalization of LDA that allows for more heterogeneous document lengths. A similar generalization could be applied to the Mixed Membership SBM. This would create a "Degree-Corrected Mixed Membership model." Under such a model, a result analogous to Corollary C.2 could be derived.

Degree-Corrected Mixed Membership. The papers which proposed the Degree-Corrected Mixed Membership model only presume that Z_i is element-wise non-negative

[Zhang et al., 2014, Jin and Ke, 2017]. As such, if the elements of Z_i are sampled in a way which satisfy the identification assumptions for Varimax, then Theorem 6.1 shows that `vsp` can estimate this model.

C.4 Latent Dirichlet Allocation

In the setting of text analysis and natural language processing, let $A \in \mathbb{N}^{n \times d}$ be a document-term matrix on n documents and d unique words,

$$A_{ij} = \text{number of times that word } j \text{ appears in document } i. \tag{19}$$

Latent Dirichlet Allocation (LDA) is a popular generative model for A that is used for modeling the topics of documents [Blei et al., 2003].

The LDA model has parameters $\xi > 0$, $\alpha \in \mathbb{R}_+^k$, and $\beta \in \mathbb{R}_+^{d \times k}$ with $\mathbf{1}_d^T \beta = \mathbf{1}_k$. The rows of β index the unique words $1, \ldots, d$. Because the elements of β are positive and each column sums to one, each column makes a probability distribution on the unique words. LDA generates a single document $i = 1, \ldots, n$ with the following steps, (1) choose $Z_i \sim Dirichlet(\alpha)$ to be the topic distribution for that document, (2) sample $N_i \sim Poisson(\xi)$ to be the number of words in the document, (3) for each of the words in the document $w = 1, \ldots, N_i$, choose the topic for that word $z_w \sim Multinomial(Z_i) \in \{1, \ldots, k\}$, and then sample the word w as multinomial with probabilities specified by the z_w column of β (i.e., w is the jth unique word with probability β_{j,z_w}).

Lemma C.1. *Under the LDA model, conditionally on the Dirichlet variables $Z_1, \ldots, Z_n$, the document-term matrix A has independent Poisson entries with*

$$\mathbb{E}(A|Z) = \xi Z \beta^T, \tag{20}$$

where $Z \in \mathbb{R}_+^{n \times k}$ has rows $Z_1, \ldots, Z_n$.

A short proof in Section F.2 relies upon the Poisson-Multinomial relationship. While Equation (21) has the form of the semi-parametric factor model (e.g. set $B = I$ and $Y = \beta$), it does not satisfy the identification assumptions for Varimax because the elements in Z_i sum to one and as such, they must be dependent. Moreover, this has the unnatural consequence of making $\mathbb{E}(A|Z)$ have rank $k-1$ or less. However, the following modification makes $\mathbb{E}(A|Z)$ have rank k and enables the application of Theorem 6.1.

In the original formulation of LDA, the number of words in document i is $N_i \sim Poisson(\xi)$, for $\xi \in \mathbb{R}_+$. About this step, Blei et al. [2003] says, "more realistic document length distributions can be used as needed." If document lengths are more heterogenous than what is modeled by Poisson(ξ), then a convenient way to increase the heterogeneity is to use Poisson overdispersion; first sampling ξ_i, then sampling $N_i \sim Poisson(\xi_i)$.

> **Natural modification to LDA:** Sample N_i, the number of words in document i, as overdispersed Poisson via (1) $\xi_i \sim Gamma(\sum_i \alpha_i, s)$ for some scale parameter $s > 0$ and (2) $N_i \sim Poisson(\xi_i)$.

This "Gamma-Poisson mixture" is a well studied model of Poisson overdispersion; under this model, N_i has the negative binomial distribution. Define $\Xi \in \mathbb{R}^{n \times n}$ as a diagonal matrix with $\Xi_{ii} = \xi_i$.

Lemma C.2. *Under the LDA model with the natural modification to N_i, conditionally on $Z_1, \ldots, Z_n$ and Ξ, the document-term matrix A has independent Poisson entries satisfying*

$$\mathbb{E}(A|\Xi, Z) = (\Xi Z)\beta^T. \tag{21}$$

Moreover, each element $(\Xi Z)_{ij}$ is independent Gamma(α_j, s) and this distribution is leptokurtic. Define Σ as a diagonal matrix with $\Sigma_{jj} = \alpha_j s^2$, the variance of Gamma$(\alpha_j, s)$. Then, the factor matrix

$$Z_* = (\Xi Z)\Sigma^{-1/2} \tag{22}$$

satisfies the identification assumptions for Varimax.

See Section F.2 for a short proof. The next result shows that `vsp` applied to the column centered version of A (i.e., $\breve{A} = A - \mathbf{1}_n(\mathbf{1}_n^T A/n)$) can estimate the LDA model with the natural modification. Similar to $\breve{A}$, define $\breve{\mathscr{A}}$ be the column centered version of $\mathscr{A}$.

Corollary C.2. *Let A be generated from the natural modification to LDA given above with k topics and let $\mathscr{A} = \mathbb{E}(A|\Xi, Z)$. Define Z_* as in Equation (22). Let $\widehat{Z}$ be the output of `vsp` using $\breve{A}$ as input (and skipping step 1). In the asymptotic regime where*

$$\Delta_n \succeq \log^{15.1} n, \quad \sigma_{\min}(\beta) \geq c_1,$$

for universal constant $c_1 \in (0, 1)$, almost surely there exists $P_n \in \mathcal{P}(k)$ s.t.

$$||\widehat{Z} - (Z_* - \mathbb{E}(Z_*))P_n||_{2\to\infty} = O_p(\Delta_n^{-.24} \log^{2.75} n). \tag{23}$$

Define the matrix $\Phi = \widehat{Z}^T \breve{A} \in \mathbb{R}^{k \times d}$ and estimate $\widehat{\beta} = (\Lambda_b^{-1}\Phi)^T \in \mathbb{R}^{d \times k}$, where Λ_b is a diagonal matrix with ith diagonal element equals to ℓ_1-norm of ith row of Φ. Under this construction,

$$||\widehat{\beta}^T - P_n^T \beta^T||_\infty = O_p(\Delta_n^{-.24} \log^{3.75} n). \tag{24}$$

The elements of Z_* are independent Gamma random variables that have been rescaled by the diagonal matrix $\Sigma^{-1/2}$ to ensure that they have unit variance. Corollary C.2 shows that `vsp` using the column-centered matrix $\breve{A}$ estimates $Z_* - \mathbb{E}(Z_*)$; similar to the previous results, this $2 \to \infty$ convergence implies that each row of $\widehat{Z}$ converges to the corresponding row of $Z_* - \mathbb{E}(Z_*)$. Using $\widehat{Z}$, the corollary constructs $\hat{\beta} \in \mathbb{R}^{d \times k}$, a simple estimator for the probability distribution of words within each of the k topics. Each of the k estimated topic distributions converges in ℓ_1 norm just a little slower than $\Delta_n^{-1/4}$. A proof of Corollary C.2 is given in Section J.2.

D Supplemental results for the Journal-Journal citation data

D.1 `bff` for $\widehat{Z}$ with $k = 10$

The list below is analogous to the list in Section 3.1.1, except this list is for $\widehat{Z}$ instead of $\widehat{Y}$. These are the largest seven elements of `bff` with $\widehat{Z}$. One noticeable difference is that arxiv is now the top term in the mathematics factor. Previously, it was the third.

1. surgery, medicine, clinical, oncology, cancer, cardiovascular, official
2. molecular, biology, cell, cancer, microbiology, immunology, cellular
3. neuroscience, psychology, psychiatry, brain, cognitive, neurology, behavior
4. materials, chemistry, physics, chemical, acs, science, physical
5. ecology, plant, biology, conservation, marine, evolution, environmental
6. earth, geology, geophysics, atmospheric, geophysical, sensing, remote
7. ieee, on, conference, transactions, processing, communications, systems
8. arxiv, mathematical, physics, mathematics, geometry, analysis, theory
9. economics, economic, finance, review, financial, business, management
10. oral, dentistry, dental, surgery, maxillofacial, orthodontics, journal

D.2 Top journals in $\widehat{Y}$ and $\widehat{Z}$ for $k = 10$

Here are the top journals in $\widehat{Y}$.

1. jama, the new england journal of medicine, the lancet, annals of internal medicine, bmj, archives of internal medicine, chest, circulation, radiology, the cochrane database of systematic reviews, annals of surgery, the american journal of medicine
2. the embo journal, cell, molecular and cellular biology, febs letters, nucleic acids research, the journal of cell biology, the journal of biological chemistry, biochimica et biophysica acta, genes development, biochemical and biophysical research communications, molecular cell, biochemistry
3. psychological bulletin, the american journal of psychiatry, biological psychiatry, archives of general psychiatry, psychological review, neuropsychologia, psychological science, journal of abnormal psychology, trends in cognitive sciences, neuroimage, journal of personality and social psychology, journal of cognitive neuroscience

4. advanced materials, journal of materials chemistry, nano letters, journal of physical chemistry c, chemistry of materials, langmuir the acs journal of surfaces and colloids, journal of the american chemical society, cheminform, applied physics letters, journal of applied physics, acs nano, chemical communications

5. ecology, oecologia, oikos, trends in ecology evolution, the american naturalist, annual review of ecology evolution and systematics, journal of applied ecology, biological conservation, ecology letters, molecular ecology, journal of ecology, conservation biology

6. earth and planetary science letters, geology, journal of geophysical research, geological society of america bulletin, tectonophysics, geophysical research letters, geological society london special publications, geochimica et cosmochimica acta, chemical geology, contributions to mineralogy and petrology, journal of petrology, geophysical journal international

7. ieee transactions on pattern analysis and machine intelligence, arxiv, ieee transactions on image processing, ieee trans pattern anal mach intell, proceedings of the ieee, ieee transactions on signal processing, international journal of computer vision, ieee communications magazine, ieee trans inf theory, ieee transactions on information theory, ieee journal on selected areas in communications, ieee transactions on communications

8. transactions of the american mathematical society, annals of mathematics, inventiones mathematicae, advances in mathematics, duke mathematical journal, mathematische annalen, communications in mathematical physics, american journal of mathematics, mathematische zeitschrift, bulletin of the london mathematical society, journal of functional analysis, proceedings of the london mathematical society

9. the american economic review, journal of political economy, quarterly journal of economics, national bureau of economic research, econometrica, the economic journal, the review of economic studies, journal of monetary economics, journal of finance, journal of econometrics, the review of economics and statistics, journal of financial economics

10. the journal of prosthetic dentistry, journal of the american dental association, journal of periodontology, oral surgery oral medicine oral pathology oral radiology and endodontics, journal of clinical periodontology, journal of dental research, journal of dentistry, journal of oral rehabilitation, dental materials official publication of the academy of dental materials, american journal of orthodontics and dentofacial orthopedics official publication of the american association of orthodontists its constituent societies and the american board of orthodontics, journal of endodontics, clinical oral implants research

Here are the top journals in $\widehat{Z}$.

1. medicine, european radiology, the journal of bone and joint surgery american volume, world journal of surgery, clinical orthopaedics and related research, spine, bmj open, skeletal radiology, annals of surgical oncology, bmc musculoskeletal disorders, circulation, european spine journal

2. the journal of biological chemistry, biochimica et biophysica acta, international journal of molecular sciences, cellular and molecular life sciences, bmc genomics, nucleic acids research, oncotarget, frontiers in immunology, cancer research, journal of cell science, biochemical and biophysical research communications, cell

3. frontiers in psychology, frontiers in human neuroscience, neuroscience biobehavioral reviews, neuropsychologia, neuroimage, psychological bulletin, journal of cognitive neuroscience, frontiers in psychiatry, cerebral cortex, behavioural brain research, psychopharmacology, experimental brain research

4. materials, journal of materials science, acs applied materials interfaces, nanomaterials, journal of nanomaterials, polymers, acs nano, nanoscale research letters, langmuir the acs journal of surfaces and colloids, journal of physical chemistry c, journal of nanoparticle research, journal of thermal analysis and calorimetry

5. oecologia, ecology and evolution, biological invasions, hydrobiologia, marine ecology progress series, biological conservation, oikos, molecular ecology, global change biology, biodiversity and conservation, behavioral ecology and sociobiology, ecology

6. earth and planetary science letters, international journal of earth sciences, journal of geophysical research, geophysical research letters, tectonophysics, earthscience reviews, geochemistry geophysics geosystems, tectonics, international geology review, arabian journal of geosciences, lithos, journal of petrology

7. ieee access, arxiv, multimedia tools and applications, neurocomputing, ieee transactions on image processing, ieee transactions on multimedia, mathematical problems in engineering, pattern recognit, ieee transactions on cybernetics, international journal of computer applications, ieee transactions on circuits and systems for video technology, ieee transactions on neural networks and learning systems

8. transactions of the american mathematical society, arxiv differential geometry, advances in mathematics, arxiv algebraic geometry, communications in mathematical physics, arxiv representation theory, arxiv geometric topology, arxiv number theory, arxiv analysis of pdes, arxiv dynamical systems, pacific journal of mathematics, mathematische annalen

9. social science research network, national bureau of economic research, the american economic review, european economic review, applied economics, economics letters, imf working papers, the economic journal, journal of economic dynamics and control, journal of public economics, review of economics and statistics, journal of banking and finance

10. the journal of contemporary dental practice, clinical oral investigations, brazilian oral research, journal of applied oral science, journal of dentistry, european journal of dentistry, international journal of dentistry, bdj, bmc oral health, the journal of prosthetic dentistry, journal of clinical and experimental dentistry, brazilian dental journal

D.3 Leading journals for $k = 100$

The list below gives the leading five journals in the $k = 100$ factors of $\widehat{Y}$. The bold font at the beginning of each line gives the first two `bff` terms for this factor; only the first appears in Table 2 of the main text.

gastroenterology–hepatology: clinical gastroenterology and hepatology the official clinical practice journal of the american gastroenterological association, alimentary pharmacology therapeutics, american journal of gastroenterology, endoscopy, journal of clinical gastroenterology

cardiovascular–cardiology: heart, international journal of cardiology, american heart journal, european heart journal, catheterization and cardiovascular interventions official journal of the society for cardiac angiography interventions

communications–ieee: ieee transactions on wireless communications, ieee communications magazine, ieee transactions on communications, ieee journal on selected areas in communications, ieee transactions on vehicular technology

pharmaceutical–drug: european journal of pharmaceutics and biopharmaceutics official journal of arbeitsgemeinschaft fur pharmazeutische verfahrenstechnik ev, international journal of pharmaceutics, journal of pharmaceutical sciences, journal of controlled release official journal of the controlled release society, pharmaceutical research

otolaryngology–neck: european archives of otorhinolaryngology, the journal of laryngology and otology, annals of otology rhinology laryngology, otolaryngology–head and neck surgery, otology neurotology official publication of the american otological society american neurotology society and european academy of otology and neurotology

rehabilitation–occupational: physical therapy, archives of physical medicine and rehabilitation, disability and rehabilitation, gait posture, clinical rehabilitation

transportation–part: transportation research record, transportation research part apolicy and practice, transportation research part bmethodological, transportation, transport reviews

communication–media: journal of communication, communication research, journalism

mass communication quarterly, journal of broadcasting electronic media, j computermediated communication
endocrinology–physiology: biology of reproduction, journal of reproduction and fertility, the journal of endocrinology, molecular reproduction and development, reproduction
environmental–water: water research, chemosphere, journal of hazardous materials, environmental science technology, the science of the total environment
ophthalmology–eye: journal of cataract and refractive surgery, eye, cornea, graefes archive for clinical and experimental ophthalmology, retina
astrophysics–physics: astronomy and astrophysics, the astrophysical journal, monthly notices of the royal astronomical society, the astronomical journal, icarus
geotechnical–engineering: journal of geotechnical and geoenvironmental engineering, geotechnique, canadian geotechnical journal, journal of geotechnical engineering, computers and geotechnics
mathematical–mathematics: inventiones mathematicae, annals of mathematics, american journal of mathematics, mathematische annalen, duke mathematical journal
mathematical–analysis: journal of differential equations, archive for rational mechanics and analysis, nonlinear analysistheory methods applications, journal of mathematical analysis and applications, arxiv analysis of pdes
microbiology–biotechnology: applied microbiology and biotechnology, journal of biotechnology, fems microbiology letters, microbiology, archives of microbiology
microbiology–plant: mycologia, fungal biology, plant disease, phytopathology, studies in mycology
neuroscience–brain: the european journal of neuroscience, trends in neurosciences, behavioural brain research, neuroreport, progress in neurobiology
parasitology–tropical: acta tropica, parasitology research, the journal of parasitology, parasitology, memorias do instituto oswaldo cruz
pharmacology–toxicology: journal of ethnopharmacology, planta medica, phytotherapy research ptr, fitoterapia, phytomedicine international journal of phytotherapy and phytopharmacology
rheumatology–arthritis: clinical and experimental rheumatology, clinical rheumatology, rheumatology, seminars in arthritis and rheumatism, arthritis care research
atmospheric–meteorology: monthly weather review, quarterly journal of the royal meteorological society, journal of the atmospheric sciences, bulletin of the american meteorological society, journal of climate
dermatology–dermatologic: dermatology, journal of the european academy of dermatology and venereology jeadv, clinical and experimental dermatology, international journal of dermatology, acta dermatovenereologica
probability–annals: stochastic processes and their applications, annals of applied probability, annals of probability, theory of probability and its applications, advances in applied probability

accounting–financial: journal of financial economics, journal of accounting and economics, journal of accounting research, journal of finance, review of financial studies
anesthesia–anaesthesia: anaesthesia, acta anaesthesiologica scandinavica, european journal of anaesthesiology, british journal of anaesthesia, canadian journal of anaesthesia
analytical–chromatography: analytical and bioanalytical chemistry, analytica chimica acta, journal of pharmaceutical and biomedical analysis, journal of chromatography b analytical technologies in the biomedical and life sciences, journal of chromatography a
entomology–insect: journal of economic entomology, environmental entomology, journal of applied entomology, entomologia experimentalis et applicata, annals of the entomological society of america
immunology–allergy: clinical and experimental allergy journal of the british society for allergy and clinical immunology, annals of allergy asthma immunology official publication of the american college of allergy asthma immunology, allergy, respiratory medicine, the european respiratory journal
immunology–cell: immunity, nature immunology, nature reviews immunology, european journal of immunology, annual review of immunology
infectious–microbiology: clinical microbiology and infection the official publication of the european society of clinical microbiology and infectious diseases, clinical infectious diseases an official publication of the infectious diseases society of america, the journal of antimicrobial chemotherapy, the lancet infectious diseases, bmc infectious diseases
management–business: academy of management journal, academy of management review, journal of management, administrative science quarterly, strategic management journal
nephrology–transplantation: clinical journal of the american society of nephrology cjasn, peritoneal dialysis international journal of the international society for peritoneal dialysis, clinical nephrology, nephrology dialysis transplantation official publication of the european dialysis and transplant association european renal association, american journal of nephrology
obstetrics–gynecology: european journal of obstetrics gynecology and reproductive biology, bjog an international journal of obstetrics and gynaecology, ultrasound in obstetrics gynecology the official journal of the international society of ultrasound in obstetrics and gynecology, acta obstetricia et gynecologica scandinavica, international journal of gynaecology and obstetrics the official organ of the international federation of gynaecology and obstetrics
psychiatry–psychiatric: acta psychiatrica scandinavica, schizophrenia bulletin, the journal of clinical psychiatry, journal of affective disorders, the british journal of psychiatry the journal of mental science
psychology–cognition: memory cognition, cognitive psychology, psychonomic bulletin review, journal of experimental psychology human perception and performance, journal of experimental psychology learning memory and cognition
psychology–social: personality and social psychology bulletin, journal of experimental

social psychology, personality social psychology bulletin, european journal of social psychology, advances in experimental social psychology
quaternary–geology: radiocarbon, quaternary science reviews, palaeogeography palaeoclimatology palaeoecology, journal of archaeological science, quaternary international
statistics–statistical: annals of statistics, journal of statistical planning and inference, biometrika, statistical science, journal of the royal statistical society series b statistical methodology
toxicology–environmental: american journal of industrial medicine, occupational and environmental medicine, scandinavian journal of work environment health, international archives of occupational and environmental health, journal of occupational and environmental medicine
veterinary–animal: journal of the american veterinary medical association, journal of veterinary internal medicine, american journal of veterinary research, the veterinary record, veterinary journal
chemistry–chemical: journal of physical chemistry c, angewandte chemie, physical chemistry chemical physics pccp, the journal of physical chemistry b, chemical communications
economics–economic: the economic journal, quarterly journal of economics, national bureau of economic research, the review of economic studies, the american economic review
education–educational: review of educational research, journal of research in science teaching, journal of educational psychology, international journal of science education, science education
geography–planning: environment and planning a, progress in human geography, environment and planning dsociety space, annals of the association of american geographers, geoforum
marketing–management: journal of marketing, journal of the academy of marketing science, journal of consumer research, journal of marketing research, journal of business research
materials–engineering: materials science and engineering astructural materials properties microstructure and processing, acta materialia, scripta materialia, metallurgical and materials transactions a, journal of materials processing technology
mechanics–structures: computers structures, international journal of solids and structures, journal of applied mechanics, journal of sound and vibration, engineering structures
neurology–neurosurgery: acta neurochirurgica, surgical neurology, neurosurgical focus, neurosurgery, clinical neurology and neurosurgery
nutrition–obesity: international journal of obesity, journal of the american dietetic association, public health nutrition, european journal of clinical nutrition, obesity
numerical–siam: numerische mathematik, siam journal on numerical analysis, siam j scientific computing, mathematics of computation, math comput
political–politics: american political science review, american journal of political science, the journal of politics, international organization, comparative political studies

radiology–imaging: european radiology, european journal of radiology, radiographics a review publication of the radiological society of north america inc, journal of magnetic resonance imaging jmri, medical physics
sociology–sociological: american sociological review, american journal of sociology, social forces, review of sociology, journal of marriage and family
circuits–math: appl math comput, comput math appl, appl math lett, communications in nonlinear science and numerical simulation, chaos solitons fractals
genetics–molecular: genome research, nature reviews genetics, plos genetics, genome biology, trends in genetics tig
language–second: the modern language journal, language learning, tesol quarterly, studies in second language acquisition, applied linguistics
language–hearing: journal of speech language and hearing research jslhr, journal of speech and hearing research, american journal of speechlanguage pathology, the journal of speech and hearing disorders, clinical linguistics phonetics
robotics–automation: autonomous robots, the international journal of robotics research, ieee transactions on robotics, robotics auton syst, 2011 ieee international conference on robotics and automation
software–trans: ieee transactions on software engineering, ieee trans software eng, ieee software, computer, commun acm
alcohol–health: addictive behaviors, journal of substance abuse treatment, addiction, drug and alcohol dependence, journal of studies on alcohol
control–decision: international journal of control, ifac proceedings volumes, autom, ieee trans automat contr, ieee transactions on automatic control
ecology–forest: forest ecology and management, journal of ecology, journal of applied ecology, ecological applications, canadian journal of forest research
ecology–evolution: animal behaviour, journal of zoology, behavioral ecology and sociobiology, behavioral ecology, behaviour
geology–earth: contributions to mineralogy and petrology, journal of petrology, lithos, tectonophysics, precambrian research
nursing–nurse: journal of advanced nursing, journal of clinical nursing, nurse education today, journal of professional nursing official journal of the american association of colleges of nursing, international journal of nursing studies
optical–optics: optics express, optics letters, ieee photonics technology letters, journal of lightwave technology, optics communications
physics–physical: journal of high energy physics, physics letters b, nuclear physics, physical review d, classical and quantum gravity
physics–fluids: physics of fluids, journal of fluid mechanics, annual review of fluid mechanics, aiaa journal, journal of computational physics
polymer–composites: journal of applied polymer science, polymer degradation and stability, composites part aapplied science and manufacturing, polymer, polymer engineering and science

sensing–remote: international journal of remote sensing, ieee trans geoscience and remote sensing, ieee transactions on geoscience and remote sensing, photogrammetric engineering and remote sensing, remote sensing of environment
surgery–orthopaedic: arthroscopy the journal of arthroscopic related surgery official publication of the arthroscopy association of north america and the international arthroscopy association, the journal of bone and joint surgery british volume, journal of shoulder and elbow surgery, the journal of arthroplasty, journal of orthopaedic trauma
surgery–surgical: surgical endoscopy, journal of the american college of surgeons, world journal of surgery, the american surgeon, american journal of surgery
surgery–plastic: annals of plastic surgery, clinics in plastic surgery, aesthetic plastic surgery, british journal of plastic surgery, the journal of craniofacial surgery
tourism–hospitality: annals of tourism research, tourism management, journal of travel research, international journal of hospitality management, international journal of contemporary hospitality management
urology–urological: journal of endourology, bju international, world journal of urology, european urology, urology
animal–science: journal of animal science, poultry science, livestock production science, animal feed science and technology, journal of dairy science
cancer–oncology: annals of oncology official journal of the european society for medical oncology, the lancet oncology, european journal of cancer, journal of clinical oncology, the oncologist
comput–j: j symb log, theor comput sci, inf comput, j acm, studia logica
energy–engineering: applied energy, energy, energy conversion and management, renewable sustainable energy reviews, applied thermal engineering
health–care: health affairs, academic medicine journal of the association of american medical colleges, journal of general internal medicine, medical care, health services research
marine–fisheries: marine ecology progress series, marine biology, journal of experimental marine biology and ecology, journal of fish biology, estuarine coastal and shelf science
nature–the: proceedings of the national academy of sciences of the united states of america, the journal of biological chemistry, nature, science, cell
sports–sport: journal of strength and conditioning research, journal of sports sciences, international journal of sports medicine, sports medicine, british journal of sports medicine
speech–processing: ieee trans speech and audio processing, ieee transactions on audio speech and language processing, ieee trans signal process, speech commun, ieee signal processing letters
vision–computer: acm trans graph, 2010 ieee computer society conference on computer vision and pattern recognition, international journal of computer vision, cvpr 2011, 2015 ieee conference on computer vision and pattern recognition cvpr
aging–gerontology: the gerontologist, the journals of gerontology series b psychological sciences and social sciences, international journal of geriatric psychiatry, international psychogeriatrics, journal of gerontology

child–psychology: child development, developmental psychology, development and psychopathology, journal of consulting and clinical psychology, child abuse neglect
fuzzy–transactions: fuzzy sets syst, inf sci, ieee transactions on fuzzy systems, ieee trans fuzzy systems, fuzzy sets and systems
plant–biology: journal of plant physiology, physiologia plantarum, plant science, journal of experimental botany, plant cell reports
comb–j: discret math, j comb theory ser b, eur j comb, journal of graph theory, siam j discret math
food–science: journal of food engineering, lwt food science and technology, food research international, journal of food science, trends in food science and technology
ieee–communications: ieee network, comput commun, ieee transactions on parallel and distributed systems, comput networks, wireless networks
ieee–power: ieee transactions on power electronics, ieee transactions on industry applications, ieee transactions on industrial electronics, ieee transactions on energy conversion, ieee transactions on power delivery
ieee–transactions: ieee transactions on microwave theory and techniques, ieee transactions on antennas and propagation, ieee antennas and wireless propagation letters, ieee microwave and wireless components letters, electronics letters
oper–res: eur j oper res, european journal of operational research, oper res, comput oper res, journal of the operational research society
oral–dentistry: journal of the american dental association, the journal of prosthetic dentistry, journal of dentistry, journal of clinical periodontology, journal of dental research
soil–water: soil science society of america journal, soil tillage research, agronomy journal, soil science, geoderma
inf–syst: j manag inf syst, inf syst res, mis quarterly, inf manag, european journal of information systems
de–enfermagem: ciencia saude coletiva, revista da escola de enfermagem da usp, revista brasileira de enfermagem, texto contexto enfermagem, revista latinoamericana de enfermagem

