Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1037–1060
https://doi.org/10.1093/jrsssb/qkad029
Advance access publication 7 July 2023
Discussion Paper

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Vintage factor analysis with Varimax performs statistical inference*

Karl Rohe and Muzhe Zeng

UW Madison Department of Statistics, 1300 University Ave, Madison, WI 53706, USA

Address for correspondence: Karl Rohe, UW Madison Department of Statistics, 1300 University Ave, Madison, WI 53706, USA. Email: karl.rohe@wisc.edu

*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

In the 1930s, Psychologists began developing Multiple-Factor Analysis to decompose multivariate data into a small number of interpretable factors without any a priori knowledge about those factors. In this form of factor analysis, the Varimax factor rotation redraws the axes through the multi-dimensional factors to make them sparse and thus make them more interpretable. Charles Spearman and many others objected to factor rotations because the factors seem to be rotationally invariant. Despite the controversy, factor rotations have remained widely popular among people analyzing data. Reversing nearly a century of statistical thinking on the topic, we show that the rotation makes the factors easier to interpret because the Varimax performs statistical inference; in particular, principal components analysis (PCA) with a Varimax rotation 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. 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. PCA with Varimax is fast, stable, and practical. Combined with Thurstone's straightforward diagnostics, this vintage approach is suitable for a wide array of modern applications.

Keywords: factor analysis, independent component analysis, Little Jiffy, orthoblique, spectral clustering

1 Introduction

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 (Kaiser, 1958; Thurstone, 1935, 1947). 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 (Jolliffe, 2002; Kaiser, 1958).

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

[1] PCA is not the preferred approach in Vintage Factor Analysis. See Remark 6.3 for a further discussion.

Received: March 31, 2023. **Accepted:** March 31, 2023

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., Bartholomew et al., 2011; Johnson & Wichern, 2007; Ramsay & Silverman, 2007). 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 (Feller, 1971 Chapter 3, Section 4; Maxwell, 1860). 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 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 5. 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, ICA, non-negative matrix decompositions, latent Dirichlet allocation, and stochastic blockmodelling.

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 2a plots a pair of principal components against one another. Each panel in Figure 2b plots these components after the Varimax rotation (i.e., with the Varimax axes). Section 3 describes this procedure in more detail. See Section 4.1 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 2b, 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 (F. Chen & Rohe, 2020) for further discussion.

Theorem 7.1, the main result of this paper, shows that `vsp` can estimate the following semi-parametric factor model.

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

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

This model is semi-parametric because it does not make parametric assumptions on the distribution of Z, Y or $A \mid Z, Y$. Section C in the online supplementary 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 $\mathbb{E}(A \mid Z, Y)$. However, if the elements of Z are independently generated from a

[2] A common point of confusion is to presume that the factors must be Gaussian if we are using PCA; see Section 6 and Remark 6.1 to see how PCA performs with non-Gaussian factors.

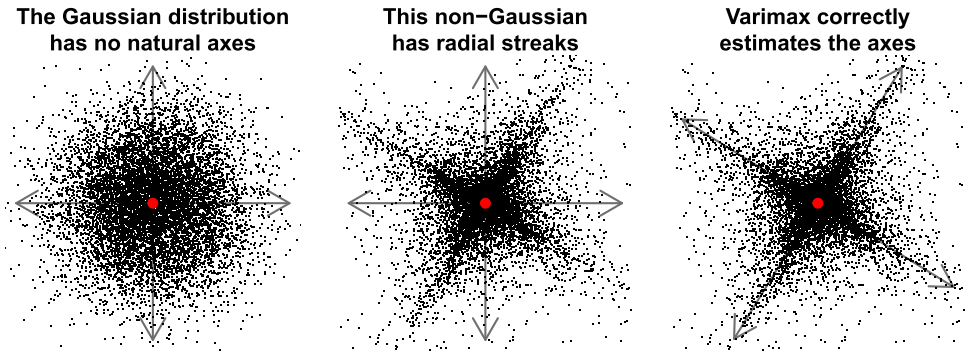

Figure 1. Maxwell's Theorem characterizes the multivariate Gaussian distribution (left panel) as the only rotationally invariant distribution of independent variables. The centre panel and the right panel give the same data; the only difference is that the right panel gives the axes that Varimax estimates.

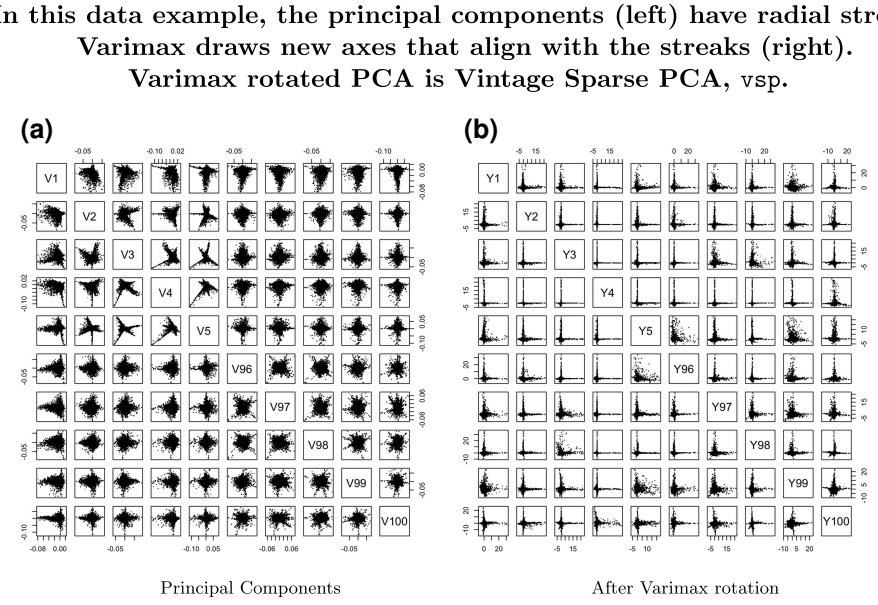

Figure 2. In this example, the data is a 22, 688 × 22, 688 matrix of citations among 22,688 academic journals. Each small panel in (a) is a scatter plot of two principal components. Each small panel in (b) is a scatter plot of two Varimax rotated components. See Section 4.1 for more details.

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 centred 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 & Zenga, 2009). See Section 5 for further discussion of leptokurtosis.

After reading Section 2 and the algorithm in Section 3, one can read Sections 4, 5, 6 and 7 in any order. Section 8 should be read after Sections 6.1 and 6.2.

Section 2 introduces Varimax and gives both algebraic and geometric intuition for why it prefers 'sparse axes'. Section 3 describes the `vsp` algorithm and some variations on the algorithm. Section 4 illustrates how to interpret the results of `vsp` by applying it to a large citation network and a large text corpus. Section 5 provides intuition for the sparsity diagnostics developed in Thurstone (1935, 1947) to show that they implicitly assess the leptokurtic assumption. Section 6 gives the population results for PCA with latent variable models and population results for Varimax applied to these population principal components. Section 7 gives the main theoretical result, Theorem 7.1. Section 8 discusses what happens when the latent variables are not independent.

1.1 Key notation

Let $\mathcal{O}(k) = \{R \in \mathbb{R}^{k\times k} : R^T R = RR^T = I_k\}$ 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 $[k] = \{1, \ldots, k\}$.

2 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 pre-processing 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 & 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 equation (2). In `R`, `varimax` is optimized via projected gradient ascent.

2.1 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

[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.

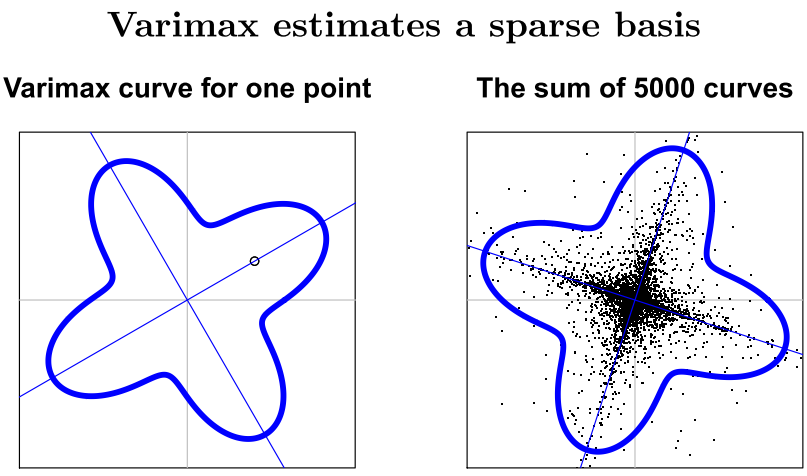

Figure 3. These curves are in polar coordinates, where the radius of the curve is the Varimax objective value for that angle. The optimal axes are displayed in blue. These axes provide an approximately sparse representation for the points because most points are close to the axes.

single point on the unit sphere in higher dimensions, $x \in \mathbb{R}^d$,

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

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.

In the right panel of Figure 3, there are 5,000 points $x_1, \ldots, x_{5,000}$ 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, $\sum_{i=1}^{5,000} v(\theta, x_i)$. The sum of the flowers is displayed as the thick blue line in the right panel. The optimal axes are the thin lines at a skewed angle; importantly, these new axes align with the radial streaks in this data.

3 `vsp`: Vintage Sparse PCA

This section describes the methodological details of Vintage Sparse PCA (`vsp`). First, the algorithm is stated. Then, Remarks 3.1 and 3.2 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.

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

Option	Motivated when …
Centring	factor modelling, topic modelling, soft-clustering.
	See Remarks 3.2 and 6.2, Section 8.1.2 and online supplementary material, Section C.4
Recentring	the factor means are desired.
	See Theorem 7.1, Remark 6.2, online supplementary material, Section F.1.
Avoid centring	hard-clustering, Stochastic Blockmodelling.
	See Section 8.1.2 and online supplementary material, Section 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 3.1.

1. **Centring** (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 centring 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 centring 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 the online supplementary material, Remark B.
3. **Varimax**. Compute the orthogonal matrices that maximize Varimax, $v(R, \widehat{U})$ and $v(R, \widehat{V})$. Define them as $R_{\hat{U}}, R_{\hat{V}} \in \mathcal{O}(k)$ respectively.
   – Output:

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

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 3.1 (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 (Amini et al., 2013; Chaudhuri et al., 2012; 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 artefact of noise in sparse matrices (Y. Zhang & Rohe, 2018). In this paper, the degree-normalization step is used for the analyses in Section 4, but it is not studied in the main theorem.

The optional centring step (step 1 of `vsp`) plays a surprising role. In particular, Proposition 6.1 in Section 6 shows that if A is centred in step 1, then `vsp` estimates the centred factors in the semi-parametric factor model (i.e., $Z - \mathbb{E}(Z)$). See Remark 6.2 and Section 8.1.2 for more discussion. To estimate Z, instead of its column centred version, the output of `vsp` can be recentred as follows.

Remark 3.2 (Optional recentring step). After running `vsp` with the centring step, it is possible to use the quantities already computed to recentre 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

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

and recentre 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 3.1 is also used, then recentre before renormalizing. Section 6 and online supplementary material, Appendix F.1 justify the estimator $\widehat{\mu}_Z$.

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

4 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 4.1 applies `vsp` to the citation patterns among academic journals. Then, in order to understand where and how 'factor analysis' is used, Section 4.2 applies `vsp` to all abstracts that contain the phrase 'factor analysis'.

4.1 `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, centring, and recentring.

In Figure 2 and in this sub-section, 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 1/1,000 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*,

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

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

	JRSS-B	*AOS*	*JASA*	*AOP*	*Nature*	*PLOS ONE*	*PNAS*	*NEJM*
Column sum	178	146	462	59	4,413	3176	3,928	3,209
Row sum	16	45	51	28	522	5,556	1,283	284

PLOS ONE, *Proceedings of the National Academy of Sciences (PNAS)*, and *The New England Journal of Medicine (NEJM)*.