E Other factors from analyzing the abstracts

E.1 Subject area factors

This section gives the 32 subject area factors in $\widehat{Y} \in \mathbb{R}^{240,331\times 50}$, when analyzing the document-term matrix of the abstracts. The term in bold is based upon our reading of the terms.
oceans: assemblages, foraminiferal, assemblage, sea, benthic, foraminifera, ocean, sediment, sediments, planktonic
soil–chemistry: humic, excitation, fluorescence, dom, parafac, dissolved, eem, emission, parallel, c1

dietary–nutrition: dietary, vegetables, meat, intake, fruits, food, diet, foods, fish, intakes
education: students, teaching, teachers, learning, teacher, student, school, academic, teach, education
metals–chemistry: zn, pb, cu, ni, cr, metals, cd, mn, fe, metal
consumer–purchase: consumers, consumer, purchase, shopping, brand, purchasing, fashion, buying, store, products
air–polution: aerosol, dust, particulate, particles, combustion, aerosols, pm2, air, burning, atmospheric
water–chemistry: groundwater, cl, hco3, na, ca2, mg2, no3, hydrochemical, so42, ions
cancer–clinical: tumor, metastasis, lymph, prognostic, survival, node, prognosis, carcinoma, cox, meier
health–measures: cholesterol, hdl, glucose, fasting, lipoprotein, triglycerides, metabolic, waist, insulin, systolic
plant–genetics: genotypes, seed, breeding, plant, yield, genetic, traits, cultivars, replications, characters
genetics: gene, genes, transcription, expression, cells, cell, dna, protein, pcr, genome
agricultural–economics: farmers, income, economic, rural, household, farm, government, farming, households, agricultural
psychiatric: disorder, symptoms, dsm, disorders, symptom, depression, diagnostic, depressive, psychiatric, compulsive
chemometrics–spectrometry: spectra, alternating, resolution, chromatography, squares, chromatographic, chemometric, calibration, spectral, mcr
patient–outcomes: care, patient, patients, medical, health, nurses, hospital, nursing, quality, hospitals
electron–photon–spectroscopy: electron, diffraction, ray, photoelectron, auger, spectroscopy, x, schmid, films, xps
speech–recognition: speaker, nist, gmm, sre, jfa, speech, gaussian, vector, recognition, ubm
geochemistry: geochemical, rocks, mineralization, ore, minerals, rock, deposits, au, quartz, geological
traditional–chinese–medicine: qi, yin, tcm, stasis, spleen, phlegm, deficiency, yang, dampness, stagnation
polution: pahs, polycyclic, hydrocarbons, aromatic, pah, pyrene, benzo, combustion, anthracene, compounds
artery–imaging: artery, imaging, myocardial, coronary, cardiac, left, brain, infarction, ventricular, heart
psychology–personalities: self, personality, negative, neuroticism, extraversion, inventory, behaviors, positively, coping, predicted
family–children–parents: children, child, mothers, parents, parent, maternal, mother, parental, birth, parenting

memory–tests: memory, verbal, neuropsychological, battery, subtests, wechsler, executive, cognitive, intelligence, abilities
tourism: tourists, tourism, tourist, destination, travel, visitors, destinations, attractions, visit, visitor
basic–chemistry–summaries: temperature, min, optimum, ph, acid, liquid, ml, conditions, ethanol, water
ecology–forest: habitat, niche, species, enfa, ecological, habitats, forest, vegetation, conservation, forests
china–econ–development: china, province, cities, regional, provinces, forward, economic, system, comprehensive, economy
finance: stock, financial, market, listed, companies, investors, returns, investment, profitability, firms
organization–business: employees, job, organizational, employee, leadership, commitment, managers, organization, organizations, satisfaction
clinical–infections: infection, risk, infections, logistic, incidence, antibiotics, staphylococcus, aureus, cases, infected

E.2 Artifacts and anomalies detected by `vsp`

Section E.2.1 lists eight factors that illustrate how `vsp` can highlight text artifacts (e.g. stop words, legal permissions, and html tags)

The final three factors in Section E.2.2 appear to be anomalies. Two of these factors found a group of six papers with identical abstracts; this was due to a parsing failure at Semantic Scholar. Each of these factors corresponds to a single page of a journal that contained all six abstracts. Semantic Scholar recognized them as separate articles, but pasted their abstracts together and assigned it to all six. The final factor appears to find papers in a Korean content farm.

E.2.1 Strange artifact factors

numbers: 4, 6, 5, 7, 9, 1, 3, 8, 2, 11
stop–words: not, they, it, or, more, but, if, many, have, what
permission–to–distribute–words: reproductions, edrs, supplied, eric, granted, reproduce, sld, oeri, permission, document
html–parse–errors: msonormaltable, tstyle, rowband, colband, noshow, 4pt, 0pt, pagination, padding, mso
french–stop–words: une, les, dans, des, pour, que, sont, factorielle, ont, etude
citations–in–abstract: j, al, et, c, l, psychology, g, 1993, his, 1990
Turkish: analizi, faktor, bu, olarak, ozet, toplam, oldugu, ile, bir, guvenirlik
legal–license: http, org, creativecommons, ltd, license, licenses, doi, copyright, www, wiley

E.2.2 Factors that correspond to anomalies

parse–error1: rostral, amygdalar, wako, riken, jst, averted, mpfc, astrocyte, 58s, s63
parse–error2: val66met, appswe, extradimensional, prepulse, presenilin, bdnf, neures, s117, k03, mayu
content–farm–korea: follows, seoul, third, spss, sports, second, frequency, amos, sport, gyeonggi

F Proofs for Proposition 5.1 and Theorem 5.1

The following is a proof of Proposition 5.1.

Proof. With input $\mathscr{A}$, recall that the row, column, and grand means are

$$\mu_r = \mathscr{A}\mathbf{1}_d/d \in \mathbb{R}^n, \quad \mu_c = \mathbf{1}_n^T\mathscr{A}/n \in \mathbb{R}^d, \quad \mu = \mathbf{1}_n^T\mathscr{A}\mathbf{1}_d/(nd) \in \mathbb{R}.$$

The centered version of $\mathscr{A}$ is defined to be $\widetilde{\mathscr{A}} = \mathscr{A} - \mu_r\mathbf{1}_d^T - \mathbf{1}_n\mu_c + \mu\mathbf{1}_n\mathbf{1}_d^T \in \mathbb{R}^{n\times d}$. Define the column means of Y and Z as

$$\mu_Y = Y^T\mathbf{1}_d/d \quad \text{and} \quad \mu_Z = \mathbf{1}_n^T Z/n.$$

Note that

$$\mu_r = ZB\mu_Y \quad \mu_c = \mu_Z BY^T \quad \text{and} \quad \mu = \mu_Z B\mu_Y.$$

Also note that $\bar{Y} = \mathbf{1}_d\mu_Y^T$ and $\bar{Z} = \mathbf{1}_n\mu_Z$. Putting the pieces together gives the result.

$$\begin{aligned}
\widetilde{\mathscr{A}} &= \mathscr{A} - \mu_r\mathbf{1}_d^T - \mathbf{1}_n\mu_c + \mu\mathbf{1}_n\mathbf{1}_d^T \\
&= ZBY^T - ZB\mu_Y\mathbf{1}_d^T - \mathbf{1}_n\mu_Z BY + \mathbf{1}_n\mu_Z B\mu_Y\mathbf{1}_d^T \\
&= ZBY^T - ZB\bar{Y}^T - \bar{Z}BY^T + \bar{Z}B\bar{Y}^T \\
&= (Z-\bar{Z})B(Y-\bar{Y})^T
\end{aligned}$$

□

The following is a proof of Theorem 5.1.

Proof. By the Assumption 1,

$$\mathbb{E}(Z_i^o) = 0, \mathbb{E}((Z_i^o)^2) = 1, \mathbb{E}((Z_i^o)^4) = \eta_i > 3, \forall i \in [k].$$

Thus,

$$\mathbb{E}(v(R, Z^o\widetilde{R}^T)) = \sum_{j=1}^{k} \mathbb{E}[Z^o\widetilde{R}^T R]_j^4.$$

To simplify notation, the proof reparameterizes the optimization parameter R as follows. For the rotation matrix R, define $O = \widetilde{R}^T R \in \mathscr{O}(k)$. We want to choose $O \in \mathscr{O}(k)$ to optimize the quantities $\sum_j \mathbb{E}[Z^o O]_j^4 = \sum_j \mathbb{E}(Z^o O_{\cdot j})^4$, where $O_{\cdot j} \in \mathbb{R}^k$ is the jth column of O. Notice elements of Z^o are independent and each has zero-mean. We have

$$\begin{aligned}\sum_j \mathbb{E}(Z^o O_{\cdot j})^4 &= \sum_{j=1}^k \left(\sum_{i=1}^k \mathbb{E}((Z_i^o)^4) O_{ij}^4 + 3 \sum_{i \neq \ell} \mathbb{E}((Z_i^o)^2 (z_\ell^o)^2) O_{ij}^2 O_{\ell j}^2 \right) \\ &= \sum_{j=1}^k \left(\sum_{i=1}^k \eta_i O_{ij}^4 + 3 \sum_{i \neq \ell} O_{ij}^2 O_{\ell j}^2 \right). \end{aligned} \tag{25}$$

The above equation only depends on the squared elements of O. Define $O^{(2)} \in \mathbb{R}^{k \times k}$ such that $O_{ij}^{(2)} = O_{ij}^2$. Because $O \in \mathscr{O}(k)$, $O^{(2)}$ is a doubly stochastic matrix, where each element is non-negative and all row and column sums are equal to one. Define

$$F_\eta(Q) = \sum_{j=1}^k \left(\sum_{i=1}^k \eta_i Q_{ij}^2 + 3 \sum_{i \neq \ell} Q_{ij} Q_{\ell j} \right). \tag{26}$$

Define $\mathcal{S}(k)$ as the set of $k \times k$ doubly stochastic matrices. Note that

$$\sum_j \mathbb{E}(Z^o O_{\cdot j})^4 = F_\mu(O^{(2)}) \leq \max_{Q \in \mathcal{S}(k)} F_\eta(Q).$$

In this way, the Varimax problem relaxes from orthonormal matrices to doubly stochastic matrices.

The rest of the proof will show that

$$\max_{Q \in \mathcal{S}(k)} F_\eta(Q) = \sum_{i=1}^k \eta_i. \tag{27}$$

Because $\sum_j \mathbb{E}(Z^o O_{\cdot j})^4$ evaluated with O as the identity matrix, is equal to $\sum_{i=1}^k \eta_i$, it follows that $O = I$ or $R = \widetilde{R}$ obtains the maximum. Moreover, for $P \in \mathcal{P}(k)$, $O = P$ (i.e. $R = \widetilde{R}P$) obtains the maximum value. It only remains to show Equation (27).

$$
\begin{aligned}
F_\eta(Q) &= \sum_{j=1}^{k}\left(\sum_{i=1}^{k}\eta_i Q_{ij}^2 + 3\sum_{i\neq\ell} Q_{ij}Q_{\ell j}\right) \\
&= \sum_{j=1}^{k}\left(\sum_{i=1}^{k}\eta_i Q_{ij}^2 + 3\left(\sum_{i=1}^{k} Q_{ij}\right)^2 - 3\sum_{i=1}^{k} Q_{ij}^2\right) \\
&= \sum_{j=1}^{k}\left(\sum_{i=1}^{k}\eta_i Q_{ij}^2 - 3\sum_{i=1}^{k} Q_{ij}^2\right) + 3k \\
&= \sum_{i=1}^{k}(\eta_i - 3)\sum_{j=1}^{k} Q_{ij}^2 + 3k \\
&\leq \sum_{i=1}^{k}(\eta_i - 3)\sum_{j=1}^{k} Q_{ij} + 3k \\
&= \sum_{i=1}^{k}(\eta_i - 3) + 3k \\
&= \sum_{i=1}^{k}\eta_i.
\end{aligned}
$$

The inequality is because $Q_{ij} \in [0,1], \forall i,j$ (this is because $Q \in \mathcal{S}(k)$).

To see that the maximum of $\sum_j \mathbb{E}(Z^o O_{\cdot j})^4$ is *only* attained by matrices in $\mathcal{P}(k)$, note that for any rotation matrix $O \notin \mathcal{P}(k)$, then

$$
\sum_j \mathbb{E}(Z^o O_{\cdot j})^4 = F_\mu(O^{(2)}) = \sum_{i=1}^{k}(\eta_i - 3)\sum_{j=1}^{k} O_{ij}^4 + 3k < \sum_{i=1}^{k}(\eta_i - 3)\sum_{j=1}^{k} O_{ij}^2 + 3k = \sum_{i=1}^{k}\eta_i,
$$

where the inequality is now strict.

□

F.1 A justification for the recentering step described in Remark 2.2

This section demonstrates how $\widehat{\mu}_Z = \sqrt{n}\widehat{\mu}_c \widehat{V}\widehat{D}^{-1}R_{\widehat{U}}$ can estimate $\mu_Z = \mathbf{1}^T Z/n$ under the Varimax assumptions on Z by studying the population behavior of $\widehat{\mu}_Z$. Define the population version of the estimator as

$$
\mu_Z^* = \sqrt{n}\mu_c V D^{-1}\widetilde{R}_U,
$$

where $\mu_c = \mathbf{1}_n^T \mathscr{A}/n = \mathbf{1}_n^T ZBY^T/n$, V and D are defined in Proposition 5.2 with the SVD of $\widetilde{\mathscr{A}}$ as

$$D = \sqrt{nd}\widetilde{D}, \quad V = d^{-1/2}\widetilde{Y}\widehat{\Sigma}_Y^{-1/2}\widetilde{R}_V^T,$$

and R_U is the population Varimax rotation $\widetilde{R}_U$ (as justified by Theorem 5.1). In the steps below, it is presumed that Z satisfies the Varimax assumptions. It is only presumed that Y is full rank. For simplicity, the $\approx$ correspond to approximating $\widehat{\Sigma}_Z$ as the identity matrix; under the Varimax assumptions, this is a reasonable approximation for large n. Recall that

$$B\widehat{\Sigma}_Y^{1/2} \approx \widetilde{R}_U^T \widetilde{D} \widetilde{R}_V.$$

Thereby,

$$\begin{aligned}
\mu_Z^* &= \sqrt{n}\mu_c V D^{-1}\widetilde{R}_U \\
&= \sqrt{n}\mu_Z(BY^T)VD^{-1}\widetilde{R}_U \\
&\approx \sqrt{n}\mu_Z(\widetilde{R}_U^T\widetilde{D}\widetilde{R}_V\widehat{\Sigma}_Y^{-1/2}Y^T)d^{-1/2}\widetilde{Y}\widehat{\Sigma}_Y^{-1/2}\widetilde{R}_V^T D^{-1}\widetilde{R}_U.
\end{aligned}$$

Then, $\widehat{\Sigma}_Y^{-1/2}Y^T\widetilde{Y}\widehat{\Sigma}_Y^{-1/2}$ is d multiplied by the identity matrix. Substituting for D and canceling out several terms yields the result,

$$\begin{aligned}
\mu_Z^* &\approx (nd)^{1/2}\mu_Z\widetilde{R}_U^T\widetilde{D}\widetilde{R}_V\widetilde{R}_V^T D^{-1}\widetilde{R}_U \\
&= (nd)^{1/2}\mu_Z\widetilde{R}_U^T\widetilde{D}\widetilde{R}_V\widetilde{R}_V^T(nd)^{1/2}\widetilde{D}^{-1}\widetilde{R}_U \\
&= \mu_Z\widetilde{R}_U^T\widetilde{D}\widetilde{R}_V\widetilde{R}_V^T\widetilde{D}^{-1}\widetilde{R}_U \\
&= \mu_Z.
\end{aligned}$$

The rigorous proof will be shown later in Proposition G.6 and Proposition G.7.

F.2 Proofs for Lemmas C.1 andC.2.

The following is a proof of Lemma C.1.

Proof. For ease of notation, refer to the topic for word w as z (instead of z_w),

$$\mathbb{P}(w = j|Z_i) = \sum_{z=1}^{k}\mathbb{P}(w = j|z, Z_i)\mathbb{P}(z|Z_i) = \sum_{z=1}^{k}\beta_{j,z}Z_{i,z} = \langle\beta_{j\cdot}, Z_i\rangle.$$

So, step 3 in the LDA model is equivalent to choosing word w to be word j with probability $[\beta Z_i]_j$. So, conditional on N_i and Z_i, the ith row of A is $Multinomial(N_i, \beta Z_i)$. Then, unconditional on N_i, due to the Poisson-Multinomial relationship, each element in the ith row of A is independent, with the distribution $A_{ij} \sim Poisson(\xi[\beta Z_i]_j)$. So, $E(A|Z) = \xi Z\beta^T$. □

The following is a proof of Lemma C.2.

Proof. There are three elements of Lemma C.2. **Part 1:** conditionally on $Z_1, \ldots, Z_n$ and Ξ, we need to show that the document-term matrix A has independent Poisson entries satisfying

$$\mathbb{E}(A|\Xi, Z) = (\Xi Z)\beta^T. \tag{28}$$

The proof of this is equivalent to the proof of Lemma C.1.

Part 2: The second part is that each element $(\Xi Z)_{ij}$ is independent Gamma(α_j, s). To see this, let $X_i \in \mathbb{R}_+^k$ have independent Gamma elements, $X_{ij} \sim$ Gamma(α_j, s). Define $\xi_i' = \sum_j X_{ij}$ and

$$Z_i' = \frac{X_i}{\xi_i'}.$$

It is well known that (1) $Z_i' \sim$ Dirichlet(α), (2) $\xi_i' \sim$ Gamma$(\sum_j \alpha_j, s)$, and (3) ξ_i' is independent of Z_i'. So,

$$(\Xi Z)_i = \xi_i Z_i \stackrel{d}{=} \xi_i' Z_i' = X_i.$$

Part 3: We need to show that $\Xi Z \Sigma^{-1/2}$ satisfies the identification assumptions for Varimax. From part 2 above, each row contains k independent random variables and each row is iid. Then, each element of ΞZ is leptokurtic because the Gamma distribution is always leptokurtic. Scaling by a constant $\Sigma^{-1/2}$ does not change this. The fourth piece of the identification assumptions for Varimax is ensured by the scaling $\Sigma^{-1/2}$.

□

G Proofs of Main Theorems

G.1 Proof of Theorem 4.1

Proof. Define the random variable $B \in \{0, 1\}$ to be equal to 1 when $X \neq 0$ and equal to 0 when $X = 0$. For some $0 < p < 1/6$, $B \sim Bernoulli(p)$. Define random variable S such that when $B = 1$, $S = X$ and when $B = 0$, then S is equal in distribution to X on the set $X \neq 0$. So,

$$X = SB.$$

Under the conditions of the theorem and the construction above, S has some arbitrary distribution with finite 4th moment and is also independent of B.

Let $\mu_i = \mathbb{E}(S^i)$. Then

$$\theta := \mathbb{E}(X) = (1-p)0 + p\mu_1 = p\mu_1, \tag{29}$$

$$\mathbb{E}[(X-\theta)^2] = p\mu_2 - p^2\mu_1^2, \tag{30}$$

$$\mathbb{E}[(X-\theta)^4] = p\mu_4 - 4p^2\mu_3\mu_1 + 6p^3\mu_2\mu_1^2 - 3p^4\mu_1^4. \tag{31}$$

So, in order to show that $\mathbb{E}[(X-\theta)^4] > 3\mathbb{E}[(X-\theta)^2]^2$, it is enough to show that

$$(\mu_4 - 3p\mu_2^2) + 6p^2\mu_1^2(\mu_2 - p\mu_1^2) > 4p\mu_3\mu_1 - 6p^2\mu_2\mu_1^2. \tag{32}$$

Using Lemma G.1 with $g = S^2, h = 2pS$, f being S's pdf, we have

$$\mu_4 - \mu_2^2 + 4p^2\mu_1^2(\mu_2 - \mu_1^2) \geq 4p\mu_3\mu_1 - 4p\mu_1^2\mu_2. \tag{33}$$

Subtract Equation (33) from Equation (32) we only need to show

$$(1-3p)\mu_2^2 + p^2\mu_1^2(2\mu_2 + 4\mu_1^2 - 6p\mu_1^2) > (4p - 6p^2)\mu_1^2\mu_2. \tag{34}$$

Notice $p < 1/6$, thus $(6p-1)(p-1) > 0 \Rightarrow 1 - 3p > 4p - 6p^2$. Thus by Jensen's Inequality

$$(1-3p)\mu_2^2 \geq (4p-6p^2)\mu_2^2 \geq (4p-6p^2)\mu_1^2\mu_2.$$

The first inequality is strict as long as $\mu_2 > \mu_1^2$. Also with $p < 1/6$ we have

$$p^2\mu_1^2(2\mu_2 + 4\mu_1^2 - 6p\mu_1^2) \geq p^2\mu_1^2(2\mu_2 + 3\mu_1^2) \geq 0.$$

The second inequality is strict as long as $p > 0, \mu_1 \neq 0$.
If $\mu_2 = \mu_1 = 0$ then $\mathbb{P}(X = 0) = 1$, contradiction. Thus X is leptokurtic.