Because the column and row sums of A have a heavy tail, we used the degree-normalization described in Remark 3.1. The sparsity in the data matrix makes `vsp` quick to compute. In `R`, on a 2.3 GHz Macbook Pro, it takes 1.3 s for $k = 10$, 13 s for $k = 50$, and 23 s 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., 2016). 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 with $k = 100$ are axis aligned. These later factors are more diffuse, suggesting that they contain more noise.[6]

4.1.1 *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 & 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
2. molecular, cell, biology, immunology, microbiology, genetics, nature
3. psychology, psychiatry, neuroscience, brain, neurology, behaviour, 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

[6] In later work, F. 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.

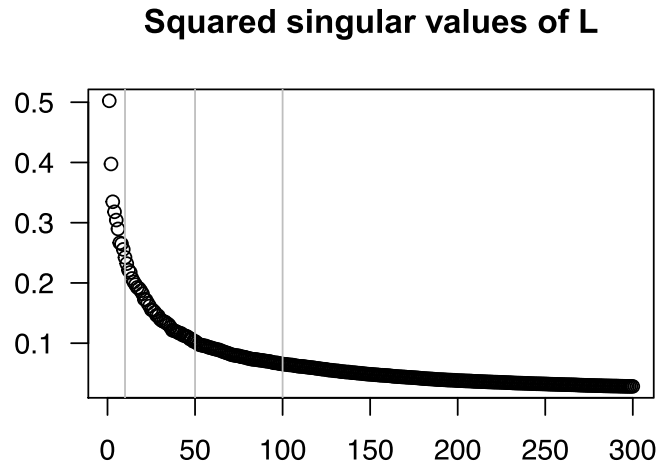

Figure 4. The first 300 squared singular values of L are plotted, along with lines at $k = 10$, 50, and 100.

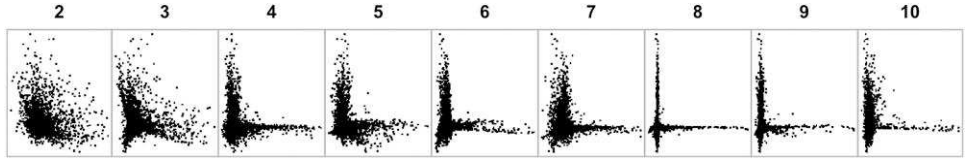

Figure 5. Each dot is a journal. The vertical axis gives factor 1, the 'medicine' factor. The horizontal axis gives the other factors, 2, ..., 10. If you squint, some panels have multiple horizontal streaks (e.g., factor 6); with a larger choice of k, these streaks reveal themselves to be buds that unfurl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

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.

The factors in $\widehat{Z}$ identify the same academic areas as $\widehat{Y}$. The leading `bff` terms for $\widehat{Z}$ are given in the online supplementary material, Section D.1 and the top 11 journals for both $\widehat{Z}$ and $\widehat{Y}$ are given in the online supplementary material, 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*, *NEJM*, *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*.

4.1.2 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 the online

[7] It does appear in Harris and Kaiser (1964); see Section 8.2 for more.

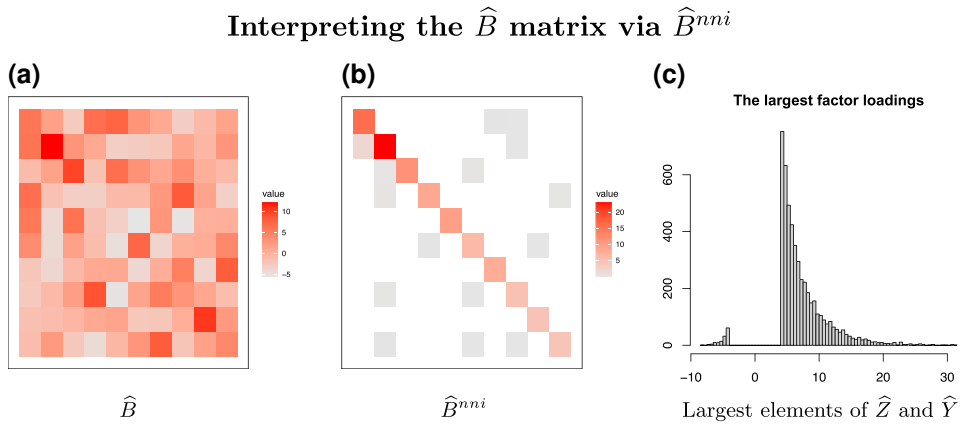

Figure 6. The $\widehat{B}$ matrix can be hard to interpret. $\widehat{B}^{nni}$ provides a clearer picture; it is constructed by thresholding away the negative values in $\widehat{Z}$, $\widehat{Y}$ as in equation (6). The right panel justifies this thresholding, by showing that the largest values in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

supplementary material, 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. Y. 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 $\widehat{B}$.

4.1.3 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 not perfectly hierarchical. This can be seen in the loadings for *AOP* 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

[8] The skewness for a random variable is $\eta_3/\eta_2^{3/2}$ where the η's are the centred 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.

Table 2. For each of the $k = 100$ factors, this table gives the top `bff` term

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	anaesthesia	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	

Note. While there are 9 `bff` terms that repeat for more than one factor (e.g., mathematical), these repeated factors identify sub-disciplines within these areas (e.g., one of the math factors finds 'applied math' journals and the other appears to find 'pure math' journals). See the online supplementary material, Appendix D.3 for the leading five journals in all 100 factors.

and 9. Meanwhile, *AOP* has its largest loading value in factor 8 (mathematics). None of these journals rank among the highest 40 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 the online supplementary material, Appendix D.3.

4.1.4 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 the online supplementary material, 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 something 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

4.2 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 et al., 2020).

4.2.1 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.

4.2.2 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 & 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 4.1, 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 s.[9] For comparison, constructing A from the 144, 136 abstracts represented as character strings requires 23 s in tidytext.

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

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 4.2.1 (e.g., Environment, Nutrition, Psychology, etc). See the online supplementary material, Section E.1 for these factors. The final 11 factors are artefacts and anomolies that are further discussed in the online supplementary material, 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–modelling(sem): equation, structural, modelling, 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

5 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 5.1 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, $R_2 \in \mathcal{O}(k)$, 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 $R_{\widehat{U}} \in \mathcal{O}(k)$ 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 5.1 describes these diagnostic practices. Section 5.2 and Theorem 5.1 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 6.1 and 7.1.

5.1 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' (Thurstone, 1947, p. 333). 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' (Thurstone, 1947, p. 334). Thurstone's five conditions are quoted in the online supplementary material, 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.

5.1.1 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 (Bartholomew et al., 2011; Johnson & Wichern, 2007; Jolliffe, 2002; Mardia et al., 1979). 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 5.1 in the next subsection shows that sparsity implies leptokurticity, the key identifying assumption for Varimax.

5.2 Kurtosis and sparsity

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

Theorem 5.1 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 5.1 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 5.1 does not provide a sharp results for the Bernoulli distribution, .83 is close to .79. Moreover, Theorem 5.1 applies to any random variable, it does not make any parametric assumptions, and the moment assumptions are only so that kurtosis is defined. See the online supplementary material, 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 the online supplementary material, Appendix G.2 for a discussion about softer forms of sparsity.

6 Mathematical intuition for `vsp` with population results

This section studies each of the three steps in `vsp` by studying their population behaviour. Statistical convergence around the population quantities is rigorously treated in Theorem 7.1 in Section 7.

6.1 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 6.1 and 6.2 study the first two steps of `vsp` applied to the population matrix

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

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 centring 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 6.2 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 6.3 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 6.1 shows that the rotation that maximizes this function is $R^T \in \mathcal{O}(k)$. So, rotating the population principal components with the population Varimax rotation yields the original factors, $(\widetilde{Z}R)R^T = \widetilde{Z}$.

6.2 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 6.1 (Step 1 of vsp). Centring $\mathscr{A}$ to get $\widetilde{\mathscr{A}}$ as in equation (3) has the effect of centring 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 the online supplementary material, 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 $\widetilde{R}_U, \widetilde{R}_V \in \mathcal{O}(k)$, 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 6.2 (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 6.1 and 6.2 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 6.1 (Relationship between PCA and the factors) Proposition 6.2 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 centred 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 6.1 below shows that Varimax can identify the nuisance matrix.

6.3 Population results for Varimax

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

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

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 equation (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 6.1) and the main theorem (Theorem 7.1).

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 6.1 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

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

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

Theorem 6.1 (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 $\widetilde{R} \in \mathcal{O}(k)$,

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

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 3.2 describes a method to recentre the factors $\widetilde{Z}$ to get Z. The Online supplementary material, Section F.1 in the Appendix gives a population justification for this recentring step.

Remark 6.2 (The role of centring). A version of Proposition 6.2 still holds for the SVD of $\mathscr{A}$ (without centring) by replacing $\widehat{\Sigma}_Z$ with Z^TZ/n and replacing $\widehat{\Sigma}_Y$ with

[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, $\mathscr{A} = ZBY^T$.

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 centring in `vsp`; centring $\mathscr{A}$ has the effect of centring 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 8.

Remark 6.3 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 $\mathscr{A} = \mathbb{E}(A \mid Z, Y) = ZBY^T$ 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 $n^{-1}\mathbb{E}(A^TA \mid Z, Y)$. 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; A. R. 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.

7 The main theorem

Theorem 7.1 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 3.2. Let $\widehat{Z}_i$ be the ith row of $\widehat{Z}$. Theorem 7.1 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 $\epsilon > 0$ neighbourhood 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 7.1 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,

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

where this expectation is conditional on Z, Y.

Assumption 3 controls the tail behaviour 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 the online supplementary material, Section C.4. This assumption is also satisfied if A contains Bernoulli random variables, as happens in Stochastic Blockmodelling. See the online supplementary material, 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 7.1. 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 7.1 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 $\Delta_n \geq \log^{11.1} n$,

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

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 7.1 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 recentring by $\mathbf{1}_n\widehat{\mu}_Z$, then a similar result holds for estimating 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 7.1 begins in the online supplementary material, Appendix G.3. online supplementary material, Corollaries C.1 and C.2 in the Appendix extend Theorem 7.1 to the Stochastic Blockmodel and Latent Dirichlet Allocation.

8 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 of a common data processing step that is not included in `vsp`.

`vsp` easily handles correlated factors; Section 8.1 gives more intuition for how and why. Then, Sections 8.1.1 and 8.1.2 describe how two data analytic choices can create the appearance of non-orthogonal factors (even when the factors are independent). Section 8.2 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., F. Chen & Rohe, 2020).

8.1 `vsp` can handle correlated factors

Proposition 6.1 and Proposition 6.2 do not make any probabilistic assumptions; both are simply results of linear algebra. Together, these propositions show that if the data matrix is not centred,

then the principal components are

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

for some rotation matrix R_U. Alternatively, if the data matrix is centred, then the principal components are a function of the centred 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 $(Z^TZ)^{-1/2}$. 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 $Z(Z^TZ)^{-1/2}$. This is good news. If underlying correlated factors had radial streaks, those radial streaks will be preserved in $Z(Z^TZ)^{-1/2}$. 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 & Kaiser, 1964; Kaiser, 1970). This differs from `vsp` only in some pre-processing 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.

8.1.1 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 sub-section describes how the common practice of scaling the eigenvectors creates the appearance of non-orthogonal factors, *even if the factors are independent*. Then, Subsection 8.1.2 explores one (necessarily unexciting) place where the remaining 1% from Kaiser's calculation might come from.

For simplicity, suppose that we are not centring 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,

$$UR^* = Z\big(R_U^TR_U\big) = Z.$$

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

[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 Section 6.3.

rotation, then the two rotation matrices cannot cancel out like they do above,

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

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 6.2). 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 & 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 the online supplementary material, Section C.3. However, if the structure in Z is not this nice, (Harris & 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 7.1 shows that rotating the principal components with Varimax *can* reveal these structures.

8.1.2 The role of centring

The appearance of non-orthogonal factors can also happen as a result of improper centring. 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 centring, 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 centred, 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 centre the data matrix, compute the principal components, and rotate with Varimax, then we can hope to recover $\widetilde{Z}$, then uncentre 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 & 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, centring *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 centring), 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 centring alone cannot fix.