Lemma G.1. *Suppose f is any distribution pdf. g, h is any integrable functions. Then*

$$\int g^2 f dx - (\int g f dx)^2 + (\int h f dx)^2 (\int h^2 f dx - (\int h f dx)^2)$$
$$\geq (\int ghf dx - \int gf dx \int hf dx) \int hf dx.$$

Let $\widetilde{g} = g - \int gfdx$, $\widetilde{h} = h - \int hfdx$. Then

$$\int g^2 f dx - (\int gfdx)^2 = \int \widetilde{g}^2 f dx,$$
$$\int h^2 f dx - (\int hfdx)^2 = \int \widetilde{h}^2 f dx,$$
$$\int ghfdx - \int gfdx \int hfdx = \int \widetilde{g}\widetilde{h} f dx.$$

By Cauchy-Schwarz Inequality,

$$\begin{aligned}
& \int g^2 f dx - (\int gfdx)^2 + (\int hfdx)^2 (\int h^2 f dx - (\int hfdx)^2) \\
= & \int \widetilde{g}^2 f dx + (\int hfdx)^2 \int \widetilde{h}^2 f dx \\
\geq & |\int hfdx| \sqrt{\int \widetilde{g}^2 f dx \int \widetilde{h}^2 f dx} \\
\geq & |\int hfdx| \int |\widetilde{g}\widetilde{h}| f dx \\
\geq & |\int hfdx| (|\int ghfdx - \int gfdx \int hfdx|) \\
\geq & (\int ghfdx - \int gfdx \int hfdx) \int hfdx.
\end{aligned}$$

□

G.2 Leptokurtosis with soft sparsity

The random variable X in Theorem 4.1 satisfies a hard sparsity condition. Imagine X as satisfying the conditions of Theorem 4.1. The next proposition studies $X + W$, where W is any independent random variable with a small variance. So, if W has a probability density, then $P(X + W = 0) \not> 0$, yet when W has expectation zero, then $X + W$ is still close to zero with high probability. In this regime, the next proposition shows that if X has a sufficiently large kurtosis, then $X + W$ is still leptokurtic, no matter the kurtosis of W.

Proposition G.1. *Let X and W be any independent random variables with four finite moments. Let $\eta_{x,j} = \mathbb{E}(X - \mathbb{E}(X))^j$ and $\eta_{w,j} = \mathbb{E}(W - \mathbb{E}(W))^j$. Let $\eta_{x,2} = 1$. For any $\epsilon > 0$, if $\eta_{w,2} < \epsilon$, and $\eta_{x,4} \geq 3(1 + \epsilon)^2$, then $X + W$ is leptokurtic.*

Note that both X and W can be rescaled to satisfy the assumption $\eta_{x,2} = 1$. In this way, it does not restrict the generality of the result. It only simplifies the notation.

Proof. Note that $\eta_{x,1} = \eta_{w,1} = 0$. Using that fact,

$$\mathbb{E}\left(X + W - \mathbb{E}(X + W)\right)^2 = \eta_{x,2} + \eta_{w,2} < 1 + \epsilon$$

and

$$\mathbb{E}\left(X + W - \mathbb{E}(X + W)\right)^4 = \eta_{x,4} + 6\eta_{x,2}\eta_{w,2} + \eta_{w,4} > 3(1 + \epsilon)^2.$$

The result follows from the definition of leptokurtic.

□

G.3 Proofs for the main results, Theorem 6.1

Proof. We need six propositions listed below to prove Theorem 6.1. Before the proof we clarify some notations. For a generic random matrix X, let R_X be its sample Varimax rotation, i.e.

$$R_X \in \underset{R \in \mathscr{O}(k)}{\arg\max}\ v(R, X),$$

where $v(R, X)$ is defined in Equation (2). Then, let R_X^* the population Varimax rotation, i.e.

$$R_X^* \in \underset{R \in \mathscr{O}(k)}{\arg\max}\ \mathcal{V}_X(R),$$

where the expectation in $\mathcal{V}_X(R) = \mathbb{E}(v(R, X\widetilde{R}))$ is defined over the distribution of X and the nuisance rotation $\widetilde{R}$ can be understood from the context. Define

$$W = \underset{W_0 \in \mathscr{O}(k)}{\arg\min} \|\widehat{U} - UW_0\|_{2\to\infty}.$$

$\mathcal{P}(k)$ is defined in Equation (13). $P_n = P_n^{(1)} P_n^{(2)} P_n^{(3)}$ where $P_n^{(i)} \in \mathcal{P}(k), i = 1, 2, 3$ are defined in Proposition G.3, G.4, G.5 respectively. Let $\mu_Z = \mathbf{1}_n^T Z/n$. J_n is n by n matrix with every entry equal to 1. $X^\dagger$ is the pseudo-inverse of X. Define $\xi = 1 + \epsilon$ for some small positive $\epsilon < 0.01$ for notation consistency with Cape et al. [2019a]. Recall $\Delta_n = n\rho_n, \bar{\Delta}_n = n\bar{\rho}_n$.

Define

$$\gamma_{ij}^{(n)} = \sup_{s \geq 2} \left(\frac{\mathbb{E}[(A_{ij} - \mathscr{A}_{ij})^s]}{s!} \right)^{1/s} \quad \text{and} \quad \gamma^{(n)} = \sup_{ij} \gamma_{ij}^{(n)}. \tag{35}$$

The $\gamma^{(n)}$ reveals the tail behaviors of sub-exponential random variables. It is useful in deriving matrix concentration results for sub-exponential random matrices later (Lemma G.4).

See Sections G.3.1 through G.3.6 for proofs to the following propositions G.2 through G.7. Several lemmas and technical details for these proofs are then delayed further into

Sections H and I.

Proposition G.2. *Let $\widehat{\Sigma}_Z = \widetilde{Z}^T\widetilde{Z}/n$. Under the settings of Theorem 6.1,*

$$\|U\widetilde{R}_U - U\widetilde{R}_U\widehat{\Sigma}_Z^{1/2}\|_{2\to\infty} = O_p(\frac{\log n}{n}). \tag{36}$$

Proposition G.3. *Under the settings of Theorem 6.1, there exists $P_n^{(1)} \in \mathcal{P}(k)$ s.t.*

$$\|\widehat{U}R_{UW}^* - U\widetilde{R}_U P_n^{(1)}\|_{2\to\infty} = O_p\left((n\rho_n)^{-1/2}n^{-1/2}\log^{\frac{5}{2}} n\right). \tag{37}$$

Proposition G.4. *Under the settings of Theorem 6.1, there exists $P_n^{(2)} \in \mathcal{P}(k)$ such that for any $\delta > 0$,*

$$\|\widehat{U}R_{UW} - \widehat{U}R_{UW}^* P_n^{(2)}\|_{2\to\infty} = O_p(n^{\delta/2-3/4}\log n). \tag{38}$$

Proposition G.5. *Under the settings of Theorem 6.1, there exists $P_n^{(3)} \in \mathcal{P}(k)$ s.t.*

$$\|\widehat{U}R_{\widehat{U}} - \widehat{U}R_{UW}P_n^{(3)}\|_{2\to\infty} = O_p\left((n\rho_n)^{-1/4}n^{-1/2}\log^{\frac{11}{4}} n\right). \tag{39}$$

Proposition G.6. *Define $P_n = P_n^{(1)}P_n^{(2)}P_n^{(3)}$ with $P_n^{(1)}, P_n^{(2)}, P_n^{(3)}$ defined in Proposition G.3, G.4, G.5 respectively. Under the settings of Theorem 6.1, for any $\delta > 0$,*

$$\|J_n(A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - \mathscr{A}VD^{-1}\widetilde{R}_U P_n)\|_{2\to\infty} = O_p\left(n^{\delta/2+1/4} + (n\rho_n)^{-1/4}n^{1/2}\log^{\frac{7}{4}} n\right). \tag{40}$$

Proposition G.7. *Under the settings of Theorem 6.1,*

$$\|J_n(\sqrt{n}\mathscr{A}VD^{-1}\widetilde{R}_U - Z)\|_{2\to\infty} = O_p(\sqrt{n}\log n). \tag{41}$$

We are going to show the bound for $\|\sqrt{n}\widehat{U}R_{\widehat{U}} - \widetilde{Z}P_n\|_{2\to\infty}$ by splitting it into four parts using triangle inequalities. Proposition G.2, G.3, G.4, G.5 give the bound for each split component. Similarly we show the bound for $\|\mathbf{1}_n\widehat{\mu}_Z - \mathbf{1}_n\mu_Z^T P_n\|_{2\to\infty}$ by decomposing it into two parts and use Proposition G.6, G.7 to give bounds. The proofs of these propositions are shown after the proof of Theorem 6.1. The propositions that justify the equalities below are numbered on the left side of the equalities.

$$
\begin{aligned}
& \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \widetilde{Z}P_n\|_{2\to\infty} \\
\text{(Proposition 5.2)} \quad &= \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \sqrt{n}U\widetilde{R}_U\widehat{\Sigma}_Z^{1/2}P_n\|_{2\to\infty} \\
&= \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \sqrt{n}\widehat{U}R_{UW}P_n^{(3)} + \sqrt{n}\widehat{U}R_{UW}P_n^{(3)} - \sqrt{n}\widehat{U}R_{UW}^*P_n^{(2)}P_n^{(3)} \\
&\quad + \sqrt{n}\widehat{U}R_{UW}^*P_n^{(2)}P_n^{(3)} - \sqrt{n}U\widetilde{R}_UP_n + \sqrt{n}U\widetilde{R}_UP_n - \sqrt{n}U\widetilde{R}_U\widehat{\Sigma}_Z^{1/2}P_n\|_{2\to\infty} \\
&\le \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \sqrt{n}\widehat{U}R_{UW}P_n^{(3)}\|_{2\to\infty} + \|\sqrt{n}\widehat{U}R_{UW}P_n^{(3)} - \sqrt{n}\widehat{U}R_{UW}^*P_n^{(2)}P_n^{(3)}\|_{2\to\infty} \\
&\quad + \|\sqrt{n}\widehat{U}R_{UW}^*P_n^{(2)}P_n^{(3)} - \sqrt{n}U\widetilde{R}_UP_n\|_{2\to\infty} + \|\sqrt{n}U\widetilde{R}_UP_n - \sqrt{n}U\widetilde{R}_U\widehat{\Sigma}_Z^{1/2}P_n\|_{2\to\infty} \\
\text{(Proposition } G.2) \quad &= \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \sqrt{n}\widehat{U}R_{UW}P_n^{(3)}\|_{2\to\infty} + \|\sqrt{n}\widehat{U}R_{UW}P_n^{(3)} - \sqrt{n}\widehat{U}R_{UW}^*P_n^{(2)}P_n^{(3)}\|_{2\to\infty} \\
&\quad + \|\sqrt{n}\widehat{U}R_{UW}^*P_n^{(2)}P_n^{(3)} - \sqrt{n}U\widetilde{R}_UP_n\|_{2\to\infty} + O_p(\frac{\log n}{\sqrt{n}}) \\
\text{(Proposition } G.3) \quad &= \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \sqrt{n}\widehat{U}R_{UW}P_n^{(3)}\|_{2\to\infty} + \|\sqrt{n}\widehat{U}R_{UW}P_n^{(3)} - \sqrt{n}\widehat{U}R_{UW}^*P_n^{(2)}P_n^{(3)}\|_{2\to\infty} \\
&\quad + O_p\left((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n\right) + O_p(\frac{\log n}{\sqrt{n}}) \\
\text{(Proposition } G.4) \quad &= \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \sqrt{n}\widehat{U}R_{UW}P_n^{(1)}\|_{2\to\infty} + O_p(n^{\delta/2-1/4}\log n) \\
&\quad + O_p\left((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n\right) + O_p(\frac{\log n}{\sqrt{n}}) \\
\text{(Proposition } G.5) \quad &= O_p\left((n\rho_n)^{-1/4}\log^{\frac{11}{4}} n\right) + O_p(n^{\delta/2-1/4}\log n) + O_p\left((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n\right) + O_p(\frac{\log n}{\sqrt{n}}) \\
&= O_p\left((n\rho_n)^{-1/4}\log^{\frac{11}{4}} n\right) + O_p(n^{\delta/2-1/4}\log n) \\
&= O_p\left(\Delta_n^{-1/4+\delta/2}\log^{\frac{11}{4}} n\right). \qquad (42)
\end{aligned}
$$

For the recentering part, by Proposition G.6, G.7,

$$
\begin{aligned}
& \|\mathbf{1}_n\widehat{\mu}_Z - \mathbf{1}_n\mu_ZP_n\|_{2\to\infty} = \frac{1}{n}\|J_n^T(\sqrt{n}A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - ZP_n)\|_{2\to\infty} \\
\le\ & \frac{1}{n}\|J_n(\sqrt{n}A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - \sqrt{n}\mathscr{A}VD^{-1}\widetilde{R}_UP_n)\|_{2\to\infty} + \frac{1}{n}\|J_n(\sqrt{n}\mathscr{A}VD^{-1}\widetilde{R}_UP_n - ZP_n)\|_{2\to\infty} \\
=\ & O_p(n^{\delta/2-1/4} + (n\rho_n)^{-1/4}\log^{\frac{7}{4}} n + \frac{\log n}{\sqrt{n}}) \\
=\ & O_p(\Delta_n^{-1/4+\delta/2}\log^{\frac{7}{4}} n). \qquad (43)
\end{aligned}
$$

Take $\delta = 0.2$. Equation (42), (43) and triangle inequality accomplish the proof. □

Before the proofs for the six propositions, two useful lemmas are given. Lemma G.2 gives bound for the maximum absolute value of Z's elements. Lemma G.3 borrows matrix $2 \to \infty$ norm's property from Cape et al. [2019b].

Lemma G.2.

$$\max_{i,j}|Z_{ij}| = O_p(\log n), \quad \max_{i,j}|\widetilde{Z}_{ij}| = O_p(\log n).$$

$$\max_{i,j}|Y_{ij}| = O_p(\log d), \quad \max_{i,j}|\widetilde{Y}_{ij}| = O_p(\log d).$$

Proof. Assumption 2 indicates Z's columns are sub-exponential variables. Thus there exists $C_0, \lambda_j > 0, j \in [k]$'s s.t.

$$\mathbb{P}(|Z_{ij} - \mathbb{E}Z_{ij}| > t) \leq C_0 \exp(-\lambda_j t) \leq C_0 \exp(-\lambda t), \tag{44}$$

with $\lambda = \min_{j\in[k]} \lambda_j$. Then

$$\begin{aligned} \mathbb{P}(\max_{i,j}|Z_{ij} - \mathbb{E}Z_{ij}| > t) &\leq \sum_{i,j} \mathbb{P}(|Z_{ij} - \mathbb{E}Z_{ij}| > t) \\ &\leq \sum_{i,j} C_0 \exp(-\lambda t) \\ &\leq knC_0 \exp(-\lambda t). \end{aligned}$$

$\Rightarrow$

$$\max_{i,j}|Z_{ij}| = O_p(\log n), \quad \max_{i,j}|\widetilde{Z}_{ij}| = O_p(\log n).$$

Similar conclusion also applies to Y. □

With Lemma G.2, it could be trivially inferred that

$$\bar{\rho}_n = O(\rho_n \log^2 n). \tag{45}$$

Lemma G.3. *Suppose $X_1 \in \mathbb{R}^{n_1 \times n_2}, X_2 \in \mathbb{R}^{n_2 \times n_3}$ are real matrices. Then*

$$\|X_1 X_2\|_{2\to\infty} \leq \|X_1\|_{2\to\infty} \|X_2\|. \tag{46}$$

This is a direct conclusion from Proposition 6.5 in Cape et al. [2019b].

G.3.1 Proof of Proposition G.2

Proof. The (i,j) entry of $\widehat{\Sigma}_Z \in \mathbb{R}^{k\times k}$ is

$$\widehat{\Sigma}_Z[i,j] = \begin{cases} \frac{1}{n}\sum_{q=1}^{n}(Z_{qi} - \widehat{\mu}_Z[i])^2 & \text{if } i = j, \\ \frac{1}{n}\sum_{q=1}^{n}(Z_{qi} - \widehat{\mu}_Z[i])(Z_{qj} - \widehat{\mu}_Z[j]) & \text{if } i \neq j. \end{cases}$$

By LLN, $\|\widehat{\Sigma}_Z - I\|_{\max} = O_p(\frac{k^2}{\sqrt{n}}) = O_p(\frac{1}{\sqrt{n}})$, thus $\|\widehat{\Sigma}_Z - I\| \leq \sqrt{k^2}\|\widehat{\Sigma}_Z - I\|_{\max} = O_p(\frac{1}{\sqrt{n}})$. Suppose eigendecomposition of $\widehat{\Sigma}_Z$ is $\widehat{\Sigma}_Z = \Psi\Lambda_Z\Psi^T$. Then

$$\|\widehat{\Sigma}_Z - I\| = \|\Lambda_Z - I\| = O_p(\frac{1}{\sqrt{n}}) \Rightarrow \|\widehat{\Sigma}_Z^{1/2} - I\| = \|\Lambda_Z^{1/2} - I\| = O_p(\frac{1}{\sqrt{n}}).$$

Also $\|\widehat{\Sigma}_Z - I\| = O_p(\frac{1}{\sqrt{n}})$ implies $\|\widehat{\Sigma}_Z^{-1/2}\| = O_p(1)$. By Proposition 5.2 and Lemma G.3, G.2,

$$\|U\|_{2\to\infty} = \frac{1}{\sqrt{n}}\|\widetilde{Z}\widehat{\Sigma}_Z^{-1/2}\|_{2\to\infty} \leq \frac{1}{\sqrt{n}}\|\widetilde{Z}\|_{2\to\infty}\|\widehat{\Sigma}_Z^{-1/2}\| = O_p(\frac{\log n}{\sqrt{n}}). \tag{47}$$

Putting the above pieces together provides a bound on the quantity of interests.

$$\begin{aligned} \|U\widetilde{R}_U - U\widetilde{R}_U\widehat{\Sigma}_Z^{1/2}\|_{2\to\infty} &\leq \|U\widetilde{R}_U\|_{2\to\infty}\|I - \widehat{\Sigma}_Z^{1/2}\| \\ &= \|U\|_{2\to\infty}\|I - \widehat{\Sigma}_Z^{1/2}\| \\ &= O_p(\frac{\log n}{n}). \end{aligned}$$

□

G.3.2 Proof of Proposition G.3

We give the statement of Lemma G.4, G.5 below and use them to prove proposition G.3. The proof of these two lemmas will be shown in Section H.

Lemma G.4. *Define the symmetrized adjacent matrix as* $\widetilde{A}_{sym} = \begin{pmatrix} 0 & \widetilde{A} \\ \widetilde{A}^T & 0 \end{pmatrix}$ *and its population version as* $\widetilde{\mathscr{A}_{sym}} = \begin{pmatrix} 0 & \widetilde{\mathscr{A}} \\ \widetilde{\mathscr{A}^T} & 0 \end{pmatrix}$. *Under the settings in Theorem 6.1,*

$$\|A_{sym} - \mathscr{A}_{sym}\| = O_p((n\rho_n\log^3 n)^{\frac{1}{2}}). \tag{48}$$

Lemma G.5. *Presume the conditions in Theorem 6.1. There exists* $W \in \mathscr{O}(k)$, *such that*

$$\|\widehat{U} - UW\|_{2\to\infty} = O_p\left((n\rho_n)^{-1/2}n^{-1/2}\log^{\frac{5}{2}} n\right). \tag{49}$$

Lemma G.5 gives a row-wise bound for the eigenvectors' fluctuations. This lemma follows from Theorem 1 in Cape et al. [2019a], which requires several conditions. Lemma G.4 is used for one of the conditions. The other conditions are either already satisfied by the assumptions of Theorem 6.1 or checked inside the proof of Lemma G.5.

Proof. Notice the fact that $2 \to \infty$ norm is invariant to rotations. From Theorem 5.1 there exist $P_n^{(1)} \in \mathcal{P}(k)$ s.t. $R_{UW}^* = W^T \widetilde{R}_U P_n^{(1)}$. Therefore

$$\begin{aligned}
\|\widehat{U} R_{UW}^* - U \widetilde{R}_U P_n^{(1)}\|_{2\to\infty} &= \|\widehat{U} W^T \widetilde{R}_U P_n^{(1)} - U \widetilde{R}_U P_n^{(1)}\|_{2\to\infty} \\
&= \|\widehat{U} W^T - U\|_{2\to\infty} \\
&= \|\widehat{U} - UW\|_{2\to\infty} \\
\text{(Lemma } G.5) &= O_p\left((n\rho_n)^{-1/2} n^{-1/2} \log^{\frac{5}{2}} n\right).
\end{aligned}$$

□

G.3.3 Proof of Proposition G.4

The proof of Proposition G.4 uses the following lemma to bound the distance between sample and population Varimax solutions (modulo permutation and sign flip).

Lemma G.6. *Recall that* $R_{\widetilde{Z}} \in \underset{R_0 \in \mathcal{O}(k)}{\arg\max}\, v(R_0, \widetilde{Z})$. *There exists* $P_n^{(2)} \in \mathcal{P}(k)$ *s.t. for* $\forall \delta > 0$

$$\|R_{\widetilde{Z}} - P_n^{(2)}\|_{2\to\infty} = O_p(n^{\delta/2 - 1/4}).$$

The proof of Lemma G.6 is in Section H.

Proof. With some previous lemmas,

$$\begin{aligned}
& \|\widehat{U} R_{UW} - \widehat{U} R_{UW}^* P_n^{(2)}\|_{2\to\infty} \\
= \; & \|\widehat{U}(R_{UW} - R_{UW}^* P_n^{(2)})\|_{2\to\infty} \\
\text{(Lemma } G.3) \; \leq \; & \|\widehat{U}\|_{2\to\infty} \|R_{UW} - R_{UW}^* P_n^{(2)}\| \\
\text{(Lemma } G.6) \; = \; & O_p(n^{\delta/2-1/4} \|\widehat{U}\|_{2\to\infty}) \\
\leq \; & O_p(n^{\delta/2-1/4} \|\widehat{U} - UW\|_{2\to\infty} + n^{\delta/2-1/4} \|UW\|_{2\to\infty}) \\
\text{(Lemma } G.5) \; = \; & O_p\left((n\rho_n)^{-1/2} n^{-3/4} \log^{\frac{5}{2}} n\right) + O_p(n^{\delta/2-1/4} \|UW\|_{2\to\infty}) \\
\leq \; & O_p\left((n\rho_n)^{-1/2} n^{-3/4} \log^{\frac{5}{2}} n\right) + O_p(n^{\delta/2-1/4} \|U\|_{2\to\infty}) \\
\text{(Equation (47))} \; \leq \; & O_p\left((n\rho_n)^{-1/2} n^{-3/4} \log^{\frac{5}{2}} n\right) + O_p(n^{\delta/2-3/4} \log n) \\
(n\rho_n \succeq \log^{2\xi} n) \; = \; & O_p(n^{\delta/2-3/4} \log n).
\end{aligned}$$

□

G.3.4 Proof of Proposition G.5

This proposition shows that $R_{\widehat{U}}$ converges to R_{UW}. The proof of Proposition G.5 is contained in Section H. This proof uses the fact that the Varimax objective function is smooth and each row of $\widehat{U}$ converges to the corresponding row of UW (i.e. $\|\widehat{U} - UW\|_{2\to\infty} \to 0$). This implies that the Varimax solution computed with $\widehat{U}$ (i.e. $R_{\widehat{U}}$) converges to the Varimax solution computed with UW (i.e. R_{UW}).

G.3.5 Proof of Proposition G.6

Proof.

$$
\begin{aligned}
& \|J_n(A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - \mathscr{A}VD^{-1}\widetilde{R}_U P_n)\|_{2\to\infty} \\
(\text{Lemma } G.3) \quad \leq \quad & \sqrt{n}\|A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - \mathscr{A}VD^{-1}\widetilde{R}_U P_n\| \\
(WR^*_{UW} = \widetilde{R}_U P_n^{(1)}) \quad = \quad & \sqrt{n}\|A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - AVD^{-1}WR_{\widehat{U}} + AVD^{-1}WR_{\widehat{U}} - AVD^{-1}WR_{UW}P_n^{(3)} \\
& \quad + AVD^{-1}WR_{UW}P_n^{(3)} - AVD^{-1}WR^*_{UW}P_n^{(2)}P_n^{(3)} \\
& \quad + AVD^{-1}WR^*_{UW}P_n^{(2)}P_n^{(3)} - \mathscr{A}VD^{-1}WR^*_{UW}P_n^{(2)}P_n^{(3)}\| \\
\leq \quad & \sqrt{n}(\|A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - AVD^{-1}WR_{\widehat{U}}\| + \|AVD^{-1}WR_{\widehat{U}} - AVD^{-1}WR_{UW}P_n^{(3)}\| \\
& \quad + \|AVD^{-1}WR_{UW} - AVD^{-1}WR^*_{UW}P_n^{(2)}\| \\
& \quad + \|AVD^{-1}WR^*_{UW} - \mathscr{A}VD^{-1}WR^*_{UW}\|). \qquad (50)
\end{aligned}
$$

The fact that $WR^*_{UW} = \widetilde{R}_U^T P_n^{(1)}$ is a direct result of Theorem 5.1. The remaining part of the proof wants to show the bounds for each term of RHS of Equation (50).