Case IV (Mean zero factors): If the latent factors already have mean zero, centring will not change anything.

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

8.2 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 Section 4.1.2 and online supplementary material, Section 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 centred, 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 & Kaiser, 1964) and (Kaiser & 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.

9 Discussion

PCA with Varimax is a vintage data analysis technique. Theorem 7.1 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 6. The leptokurtosis condition is satisfied if the factors are sparse and this condition can be examined in the data. In fact, Section 5 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.

Acknowledgments

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 Dongzhou Huang for helpful feedback. 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.

Conflict of interest: None declared.

Funding

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.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Amini A. A., Chen A., Bickel P. J., & Levina E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, *41*(4), 2097–2122.

Ammar W., Groeneveld D., Bhagavatula C., Beltagy I., Crawford M., Downey D., Dunkelberger J., Elgohary A., Feldman S., & Ha V., *et al.* (2018). 'Construction of the literature graph in semantic scholar', arXiv, arXiv:1805.02262, preprint: not peer reviewed.

Anandkumar A., Ge R., Hsu D., Kakade S. M., & Telgarsky M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, *15*(1), 2773–2832.

Anderson T. W., & Rubin H. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability* (Vol. 5, pp. 111–150). University of California Press.

Bartholomew D. J., Knott M., & Moustaki I. (2011). *Latent variable models and factor analysis: A unified approach*. Wiley Series in Probability and Statistics. Wiley.

Bertsimas D., Copenhaver M. S., & Mazumder R. (2017). Certifiably optimal low rank factor analysis. *The Journal of Machine Learning Research*, *18*(1), 907–959.

Bickel P. J., & Chen A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, *106*(50), 21068–21073. https://doi.org/10.1073/pnas.0907096106

Chaudhuri K., Chung F., & Tsiatas A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory* (pp. 35–1).

Chen F., Roch S., Rohe K., & Yu S. (2021). 'Estimating graph dimension with cross-validated eigenvalues', arXiv, arXiv:2108.03336, preprint: not peer reviewed.

Chen F., & Rohe K. (2020). 'A new basis for sparse pca', arXiv, arXiv:2007.00596, preprint: not peer reviewed.

Feller W. (1971). *An introduction to probability theory and its applications* (Vol. 2). John Wiley and Sons, Inc.

Fiori A. M., & Zenga M. (2009). Karl Pearson and the origin of kurtosis. *International Statistical Review*, *77*(1), 40–50. https://doi.org/10.1111/j.1751-5823.2009.00076.x

Harris C. W., & Kaiser H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, *29*(4), 347–362. https://doi.org/10.1007/BF02289601

Holland P. W., Laskey K. B., & Leinhardt S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*(2), 109–137. https://doi.org/10.1016/0378-8733(83)90021-7

Hyvärinen A., Karhunen J., & Oja E. (2004). *Independent component analysis* (Vol. 46). John Wiley & Sons.

Jin J., & Ke Z. T. (2017). 'A sharp lower bound for mixed-membership estimation', arXiv, arXiv:1709.05603, preprint: not peer reviewed.

Johnson R. A., & Wichern D. W. (2007). *Applied multivariate statistical analysis*. Pearson Education International, Pearson Prentice Hall.

Jolliffe I. T. (2002). *Principal component analysis*. Springer Series in Statistics. Springer.

Kaiser H. F. (1958). The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. https://doi.org/10.1007/BF02289233

Kaiser H. F. (1970). A second generation Little Jiffy. *Psychometrika*, *35*(4), 401–415. https://doi.org/10.1007/BF02291817
Kaiser H. F., & Rice J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115
Le C. M., Levina E., & Vershynin R. (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms*, *51*(3), 538–561. https://doi.org/10.1002/rsa.20713
Lei L., Li X., & Lou X. (2020). 'Consistency of spectral clustering on hierarchical stochastic block models', arXiv, arXiv:2004.14531, preprint: not peer reviewed.
Mardia K. V., Kent J. T., & Bibby J. M. (1979). *Multivariate analysis*. Probability and Mathematical Statistics. 10th printing in 1995. Academic Press.
Maxwell J. C. (1860). 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. https://doi.org/10.1080/14786446008642818
Ramsay J. O, & Silverman B. W. (2005). *Functional data analysis*. Springer Series in Statistics. Springer.
Ramsay J. O., & Silverman B. W. (2007). *Applied functional data analysis: Methods and case studies*. Springer.
Revelle W. (2017). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, https://CRAN.R-project.org/package=psych. R package version 1.7.8.
Rohe K., Qin T., & Yu B. (2016). Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, *113*(45), 12679–12684. https://doi.org/10.1073/pnas.1525793113
Salcianu A. *et al.* (2020). *Google compact language detector v3 (cld3)*. https://doi.org/10.21105/joss.00037
Silge J., & Robinson D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *The Journal of Open Source Software*, *1*(3), 37. https://doi.org/10.21105/joss.00037
Thurstone L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press.
Thurstone L. L. (1947). *Multiple factor analysis*. University of Chicago Press.
Vu V. Q., & Lei J. (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, *41*(6), 2905–2947. https://doi.org/10.1214/13-AOS1151
Wang S, & Rohe K. (2016). Coauthorship and citation networks for statisticians: Comment. *The Annals of Applied Statistics*, *10*(4), 1779–1812. doi:10.1214/16- AOAS977
Zhang A. R., Cai T. T., & Wu Y. (2018). 'Heteroskedastic PCA: Algorithm, optimality, and applications', arXiv, arXiv:1810.08316, preprint: not peer reviewed.
Zhang Y., Levina E., & Zhu Ji (2014). 'Detecting overlapping communities in networks using spectral methods' arXiv, arXiv:1412.3432, preprint: not peer reviewed.
Zhang Y., & Rohe K. (2018). 'Understanding regularized spectral clustering via graph conductance', arXiv, arXiv:1806.01468, preprint: not peer reviewed.

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1061–1098

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Discussion Paper Contribution

Proposer of the vote of thanks to Rohe & Zeng and contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference'

Peter Hoff

Department of Statistical Science, Duke University

Address for correspondence: Peter Hoff, Department of Statistical Science, Duke University, Box 90251, Duke University, Durham NC 27708-0251, USA. Email: peter.hoff@duke.edu

I congratulate the authors on their article, which brings together theory and methods from the established areas of factor analysis and independent components analysis, combines these with new results, and applies the combination to the developing research area of matrix-variate data analysis.

To many statisticians, factor analysis is simply a model that specifies that the covariance matrix Σ of a d-variate population is equal to a rank-k matrix $M = HH^\top$ (with $k < d$) plus a diagonal matrix D. An appealing feature of such a model is that it can have many fewer parameters than has an unrestricted covariance model. But to many researchers, the appeal of the factor analysis model is that it might actually be true! As is well known, if the d-vector of variables y is equal to some unknown linear combination H of k uncorrelated common factors plus uncorrelated noise, that is, $y = Hf + D^{1/2}e$ where $\text{Var}[f] = I_k$ and $\text{Var}[e] = I_d$, then indeed the variance of y is $M + D$ where $M = HH^\top$.

In the literature that I am familiar with, the conundrum of classical factor analysis has been that interest is really in the factor loading matrix H and not M, but M is typically estimated from only the first two moments of the data, from which H is not recoverable. The VARIMAX rotation selects an H which may be preferred on aesthetic grounds, or because of some assumptions about the process being studied, or because of the invariance properties of VARIMAX—the latter being the original motivation in Kaiser (1958).

While the focus of factor analysis is primarily on the factor loading matrix, or 'column effects', in independent components analysis (ICA) the goal is to recover the uncorrelated latent factors, or the 'row effects.' This is done by whitening the data by multiplying each data vector by a matrix H^{-1} for any H such that $\text{Var}[y] = HH^\top$. But which whitening matrix to use? If the distribution of the latent factors is invariant to orthogonal rotation (i.e., the distribution is Gaussian), then the correct H cannot be determined. The assumption of ICA is that the factors (or all but one of them) are non-Gaussian, and if this is correct, the latent factors may well estimated by choosing a whitening matrix that makes the estimated factors maximally non-Gaussian by some measure.

Standard multivariate statistical procedures, like factor analysis or ICA, typically focus on describing heterogeneity along only one of the two index sets of the data matrix. However, there has been rapid growth in the number of applications where the data are 'matrix-variate' and inferences about the row objects *and* the column objects are desired. For example, the rows and columns could represent biological samples and genes, or consumers and products, or exporters and

Received: March 31, 2023. **Accepted:** April 3, 2023

importers. Analyses of such data are often based on matrix factorisation models like the one presented by Rohe and Zeng:

$$A \approx ZBY^{\top}$$

$$a_{i,j} = z_i^{\top} B y_j,$$

where A is the data matrix, the values of $z_1, \ldots, z_n$ represent heterogeneity along the rows, and $y_1, \ldots, y_d$ represent heterogeneity along the columns. As with the identification of the loading matrix in factor analysis, the conundrum of matrix-variate data analysis is that there are infinitely many matrix factorizations that give the same low-rank least-squares approximation to A. How to select from among them? One approach is to abandon least-squares and instead infer the correct factorisation based on specific distributional assumptions about the latent row and column factors, often using a parametric statistical model. Such approaches can incorporate subject-matter knowledge about the factors into the estimation procedure, but are typically very computationally intensive. Alternatively, standard matrix factorisation methods, such as the singular value decomposition, are relatively inexpensive computationally, but they select a factorisation using arbitrary identifiability constraints that are not derived from the data.

What has been needed is a data analysis method that is computationally inexpensive, but also identifies the factors using information from the data. I propose a vote of thanks to Rohe and Zeng for providing just such a method. As they show in their article, the classic VARIMAX criterion, applied to rows or columns of a data matrix, can identify rotations that recover non-Gaussian latent factors. Their results unify several multivariate statistical methods, and highlight that much of what might be thought of as multivariate data analysis should really be considered as matrix-variate data analysis.

Conflict of interest: None declared.

References

Kaiser H. F. (1958). The VARIMAX criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. https://doi.org/10.1007/BF02289233

https://doi.org/10.1093/jrsssb/qkad030
Advance access publication 4 April 2023

Seconder of the vote of thanks to Rohe & Zeng and contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference'

Marianna Pensky

Department of Mathematics, University of Central Florida, Orlando, United States

Address for correspondence: Marianna Pensky, Department of Mathematics, University of Central Florida, Orlando, FL 32817, USA. Email: marianna.pensky@ucf.edu

We would like to congratulate the authors on publication of a truly seminal paper. Indeed, they managed to accomplish a rare and extremely valuable task: take a technique, Varimax, that has been used for half a century for generating sparse PCA, provide conditions for its applicability

Table 1. Δ_Z for the spectral clustering in Lei and Rinaldo (2015), vsp and adjusted vsp, averaged over 1,000 runs (standard deviations in parentheses)

Estimation in the stochastic block model					
n	a	w	Clustering	vsp	Adjusted vsp
100	0.5	0.6	0.1840 (0.0554)	0.2538 (0.0202)	0.1834 (0.0577)
200	0.5	0.6	0.0396 (0.0421)	0.1776 (0.0095)	0.0404 (0.0428)
300	0.5	0.6	0.0052 (0.0170)	0.1448 (0.0065)	0.0052 (0.0170)
400	0.5	0.6	0.0004 (0.0047)	0.1248 (0.0045)	0.0004 (0.0047)
500	0.5	0.6	0.0000 (0.0014)	0.1118 (0.0037)	0.0000 (0.0000)
100	0.5	0.8	0.5904 (0.0723)	0.5732 (0.0775)	0.5893 (0.0732)
200	0.5	0.8	0.3940 (0.0449)	0.3805 (0.0316)	0.3926 (0.0447)
300	0.5	0.8	0.2772 (0.0321)	0.3012 (0.0155)	0.2765 (0.0321)
400	0.5	0.8	0.2004 (0.0257)	0.2573 (0.0105)	0.2002 (0.0258)
500	0.5	0.8	0.1476 (0.0224)	0.2285 (0.0077)	0.1479 (0.0225)
100	0.25	0.6	0.4877 (0.0787)	0.4736 (0.0748)	0.4848 (0.0795)
200	0.25	0.6	0.2678 (0.0383)	0.3041 (0.0184)	0.2667 (0.0383)
300	0.25	0.6	0.1600 (0.0301)	0.2431 (0.0106)	0.1603 (0.0306)
400	0.25	0.6	0.0989 (0.0275)	0.2097 (0.0076)	0.0996 (0.0271)
500	0.25	0.6	0.0600 (0.0275)	0.1868 (0.0058)	0.0601 (0.0274)
100	0.25	0.8	0.6633 (0.0335)	0.6655 (0.0377)	0.6626 (0.0338)
200	0.25	0.8	0.6502 (0.0396)	0.6421 (0.0461)	0.6502 (0.0393)
300	0.25	0.8	0.6010 (0.0562)	0.5824 (0.0632)	0.6009 (0.0564)
400	0.25	0.8	0.5112 (0.0526)	0.4858 (0.0528)	0.5112 (0.0527)
500	0.25	0.8	0.4345 (0.0319)	0.4138 (0.0243)	0.4340 (0.0318)