First term of Equation (50) is $\|A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - AVD^{-1}R_{\widehat{U}}\|$. By Lemma G.5

$$\|\widehat{U} - UW\|_{2\to\infty} = O_p\left((n\rho_n)^{-1/2}n^{-1/2}\log^{\frac{5}{2}} n\right),$$

and by the same virtue (notice Y also satisfies Assumption 2, it could be shown by transposing the adjacency matrix) there exists $W_2 \in \mathscr{O}(k)$ s.t.

$$\|\widehat{V} - VW_2\|_{2\to\infty} = O_p\left((n\rho_n)^{-1/2}n^{-1/2}\log^{\frac{5}{2}} n\right).$$

By assumptions and Lemma G.4,

$$\|D^{-1}\| = O_p((n\rho_n)^{-1}), \|A - \mathscr{A}\| = O_p((n\rho_n \log^3 n)^{\frac{1}{2}}).$$

Notice that $\|X\| \leq \sqrt{m_1}\|X\|_{2\to\infty}$ for $\forall X \in \mathbb{R}^{m_1\times m_2}$ and $\|V\| = 1$. Therefore

$$
\begin{aligned}
& \|A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - AVD^{-1}WR_{\widehat{U}}\| \\
= \;& \|A\widehat{V}\widehat{D}^{-1} - AVD^{-1}W\| \\
\leq \;& \|A\|\|\widehat{V}\widehat{D}^{-1} - VD^{-1}W\| \\
= \;& \|A - \mathscr{A} + \mathscr{A}\|\|\widehat{V}\widehat{D}^{-1} - VW_2\widehat{D}^{-1} + VW_2\widehat{D}^{-1} - VD^{-1}W\| \\
\leq \;& (\|A - \mathscr{A}\| + \|\mathscr{A}\|)(\|\widehat{V}\widehat{D}^{-1} - VW_2\widehat{D}^{-1}\| + \|VW_2\widehat{D}^{-1} - VD^{-1}W\|) \\
\leq \;& (\|A - \mathscr{A}\| + \|\mathscr{A}\|)(\|\widehat{V} - VW_2\|\|\widehat{D}^{-1}\| + \|V\|\|W_2\widehat{D}^{-1} - D^{-1}W\|) \\
= \;& O_p(\|\mathscr{A}\| \times \|D^{-1}\|) \times [O_p\left((n\rho_n)^{-1/2}n^{-1/2}\log^{\frac{5}{2}} n\right) + O_p\left((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n\right)] \\
= \;& O_p\left((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n\right). \qquad (51)
\end{aligned}
$$

The third equation employs the bound of $\|W_2\widehat{D}^{-1} - D^{-1}W\|$ from the following deduction:

$$
\begin{aligned}
& \|W_2\widehat{D}^{-1} - D^{-1}W\| \\
= \;& \|\widehat{D}^{-1} - W_2^T D^{-1}W\| \\
= \;& \|\widehat{V}\widehat{D}^{-1}\widehat{U}^T - \widehat{V}W_2^T D^{-1}W\widehat{U}^T\| \\
= \;& \|\widehat{V}\widehat{D}^{-1}\widehat{U}^T - VD^{-1}U^T + (V - \widehat{V}W_2^T)D^{-1}WU + \widehat{V}W_2^T D^{-1}(U^T - W\widehat{U}^T)\| \\
\leq \;& \|\widehat{V}\widehat{D}^{-1}\widehat{U}^T - VD^{-1}U^T\| + \|(V - \widehat{V}W_2^T)D^{-1}WU\| + \|\widehat{V}W_2^T D^{-1}(U^T - W\widehat{U}^T)\| \\
\leq \;& \|A^\dagger - \mathscr{A}^\dagger\| + \sqrt{d}\|V - \widehat{V}W_2^T\|_{2\to\infty}\|D^{-1}\| + \sqrt{n}\|D^{-1}\|\|U^T - W\widehat{U}^T\|_{2\to\infty} \\
\leq \;& \|A^\dagger\|\|A - \mathscr{A}\|\|\mathscr{A}^\dagger\| + \sqrt{d}\|V - \widehat{V}W_2^T\|_{2\to\infty}\|D^{-1}\| + \sqrt{n}\|D^{-1}\|\|U^T - W\widehat{U}^T\|_{2\to\infty} \\
= \;& O_p(\|D^{-1}\|) \times \left(O_p\left((n\rho_n)^{-1/2}\log^{\frac{3}{2}} n\right) + O_p\left((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n\right)\right) \\
= \;& O_p(\|D^{-1}\|) \times O_p\left((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n\right).
\end{aligned}
$$

Second term of Equation (50) is $\|AVD^{-1}R_{\widehat{U}} - AVD^{-1}WR_{UW}P_n^{(3)}\|$. According to Equation (98) (in the proof of Proposition G.5) there exists a $P_n^{(3)} \in \mathcal{P}(k)$ s.t.

$$
\|R_{\widehat{U}} - R_{UW}P_n^{(3)}\|_{2\to\infty} = O_p\left((n\rho_n)^{-1/4}\log^{\frac{7}{4}} n\right).
$$

Therefore,

$$
\begin{aligned}
& \|AVD^{-1}WR_{\widehat{U}} - AVD^{-1}WR_{UW}P_n^{(3)}\| \\
\leq \;& \|A\|\|V\|\|D^{-1}W\|\sqrt{k}\|R_{\widehat{U}} - R_{UW}P_n^{(3)}\|_{2\to\infty} \\
= \;& O_p(1) \times \|R_{\widehat{U}} - R_{UW}P_n^{(3)}\|_{2\to\infty} \\
= \;& O_p\left((n\rho_n)^{-1/4}\log^{\frac{7}{4}} n\right). \qquad (52)
\end{aligned}
$$

Third term of Equation (50) is $\|AVD^{-1}WR_{UW} - AVD^{-1}WR^*_{UW}P^{(2)}_n\|$. Recall Proposition G.4, Theorem 5.1, there is $P^{(2)}_n \in \mathcal{P}(k)$, s.t. for any $\delta > 0$,

$$\|R_{UW} - R^*_{UW}P^{(2)}_n\|_{2\to\infty} = O_p(n^{\delta/2-1/4}).$$

Therefore,

$$\begin{aligned}\|AVD^{-1}WR_{UW} - AVD^{-1}WR^*_{UW}P^{(2)}_n\| &\leq \|A\|\|V\|\|D^{-1}W\|\sqrt{k}\|R_{UW} - R^*_{UW}P^{(2)}_n\|_{2\to\infty} \\ &= O_p(1) \times \|R_{UW} - R^*_{UW}P^{(2)}_n\|_{2\to\infty} \\ &= O_p(n^{\delta/2-1/4}). \end{aligned} \tag{53}$$

Fourth term of Equation (50) is $\|AVD^{-1}WR^*_{UW} - \mathscr{A}VD^{-1}WR^*_{UW}P^{(1)}_n\|$. Reusing Lemma G.4, we have

$$\begin{aligned}\|AVD^{-1}WR^*_{UW} - \mathscr{A}VD^{-1}WR^*_{UW}\| &= \|AVD^{-1} - \mathscr{A}VD^{-1}\| \\ &\leq \|A - \mathscr{A}\|\|V\|\|D^{-1}\| \\ &= O_p((n\rho_n)^{-\frac{1}{2}}\log^{\frac{3}{2}} n). \end{aligned} \tag{54}$$

Plugging (51), (52), (53), (54) into (50) arrives our combined bound:

$$\begin{aligned}&\|J_n(A\widehat{V}\widehat{D}^{-1}R_{\widehat{U}} - \mathscr{A}VD^{-1}\widetilde{R}_U P_n)\|_{2\to\infty} \\ =\ & O_p(\sqrt{n} \times ((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n + (n\rho_n)^{-1/4}\log^{\frac{7}{4}} n + n^{\delta/2-1/4} + (n\rho_n)^{-1/2}\log^{\frac{3}{2}} n)) \\ =\ & O_p\left(n^{\delta/2+1/4} + (n\rho_n)^{-1/4}n^{1/2}\log^{\frac{7}{4}} n\right). \end{aligned} \tag{55}$$

□

G.3.6 Proof of Proposition G.7

Proof. With Lemma G.3 and Proposition 5.2,

$$\begin{aligned}
& \|J_n(\sqrt{n}\mathscr{A}VD^{-1}\widetilde{R}_U - Z)\|_{2\to\infty} \\
\leq\ & \sqrt{n}\|\sqrt{n}\mathscr{A}VD^{-1}\widetilde{R}_U - Z\| \\
=\ & \sqrt{n}\|\sqrt{n}ZBY^TVD^{-1}\widetilde{R}_U - Z\| \\
=\ & \sqrt{n}\|ZB(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1/2}\widetilde{R}_V^T\widetilde{D}^{-1}\widetilde{R}_U - Z\| \\
=\ & \sqrt{n}\|ZB(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1}B^{-1}\widetilde{\Sigma}_Z^{-1/2} - Z\| \\
\leq\ & \sqrt{n}\|Z\|(\|B(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1}B^{-1}\widetilde{\Sigma}_Z^{-1/2} - I\|) \\
\leq\ & \sqrt{n}\|Z\|(\|B(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1}B^{-1}\widetilde{\Sigma}_Z^{-1/2} - B(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1}B^{-1}\| + \|B(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1}B^{-1} - I\|) \\
\leq\ & \sqrt{n}\|Z\|(\|B\|\|B^{-1}\|\|Y^T\widetilde{Y}/d\|\|\widetilde{\Sigma}_Y^{-1}\|\|\widetilde{\Sigma}_Z^{-1/2} - I\| + \|B\|\|B^{-1}\|\|(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1} - I\|).
\end{aligned}$$

By Lemma G.2, $\|Z\| \leq \sqrt{nk}\max|Z_{ij}| = O_p(\sqrt{n}\log n)$. Conditions in main theorem statement imply $\|B\|\|B^{-1}\| = O_p(1)$. Using LLN results (similar to proofs in Proposition G.2) there are $\|\widetilde{\Sigma}_Y^{-1}\| = O_p(1), \|\widetilde{\Sigma}_Z^{-1/2} - I\| = O_p(1/\sqrt{n})$.

Notice that the (i,j) entry of $\bar{Y}^T\widetilde{Y}/d$ is $\frac{1}{d}\widehat{\mu}_Y[i]\sum_{q=1}^n(Y_{qj} - \widehat{\mu}_Y[j]) = 0$. By LLN

$$\|Y^T\widetilde{Y}/d\| \leq \|\widetilde{Y}^T\widetilde{Y}/d\| + \|\bar{Y}^T\widetilde{Y}/d\| = O_p(1)$$

and

$$(Y^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1} - I = (\bar{Y}^T\widetilde{Y}/d)\widetilde{\Sigma}_Y^{-1},$$

is the zero matrix. Summarize these results and simplify the bounds give the desired conclusion,

$$\|J_n(\sqrt{n}\mathscr{A}VD^{-1}\widetilde{R}_U - Z)\|_{2\to\infty} = O_p(\sqrt{n}\log n).$$

□

H Technical Proofs

Proof of Lemma G.4

This part of proof needs a matrix concentration bound for sub-exponential random variables. Here we cite an existing result shown below.

Lemma H.1 (Tropp [2012])**.** *Let $X_1, X_2, ..., X_n$ be independent random $N \times N$ self-adjoint matrices. Assume that $\mathbb{E}(X_i) = 0$ for all i, and $\mathbb{E}(X_i^p) \preceq \frac{p!}{2}R^{p-2}A_i^2$ for $p \geq 2$. Compute the variance parameter*

$$\sigma^2 := \|\sum_k A_k^2\|.$$

Then for any $t > 0$,

$$\mathbb{P}(\|\sum_{i=1}^{n} X_i\| \geq t) \leq N \times \exp(-\frac{t^2}{2\sigma^2 + 2Rt}). \tag{56}$$

Now we make use of Lemma H.1 to prove Lemma G.4.

Proof. Let $E^{i,j}$ be the $(n+d) \times (n+d)$ matrix with 1 in the (i,j) and (j,i) entries and 0 elsewhere. γ_{ij}, γ are defined in Equation (35) (for simplicity we ignore the (n)-superscripts). To utilize Lemma H.1, we express $\widetilde{A}_{sym} - \widetilde{\mathscr{A}_{sym}}$ as the sum of matrices,

$$Y_{i,n+j} = (A_{ij} - \mathscr{A}_{ij})E^{i,n+j}, i = 1, ..., n, j = 1, ..., d.$$

Noice that

$$\|\widetilde{A}_{sym} - \widetilde{\mathscr{A}_{sym}}\| = \|\sum_{i=1}^{n}\sum_{j=1}^{d} Y_{i,n+j}\|,$$

and $\mathbb{E}(Y_{i,n+j}) = 0$. Moreover,

$$(E^{i,n+j})^p = E^{i,i} + E^{n+j,n+j}, p = 2, 4, ...,$$

$$(E^{i,n+j})^p = E^{i,n+j}, p = 3, 5, 7, ...,$$

and $\mathbb{E}[(A_{ij} - \mathscr{A}_{ij})^p] \leq \gamma_{ij}^p p! \leq \gamma^p p!$, for $\forall i, j, p \geq 2$. These relations indicate

$$\mathbb{E}(Y_{i,n+j}^p) \preceq \frac{p!}{2} \cdot \gamma_{ij}^{p-2} \cdot (\frac{\gamma_{ij}^2}{2}(E^{i,i} + E^{n+j,n+j})) \preceq \frac{p!}{2} \cdot \gamma^{p-2} \cdot (\frac{\gamma^2}{2}(E^{i,i} + E^{n+j,n+j})), \forall p \geq 2. \tag{57}$$

When $\bar{\rho}_n \geq 1$, we can treat A_i's in Lemma H.1 as $\frac{\gamma}{2}(E^{i,i} + E^{n+j,n+j})$ in our scenario. Therefore,

$$\begin{aligned}
\sigma^2 &= \frac{\gamma}{2}\|\sum_{i=1}^{n}\sum_{j=1}^{d}(E^{i,i} + E^{n+j,n+j})\| \\
&= \frac{\bar{\rho}_n}{4}\|\sum_{i=1}^{n}\sum_{j=1}^{d}(E^{i,i} + E^{n+j,n+j})\| \\
&= \frac{\bar{\rho}_n}{4}\|\sum_{i=1}^{n}[\sum_{j=1}^{d} E^{i,i}] + \sum_{j=1}^{d}[\sum_{i=1}^{n} E^{n+j,n+j}]\| \\
&\leq \frac{\bar{\rho}_n}{4}(\|\sum_{i=1}^{n}[\sum_{j=1}^{d} E^{i,i}]\| + \|\sum_{j=1}^{d}[\sum_{i=1}^{n} E^{n+j,n+j}]\|) \\
&= \frac{(n+d)\bar{\rho}_n}{4}.
\end{aligned}$$

When $\bar{\rho}_n \leq 1$, Assumption 3 suggests

$$\mathbb{E}[(A_{ij} - \mathscr{A}_{ij})^p] \leq (p-1)!\bar{\rho}_n \leq \frac{p!}{2} \cdot 1^{p-2} \cdot \bar{\rho}_n,$$

thus

$$\mathbb{E}(Y_{i,n+j}^p) \preceq \frac{p!}{2} \cdot (E^{i,i} + E^{n+j,n+j}), \forall p \geq 2. \tag{58}$$

Then a similar bound could be derived:

$$\sigma^2 \leq \bar{\rho}_n \| \sum_{i=1}^{n} \sum_{j=1}^{d} (E^{i,i} + E^{n+j,n+j}) \| \leq (n+d)\bar{\rho}_n.$$

Therefore, by Lemma H.1, the bound for $\|\widetilde{A}_{sym} - \widetilde{\mathscr{A}_{sym}}\|$ is obtained.

$$\mathbb{P}(\|\widetilde{A}_{sym} - \widetilde{\mathscr{A}_{sym}}\| \geq t) \leq \begin{cases} (n+d)\exp(-\frac{t^2}{\frac{(n+d)\bar{\rho}_n}{4} + 2\gamma t}) & \bar{\rho}_n \geq 1, \\ (n+d)\exp(-\frac{t^2}{(n+d)\bar{\rho}_n + 2t}) & \bar{\rho}_n < 1. \end{cases}$$

With Assumption 3 and Equation (45) this also implies,

$$\|\widetilde{A}_{sym} - \widetilde{\mathscr{A}_{sym}}\| = O((n\rho_n \log^3 n)^{\frac{1}{2}}).$$

□

Before we show the proof of Lemma G.5, we illustrate the following lemma that shows important property of matrix with special structure and could be utilized to convert bounds of eigenvectors' perturbation of symmeticed matrices to original adjacency matrices'.

Lemma H.2. *Suppose* $M = \begin{pmatrix} M_1 & M_2 \\ M_2 & M_1 \end{pmatrix}$ *is a blockwise symmetric matrix* ($M_1, M_2 \in \mathbb{R}^{k \times k}$). *Let* $M = U_M D_M V_M^T$ *be* M*'s singular vector decomposition. Then* $N = U_M V_M^T$ *has the same blockwise symmetric structure:* $N = \begin{pmatrix} N_1 & N_2 \\ N_2 & N_1 \end{pmatrix}$.

Proof. Let $M_1 + M_2 = S_1\Sigma_1 T_1^T, M_1 - M_2 = S_2\Sigma_2 T_2^T$ be the singular decompositions of them. Then

$$M_1 = \frac{1}{2}(S_1\Sigma_1 T_1^T + S_2\Sigma_2 T_2^T), M_2 = \frac{1}{2}(S_1\Sigma_1 T_1^T - S_2\Sigma_2 T_2^T).$$

Plug in the equations,

$$\begin{aligned} M &= \begin{pmatrix} M_1 & M_2 \\ M_2 & M_1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} S_1\Sigma_1 T_1^T + S_2\Sigma_2 T_2^T & S_1\Sigma_1 T_1^T - S_2\Sigma_2 T_2^T \\ S_1\Sigma_1 T_1^T - S_2\Sigma_2 T_2^T & S_1\Sigma_1 T_1^T + S_2\Sigma_2 T_2^T \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sqrt{2}}{2}S_1 & \frac{\sqrt{2}}{2}S_2 \\ \frac{\sqrt{2}}{2}S_1 & -\frac{\sqrt{2}}{2}S_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2}T_1 & \frac{\sqrt{2}}{2}T_2 \\ \frac{\sqrt{2}}{2}T_1 & -\frac{\sqrt{2}}{2}T_2 \end{pmatrix}^T. \end{aligned}$$

Then $U_M = \frac{\sqrt{2}}{2}\begin{pmatrix} S_1 & S_2 \\ S_1 & -S_2 \end{pmatrix}, V_M = \frac{\sqrt{2}}{2}\begin{pmatrix} T_1 & T_2 \\ T_1 & -T_2 \end{pmatrix}, D_M = \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix}$. We prove the result by the following observation.

$$U_M V_M^T = \frac{1}{2}\begin{pmatrix} S_1T_1^T + S_2T_2^T & S_1T_1^T - S_2T_2^T \\ S_1T_1^T - S_2T_2^T & S_1T_1^T + S_2T_2^T \end{pmatrix}.$$

□

Lemma G.4 obtains the bound of spectral norm of $\widetilde{A}_{sym} - \widetilde{\mathscr{A}_{sym}}$. Next lemma makes use of this result to show the bound for distance between eigen-spaces of $\widetilde{A}_{sym}$ and $\widetilde{\mathscr{A}_{sym}}$. Some theoretical results from Cape et al. [2019a] is borrowed to show row-wise bounds.

Proof of Lemma G.5

Proof. Let $\mu_* = \mu_r \mathbf{1}_d^T + \mathbf{1}_n \mu_c^T - \mu \mathbf{1}_n \mathbf{1}_d^T, \widehat{\mu}_* = \widehat{\mu}_r \mathbf{1}_d^T + \mathbf{1}_n \widehat{\mu}_c^T - \widehat{\mu} \mathbf{1}_n \mathbf{1}_d^T$, we symmetrize centered adjacent matrix as before: $\widetilde{A}_{sym} = \begin{pmatrix} 0 & \widetilde{A} \\ \widetilde{A}^T & 0 \end{pmatrix}$, $\widetilde{\mathscr{A}_{sym}} = \begin{pmatrix} 0 & \widetilde{\mathscr{A}} \\ \widetilde{\mathscr{A}^T} & 0 \end{pmatrix}$.

$$\begin{aligned}
\widetilde{A}_{sym} - \widetilde{\mathscr{A}_{sym}} &= \begin{pmatrix} 0 & \widetilde{A} \\ \widetilde{A}^T & 0 \end{pmatrix} - \begin{pmatrix} 0 & \widetilde{\mathscr{A}} \\ \widetilde{\mathscr{A}^T} & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & A - \widehat{\mu}_* \\ (A - \widehat{\mu}_*)^T & 0 \end{pmatrix} - \begin{pmatrix} 0 & \mathscr{A} - \mu_* \\ (\mathscr{A} - \mu_*)^T & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & A - \widehat{\mu}_* \\ (A - \widehat{\mu}_*)^T & 0 \end{pmatrix} - \begin{pmatrix} 0 & \mathscr{A} - \widehat{\mu}_* \\ (\mathscr{A} - \widehat{\mu}_*)^T & 0 \end{pmatrix} + \\
&\qquad \begin{pmatrix} 0 & \mathscr{A} - \widehat{\mu}_* \\ (\mathscr{A} - \widehat{\mu}_*)^T & 0 \end{pmatrix} - \begin{pmatrix} 0 & \mathscr{A} - \mu_* \\ (\mathscr{A} - \mu_*)^T & 0 \end{pmatrix} \\
&:= A_1 - A_2 + A_2 - A_3.
\end{aligned}$$

From Proposition 5.2 we know $\mathscr{A} - \mu_*$ is of rank k. Suppose the eigen-decomposition of A_3 is $U_3 D_3 U_3^T$. Then $U_3 \in \mathbb{R}^{(n+d)\times 2k}$. D_3 is diagonal matrix with $2k$ non-zero elements. Let $U_i \in \mathbb{R}^{(n+d)\times 2k}$ be A_i's eigenvectors corresponding to A_i's k largest and k smallest eigenvalues and D_i being diagonal matrix contains these eigenvalues, $i = 1, 2$.

Define

$$\begin{aligned}
W_{(3\to 1)} &:= \underset{W_0 \in \mathscr{O}(2k)}{\arg\min} \|U_1 - U_3 W_0\|_{2\to\infty}, \\
W_{(2\to 1)} &:= \underset{W_0 \in \mathscr{O}(2k)}{\arg\min} \|U_1 - U_2 W_0\|_{2\to\infty}, \\
W_{(3\to 2)} &:= \underset{W_0 \in \mathscr{O}(2k)}{\arg\min} \|U_2 - U_3 W_0\|_{2\to\infty},
\end{aligned}$$

then

$$\begin{aligned}\|U_1 - U_3 W_{(3\to 1)}\|_{2\to\infty} &\leq \|U_1 - U_3 W_{(3\to 2)} W_{(2\to 1)}\|_{2\to\infty} \\ &= \|U_1 - U_2 W_{(2\to 1)} + U_2 W_{(2\to 1)} - U_3 W_{(3\to 2)} W_{(2\to 1)}\|_{2\to\infty} \\ &\leq \|U_1 - U_2 W_{(2\to 1)}\|_{2\to\infty} + \|U_2 W_{(2\to 1)} - U_3 W_{(3\to 2)} W_{(2\to 1)}\|_{2\to\infty} \\ &= \|U_1 - U_2 W_{(2\to 1)}\|_{2\to\infty} + \|U_2 - U_3 W_{(3\to 2)}\|_{2\to\infty}. \end{aligned} \tag{59}$$

In the following steps, we deal with the two terms in (59) separately.