and produce the error bounds. They name this new version Vintage Sparse PCA (vsp). In particular, if $X = ZBY^T$, where components of $Z = \{Z_{i,j}\}$ and $Y = \{Y_{i,j}\}$ are independent zero mean unit variance leptokurtic random variables, and rows of matrices Z and Y are identically distributed, then matrices Z and Y are identifiable, and Varimax allows one to do this. The paper provides very elegant arguments why kurtosis $\kappa > 3$ leads to identifiability of matrices Z and Y. Applications of vsp include, among others, Independent Component analysis, Stochastic Block Model (SBM), Degree-Corrected Stochastic Block Model (DCBM), Overlapping, Mixed Membership and Degree-Corrected Mixed Membership Stochastic Block Models, and sparse dictionary learning.

Since each of the above research areas developed its own techniques, it would be interesting to see how vsp performs for specific types of problems. The authors do not provide any numerical examination of the precision of the vsp in various scenarios (due to their sheer multitude and the fact that the complete paper is already over 100 pages). Therefore, we carry out a limited simulation study that complements the paper.

Specifically, we study three simulations scenarios. Scenario 1 considers SBM with $k = 2$ communities, where Z is a clustering matrix with exactly one 1 per row and $Y = Z$. Scenario 2 examines DCBM, where again $Z = Y$ and matrix $Z = \Theta W$, Θ is a diagonal and W is a clustering matrix. Scenario 3 deals with matrices $Z \in \mathbb{R}^{n\times k}$ and $Y \in \mathbb{R}^{n\times d}$ comprised of independent T random variables with ν degrees of freedom. In the first two scenarios, we generated clusters using multinomial distribution with equal probabilities. Elements of Θ are generated as Uniform on [0, 1]. For SBM and DCBM, the diagonal and nondiagonal elements of matrix B are, respectively, equal to a and wa. For Scenario 3, elements of B are Uniform on [0, 1].

In order to make matrices Z and Y identifiable, we renormalise Z to have column norms $\sqrt{n}$ with the respective readjustment of matrix B. We choose $k = d = 2$, vary n, w, a, and set $X = ZBY^T$. Since $\mathbb{E}(\Theta_{i,i}) = 0.5$, values $a = 0.5$ and $a = 0.25$ for SBM corresponds to $a = 1.0$ and $a = 0.5$ for

Table 2. Δ_Z for the spectral clustering in Gao et al. (2018), vsp and adjusted vsp, averaged over 1,000 runs (standard deviations in parentheses)

Estimation in the degree corrected stochastic block model					
n	a	w	Clustering	vsp	Adjusted vsp
100	1.0	0.6	0.4349 (0.1014)	0.4527 (0.0627)	0.4374 (0.0806)
200	1.0	0.6	0.2277 (0.0462)	0.3138 (0.0251)	0.2683 (0.0327)
300	1.0	0.6	0.1543 (0.0280)	0.2555 (0.0161)	0.2055 (0.0188)
400	1.0	0.6	0.1185 (0.0176)	0.2209 (0.0117)	0.1718 (0.0122)
500	1.0	0.6	0.0973 (0.0131)	0.1975 (0.0093)	0.1505 (0.0091)
100	1.0	0.8	0.8816 (0.0941)	0.8611 (0.0963)	0.8857 (0.0981)
200	1.0	0.8	0.7075 (0.1113)	0.6810 (0.1061)	0.7026 (0.1125)
300	1.0	0.8	0.5224 (0.0687)	0.5253 (0.0446)	0.5267 (0.0561)
400	1.0	0.8	0.4108 (0.0510)	0.4488 (0.0286)	0.4321 (0.0385)
500	1.0	0.8	0.3392 (0.0404)	0.3991 (0.0215)	0.3706 (0.0290)
100	0.5	0.6	0.8321 (0.1152)	0.8077 (0.1215)	0.8318 (0.1232)
200	0.5	0.6	0.5725 (0.0862)	0.5545 (0.0622)	0.5691 (0.0744)
300	0.5	0.6	0.4073 (0.0546)	0.4382 (0.0301)	0.4255 (0.0416)
400	0.5	0.6	0.3116 (0.0406)	0.3767 (0.0206)	0.3468 (0.0291)
500	0.5	0.6	0.2489 (0.0327)	0.3366 (0.0161)	0.2967 (0.0221)
100	0.5	0.8	0.9418 (0.0566)	0.9393 (0.0580)	0.9579 (0.0577)
200	0.5	0.8	0.9407 (0.0530)	0.9291 (0.0585)	0.9539 (0.0570)
300	0.5	0.8	0.9148 (0.0638)	0.8978 (0.0743)	0.9263 (0.0710)
400	0.5	0.8	0.8718 (0.0779)	0.8467 (0.0873)	0.8780 (0.0845)
500	0.5	0.8	0.7981 (0.0861)	0.7660 (0.0920)	0.7991 (0.0913)

Table 3. Δ_Z and Δ_Y, averaged over 1,000 runs (standard deviations in parentheses), for the vsp in the case of T distribution with ν degrees of freedom and no random errors