We bound $\|U_1 - U_2 W_{(2\to 1)}\|_{2\to\infty}$ by using Theorem 1 in Cape et al. [2019a]. There are four conditions of Cape's Theorem that should be evaluated. The first two conditions are followed trivially from the statements of Theorem 6.1. Let $\widetilde{\sigma}_{\max}$ and $\widetilde{\sigma}_{\min}$ represent the largest and kth largest singular values of $\widetilde{\mathscr{A}}$. Then with Equation (10), it is obvious to show

$$\widetilde{\sigma}_{\min} \geq C_1 \Delta_n = C_1 n \rho_n, \quad \widetilde{\sigma}_{\max}/\widetilde{\sigma}_{\min} \leq C_2, \tag{60}$$

for some positive constants C_1, C_2.

The following part checks the fourth condition stated below. After confirming the Cape's fourth condition, the proof will address the third condition.

Cape's 4th Condition: Write $E_1 = A_1 - A_2$. There exist constants $C_{E_1}, v > 0, \nu > 0, \xi > 1$, such that for all integers $1 \leq s \leq s(n) := \lceil \log n / \log(n\bar{\rho}_n) \rceil$, for each fixed standard basis vector e_i and any fixed unit vector u, with probability at least $1 - \exp(-\nu \log^{\xi} n)$ (provided $n \geq n_0(C_E, \nu, \xi)$)),

$$|\langle E_1^s u, e_i \rangle| \leq C_{E_1}^s (n\bar{\rho}_n)^{s/2} (\log n)^{s\xi} \|u\|_{\infty}. \tag{61}$$

Using an argument in Lemma 7.10 of Erdős et al. [2013], Mao et al. [2017] shows that the following Upper Bound Condition is sufficient for Cape's 4th Condition. That appears as Lemma 5.5 of Mao et al. [2017]. The key idea to show Cape's 4th assumption is applying the inequality of Upper Bound Condition to upper bound the number of non-zero tems in the summation via a multigraph construction for paths counting. The difference is that our main theorem allows sub-exponential random variables, which is possibly unbounded. Thus, the only thing that needs to be checked is the following upper bound condition, with the order of magnitude m ranges from 2 to the number of vertices in constructed multigraph.

Upper Bound Condition Let $H = \frac{A - \mathscr{A}}{\sqrt{n\bar{\rho}_n}}$ and H_{ij} represents its element on (i,j)th entry. Then there exists a positive constant C_{E_1}, such that eventually in n,

$$\mathbb{E}(|H_{ij}|^m) \leq \frac{C_{E_1}}{n}, \quad \forall\ 2 \leq m \leq \log^{\xi} n. \tag{62}$$

For any positive even number $m \leq \log^{\xi} n$. If $\bar{\rho}_n \leq 1$, by Assumption 3,

$$\mathbb{E}(|H_{ij}|^m) = \mathbb{E}(H_{ij}^m) \leq \frac{(m-1)!\bar{\rho}_n}{(n\bar{\rho}_n)^{m/2}} = \frac{(m-1)!}{(n\bar{\rho}_n)^{m/2-1}} \times \frac{1}{n},$$

therefore, Equation (62) is true for $m = 2$ when choosing $C_{E_1} = 1$. When $m \geq 4$, we want to show

$$(m-1)! \leq (n\bar{\rho}_n)^{m/2-1}. \tag{63}$$

For any positive integer N, we have $N! \leq N^{N+1/2}\exp\{-N+1\}$. Thus $(m-1)! \leq (m-1)^{m-1/2}\exp\{2-m\} \leq m^m \exp\{2-m\}$. Then to show Equation (63) we only need to show

$$m\log m + 2 - m \leq (\frac{m}{2} - 1)\log(n\bar{\rho}_n). \tag{64}$$

Denote $F(m) = m\log m + 2 - m - (\frac{m}{2} - 1)\log(n\bar{\rho}_n)$. Then $dF(m)/dm$ equals to $\log m - \log(n\bar{\rho}_n)/2$. Since $m \leq \log^{\xi} n$ and $n\bar{\rho}_n \geq n\rho_n \geq C_0 \log^{10} n$ for a positive C_0. $F(m)$ obtains its maximum at $m = 4$ with proper selection of ξ and large enough n. Equation (64) will be trivial with $m = 4$ and large enough n.

If $\bar{\rho}_n \geq 1$, by Assumption 3,

$$\mathbb{E}(|H_{ij}|^m) = \mathbb{E}(H_{ij}^m) \leq \frac{(m-1)!\bar{\rho}_n^{m/2}}{(n\bar{\rho}_n)^{m/2}} = \frac{(m-1)!}{n^{m/2}}.$$

The target boils down to $(m-1)! \leq C_{E_1} n^{m/2-1}$. Again, since for all positive integers N, there is $N! \leq N^{N+1/2}\exp\{-N+1\}$. Then,

$$\begin{aligned}
(m-1)! \leq n^{m/2-1} &\Leftarrow (m-1)^{m-1/2}\exp\{-m+2\} \leq n^{m/2-1}C_{E_1} \\
&\Leftrightarrow (m-\frac{1}{2})\log(m-1) + 2 - m \leq (\frac{m}{2}-1)\log n + \log(C_{E_1}) \\
(\text{since } \log n^{\xi} \geq m) &\Leftarrow (m-\frac{1}{2})\log(m-1) + 2 - m \leq (\frac{m}{2}-1)m^{1/\xi} + \log(C_{E_1}). \quad (65)
\end{aligned}$$

Since $1/\xi > 0$, there exists a positive integer M, such that $(\frac{m}{2}-1)m^{1/\xi} > (m-\frac{1}{2})\log(m-1)+2-m$ for all integer $m > M$. Choose C_{E_1} such that $\log(C_{E_1}) > (m-\frac{1}{2})\log(m-1)+2-m$ for all integer $2 \leq m \leq M$. Then Equation (65) is proved.

Hence the upper bound condition holds for even m. For odd number $m \geq 3$, by Cauchy-Schwarz inequality,

$$(\mathbb{E}(|H_{ij}|^m))^2 \leq \mathbb{E}(|H_{ij}|^{m-1})\mathbb{E}(|H_{ij}|^{m+1}) \leq \frac{1}{n^2}.$$

Thus the upper bound condition holds for all integer $m \geq 2$ and therefore Cape's fourth condition is valid.

We claim that Cape's third condition could be relaxed from $\|E\| = O_p((n\rho_n)^{\frac{1}{2}})$ to $\|E\| = O_p((n\rho_n \log^3 n)^{\frac{1}{2}})$ as inferred from Lemma G.4, with only slight modifications in Cape's converge rate result. In the proof of Theorem 1 of Cape et al. [2019a], the bound of LHS of (5) comes from three quantities: $\|E\widehat{U}\widehat{\Lambda}^{-1}\|_{2\to\infty}, \|R^{(1)}\|_{2\to\infty}, \|R^{(2)}_W\|_{2\to\infty}$ (these three terms' notations are from Cape et al. [2019a]). Our relaxation of $\|E\|$ adds an extra $\log^{\frac{3}{2}} n$ term to $\|R^{(1)}\|_{2\to\infty}$'s bound, while the third term remains the same because of Cape's 4th assumption (we have already checked). Therefore by Theorem 1 of Cape et al. [2019a], we arrives the following conclusion.

$$\begin{aligned}\|U_1 - U_2 W_{(2\to1)}\|_{2\to\infty} &= O_p\left(((n\bar{\rho}_n)^{-1/2}\log^{\xi} n + (n\rho_n)^{-1/2}\log^{\frac{3}{2}} n) \times \|U_2\|_{2\to\infty}\right) \\ &= O_p\left((n\rho_n)^{-1/2}\log^{\frac{3}{2}} n \times \|U_2\|_{2\to\infty}\right). \end{aligned} \tag{66}$$

To bound $\|U_2 - U_3 W_{(3\to2)}\|_{2\to\infty}$, we employ Theorem 4.2 in Cape et al. [2019b].

Theorem H.1. *[Theorem 4.2 in Cape et al. [2019b]] Suppose the diagonal elements of D_3 are sorted in descending order. If $|D_3[k]| > 4\|E_2\|_\infty$, where $E_2 = A_2 - A_3$, $D_3[j]$ is the j-th diagonal element of D_3. Then there exists $W_3 \in \mathscr{O}(k)$ such that*

$$\|U_2 - U_3 W_{(3\to2)}\|_{\max} \le 14(\frac{\|E_2\|_\infty}{|D_3[k]|})\|U_3\|_{2\to\infty}. \tag{67}$$

Before applying Theorem H.1, we should check its assumptions. Reuse the notation for the singular values of $\widetilde{\mathscr{A}}$ as in Equation (60). Notice $\widetilde{\sigma}_{\min} \succeq c_1 n\rho_n$ and $\mu_* = \mu_r \mathbf{1}_d^T + \mathbf{1}_n\mu_c^T - \mu\mathbf{1}_n\mathbf{1}_d^T$. Denotes $\mu_r = \mathscr{A}\mathbf{1}_d/d := T/d$. Similarly define $\widehat{T} := A\mathbf{1}_d$. By Hoeffding's Concentration Inequality,

$$\mathbb{P}(|\widehat{T}_i - T_i| \ge a) \le 2\exp\{-\frac{a^2}{2\sum_{j=1}^d var(A_{ij})}\} \le 2\exp\{-\frac{a^2}{d\bar{\rho}_n}\}.$$

Therefore, $\widehat{T}_i - T_i = O_p((n\bar{\rho}_n)^{\frac{1}{2}})$. The same bound applies to $\|\widehat{\mu}_c - \mu_c\|, \|\widehat{\mu} - \mu\|$. These imply that any entry of E_2 could be bounded by $O_p(\sqrt{n\bar{\rho}_n}/n)$. Thus $\|E_2\|_\infty = O_p((n\rho_n \log^2 n)^{\frac{1}{2}})$. Then with large enough n, there must be $|D_3[k]| > 4\|E_2\|_\infty$.

With Theorem H.1,

$$\|U_2 - U_3 W_{(3\to2)}\|_{2\to\infty} \le \sqrt{k}\|U_2 - U_3 W_{(3\to2)}\|_{\max} = O_p((n\rho_n)^{-\frac{1}{2}}\log n\|U_3\|_{2\to\infty}). \tag{68}$$

Combining (59), (66), (68) gives

$$\begin{aligned}
& \|U_1 - U_3 W_{(3\to1)}\|_{2\to\infty} \\
\leq\ & \|U_1 - U_2 W_{(2\to1)}\|_{2\to\infty} + \|U_2 - U_3 W_{(3\to2)}\|_{2\to\infty} \\
=\ & O_p\left((n\rho_n)^{-1/2}\log^{\frac{3}{2}} n \times \|U_2\|_{2\to\infty} + (n\rho_n)^{-\frac{1}{2}}\log n \|U_3\|_{2\to\infty}\right). \\
\preceq\ & O_p\left((n\rho_n)^{-1/2}\log^{\frac{3}{2}} n \times (\|U_2 - U_3 W_{(3\to2)}\|_{2\to\infty} + \|U_3\|_{2\to\infty}) + (n\rho_n)^{-\frac{1}{2}}\log n \|U_3\|_{2\to\infty}\right) \\
=\ & O_p\left((n\rho_n)^{-1/2}\log^{\frac{3}{2}} n \times \|U_3\|_{2\to\infty}\right). \qquad (69)
\end{aligned}$$

The next step is to convert the symmetrized adjacency matrices' eigenvectors' perturbation bound to that of original adjacency matrices. Suppose the k-rank singular value decompositions (only retains the top k singular values and corresponding eigenvectors) of $A - \widehat{\mu}_*, \mathscr{A} - \widehat{\mu}_*$ and $\mathscr{A} - \mu_*$ are $F_1\Lambda_1 L_1^T, F_2\Lambda_2 L_2^T, F_3\Lambda_3 L_3^T$ repectively. Then, for $i = 1, 2, 3$ there must be

$$U_i = \frac{1}{\sqrt{2}}\begin{pmatrix} F_i & -F_i \\ L_i & L_i \end{pmatrix}.$$

Suppose

$$W_{(3\to1)} = \begin{pmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{pmatrix}$$

with each sub-matrix having $k \times k$ dimension.

Arguments in Cape et al. [2019a] (proof of Theorem 1) implies that if we have singular value decomposition $U_3^T U_1 = U_o D_o V_o^T$, then $W_{(3\to1)} = U_o V_o^T$. It is trivial to see that $U_3^T U_1$ has special block-wise structure

$$U_3^T U_1 = \frac{1}{2}\begin{pmatrix} F_3^T F_1 + L_3^T L_1 & -F_3^T F_1 + L_3^T L_1 \\ -F_3^T F_1 + L_3^T L_1 & F_3^T F_1 + L_3^T L_1 \end{pmatrix}.$$

Thus Lemma H.2 indicates

$$W^{11} = W^{22}, \quad W^{12} = W^{21}. \qquad (70)$$

Notice $W_{(3\to1)}$ is orthogonal matrix, therefore

$$\begin{aligned}
W_{(3\to1)}W_{(3\to1)}^T = I \ &\Rightarrow\ \begin{pmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{pmatrix}\begin{pmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{pmatrix}^T = I \\
&\Rightarrow\ \begin{pmatrix} W^{11} & W^{12} \\ W^{12} & W^{11} \end{pmatrix}\begin{pmatrix} W^{11} & W^{12} \\ W^{12} & W^{11} \end{pmatrix}^T = I \\
&\Rightarrow\ W^{11}W^{11T} + W^{12}W^{12T} = I, \quad W^{11}W^{12T} + W^{12}W^{11T} = 0, \\
\text{(Equation (70))} \ &\Rightarrow\ (W^{21} - W^{22})(W^{21} - W^{22})^T = I, \quad (W^{11} - W^{12})(W^{11} - W^{12})^T = I.
\end{aligned}$$

These indicate $W^{21} - W^{22}$ and $W^{11} - W^{12}$ are both orthogonal matrices.

Notice $\|U\|_{2\to\infty} = O_p(\log n/\sqrt{n})$ from Equation (47), similarly there is same upper bound for $\|V\|_{2\to\infty}$. Therefore

$$\begin{aligned}
\inf_{W\in\mathscr{O}(k)} \|\widehat{U} - UW\|_{2\to\infty} &= \inf_{R\in\mathscr{O}(k)} \|F_1 - F_3 R\|_{2\to\infty} \\
&\leq \max\{\|F_1 - F_3(W^{11} - W^{21})\|_{2\to\infty}, \|F_1 - F_3(W^{12} - W^{22})\|_{2\to\infty}\} \\
&\leq \| \begin{pmatrix} F_1 & , -F_1 \end{pmatrix} - \begin{pmatrix} F_3 & , -F_3 \end{pmatrix} W_{(3\to 1)}\|_{2\to\infty} \\
&\leq \|U_1 - U_3 W_{(3\to 1)}\|_{2\to\infty} \\
&= O_p\left((n\rho_n)^{-1/2} \log^{\frac{3}{2}} n \times \|U_3\|_{2\to\infty}\right) \\
&= O_p\left((n\rho_n)^{-1/2} \log^{\frac{3}{2}} n \times (\|U\|_{2\to\infty} + \|V\|_{2\to\infty})\right) \\
&= O_p\left((n\rho_n)^{-1/2} n^{-1/2} \log^{\frac{5}{2}} n\right).
\end{aligned}$$

□

Proof for requisite results of Lemma G.6

The proof of Lemma G.6 is decomposed into Lemmas H.3, H.4, H.5, H.6, and H.7. All of which are stated below.

Lemmas H.3, H.4, H.5 bound the tail behavior of the sample Varimax objective function. Lemmas H.6 and H.7 bound the difference between the sample and population versions of the Varimax objective function, uniformly over the space of orthogonal matrices. Lemma G.6 puts these pieces together with the first and second order conditions for Varimax described in Section I to show that the optimum of the sample Varimax objective function must be close to the optimum of the population Varimax function (modulo permutation and sign-flip).

The next few lines show the existence of moment generating function (MGF) of linear inner-product of sub-exponential random vectors. Recall $Z_i \in \mathbb{R}^k$ contains sub-exponential random variables. Following the notations in the proof of Lemma G.2 (Equation (44)) the tail property of Z could be shown as

$$\mathbb{P}(Z_{ij} - \mathbb{E}Z_{ij} > t) \leq C_0 \exp(-\lambda t), \tag{71}$$

then for $\forall r \in \mathbb{R}^k$, by independency

$$\mathbb{E}\exp(t\langle \widetilde{Z}_i, r\rangle) = \mathbb{E}\exp(t\sum_{j=1}^{k} \widetilde{Z}_{ij} r_j) = \Pi_{j=1}^{k} \mathbb{E}\exp(t\widetilde{Z}_{ij} r_j),$$

$\Rightarrow \langle \widetilde{Z}_i, r\rangle$ also has MGF.

Lemma H.3. *$\widetilde{Z}_i \in \mathbb{R}^k$ is i-th row of $\widetilde{Z}$. $r \in \mathbb{R}^k$ is arbitrary. λ is defined in Equation (71). Let*

$$J_i = \langle \widetilde{Z}_i, r\rangle^4 - \mathbb{E}[\langle \widetilde{Z}_i, r\rangle^4],$$

then for $\forall t > 0$, there exists a positive constant C_1 s.t.

$$\mathbb{P}(J_i > t) \leq C_1 \exp(-\lambda t^{\frac{1}{4}}). \tag{72}$$

Proof. Write $X_i = \langle \widetilde{Z}_i, r\rangle$. By Markov Inequality,

$$\begin{aligned}
\mathbb{P}(J_i > t) &= \mathbb{P}(X_i^4 - \mathbb{E}(X_i^4) > t) \\
&= \mathbb{P}(X_i^4 > t + \mathbb{E}(X_i^4)) \\
&= \mathbb{P}(X_i - \mathbb{E}X_i > (t + \mathbb{E}(X_i^4))^{\frac{1}{4}} - \mathbb{E}X_i) \\
&\leq C_0' \exp(-\lambda((t + \mathbb{E}[X_i^4])^{\frac{1}{4}} - \mathbb{E}X_i)) \\
&\leq C_0' \exp(\lambda \mathbb{E}X_i) \exp(-\lambda t^{\frac{1}{4}}) \\
&:= C_1 \exp(-\lambda t^{\frac{1}{4}}).
\end{aligned}$$

□

The following lemma makes use of Lemma H.3 and gives bound to the sum of sequence $|\sum_{i=1}^n J_i|$.

Lemma H.4. *With previous definitions, then for any $\delta > 0$,*

$$|\sum_{i=1}^n J_i| = O_p(n^{\frac{1}{2}+\delta}).$$

Proof. For sequence $\alpha_n \uparrow \infty$. Define $\widetilde{J}_i = J_i \mathbf{1}(J_i \leq \alpha_n)$. Then $\widetilde{J}_i$'s are independent bounded random variables. Let $\mathfrak{A} = \{\bigcap_{i=1}^n \{J_i = \widetilde{J}_i\}\}$ and $\mathfrak{B} = \{|\sum_{i=1}^n J_i| > t\}$. Then

$$\begin{aligned}
\mathbb{P}(\mathfrak{B}) &= \mathbb{P}(\mathfrak{B} \cap \mathfrak{A}) + \mathbb{P}(\mathfrak{B} \cap \mathfrak{A}^c) \\
&\leq \mathbb{P}(\{|\sum_{i=1}^n \widetilde{J}_i| > t\}) + \mathbb{P}(\mathfrak{A}^c).
\end{aligned} \tag{73}$$

Notice $\widetilde{J}_i \in [-c, \alpha_n]$ with $c = -\mathbb{E}(X_1^4)$ is a bounded random variable. Therefore it is sub-gaussian with domain interval length $\sigma \leq \alpha_n + c \leq 2\alpha_n$ for large enough n. By Hoeffding Concentration Inequality,

$$\mathbb{P}(\{|\sum_{i=1}^n \widetilde{J}_i| > t\}) \leq 2\exp(-\frac{2t^2}{\sum_i \sigma^2}) \leq 2\exp(-\frac{t^2}{2n\alpha_n^2}). \tag{74}$$

By Lemma H.3 there is,

$$\begin{aligned}\mathbb{P}(\mathfrak{A}^c) &= \mathbb{P}(\{\cap_i\{J_i \le \alpha_n\}\}^c)\\ &= \mathbb{P}(\cup_i\{J_i > \alpha_n\})\\ &\le n\mathbb{P}(J_i > \alpha_n)\\ &\le nC_2\exp(-\lambda\alpha_n^{\frac{1}{4}}). \end{aligned} \tag{75}$$

Plugging (74)(75) in (73) and choosing $\varepsilon > \delta, t = n^{1/2+\varepsilon}, \alpha_n = n^\delta$ gives

$$|\sum_{i=1}^n J_i| = O_p(n^{\frac{1}{2}+\delta}). \tag{76}$$

□

Similar conclusion applies to second moment terms.

Lemma H.5. *With same notations as Lemma H.3. Define* $Y_i = \langle\widetilde{Z}_i, r\rangle^2 - \mathbb{E}[\langle\widetilde{Z}_i, r\rangle^2]$*, then*

$$|\sum_{i=1}^n Y_i| = O_p(n^{\frac{1}{2}+\delta}). \tag{77}$$

Proof. This part employs the same strategy as the proof of Lemma H.4. The only difference is the bound of (75). But the dominating bound (74) is the same. Thus we obtain the similar bound for $|\sum_{i=1}^n Y_i|$. □

Lemma H.6. *Suppose* $r_1, r_2, ...r_{n_0} \in \mathbb{R}^k$*. Denote*

$$\mathbb{J}_\ell = |\sum_{i=1}^n \langle\widetilde{Z}_i, r_\ell\rangle^4 - n\mathbb{E}[\langle\widetilde{Z}_i, r_\ell\rangle^4]|.$$

Assume $n_0 = an^b$ *with some positive constant* a, b*. Then for any* $\delta > 0$*,*

$$\max_\ell \mathbb{J}_\ell = O_p(n^{\frac{1}{2}+\delta}). \tag{78}$$

Similarly if we define

$$\mathbb{Y}_\ell = \langle\widetilde{Z}_i, r_\ell\rangle^2 - \mathbb{E}[\langle\widetilde{Z}_i, r_\ell\rangle^2],$$

we have

$$\max_\ell \mathbb{Y}_\ell = O_p(n^{\frac{1}{2}+\delta}). \tag{79}$$

Proof. Take $\varepsilon > \delta, t = n^{1/2+\varepsilon}, \alpha_n = n^{\delta}$. Applying the same strategy in Lemma H.4 (Equation (74), (75)) and basic probability rules,

$$\begin{aligned}\mathbb{P}(\max_{\ell} \mathbb{J}_\ell > t) &\leq 2n_0 \exp(-\frac{t^2}{2n\alpha_n^2}) + n_0 n C_2 \exp(-\lambda \alpha_n^{\frac{1}{4}}) \\ &= 2n_0 \exp(-\frac{n^{1+2\varepsilon}}{2n^{1+2\delta}}) + n_0 n C_2 \exp(-\lambda n^{\delta/4})\end{aligned}$$

Since

$$\log(2n_0) - \frac{1}{2}n^{2\varepsilon-2\delta} = \log 2a + b\log n - n^{2\varepsilon-2\delta} \to -\infty,$$

$$\log(C_2 n_0 n) - \lambda n^{\delta/4} = \log C_2 \log a + (b+1)\log n - \lambda n^{\delta/4} \to -\infty,$$

they could be reduced to $\max_\ell \mathbb{J}_\ell = O_p(n^{\frac{1}{2}+\delta})$. Similar approach gives $\max_\ell \mathbb{Y}_\ell = O_p(n^{\frac{1}{2}+\delta})$. □

Lemma H.7. *Recall notations in equation* (2)*, Section 2.1 and Section 1. For readability we slightly abuse the notations and write* $\widehat{V}(R) = v(R, \widetilde{Z})$*. And*

$$V(R) := \mathcal{V}_{\widetilde{U}}(R) = \sum_{j=1}^{k} Var([\widetilde{Z}_i \widetilde{U} R]_j).$$

Then for $\forall \delta > 0$*, there is a uniform bound between these two quantities,*

$$\sup_{O \in \mathscr{O}(k)} |\widehat{V}(O) - V(O)| = O_p(n^{\delta - 1/2}). \tag{80}$$

Proof. This part of the proof adapts the covering balls strategy to give this uniform bound. Let $\mathbb{R} = \{R_1, R_2, ..., R_N\}$ be a ε-cover for orthogonal matrices. This means for $\forall O \in \mathcal{O}(k)$, there exists ℓ s.t. $d(O, R_\ell) < \varepsilon$ where $d(X, Y)$ is the $\sin\Theta$ distance. Let $D(\varepsilon, \mathcal{O}(k), d)$ be the ε-packing number. Using the notes in Van de Geer [2000] and Lemma 4.1 in Pollard [1990] we have,

$$N \leq D(\varepsilon, \mathcal{O}(k), d) \leq D(\varepsilon, \mathcal{O}(k), d_F) \leq (\frac{6}{\varepsilon})^{k^2} := N_0. \tag{81}$$

d_F is Frobenius norm distance. The second inequality is true because for any $O_1, O_2 \in \mathcal{O}(k)$,

$$\|\sin(O_1, O_2)\|_F^2 \leq \inf_{Q \in \mathcal{O}(k)} \|O_1 - O_2 Q\|_F^2 \leq \|O_1 - O_2\|_F^2$$

Let $\varepsilon = 1/n$ then $N \leq N_0 = (6n)^{k^2}$. For $\forall O \in \mathcal{O}(k)$, choose $R_\ell \in \mathbb{R}$ such that $d(O, R_\ell) < \varepsilon$. Then by Triangle Inequality,

$$|\widehat{V}(O) - V(O)| \leq |\widehat{V}(O) - \widehat{V}(R_\ell)| + |\widehat{V}(R_\ell) - V(R_\ell)| + |V(O) - V(R_\ell)|. \tag{82}$$

Lemma H.4, H.6 indicates

$$|\widehat{V}(R_\ell) - V(R_\ell)| \leq \max_j |\widehat{V}(R_j) - V(R_j)| = O_p(n^{\delta - \frac{1}{2}}). \tag{83}$$

Notice $\widehat{V}(R) = \sum_{j=1}^k \left(\frac{1}{n} \sum_{i=1}^n [\widetilde{Z}R]_{ij}^4 - \left(\frac{1}{n} \sum_{i=1}^n [\widetilde{Z}R]_{ij}^2 \right)^2 \right)$ and $\max_{ij} |\widetilde{Z}_{ij}| = O(\log n)$ (Lemma G.2). Also for $\forall O, R \in \mathcal{O}(k)$ such that $d(O, R) < \varepsilon$, there is fact that

$$d(O, R) \geq \frac{1}{\sqrt{2}} \|O - R\|_F, \forall O, R \in \mathscr{O}(k),$$

then

$$\begin{aligned}
|\sum_{ij} ([\widetilde{Z}O]_{ij}^4 - [\widetilde{Z}R]_{ij}^4)| &= |\sum_{ij} ([\widetilde{Z}O]_{ij}^2 + [\widetilde{Z}R]_{ij}^2)([\widetilde{Z}O]_{ij} + [\widetilde{Z}R]_{ij})([\widetilde{Z}O]_{ij} - [\widetilde{Z}R]_{ij})| \\
&\leq \sum_{ij} |([\widetilde{Z}O]_{ij}^2 + [\widetilde{Z}R]_{ij}^2)([\widetilde{Z}O]_{ij} + [\widetilde{Z}R]_{ij})([ZO]_{ij} - [\widetilde{Z}R]_{ij})| \\
&\leq O_p(\log^2 n) \sum_{ij} |([\widetilde{Z}O]_{ij} + [\widetilde{Z}R]_{ij})([\widetilde{Z}O]_{ij} - [\widetilde{Z}R]_{ij})| \\
&\leq O_p(\log^3 n) \sum_{ij} |[\widetilde{Z}O]_{ij} - [\widetilde{Z}R]_{ij}| \\
&\leq O_p(n \log^3 n) \|\widetilde{Z}(O - R)\|_{1 \to \infty} \\
&\leq O_p(n \log^3 n) \|\widetilde{Z}\|_{\max} (\sum_{ij} |O_{ij} - R_{ij}|) \\
&= O_p(n \log^4 n) (\sum_{ij} |O_{ij} - R_{ij}|) \\
&\leq O_p(n \log^4 n) \times k \sqrt{\sum_{ij} |O_{ij} - R_{ij}|^2} \\
&= O_p(n \log^4 n) \|O - R\|_F \\
&\leq O_p(n \log^4 n) \times \sqrt{2} d(O, R) \\
&= O_p(\varepsilon n \log^4 n) \\
&= O_p(\log^4 n),
\end{aligned} \tag{84}$$

and

$$
\begin{aligned}
& |\sum_j [(\sum_i [\widetilde{Z}O]_{ij}^2)^2 - (\sum_i [\widetilde{Z}R]_{ij}^2)^2]| \\
= \;& |\sum_j [(\sum_i ([\widetilde{Z}O]_{ij}^2 + [\widetilde{Z}R]_{ij}^2))(\sum_i ([\widetilde{Z}O]_{ij}^2 - [\widetilde{Z}R]_{ij}^2))]| \\
\leq \;& \sum_j |(\sum_i ([\widetilde{Z}O]_{ij}^2 + [\widetilde{Z}R]_{ij}^2))(\sum_i ([\widetilde{Z}O]_{ij}^2 - [\widetilde{Z}R]_{ij}^2))| \\
\leq \;& \sum_j |(2n\|\widetilde{Z}\|_{\max}^2)(\sum_i ([\widetilde{Z}O]_{ij}^2 - [\widetilde{Z}R]_{ij}^2))| \\
\leq \;& O_p(n\log^2 n)\sum_j |\sum_i ([\widetilde{Z}O]_{ij} + [\widetilde{Z}R]_{ij})([\widetilde{Z}O]_{ij} - [\widetilde{Z}R]_{ij})| \\
\leq \;& O_p(n\log^3 n)\sum_j |\sum_i [\widetilde{Z}O]_{ij} - [\widetilde{Z}R]_{ij}| \\
\leq \;& O_p(n^2\log^3 n)\|\widetilde{Z}(O-R)\|_{1\to\infty} \\
\leq \;& O_p(n^2\log^3 n)\|\widetilde{Z}\|_{\max}(\sum_{ij}|O_{ij} - R_{ij}|) \\
\leq \;& O_p(n^2\log^4 n) \times k\|O-R\|_F \\
\leq \;& O_p(n^2\log^4 n) \times \sqrt{2} \times d(O,R) \\
= \;& O_p(n^2\varepsilon\log^4 n) \\
= \;& O_p(n\log^4 n). \qquad (85)
\end{aligned}
$$

With Equation (84), (85),

$$
\begin{aligned}
& |\widehat{V}(O) - \widehat{V}(R_\ell)| \\
\leq \;& \frac{1}{n}|\sum_{ij}([\widetilde{Z}O]_{ij}^4 - [\widetilde{Z}R_\ell]_{ij}^4|) + \frac{1}{n^2}|\sum_j [(\sum_i [\widetilde{Z}O]_{ij}^2)^2 - (\sum_i [\widetilde{Z}R_\ell]_{ij}^2)^2]| \\
= \;& O_p(\frac{\log^4 n}{n}).
\end{aligned}
$$

Assume $\mathbb{E}(\widetilde{Z}_{1\ell}^j) = \mu_j^{(\ell)}$ refers to j-th moment of Z's ℓ-th column, similar to the proof of Theorem 5.1 the population Varimax function could be expressed as,

$$V(Q) = \sum_j (\mathbb{E}([\widetilde{Z}Q]_j^4) - \mathbb{E}([\widetilde{Z}Q]_j^2)^2) = \sum_{i=1}^k (\mu_4^{(i)} - 3)\|Q_{i\cdot}\|_4^4 + 3k.$$

Write $\xi_i = \mu_4^{(i)} - 3\mu_2^{(i)2} = \mu_4^{(i)} - 3$ (is positive by leptokurtic assumption) and $\xi_0 = \max_i \xi_i$ (is a finite positive constant). For $\forall O, R \in \mathcal{O}(k)$ such that $d(O, R) < \varepsilon$, there is

$$\begin{aligned}
|V(O) - V(R)| &= |\sum_{i=1}^k \xi_i(\|O_{i\cdot}\|_4^4 - \|R_{i\cdot}\|_4^4)| \\
&\leq \sum_{i=1}^k |\xi_i(\|O_{i\cdot}\|_4^4 - \|R_{i\cdot}\|_4^4)| \\
&\leq \xi_0 \sum_{i=1}^k |\|O_{i\cdot}\|_4^4 - \|R_{i\cdot}\|_4^4| \\
&= \xi_0 \sum_{i=1}^k |\sum_{j=1}^k (O_{ij}^2 + R_{ij}^2)(O_{ij} + R_{ij})(O_{ij} - R_{ij})| \\
&\leq \xi_0 \sum_{i=1}^k \sum_{j=1}^k |(O_{ij}^2 + R_{ij}^2)(O_{ij} + R_{ij})(O_{ij} - R_{ij})| \\
&\leq 4\xi_0 \sum_{i=1}^k \sum_{j=1}^k |O_{ij} - R_{ij}| \\
&\leq 4\xi_0 k \|O - R\|_F \\
&= O_p(\frac{1}{n}).
\end{aligned}$$

Summing up three bounds of Equation (82) obtains a uniform bound of $|\widehat{V}(O) - V(O)|$:

$$\sup_{O \in \mathcal{O}(k)} |\widehat{V}(O) - V(O)| = O_p(n^{\delta - \frac{1}{2}}). \tag{86}$$

□

Proof of Lemma G.6

This part of proof needs first/second order condition of population varimax function here. These two conditions are stated as Corollary H.1, H.2 below. For now the notations of sample & population varimax functions follow the proofs of Lemma H.7.

Corollary H.1 (FOC)**.** *If identity matrix I is a stationary point of $\mathcal{V}_I(R_0)$ then Z satisfies the following condition,*

$$\mathbb{E}Z_{1i}^2\mathbb{E}(Z_{1i}Z_{1j}) - \mathbb{E}(Z_{1i}^3Z_{1j}) = \mathbb{E}Z_{1j}^2\mathbb{E}(Z_{1i}Z_{1j}) - \mathbb{E}(Z_{1j}^3Z_{1i}), \forall i \neq j. \tag{87}$$

Corollary H.2 (SOC)**.** *Notate $O = \widetilde{Z}_1^T\widetilde{Z}_1$, if I is a local maxima of the population Varimax $\mathcal{V}_I(R_0)$, then the following condition is true,*

$$3\mathbb{E}[tr(diag(OK)^2)] \leq \mathbb{E}\langle OdiagO, KK^T\rangle, \tag{88}$$

for any skew-symmetric matrix K.

The proof of Corollary H.1 is trivial. Corollary H.2 is a direct result of Theorem I.2 in Section I. Now we could proceed to the proof of Lemma G.6.

Proof. By Proposition G.4, $R_{\widetilde{Z}}$ is converging to elements of $\mathcal{P}(k)$, WLOG we may assume $P_n^{(2)} = I$ and $R_{\widetilde{Z}} \to I$ (i.e. let $P_n^{(2)} = \widetilde{R}_U^T$), since elements of $\mathcal{P}(k)$ are isolated to each other ($\forall P_1 \neq P_2 \in \mathcal{P}(k)$, $\|P_1 - P_2\| \geq 2/\sqrt{k}$). We may constrain our analysis on a fixed neighborhood of I, $B(I, \delta_c)$ s.t. $\{I\} = B(I, \delta_c) \cap \mathcal{P}(k)$. Now we want to show,

$$\|R_{\widetilde{Z}} - I\|_{2\to\infty} = O_p(n^{\delta/2-1/4}).$$

By Lie algebra theory there is a $k \times k$ skew-symmetric matrix K s.t. $R_{\widetilde{Z}} = \exp(K)$. Define

$$\gamma(t) = \exp(tK).$$

Then $\gamma(0) = I$, $\gamma(1) = R_{\widetilde{Z}}$. We want to evaluate population varimax function's first order and second order condition at I (global optimal solution). To achieve that we should show: $R \to I \Rightarrow K \to 0$. This could be proved by using matrix logarithm algebra.

$$\|K\| = \|\log R_{\widetilde{Z}} - \log I\| \leq \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i}\|R_{\widetilde{Z}} - I\|^i \to 0.$$

Differential calculations indicates:

$$\frac{d}{dt}V(\gamma(t))|_{t=0} = \nabla V(\gamma(t))^T\frac{d\gamma}{dt}|_{t=0} = \langle \nabla V(I), K\rangle, \tag{89}$$

$$\begin{aligned}\frac{d^2}{dt^2}V(\gamma(t))|_{t=0} &= \langle\nabla^2 V(\frac{d\gamma(t)}{dt})|_{t=0}, K\gamma(t)\rangle + \langle\nabla V, K^2\gamma(0)\rangle \\ &= \langle\nabla^2 V\cdot K, K\rangle + \langle\nabla V, K^2\rangle.\end{aligned}$$

Equation (2) of Chu and Trendafilov [1998] is a reformulated version of Varimax function. Taking expectation of it (Lemma I.1 allows exchanging differential and expectation), and notate $E = \widetilde{Z}^T\widetilde{Z} - n\widetilde{Z}_1^T\widetilde{Z}_1$. The varimax function could be rewritten as:

$$V(Q) := \mathbb{E}[trace(diag(Q^TEQ)^2)] \Rightarrow \nabla V(Q) = 4\mathbb{E}[EQdiag(Q^TEQ], Q \in \mathscr{O}(k). \tag{90}$$

Let $O = \widetilde{Z}_1^T\widetilde{Z}_1$, then

$$\nabla V(I) = 4\mathbb{E}[(I-O)diag(I-O)] = 4\mathbb{E}(Odiag(O)) - 4I. \tag{91}$$

Thus by Corollary H.1, $\nabla V(I)$ is symmetric $\Rightarrow \langle\nabla V(I), K\rangle = 0$.

By Frechet derivatives, for any $H \in \mathbb{R}^{k\times k}$ and $t > 0$,

$$\begin{aligned}\nabla V(Q+tH) - \nabla V(Q) &= 4\mathbb{E}[E(Q+tH)diag((Q+tH)^TE(Q+tH)] - 4\mathbb{E}[EQdiag(Q^TEQ)] \\ &= 4t\mathbb{E}[EHdiag(Q^TEQ) + 2EQdiag(H^TEQ)] + O(t^2),\end{aligned}$$

choose $H = K, Q = I$, which means the derivative of $\nabla V(Q)$ evaluated at I in the direction of K is

$$\nabla^2 V(I)\cdot K = -4K + 4\mathbb{E}[OKdiag(O)] + 8\mathbb{E}[Odiag(OK)]. \tag{92}$$

Theorem 5.1 indicates identity matrix I is one of the global maximas of population Varimax function. Applying Second Order Condition result (Corollary H.2), Lemma I.2 and reusing $\langle\nabla V(I), K\rangle = 0$ obtains

$$\begin{aligned}\langle\nabla^2 V(I)\cdot K, K\rangle &= -4\langle K, K\rangle + 4\mathbb{E}\langle OKdiag(O), K\rangle + 8\mathbb{E}\langle Odiag(OK), K\rangle \\ &= -4\|K\|_F^2 + 12\mathbb{E}[trace(diag(OK)^2)] \\ &\leq -4\|K\|_F^2 + 4\mathbb{E}\langle Odiag(O), KK^T\rangle,\end{aligned}$$

and

$$\langle\nabla V(I), K^2\rangle = -4\mathbb{E}\langle Odiag(O), KK^T\rangle + 4\|K\|_F^2. \tag{93}$$

Let $K^u = K/\|K\|$, then

$$\begin{aligned}\frac{\partial^2 V}{\partial t^2}|_{t=0} &= \langle\nabla^2 V(I)\cdot K, K\rangle + \langle\nabla V(I), K^2\rangle \\ &= -4\|K\|_F^2 + 12\mathbb{E}[trace(diag(OK)^2)] - 4\mathbb{E}\langle Odiag(O), KK^T\rangle + 4\|K\|_F^2 \\ &= 4\times(3\mathbb{E}[trace(diag(OK)^2) - \mathbb{E}\langle Odiag(O), KK^T\rangle]) \\ &= 4\times(3\mathbb{E}[trace(diag(OK^u)^2) - \mathbb{E}\langle Odiag(O), K^uK^{uT}\rangle])\|K\|^2 \\ &\leq -C_s\|K\|^2.\end{aligned}$$

Here

$$-C_s = \max_{\|K^u\|=1, K^u \in \mathcal{O}(k)^\perp} 4 \times (3\mathbb{E}[trace(diag(OK^u)^2) - \mathbb{E}\langle Odiag(O), K^u K^{uT}\rangle]), \quad (94)$$

is a negative constant (thus C_s is a positive constant) since RHS of Equation (94) is upper-bounded and the set of skewed symmetric matrix with unit Frobenius norm is a bounded, closed and compact space. With derived results and Taylor expansion,

$$\begin{aligned} V(R_{\widetilde{Z}}) &= V(I) + \langle \nabla V(I), K\rangle + \langle \nabla^2 V(I)\cdot K, K\rangle + \langle \nabla V(I), K^2\rangle + o(\|K\|^2) \\ &= V(I) + \langle \nabla^2 V(I)\cdot K, K\rangle + \langle \nabla V(I), K^2\rangle + o(\|K\|^2) \\ &\leq V(I) - C_s\|K\|^2 + o(\|K\|^2). \end{aligned}$$

By Lemma H.7 there exists $\varepsilon_0 = O_p(n^{\delta-1/2})$ s.t.

$$|\widehat{V}(R_{\widetilde{Z}}) - V(R_{\widetilde{Z}})| < \varepsilon_0, \quad |\widehat{V}(I) - V(I)| < \varepsilon_0,$$

$\Rightarrow$

$$\widehat{V}(R_{\widetilde{Z}}) - \varepsilon_0 < V(R_{\widetilde{Z}}) < V(I) < \widehat{V}(I) + \varepsilon_0,$$

$\Rightarrow$

$$V(I) - V(R_{\widetilde{Z}}) < 2\varepsilon_0.$$

These implies

$$\|K\|^2 < \frac{2\varepsilon_0 + o(\|K\|^2)}{C_s}. \quad (95)$$

Therefore $\|K\| = O_p(n^{\delta/2-1/4})$. By matrix exponential algebra,

$$\|R_{\widetilde{Z}} - I\|_{2\to\infty} \leq \|R_{\widetilde{Z}} - I\| = \|\sum_{i=1}^{\infty} \frac{K^i}{i!}\| \leq \sum_{i=1}^{\infty} \frac{\|K\|^i}{i!} = O_p(n^{\delta/2-1/4}). \quad (96)$$

□

Detailed Proof of Proposition G.5

Proof. Write

$$V_1(O) = \sum_{\ell=1}^{k} \left(\frac{1}{n}\sum_{i=1}^{n} [\sqrt{n}\widehat{U}O]_{i\ell}^4 - \left(\frac{1}{n}\sum_{i=1}^{n} [\sqrt{n}\widehat{U}O]_{i\ell}^2\right)^2 \right),$$

$$V_2(O) = \sum_{\ell=1}^{k} \left(\frac{1}{n}\sum_{i=1}^{n} [\sqrt{n}UWO]_{i\ell}^4 - \left(\frac{1}{n}\sum_{i=1}^{n} [\sqrt{n}UWO]_{i\ell}^2\right)^2 \right).$$

To be specific, V_1 is the sample version of Varimax function with perturbed eigenvectors as input. V_2 is sample version of Varimax function with true eigenvectors rotated with W (specified in Lemma G.5). The proof of Proposition G.5 could be described as two parts. First part shows the uniform upper bound for difference between V_1, V_2 (Equation (97)). Similar to the proof of Lemma G.6, the second part explores the first and second order condition of Equation (99) to obtain the bound for the difference between solutions of V_1 and V_2 (modulo permutation and sign-flip).

Mathematically speaking, the first part (uniform upper bound for difference between V_1, V_2) is equivalent to

$$\sup_{O\in\mathcal{O}(k)} |V_1(O) - V_2(O)| \leq O_p\left((n\rho_n)^{-1/2}\log^{\frac{7}{2}} n\right). \tag{97}$$

In the proof of Proposition G.5 let X_i be the ith row of $\sqrt{n}\widehat{U}$, and ith row of $\sqrt{n}UW$ be $X_i + \epsilon_i$. From Lemma G.5, for any unit length vector $r \in \mathbb{R}^k$, we have

$$|(\langle X_i, r\rangle - \langle X_i + \epsilon_i, r\rangle)| \leq \|\epsilon_i\|\|r\| \leq \sqrt{n}\|\widehat{U} - UW\|_{2\to\infty} = O_p((n\rho_n)^{-1/2}\log^{\frac{5}{2}} n).$$

Therefore,

$$\begin{aligned}
& |\sum_{i=1}^n (\langle X_i, r\rangle^4 - \langle X_i + \epsilon_i, r\rangle^4)| \\
\leq & \sum_{i=1}^n |(\langle X_i, r\rangle^2 + \langle X_i + \epsilon_i, r\rangle^2)(\langle X_i, r\rangle + \langle X_i + \epsilon_i, r\rangle)(\langle X_i, r\rangle - \langle X_i + \epsilon_i, r\rangle)| \\
\leq & \sum_{i=1}^n (\langle X_i, r\rangle^2 + \langle X_i + \epsilon_i, r\rangle^2)(\|X_i\|_2 + \|X_i + \epsilon_i\|_2)|(\langle X_i, r\rangle - \langle X_i + \epsilon_i, r\rangle)| \\
\leq & \sum_{i=1}^n (\langle X_i, r\rangle^2 + \langle X_i + \epsilon_i, r\rangle^2)O_p(\log n)|(\langle X_i, r\rangle - \langle X_i + \epsilon_i, r\rangle)| \\
\leq & \sum_{i=1}^n (\langle X_i, r\rangle^2 + \langle X_i + \epsilon_i, r\rangle^2)O_p(\log n)\sqrt{n}\|\widehat{U} - UW\|_{2\to\infty} \\
\leq & \sum_{i=1}^n (\langle X_i, r\rangle^2 + \langle X_i + \epsilon_i, r\rangle^2)O_p\left((n\rho_n)^{-1/2}\log^{\frac{7}{2}} n\right).
\end{aligned}$$

Notice that columns of $\widehat{U}$ and UW have unit length and R is an orthogonal matrix. Thus the columns of $\widehat{U}R$ and UWR are all of unit length. Therefore

$$\left(\frac{1}{n}\sum_{i=1}^n [\sqrt{n}\widehat{U}R]_{i\ell}^2\right)^2 = \left(\sum_{i=1}^n [\widehat{U}R]_{i\ell}^2\right)^2 = 1^2 = \left(\sum_{i=1}^n [UWR]_{i\ell}^2\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n [\sqrt{n}UWR]_{i\ell}^2\right)^2.$$

Let O_ℓ be the ℓth column of O. Then for any $O \in \mathcal{O}(k)$,

$$\begin{aligned}
& |V_1(O) - V_2(O)| \\
\leq & \sum_{\ell=1}^{k} | \left(\frac{1}{n} \sum_{i=1}^{n} [\sqrt{n}\widehat{U}O]_{i\ell}^4 - \left(\frac{1}{n} \sum_{i=1}^{n} [\sqrt{n}\widehat{U}O]_{i\ell}^2 \right)^2 \right) \\
& \qquad - \left(\frac{1}{n} \sum_{i=1}^{n} [\sqrt{n}UWO]_{i\ell}^4 - \left(\frac{1}{n} \sum_{i=1}^{n} [\sqrt{n}UWO]_{i\ell}^2 \right)^2 \right) | \\
\leq & \frac{1}{n} \sum_{\ell=1}^{k} | \sum_{i=1}^{n} ([\sqrt{n}\widehat{U}O]_{i\ell}^4 - [\sqrt{n}UWO]_{i\ell}^4) | \\
\leq & \frac{1}{n} \sum_{\ell=1}^{k} | \sum_{i=1}^{n} (\langle X_i, O_\ell \rangle^4 - \langle X_i + \epsilon_i, O_\ell \rangle^4) | \\
\leq & \frac{1}{n} \sum_{\ell=1}^{k} \sum_{i=1}^{n} (\langle X_i, O_\ell \rangle^2 + \langle X_i + \epsilon_i, O_\ell \rangle^2) O_p \left((n\rho_n)^{-1/2} \log^{\frac{7}{2}} n \right) \\
= & \sum_{\ell=1}^{k} O_p \left((n\rho_n)^{-1/2} \log^{\frac{7}{2}} n \right) \\
= & O_p \left((n\rho_n)^{-1/2} \log^{\frac{7}{2}} n \right).
\end{aligned}$$

Since the orthogonal matrix O here is arbitrary, therefore the Equation (97) is proved.

For the next step, we want to show the upper bound of $2 \to \infty$ norm distance between $R_{\widehat{U}}$ and $R_{UW} P_n^{(3)}$ ($P_n^{(3)} \in \mathcal{P}(k)$ is defined in Proposition G.5),

$$\| R_{\widehat{U}} - R_{UW} P_n^{(3)} \|_{2 \to \infty} = O_p \left((n\rho_n)^{-1/4} \log^{\frac{7}{4}} n \right). \tag{98}$$

For simplicity notate $R_1 = R_{\widehat{U}}$, $R_2 = R_{UW}$. There are $k \times k$ skew-symmetric matrices K_1, K_2 s.t. $R_1 = \exp(K_1)$, $R_2 = \exp(K_2)$. Define

$$\gamma_2(t) = \exp((1-t)K_2 + tK_1), \tag{99}$$

then $\gamma_2(0) = R_2$, $\gamma_2(1) = R_1$. Again, as in the proof of Lemma G.6 we assume

$$I = \underset{P_0 \in \mathcal{P}(k)}{\arg\min} \|R_1 - R_2 P_0\|_{2\to\infty},$$

and we constrain our analysis on a neighborhood of R_2: $B(R_2, \delta_p) := \{P \in \mathscr{O}(k) | \|P - R_2\| < \delta_p\}$, such that

$$B(R_2, \delta_p) \cap \{R_2 P_0 | P_0 \in \mathcal{P}(k)\} = \{R_2\}.$$

This indicates for any $R \in B(R_2, \delta_p)$ there is $V_2(R) \leq V_2(R_2)$.

Before Taylor expansion analysis, we should check that $\|R_1 - R_2\| \xrightarrow{p} 0$ is true. After that we should show $\|K_1 - K_2\| \xrightarrow{p} 0$ is also true. By definition,

$$V_1(R_1) \geq V_1(R_2) - o_p(1), |V_1(R_2) - V_2(R_2)| \xrightarrow{p} 0 \Rightarrow V_1(R_1) \geq V_2(R_2) - o_p(1). \tag{100}$$

Then,

$$\begin{aligned} V_2(R_2) - V_2(V_1) &\leq V_1(R_1) - V_2(R_1) + o_p(1) && (101)\\ &\leq \sup_{R \in \mathcal{O}(k)} |V_1(R) - V_2(R)| + o_p(1) \xrightarrow{p} 0. && (102) \end{aligned}$$

By conditions, for any $\epsilon_0 > 0, \eta_0 > 0$ such that $V_2(R) < V_2(R_2) - \eta_0$ for every $R \in \mathscr{O}(k)$ with $\|R - R_2\| \geq \epsilon_0$. Thus the event $\{\|R_2 - R_1\|\}$ is contained in the event $\{V_2(R_1) < V_2(R_2) - \eta_0\}$. The probability of the latter event goes to 0. Therefore $\|R_1 - R_2\| \xrightarrow{p} 0$. By Lemma G.6, with high probability, $R_1 R_2^T$ is converging to I as n grows. Variant of Baker-Cambell-Hausdorff formula gives

$$\begin{aligned} \|K_1 - K_2\| &= \|\log R_1 R_2^T R_2 - \log R_2\| \\ &= \|\log R_1 R_2^T + \frac{1}{2}[\log R_1 R_2^T, \log R_2] + \cdots\| \\ &\leq \|\log(I + R_1(R_2^T - R_1^T))\|_F + o_p(\|\log R_2\|) \\ &\xrightarrow{p} 0. \end{aligned}$$

With differential calculation results in Chu and Trendafilov [1998],

$$\frac{d}{dt} V_2(\gamma_2(t))|_{t=0} = \nabla V_2(\gamma_2(t))^T \frac{d\gamma_2}{dt}|_{t=0} = \langle \nabla V_2(R_2), R_2(K_1 - K_2) \rangle, \tag{103}$$

$$\begin{aligned}\frac{d^2}{dt^2}V_2(\gamma_2(t))|_{t=0} &= \langle\nabla^2 V_2(\frac{d\gamma_2(t)}{dt})|_{t=0}, \gamma_2(0)(K_1-K_2)\rangle + \langle\nabla V, \gamma_2(0)(K_1-K_2)^2\rangle \\ &= \langle\nabla^2 V_2 \cdot R_2(K_1-K_2), R_2(K_1-K_2)\rangle + \langle\nabla V_2, R_2(K_1-K_2)^2\rangle.\end{aligned}$$

By Equation (7),(8) of Chu and Trendafilov [1998],

$$V_2(Q) = n^{-3} trace\left[\sum_{i=1}^{n} diag(Q^T E_i Q)^2\right], \tag{104}$$

where $E_i = (UW)^T UW - n(X_i+\epsilon_i)(X_i+\epsilon_i)^T$. And

$$\nabla V_2(Q) = 4n^{-3}\left[\sum_{i=1}^{n} E_i Q diag(Q^T E_i Q)\right]. \tag{105}$$

Theorem 3.1 in Chu and Trendafilov [1998] implies

$$R_2^T \nabla V_2(R_2) = \sum_{i=1}^{n} R_2^T E_i R_2 diag(R_2^T E_2 Q_2)$$

is symmetric, thus

$$\begin{aligned}\langle\nabla V_2(R_2), (K_1-K_2)R_2\rangle &= trace[\nabla V_2(K_1-K_2)^T R_2^T] \\ &= trace[R_2^T \nabla V_2(K_1-K_2)^T] \\ &= \langle R_2^T \nabla V_2, K_1-K_2\rangle \\ &= 0. \end{aligned}\tag{106}$$

The last equality is because K is skew-symmetric and $R_2^T \nabla V_2(R_2)$ is symmetric.

By Frechet derivatives, for any $H \in \mathbb{R}^{k\times k}$ and $t > 0$, we have

$$\begin{aligned}&\nabla V_2(Q+tH) - \nabla V_2(Q) \\ =\ & 4n^{-3}\left[\sum_{i=1}^{n} E_i(Q+tH) diag((Q+tH)^T E_i (Q+tH))\right] - 4n^{-3}\left[\sum_{i=1}^{n} E_i Q diag(Q^T E_i Q)\right] \\ =\ & 4n^{-3}\left[\sum_{i=1}^{n} (E_i H diag(Q^T E_i Q + E_i Q(diag(H^T E_i Q) + diag(Q^T E_i H))\right] + O(t^2).\end{aligned}$$

Choosing $H = R_2(K_1 - K_2), Q = R_2$ gives the derivative of $\nabla V_2(Q)$ evaluated at R_2 in the direction of $R_2(K_1 - K_2)$

$$\begin{aligned}\nabla^2 V_2(R_2) \cdot R_2(K_1 - K_2) &= 4n^{-3} \sum_{i=1}^{n} [E_i R_2(K_1 - K_2) diag(R_2^T E_i R_2) \\ &+ E_i R_2 diag((K_1 - K_2)^T R_2^T E_i R_2) \\ &+ E_i R_2 diag(R_2^T E_i R_2(K_1 - K_2))].\end{aligned}$$

Applying Corollary 3.3 of Chu and Trendafilov [1998],

$$\begin{aligned}&\langle \nabla^2 V_2(R_2) \cdot R_2(K_1 - K_2), R_2(K_1 - K_2)\rangle + \nabla V, R_2(K_1 - K_2)^2\rangle \\ =\ & 4n^{-3} \sum_{i=1}^{n} [\langle E_i R_2(K_1 - K_2) diag(R_2^T E_i R_2), R_2(K_1 - K_2)\rangle \\ +\ & \langle E_i R_2 diag((K_1 - K_2)^T R_2^T E_i R_2), R_2(K_1 - K_2)\rangle \\ +\ & \langle E_i R_2 diag(R_2^T E_i R_2(K_1 - K_2)), R_2(K_1 - K_2)\rangle \\ +\ & \langle E_i R_2 diag(R_2^T E_i R_2), R_2(K_1 - K_2)^2\rangle] \\ =\ & 4n^{-3} \sum_{i=1}^{n} [\langle R_2^T E_i R_2(K_1 - K_2) diag(R_2^T E_i R_2), (K_1 - K_2)\rangle \\ +\ & 2\langle R_2^T E_i R_2 diag(R_2^T E_i R_2(K_1 - K_2)), K_1 - K_2\rangle \\ +\ & \langle R_2^T E_i R_2 diag(R_2^T E_i R_2), (K_1 - K_2)^2\rangle] \le 0. \end{aligned} \tag{107}$$

Let $K_o^u = (K_1 - K_2)/\|K_1 - K_2\|_F$, then

$$\begin{aligned}&\langle \nabla^2 V_2(R_2) \cdot R_2(K_1 - K_2), R_2(K_1 - K_2)\rangle + \nabla V, R_2(K_1 - K_2)^2\rangle \\ =\ & 4n^{-3} \|K_1 - K_2\|_F \sum_{i=1}^{n} [\langle R_2^T E_i R_2 K_o^u diag(R_2^T E_i R_2), K_o^u\rangle \\ +\ & 2\langle R_2^T E_i R_2 diag(R_2^T E_i R_2 K_o^u), K_o^u\rangle \\ +\ & \langle R_2^T E_i R_2 diag(R_2^T E_i R_2), (K_o^u)^2\rangle] \\ \le\ & -C_{ss} \|K_1 - K_2\|_F. \end{aligned} \tag{108}$$

Here

$$\begin{aligned}-C_{ss} = \max_{\|K\|_F = 1, K \in \mathcal{O}(k)^\perp} & 4n^{-3} \sum_{i=1}^{n} [\langle R_2^T E_i R_2 K diag(R_2^T E_i R_2), K\rangle \\ & + 2\langle R_2^T E_i R_2 diag(R_2^T E_i R_2 K), K\rangle + \langle R_2^T E_i R_2 diag(R_2^T E_i R_2), K^2\rangle]\end{aligned}$$

is a negative constant. With Taylor expansion and Equation (106), (107), (108), there is

$$
\begin{aligned}
V_2(R_1) &= V_2(R_2) + \langle \nabla V_2(R_2), (K_1 - K_2)R_2 \rangle + \langle \nabla^2 V_2(R_2) \cdot R_2(K_1 - K_2), R_2(K_1 - K_2) \rangle + \\
&\quad \langle \nabla V, R_2(K_1 - K_2)^2 \rangle + o(\|K_1 - K_2\|_F^2) \\
&= V_2(R_2) + \langle \nabla^2 V_2(R_2) \cdot R_2(K_1 - K_2), R_2(K_1 - K_2) \rangle + \\
&\qquad \nabla V, R_2(K_1 - K_2)^2 \rangle + o(\|K_1 - K_2\|_F^2) \\
&\leq V_2(R_2) - C_{ss}\|K_1 - K_2\|_F^2 + o(\|K_1 - K_2\|_F^2). \qquad (109)
\end{aligned}
$$

With Equation (97), there exists $\varepsilon_1 = O_p((n\rho_n)^{-1/2} \log^{\frac{7}{2}} n)$, s.t.

$$
|V_2(R_1) - V_1(R_1)| < \varepsilon_1, \quad |V_2(R_2) - V_1(R_2)| < \varepsilon_1,
$$

$\Rightarrow$

$$
V_1(R_1) - \varepsilon_1 < V_2(R_1) < V_2(R_2) < V_1(R_2) + \varepsilon_1,
$$

$\Rightarrow$

$$
V_2(R_2) - V_2(R_1) < 2\varepsilon_1.
$$

Then from Equation (109) there is

$$
\|K_1 - K_2\|_F^2 < \frac{2\varepsilon_1 + o(\|K_1 - K_2\|_F^2)}{C_{ss}}. \qquad (110)
$$

Thus $\|K_1 - K_2\|_F = O_p\left((n\rho_n)^{-1/4} \log^{\frac{7}{4}} n\right)$. By Lie Product Formula, for any $k \times k$ matrices S_1, S_2 the exponential of their sum could be expressed as

$$
\exp(S_1 + S_2) = \lim_{m \to \infty} (\exp(\frac{S_1}{m}) \exp(\frac{S_2}{m}))^m.
$$

Thus

$$
\begin{aligned}
& R_1 - R_2 \\
= \;& \exp(K_2 + K_1 - K_2) - \exp(K_2) \\
= \;& \lim_{m \to \infty} \{[\exp(\frac{K_2}{m}) \exp(\frac{K_1 - K_2}{m})]^m - (\exp(\frac{K_2}{m}))^m\} \\
= \;& \lim_{m \to \infty} [\exp(\frac{K_2}{m}) \exp(\frac{K_1 - K_2}{m}) - \exp(\frac{K_2}{m})] \\
& \times \; [\sum_{i=1}^{m} (\exp(\frac{K_2}{m}) \exp(\frac{K_1 - K_2}{m}))^i (\exp(\frac{K_2}{m}))^{m-i}].
\end{aligned}
$$

Since K_1, K_2 are skew-symmetric matrices, we have $\|\exp(\frac{K_1-K_2}{m})\| = \|\exp(\frac{K_2}{m})\| = 1$, and

$$
\begin{aligned}
& \|R_1 - R_2\|_{2\to\infty} \\
\leq\ & \|R_1 - R_2\| \\
=\ & \|\exp(K_2 + K_1 - K_2) - \exp(K_2)\| \\
=\ & \lim_{m\to\infty} \|[\exp(\frac{K_2}{m})\exp(\frac{K_1-K_2}{m}) - \exp(\frac{K_2}{m})][\sum_{i=1}^{m}(\exp(\frac{K_2}{m})\exp(\frac{K_1-K_2}{m}))^i(\exp(\frac{K_2}{m}))^{m-i}]\| \\
\leq\ & \lim_{m\to\infty} \|\exp(\frac{K_2}{m})\| \cdot \|\exp(\frac{K_1-K_2}{m}) - I\|(\sum_{i=1}^{m}\|\exp(\frac{K_2}{m})\|^{m-i} \cdot \|\exp(\frac{K_1-K_2}{m})\|^i) \\
=\ & \lim_{m\to\infty} \|\exp(\frac{K_1-K_2}{m}) - I\|(m-1) \\
=\ & \lim_{m\to\infty} \|\sum_{i=1}^{\infty}(\frac{K_1-K_2}{m})^i\|(m-1) \\
\leq\ & \lim_{m\to\infty} \sum_{i=1}^{\infty}\|\frac{K_1-K_2}{m}\|^i(m-1) \\
=\ & O_p(\|K_1 - K_2\|) \\
\leq\ & O_p(\|K_1 - K_2\|_F) \\
=\ & O_p\left((n\rho_n)^{-1/4}\log^{\frac{7}{4}} n\right). \qquad (111)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \|\sqrt{n}\widehat{U}R_{\widehat{U}} - \sqrt{n}\widehat{U}R_{UW}P_n^{(3)}\|_{2\to\infty} \\
(\text{Lemma } G.3)\quad \leq\ & \sqrt{n}\|\widehat{U}\|_{2\to\infty}\|R_{\widehat{U}} - R_{UW}P_n^{(3)}\| \\
\leq\ & \sqrt{n}(\|\widehat{U} - UW\|_{2\to\infty} + \|UW\|_{2\to\infty})\|R_{\widehat{U}} - R_{UW}P_n^{(3)}\| \\
(\text{Equation } (47))\quad =\ & O_p(\log n) \times \|R_{\widehat{U}} - R_{UW}P_n^{(3)}\| \\
(\text{Equation } (111))\quad =\ & O_p\left((n\rho_n)^{-1/4}\log^{\frac{11}{4}} n\right).
\end{aligned}
$$

□

I First and Second Order Condition for Population Varimax

This section exploits the first and second order condition of the population Varimax function based on the similar results of the sample Varimax function (Sherin [1966], Neudecker [1981], Chu and Trendafilov [1998]). This section is self-contained and only reuses the notations of (2) and the definition of Assumption 1. We redefine some notations here.

Assumption 4. *$U = ZR_U^T$ with $R_U \in \mathcal{O}(k)$, $Z \in \mathbb{R}^{n\times k}$ with Z satisfying Assumption 1. Let z_0 represents first row of Z. $O = z_0 z_0^T$. z_i is the ith element of z_0 with $\mathbb{E}(z_i) = 0, \forall i \in [k]$. Population Varimax function is $\mathcal{V}(R) = \mathbb{E}(v(R, U))$.*

Optimization conditions for population Varimax function borrows conclusions from Chu and Trendafilov [1998]. The math requires switching the order of expectation and differential operations. Lemma I.1 shows that this is valid for Varimax function. The proof of the lemmas and Theorems in current section are all contained in Section I.3.

Lemma I.1. *Under Assumption 4, the expectation operator and differential operator of Varimax function are exchangeable,*

$$\frac{\partial \mathbb{E}v(R,U)}{\partial R} = \mathbb{E}\frac{\partial v(R,U)}{\partial R}.$$

I.1 First Order Condition (FOC)

The First Order Condition for sample varimax function is (Sherin [1966], Neudecker [1981]):

$$(U^T URD - U^T H)R^T = R(U^T URD - U^T H)^T \tag{112}$$

Where R is a stationary point of $v(R, U)$. D is a diagonal matrix with j-th element equals to $\frac{1}{n}\sum_{i=1}^n (UR)_{ij}^2$. H is $n \times k$ matrix with $(H)_{ij} = (UR)_{ij}^3$.

Theorem I.1 (FOC)**.** *Under Assumption 4, if $R \in \mathcal{O}(k)$ is a stationary point of $\mathcal{V}(R)$, then*

$$\mathbb{E}z_i^2\mathbb{E}(z_i z_j) - \mathbb{E}(z_i^3 z_j) = \mathbb{E}z_j^2\mathbb{E}(z_i z_j) - \mathbb{E}(z_j^3 z_i), \forall i \neq j. \tag{113}$$

With Assumption 4, Theorem I.1 is a trivial result. The FOC only tells about local stationary points. To consider curvature information and ensure the stationary point is local maxima, we also need to figure out the Second Order Condition.

I.2 Second Order Condition (SOC)

Chu and Trendafilov [1998] shows SOC result for sample Varimax function. The current subsection is deriving counterpart results of the population Varimax function. To describe the SOC on sample data, Chu and Trendafilov [1998] reformulate the Varimax criterion and express the problem in a simultaneously diagonalizing symmetric matrices form (ten Berge [1984]). The detailed SOC statement for sample Varimax is shown below.

Write $E_i = U^T U - n u_i u_i^T$ with u_i^T being U's i-th row. The sufficient (necessary) SOC of $v(R, U)$ is:

$$\sum_{i=1}^{n} (\langle U^T E_i R diag(R^T E_i R), K^2 \rangle + \langle R^T E_i R K diag(R^T E_i R), K \rangle + 2\langle R^T E_i R diag(R^T E_i R K), K \rangle) < (\leq) 0, \tag{114}$$

for any non-zero skew-symmetric matrix K. Since Varimax condition gives us a special covariance structure of z_0 (e.g. $Cov(z_0) = I$), we could derive SOC for population Varimax function from (114).

Theorem I.2 (SOC)**.** *Under Assumption 4, a sufficient (necessary) condition for R_U to be one of the maximas of the population Varimax is*

$$3\mathbb{E}[tr(diag(OK)^2)] < (\leq)\mathbb{E}\langle O diag O, K K^T \rangle. \tag{115}$$

I.3 Proofs in Section I

I.3.1 Proof of Lemma I.1

Proof. The main idea of the proof is applying Dominant Converge Theorem (DCT). For simplicity, write $\mathbb{E}[U_{ij}^q] = \mu_q^{(j)}$ as q-th moment of U's j-th column and $G_i = U^T U - n u_i u_i^T$, $i \in [n]$, with u_i^T being the i-th row of U. By (8) of Chu and Trendafilov [1998],

$$\frac{\partial v(R, U)}{\partial R} = \frac{4}{n^3} \sum_{i=1}^{n} G_i R diag(R^T G_i R). \tag{116}$$

The goal is to bound the spectral norm of RHS of Equation (116). Notice for $\forall i \in [n]$,

$$\begin{aligned} \|G_i R diag(R^T G_i R)\| &\leq \|G_i R\| \cdot \|diag(R^T G_i R)\| \\ &= \|G_i\| \cdot \|diag(R^T G_i R)\| \\ &= \|U^T U - n u_i u_i^T\| \cdot \|diag(R^T U^T U R) - diag(n R^T u_i u_i^T R)\| \\ &\leq (\|U^T U\| + n\|u_i u_i^T\|) \times \\ &\quad (\|diag(R^T U^T U R)\| + n\|diag(R^T u_i u_i^T R)\|). \end{aligned} \tag{117}$$

Basic matrix algebra implies

$$\|U^T U\| \leq \|U^T\| \cdot \|U\| = \|U\|^2 \leq \|U\|_F^2, \quad \|u_i u_i^T\| \leq \|u_i\|^2.$$

Notice that the i-th diagonal element of $R^T U^T U R$ is

$$\begin{aligned} \sum_{t=1}^{k}[(\sum_{s=1}^{k} U_{is} R_{st})^2] &\leq \sum_{t=1}^{k}[(\sum_{s=1}^{k} U_{is}^2)(\sum_{s=1}^{k} R_{st}^2)] \\ &= k \sum_{s=1}^{k} U_{is}^2 \end{aligned}$$

$\Rightarrow$

$$\|diag(R^T U^T U R)\| \leq \sum_{i=1}^{n} k(\sum_{s=1}^{k} U_{is}^2) = k\|U\|_F^2. \tag{118}$$

Similarly,

$$\|diag(R^T u_i u_i^T R)\| \leq k\|u_i\|^2. \tag{119}$$

Plugging Equation (118), (119) into Equation (117) yields

$$\begin{aligned} \|G_i R diag(R^T G_i R)\| &\leq (\|U\|_F^2 + n\|u_i\|^2)(k\|U\|_F^2 + nk\|u_i\|^2) \\ &= k\|U\|_F^4 + 2kn\|u_i\|^2\|U\|_F^2 + kn^2\|u_i\|^4. \end{aligned}$$

Getting back to Equation (116), we have

$$\begin{aligned} \|\frac{\partial v(R,U)}{\partial R}\| &\leq \frac{4}{n^3} \sum_{i=1}^{n} \|G_i R diag(R^T G_i R)\| \\ &\leq \frac{4}{n^3} \sum_{i=1}^{n} (k\|U\|_F^4 + 2kn\|u_i\|^2\|U\|_F^2 + kn^2\|u_i\|^4) \\ &:= F_n. \end{aligned}$$

The F_n is a random variable (depends on norms of random matrix and vectors). It will be

sufficient to give constant bound to the expectation of each term of F_n. Notice

$$\begin{aligned}
\mathbb{E}\|U\|_F^4 &= \mathbb{E}[(\sum_{ij} U_{ij}^2)^2] \\
&= n\sum_{j=1}^{k}\mu_4^{(j)} + \binom{n}{2}\sum_{j=1}^{k}\mu_2^{(j)2} + n^2\sum_{1\le\ell\neq j\le k}\mu_2^{(\ell)}\mu_2^{(j)}, \\
\mathbb{E}(\|u_i\|^2\|U\|_F^2) &= \mathbb{E}[(\sum_{j=1}^{k} U_{ij}^2)(\sum_{i,j} U_{ij}^2)] \\
&= \sum_{j=1}^{k}\mu_4^{(j)} + (n-1)\sum_{j=1}^{k}\mu_2^{(j)2} + n\sum_{1\le\ell\neq j\le k}\mu_2^{(\ell)}\mu_2^{(j)}, \\
\mathbb{E}(\|u_i\|^4) &= \mathbb{E}[(\sum_{j=1}^{k} U_{ij}^2)^2]\sum_{j=1}^{k}\mu_4^{(j)} + \sum_{1\le\ell\neq j\le k}\mu_2^{(\ell)}\mu_2^{(j)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}(F_n) &= \frac{4k}{n^2}\mathbb{E}\|U\|_F^4 + \frac{8k}{n^2}\sum_{i=1}^{n}\mathbb{E}(\|u_i\|^2\|U\|_F^2) + \frac{4k}{n}\sum_{i=1}^{n}\mathbb{E}(\|u_i\|^4) \\
&= (4k + \frac{12k}{n})M_1 + \frac{17k(n-1)}{n}M_2 + 16kM_3 \\
&\le 5kM_1 + 17kM_2 + 16kM_3 \\
&< \infty.
\end{aligned}$$

Here $M_1 = \sum_{j=1}^{k}\mu_4^{(j)}$, $M_2 = \sum_{j=1}^{k}\mu_2^{(j)2}$, $M_3 = \sum_{1\le\ell\neq j\le k}\mu_2^{(\ell)}\mu_2^{(j)}$ are all constants in our settings. Then DCT accomplishes our proof.

□

I.3.2 Proof of Theorem I.2

The following Lemma is useful in the proof of Theorem I.2.

Lemma I.2. *For any symmetric matrix $S = vv^T$ where v is a k-dimension vector. Any $k \times k$ matrix P. We have*

$$\langle Sdiag(SP), P\rangle = \langle SPdiag(S), P\rangle. \tag{120}$$

Proof. Let $(S)_{i,j} = v_i v_j$, $(P)_{i,j} = P_{ij}$. We only need to prove $Sdiag(SP) = SPdiag(S)$. For $diag(SP)$. Its j-th diagonal element is $v_j\sum_k v_k P_{kj}$. Multiplying a diagonal matrix on right side is equal to multiplying i-th diagonal element to i-th column. Thus

$(Sdiag(SP))_{ij} = v_i v_j^2 \sum_k v_k P_{kj}$.

For $SPdiag(S)$ we have $(SP)_{ij} = v_i \sum_k v_k P_{kj} \Rightarrow (SPdiag(S))_{ij} = v_i v_j^2 \sum_k v_k P_{kj}$.

□

Now we return to the proof of Theorem I.2

Proof. By Lemma I.1 and Slutsky Theorem,

$$\begin{aligned}
\mathbb{E}[R_U^T E_i R_U diag(R_U^T E_i R_U)] &= n^2 \mathbb{E}[(I-O)diag(I-O)], \\
\mathbb{E}[R_U^T E_i R_U K diag(R_U^T E_i R_U)] &= n^2 \mathbb{E}[(I-O)K diag(I-O)], \\
\mathbb{E}[R_U^T E_i R_U diag(R_U^T E_i R_U K)] &= n^2 \mathbb{E}[(I-O)diag(K-OK)].
\end{aligned}$$

The expectation of the first term of Equation (114) equals to

$$\begin{aligned}
n^2 \mathbb{E}\langle((I-O)diag(I-O), K^2\rangle &= n^2 \mathbb{E}\langle I - O - diagO + OdiagO, K^2\rangle \\
&= n^2(\mathbb{E}\langle Odiag(O), K^2\rangle - \langle I, K^2\rangle). \qquad (121)
\end{aligned}$$

Similarly, the expectation of the second term of Equation (114) is

$$n^2 \mathbb{E}\langle (I-O)K diag(I-O), K\rangle = n^2(\mathbb{E}\langle OK diag(O), K\rangle - \langle K, K\rangle), \qquad (122)$$

and the expectation of the third term of Equation (114) is

$$n^2 \mathbb{E}\langle (I-O)diag((I-O)K), K\rangle = n^2(\mathbb{E}\langle Odiag(OK), K\rangle - \langle diag(K), K\rangle). \qquad (123)$$

Notice K is skew-symmetric, there are

$$\langle I, K^2\rangle = tr(K^2) = -tr(KK^T) = -\langle K, K\rangle, \langle diag(K), K\rangle = 0. \qquad (124)$$

By Lemma I.2,

$$\langle Odiag(OK), K\rangle = \langle OK diag(O), K\rangle. \qquad (125)$$

Properties of trace operator indicate that $tr(Ydiag(Y)) = tr((diag(Y)^2)$ for any square matrix Y. Then with Equation (121), (122), (123), (124), (125), the second order condition (114) boils down to

$$3\mathbb{E}[tr(diag(OK)^2)] < (\leq) \mathbb{E}\langle OdiagO, KK^T\rangle, \qquad (126)$$

for any non-zero skewed symmetric matrix K. □

J Proofs of Corollaries C.1 and C.2

As we pointed out in Section C.3, independent columns assumption does not hold in the degree-corrected stochastic block model (DC-SBM). We could still make use of the first and second order condition to show that vsp could estimate Z correctly. Similar to Proposition 5.2 we have

$$U = \frac{1}{\sqrt{n}} Z(Z^T Z/n)^{-\frac{1}{2}} \widetilde{R}_U, \quad \widetilde{R}_U \in \mathscr{O}(k). \tag{127}$$

The proof of Corollary C.1, C.2 will focus on validating some key conditions and assumptions.

J.1 Proof of Corollary C.1

Proof. To borrow the conclusion from Theorem 6.1, it will be sufficient to check some results. Notice we don't have the centering step and symmetrized adjacency matrices in SBM, such difference only simplifies the proof without introducing extra layers of perturbation. The only things we have to check (because of the dependency of Z's columns) are conclusions of Theorem 5.1, Theorem I.1, I.2, Lemma G.4, Assumption 3 and the arguments in the proof of Lemma G.6 that shows the existence of moment generating function of linear inner-product of Z_i (Z's ith row).

J.1.1 Theorem 5.1 Under DC-SBM

Recall that in current setting,

$$\mathcal{V}_{\widetilde{R}_U}(Q) = \sum_j (\mathbb{E}([Z\widetilde{R}_U Q]_j^4) - \mathbb{E}([Z\widetilde{R}_U Q]_j^2)^2).$$

We want to show

$$\arg\max_{Q \in \mathscr{O}(k)} \mathcal{V}_{\widetilde{U}}(Q) = \{\widetilde{R}_U^T P | P \in \mathcal{P}(k)\}.$$

Let $X = Z_1 - \mathbb{E}(Z_1)$, $\mathbb{E}(X_i^j) = \mu_j^{(i)}$. Since there is exactly one non-zero entry in X's elements, we have

$$\sum_{j=1}^k \mathbb{E}([XQ]_j^2)^2 = \sum_{j=1}^k \mathbb{E}(\sum_{i=1}^k X_i^2 Q_{ij}^2)^2 = \sum_{j=1}^k \sum_{i=1}^k \mu_2^{(i)2} Q_{ij}^4,$$

and

$$\sum_{j=1}^k \mathbb{E}([XQ]_j^4) = \sum_{j=1}^k \sum_{i=1}^k \mathbb{E}(X_i^4 Q_{ij}^4) = \sum_{i=1}^k \mu_4^{(i)} Q_{ij}^4.$$

Therefore

$$\begin{aligned}\mathcal{V}_I(Q) &= \sum_j (\mathbb{E}([XQ]_j^4) - \mathbb{E}([XQ]_j^2)^2) \\ &= \sum_{i=1}^{k} (\mu_4^{(i)} - \mu_2^{(i)2}) \|Q_{i\cdot}\|_4^4.\end{aligned}$$

Jensen Inequality indicates that $\mu_4^{(i)} - \mu_2^{(i)2} > 0$ for $\forall i \in [k]$. The remaining part follows the same approach as in the proof of Theorem 5.1.

J.1.2 Theorem I.1 Under DC-SBM

For $\forall i \in [n]$ and $j, \ell \in [k], j \neq \ell$, we have

$$Z_{ij}Z_{i\ell} = 0, Z_{ij}^3 Z_{i\ell} = 0 \Rightarrow \mathbb{E}[Z_{ij}Z_{i\ell}] = 0, \mathbb{E}[Z_{ij}^3 Z_{i\ell}] = 0.$$

J.1.3 Theorem I.2 Under DC-SBM

To show SOC. Let $E_i = U^T U - n u_i^T u_i, u_i$ is the ith row of U. It is sufficient to show

$$\begin{aligned}\mathbb{E}[\langle \widetilde{R}_U E_i \widetilde{R}_U^T diag(\widetilde{R}_U E_i \widetilde{R}_U^T), K^2\rangle + \langle \widetilde{R}_U E_i \widetilde{R}_U^T K diag(\widetilde{R}_U E_i \widetilde{R}_U^T), K\rangle \\ +2\langle \widetilde{R}_U E_i \widetilde{R}_U^T diag(\widetilde{R}_U E_i \widetilde{R}_U^T K), K\rangle] \leq 0,\end{aligned} \tag{128}$$

Let $E^{i,j}$ be k by k matrix with 1 in (i,j) entry and 0's elsewhere. Then

$$\widetilde{R}_U E_i \widetilde{R}_U^T = \widetilde{R}_U \widetilde{R}_U^T - n\widetilde{R}_U u_i^T u_i \widetilde{R}_U^T = I - n Z_i^T Z_i E^{z(i),z(i)}. \tag{129}$$

Let $K = \begin{pmatrix} 0 & K_{1,2} & ... & K_{1,k} \\ K_{2,1} & 0 & ... & K_{2,k} \\ ... & ... & ... & ... \\ K_{k,1} & K_{k,2} & ... & 0 \end{pmatrix}$ with $K_{i,j} = -K_{j,i}$ for $i > j$. Then $diag(K^2) = diag(-\sum_{i\neq 1} K_{1i}^2, -\sum_{i\neq 2} K_{2i}^2, ... - \sum_{i\neq k} K_{ki}^2)$. Now we examine each term of (128),

$$
\begin{aligned}
& \langle \widetilde{R}_U E_i \widetilde{R}_U^T diag(\widetilde{R}_U E_i \widetilde{R}_U^T), K^2 \rangle \\
= \; & \langle \langle diag(1,1,...,(1-n\theta_{i,z(i)}^2)^2,...,1), K^2 \rangle \\
= \; & \sum_{\ell,j} K_{\ell j}^2 - [(n\theta_{i,z(i)}^2)^2 - 2n\theta_{i,z(i)}^2] \sum_{\ell=1}^{k} K_{z(i)\ell}^2, \qquad (130)
\end{aligned}
$$

and (131)

$$
\begin{aligned}
& \langle \widetilde{R}_U E_i \widetilde{R}_U^T K diag(\widetilde{R}_U E_i \widetilde{R}_U^T), K \rangle \\
= \; & \langle diag(1,1,...,1-n\theta_{i,z(i)}^2,...,1) K diag(1,1,...,1-n\theta_{i,z(i)}^2,...,1), K \rangle \\
= \; & \sum_{\ell,j} K_{\ell j}^2 - 2n\theta_{i,z(i)}^2 \sum_{\ell}^{k} K_{z(i)\ell}^2. \qquad (132)
\end{aligned}
$$

Since K's diagonal elements are all zero's. $diag(\widetilde{R}_U E_i \widetilde{R}_U^T K)$ will be zero matrix.

$$
\langle \widetilde{R}_U E_i \widetilde{R}_U^T diag(\widetilde{R}_U E_i \widetilde{R}_U^T K), K \rangle = 0 \qquad (133)
$$

From Equations (130), (132), (133), to prove Equation (128) it will be suffice to show:

$$
\sum_i (\sum_{\ell j}^{k} K_{\ell j}^2 + [(n\theta_{i,z(i)}^2)^2 - 2n\theta_{i,z(i)}^2] \sum_{\ell=1}^{k} K_{z(i)\ell}^2) \geq \sum_i (\sum_{\ell,j} K_{\ell j}^2 - 2n\theta_i^2 \sum_j K_{z(i)\ell}^2),
$$

$\Leftrightarrow$

$$
\sum_i ([(n\theta_{i,z(i)}^2)^2 - 2n\theta_{i,z(i)}^2] \sum_{\ell=1}^{k} K_{z(i)\ell}^2 + 2n\theta_{i,z(i)}^2 \sum_j K_{z(i)\ell}^2) \geq 0,
$$

$\Leftrightarrow$

$$
\sum_i ((n\theta_{i,z(i)}^2)^2 \sum_{\ell=1}^{k} K_{z(i)\ell}^2) \geq 0. \qquad (134)
$$

The last inequality is strict as long as K is not zero matrix and θ_i's are all positive. We conclude that (128) is true.

J.1.4 Lemma G.4 Under DC-SBM

Under DC-SBM, elements of A are sub-gaussian variables. Thus we could utilize a simpler concentration matrix inequality than Lemma H.1. We apply the following lemma to show the bound for perturbation between A and $\mathscr{A}$.

Lemma J.1 ((Matrix Bernstein Inequality, Tropp [2012])). *Let $X_1, X_2, ..., X_m$ be independent random $N \times N$ symmetric matrix. Assume $\|X_i - \mathbb{E}(X_i)\| \leq M, \forall i$. Write $v^2 = \|\sum_i var(X_i)\|, X = \sum_i X_i$. Then for any $a > 0$,*

$$\mathbb{P}(\|X - \mathbb{E}(X)\| \geq a) \leq 2N \exp(-\frac{a^2}{2v^2 + 2Ma/3}).$$

Let E^{ij} be a n by n matrix with 1 in the (i,j) and (j,i) entries and 0 elsewhere. Write $p_{ij} = \mathscr{A}_{ij}$. Then we could express $A - \mathscr{A}$ as sum of matrices,

$$Y_{ij} = (A_{ij} - p_{ij})E^{ij}, i < j.$$

Notice that

$$\|A - \mathscr{A}\| = \|\sum_{1 \leq i < j \leq n} Y_{ij}\|,$$

and

$$\|Y_{ij}\| \leq \|E^{ij}\| = 1.$$

Moreover,

$$\mathbb{E}(Y_i j) = 0 \text{ and } \mathbb{E}(Y_{ij}^2) = (p_{ij} - p_{ij}^2)(E^{ii} + E^{jj}), \forall i < j.$$

Then we could get an upper bound for v^2,

$$\begin{aligned}
v^2 &= \|\sum_{1 \leq i < j \leq n} \mathbb{E}[Y_{ij}^2]\| \\
&= \|\sum_{1 \leq i < j \leq n} (p_{ij} - p_{ij}^2)(E^{ii} + E^{jj})\| \\
&= \frac{1}{2}\|\sum_{1 \leq i,j \leq n} (p_{ij} - p_{ij}^2)(E^{ii} + E^{jj})\| \\
&= \frac{1}{2}\|\sum_{i=1}\sum_{j \neq i} (p_{ij} - p_{ij}^2)E^{ii}\| \\
&\leq \frac{1}{2}\max_{1 \leq i \leq n}(\sum_{j \neq i}(p_{ij} - p_{ij}^2)) \\
&\leq \frac{1}{2}\max_{1 \leq i \leq n}(\sum_{j \neq i} p_{ij}) \\
&\leq \frac{n}{2}.
\end{aligned}$$

From Lemma J.1 we obtain,

$$\mathbb{P}(\|A - \mathscr{A}\| > a) \leq 2N \exp(-\frac{a^2}{n + 2a/3}). \tag{135}$$

J.1.5 Assumption 3 with Bernoulli Random Variables

Suppose A_{ij} has m-th central moment being $\mu_{m,ij}$ and m-th moment being $\mu'_{m,ij}$. Since $A_{ij} \sim Bernoulli(\mathscr{A}_{ij})$, then $\bar{\rho}_n \leq 1$. For any m,

$$\mathbb{E}[(A_{ij} - \mathscr{A}_{ij})^m] = \mu_{m,ij} \leq |\mu'_{m,ij}| = |\mathscr{A}_{ij}| \leq \bar{\rho}_n, \forall i, j. \tag{136}$$

J.1.6 Arguments of Lemma G.6 under DC-SBM

Since each $Z_i, i \in [k]$ has only one non-zero entry, for $\forall r \in \mathbb{R}^k, i \in [k]$, we have

$$\mathbb{E}\exp(t\langle Z_i, r\rangle) = \mathbb{E}\exp(t\sum_{j=1}^{k} Z_{ij}r_j) = \mathbb{E}\exp(tZ_{i,z(i)}r_{z(i)}) = \Pi_{j=1}^{k}\mathbb{E}\exp(tZ_{ij}r_j).$$

□

J.2 Proofs for Corollary C.2

Proof. In current LDA settings, we need Assumption 2 in Theorem 6.1 on $\widetilde{Z}_* = Z_* - \mathbb{E}(Z_*)$, which is already implied in Corollary C.2 setup. Recall that:

$$\mathbb{E}(\breve{A}|\Xi, Z) = \widetilde{Z}_*(\sqrt{n}\Sigma^{1/2})(n^{-1/2}\beta^T). \tag{137}$$

Compared with the semi-parametric factor model in Definition 1, $\sqrt{n}\Sigma^{1/2}$ plays the role of block matrix B and satisfies all the conditions in Theorem 6.1. Other than that Assumption 3 needs to be checked to prove Equation (23) and the $2 \to \infty$ norm of Y (in Equation (137), this is $(n\rho_n)^{-1}\Sigma^{1/2}\beta^T$) needs to be bounded by $O(\log d) \asymp O(\log n)$. After that we will show the error bound for topics estimation.

Notice that s controls the scaling of the term-document matrix, the following inference reflects its relation to Δ_n. Recall that

$$\rho_n = \frac{1}{nd}\sum_{i,j}\mathscr{A}_{ij}, \quad \Delta = n\rho_n.$$

And,

$$\mathbf{1}_n^T\mathscr{A}\mathbf{1}_d = \mathbf{1}_n^T\Xi Z\beta^T\mathbf{1}_d = \mathbf{1}_n^T\Xi Z\mathbf{1}_k = \mathbf{1}_n^T\Xi\mathbf{1}_n. \tag{138}$$

Equation (138) implies that

$$\Delta_n = n\rho_n = \frac{n}{nd}\mathbf{1}_n^T\mathscr{A}\mathbf{1}_d = \frac{1}{d}\mathbf{1}_n^T\Xi\mathbf{1}_n \asymp s. \tag{139}$$

J.2.1 Assumption 3 Under Poisson Random Variables

Suppose A_{ij} has m-th central moment being $\mu_{m,ij}$. Since $A_{ij} \sim Poisson(\mathscr{A}_{ij})$, recall the recurrence relation of poisson distribution (Riordan [1937]),

$$\mu_{m+1,ij} = \mathscr{A}_{ij}(\frac{d\mu_{m,ij}}{d\mathscr{A}_{ij}} + m\mu_{m-1,ij}), \quad \mu_1 = 0, \mu_2 = \mathscr{A}_{ij}, \forall i, j.$$

It could be shown by induction that

$$\mu_{m,ij} \leq (m-1)! \times \max\{\mathscr{A}_{ij}^{[\frac{m}{2}]}, \mathscr{A}_{ij}\}. \tag{140}$$

Thus, Assumption 3 is satisfied for Poisson.

J.2.2 Upper Bound for $\|n^{-1/2}\beta^T\|_{2\to\infty}$

Notice for arbitrary j-th row of $n^{-1/2}\beta^T$, it has ℓ_2-norm

$$n^{-1/2}\sqrt{\sum_{\ell=1}^{k}\beta_{\ell j}^2} \leq n^{-1/2}\sqrt{\sum_{\ell=1}^{k}\beta_{\ell j}} = n^{-1/2}.$$

Therefore, $\|n^{-1/2}\beta^T\|_{2\to\infty} = O(n^{-\frac{1}{2}})$, which is much smaller than $O(\log n)$.

J.2.3 Topics Estimation

For technical convenience, this proof uses an equivalent construction of $\widehat{\beta}$. Define $\Omega = (\widehat{Z}^T\widehat{Z})^{-1}\widehat{Z}^T\breve{A} = \Phi/n$ and $\widehat{\beta} = (\Lambda_o^{-1}\Omega)^T \in \mathbb{R}^{d\times k}$, where Λ_o is a diagonal matrix with ith diagonal element equals to ℓ_1-norm of ith row of Ω.

For the topic estimation $\widehat{\beta}$, from Equation (137) there is,

$$\breve{\mathscr{A}}^T\breve{\mathscr{A}} = n\beta\Sigma^{1/2}(\widetilde{Z}_*^T\widetilde{Z}_*/n)\Sigma^{1/2}\beta^T.$$

By LLN we have $(\widetilde{Z}_*^T\widetilde{Z}_*/n)[i,j] = \mathbb{1}\{i=j\} + O(1/\sqrt{n})$. Notice that the jth diagonal element of $\Sigma_{jj} = \alpha_j s^2 \succeq n\rho_n$. Also $\sigma_{\min}(\beta) > c_1 > 0$, and

$$\sigma_{\max}(\beta) = \|\beta\| < \|\beta\|_F = (\sum_{ij}\beta_{ij}^2)^{\frac{1}{2}} \leq (\sum_{ij}\beta_{ij})^{\frac{1}{2}} = k^{\frac{1}{2}}$$

is upper bounded. Therefore

$$\sigma_{\min}(\breve{\mathscr{A}}) \asymp \sigma_{\min}((n\beta\Sigma\beta^T)^{\frac{1}{2}}) \asymp \sqrt{n}s \succeq n\rho_n.$$

With conclusions of Proposition G.2, Lemma G.3, G.4, Davis-Kahan $\sin\Theta$ Theorem, Equation (23) and triangle inequality, there exists $P_n \in \mathcal{P}(k)$ (similar to the P_n in Equation (23)) s.t. for any $\delta, \epsilon > 0$,

$$\begin{aligned}
\|\Omega - P_n^T\Sigma^{1/2}\beta^T\|_{2\to\infty} &\leq \frac{1}{n}\left[\|(\widehat{Z}^T(\check{A} - \check{\mathscr{A}})\|_{2\to\infty} + \|(\widehat{Z}^T\widetilde{Z}_* - P_n^T)\Sigma^{1/2}\beta^T\|_{2\to\infty}\right] \\
&\leq \frac{1}{n}\left[\|\widehat{Z}^T\|\|\check{A} - \check{\mathscr{A}}\| + \|\widehat{Z}^T\|\|\widetilde{Z}_* - \widehat{Z}P_n^T\|\|\Sigma^{1/2}\beta^T\|\right] \\
&= \frac{1}{n}\left[\|\widehat{Z}^T\|\|\check{A} - \check{\mathscr{A}}\| + \|\widehat{Z}^T\|\|\widehat{Z} - \widetilde{Z}_*P_n\|\|\Sigma^{1/2}\beta^T\|\right] \\
&\leq \frac{1}{n}\left[\|\widehat{Z}^T\|\|\check{A} - \check{\mathscr{A}}\| + \sqrt{n}\|\widehat{Z}^T\|\|\widehat{Z} - \widetilde{Z}_*P_n\|_{2\to\infty}\|\Sigma^{1/2}\beta^T\|\right] \\
&= O_p(\frac{\Delta_n^{1/2}\log^{5/2} n}{n}) + O_p(\frac{\Delta_n^{3/4+\delta/2}\log^{15/4} n}{\sqrt{n}}) \\
&= O_p(\frac{\Delta_n^{3/4+\delta/2}\log^{15/4} n}{\sqrt{n}}).
\end{aligned}$$

Let Ω_ℓ be the ℓth row of Ω, ζ_ℓ be the ℓth row of $P_n^T\Sigma^{1/2}\beta^T$. Then for $\forall \ell \in [k]$ there exists $\varepsilon_n = O_p((\Delta_n^{3/4+\delta/2}\log^{15/4} n)/\sqrt{n})$ s.t. with high probability,

$$\|\Omega_\ell - \zeta_\ell\| \leq \varepsilon_n \Rightarrow \|\Omega_\ell - \zeta_\ell\|_1 \leq \sqrt{d}\varepsilon_n. \tag{141}$$

Notice any ℓ-th column of β has unit norm: $\|\beta_\ell\|_1 = 1, \forall \ell \in [k]$. Denotes $\alpha_{\min} = \min_j \alpha_j, \alpha_{\max} = \max_j \alpha_j$, then RHS of Equation (141) reflects

$$s\sqrt{\alpha_{\min}} - \sqrt{d}\varepsilon_n \leq \|\Omega_\ell\|_1 \leq s\sqrt{\alpha_{\max}} + \sqrt{d}\varepsilon_n. \tag{142}$$

With Equation (142) and notice the j-th diagonal element of $\Sigma^{1/2}$ is $s\sqrt{\alpha_j}$, we also have

$$\max_{j,\ell\in[k]} |\Sigma_{jj}^{1/2} - \|\Omega_\ell\|_1| \leq \sqrt{d}\varepsilon_n. \tag{143}$$

Since LHS of (142) is greater than 0 with large n. Let $[X]_\ell$ represents the ℓ-th row of

matrix X. Then for $\forall \ell \in [k]$,

$$\begin{aligned}
\|\widehat{\beta}_\ell^T - [P_n^T \beta^T]_\ell\|_1 &= \|\frac{\Omega_\ell}{\|\Omega_\ell\|_1} - [P_n^T \beta^T]_\ell\|_1 \\
&\leq \frac{1}{\|\Omega_\ell\|_1}\|\Omega_\ell - \zeta_\ell\|_1 + \|[P_n^T(\frac{\Sigma^{1/2}}{\|\Omega_\ell\|_1} - 1)\beta^T]_\ell| \\
&\leq \frac{\sqrt{d}\varepsilon_n}{s\sqrt{\alpha_{\min}} - \sqrt{d}\varepsilon_n} + \frac{1}{\|\Omega_\ell\|_1} \times \max_{j\in[k]} |\Sigma_{jj}^{1/2} - \|\Omega_\ell\|_1| \\
&\leq \frac{2\sqrt{d}\varepsilon_n}{s\sqrt{\alpha_{\min}} - \sqrt{d}\varepsilon_n} \\
&= O_p(\sqrt{d}\varepsilon_n/s) \\
&= O_p(\Delta_n^{-1/4+\delta/2} \log^{15/4} n).
\end{aligned}$$

□