Estimation for T-random matrices, no noise					
n	d	ν	κ	Δ_Z	Δ_Y
100	200	5	9	0.1712 (0.1415)	0.1197 (0.1058)
200	400	5	9	0.1229 (0.1091)	0.0869 (0.0753)
300	600	5	9	0.0984 (0.0830)	0.0688 (0.0541)
400	800	5	9	0.0839 (0.0726)	0.0589 (0.0516)
500	1,000	5	9	0.0772 (0.0626)	0.0540 (0.0456)
100	200	10	4	0.2586 (0.1867)	0.2102 (0.1699)
200	400	10	4	0.2068 (0.1663)	0.1508 (0.1399)
300	600	10	4	0.1743 (0.1575)	0.1332 (0.1296)
400	800	10	4	0.1519 (0.1451)	0.1228 (0.1322)
500	1,000	10	4	0.1338 (0.1245)	0.1182 (0.1419)
100	200	16	3.5	0.2886 (0.2033)	0.2616 (0.1924)
200	400	16	3.5	0.2718 (0.2025)	0.2253 (0.1889)

(*continued*)

Table 3. Continued

Estimation for T-random matrices, no noise					
n	d	v	κ	Δ_Z	Δ_Y
300	600	16	3.5	0.2439 (0.1925)	0.2279 (0.1968)
400	800	16	3.5	0.2253 (0.1878)	0.2375 (0.2175)
500	1,000	16	3.5	0.2255 (0.1944)	0.2762 (0.2412)
100	200	28	3.25	0.3378 (0.2094)	0.3168 (0.2130)
200	400	28	3.25	0.3287 (0.2125)	0.2902 (0.2104)
300	600	28	3.25	0.3045 (0.2157)	0.3055 (0.2193)
400	800	28	3.25	0.2932 (0.2089)	0.3616 (0.2345)
500	1,000	28	3.25	0.2966 (0.2153)	0.4184 (0.2430)

Table 4. Δ_Z and Δ_Y, averaged over 1,000 runs (standard deviations in parentheses), for the vsp in the case of T distribution with $v = 5$ degrees of freedom and iid Gaussian random errors with zero mean and standard deviation σ

Estimation for T-random matrices, $v = 5$, noise level σ				
n	d	σ	Δ_Z	Δ_Y
100	200	0.1	0.2209 (0.2064)	0.2022 (0.2090)
200	400	0.1	0.1721 (0.1799)	0.1399 (0.1753)
300	600	0.1	0.1355 (0.1634)	0.1198 (0.1614)
400	800	0.1	0.1246 (0.1626)	0.1104 (0.1657)
500	1,000	0.1	0.1027 (0.1241)	0.0926 (0.1317)
100	200	0.2	0.2850 (0.2540)	0.2665 (0.2553)
200	400	0.2	0.1870 (0.1934)	0.1757 (0.1918)
300	600	0.2	0.1641 (0.1968)	0.1576 (0.2020)
400	800	0.2	0.1486 (0.1912)	0.1446 (0.1959)
500	1,000	0.2	0.1376 (0.1811)	0.1333 (0.1873)
100	200	0.3	0.3178 (0.2716)	0.3077 (0.2718)
200	400	0.3	0.2342 (0.2454)	0.2306 (0.2486)
300	600	0.3	0.1923 (0.2262)	0.1960 (0.2337)
400	800	0.3	0.1701 (0.2061)	0.1736 (0.2129)
500	1,000	0.3	0.1503 (0.1868)	0.1563 (0.2009)
100	200	0.4	0.3604 (0.2886)	0.3537 (0.2940)
200	400	0.4	0.2624 (0.2659)	0.2633 (0.2684)
300	600	0.4	0.2292 (0.2521)	0.2331 (0.2617)
400	800	0.4	0.1959 (0.2368)	0.1991 (0.2424)
500	1,000	0.4	0.1725 (0.2184)	0.1822 (0.2262)

DCBM. We generate A as symmetric matrix with independent Bernoulli entries for SBM and DCBM, while $A = X + \sigma\Xi$ where Ξ has iid standard Gaussian entries for Scenario 3.

In Scenarios 1 and 2, we compare vsp with the spectral clustering algorithms in Lei and Rinaldo (2015) and Gao et al. (2018), respectively, where matrix $\hat{Z}$ is based on clustering assignment and estimator $\hat{\Theta}$ of Θ. We add the third estimator, adjusted vsp, which leaves only the largest (in absolute value) element of the vsp estimator $\hat{U}$ in each row and renormalise $\hat{Z}$ accordingly. For DCBM, we adjust $\hat{Z}$ to the column norms $\sqrt{n}$. Note that, for Scenarios 1 and 2, all three algorithms

recover Z perfectly if matrix X is available. Scenario 3 is remarkably different from 1 and 2 since vsp does not recover Z and Y exactly from matrix X, and there is no 'yardstick' algorithm for comparison. Hence, for Scenario 3, we study performance of vsp only, for both X and A, which corresponds to $\sigma = 0$ and $\sigma > 0$.

Results of simulations are presented in Tables 1–4. The errors are measured as Frobenius norms $\Delta_Z = \|\hat{Z} - Z\|_F/\sqrt{n\,k}$ and $\Delta_Y = \|\hat{Y} - Y\|_F/\sqrt{n\,d}$, averaged over 1,000 runs. The standard deviations of the means are reported in parentheses.

Tables 1 and 2 confirm that the algorithms designed specifically for SBM and DCBM have better precision that vsp since they 'know' that matrix Z has only one nonzero element per row. However, adjusted vsp, which makes use of this information, performs very similarly to algorithms specifically designed for SBM and DCBM, with clustering algorithm of Gao et al. (2018) being slightly more precise in the case of DCBM. Hence, adjusted vsp can be used for clustering in the SBM (with average miss-classification proportion Δ_Z^2). The errors grow as a decreases and w increases due, respectively to sparsity increase and decline of the signal-to-noise ratio.

Table 3 shows that, as ν grows and kurtosis $\kappa = 3 + 6/(\nu - 4)$ decreases, precision of the vsp declines, even when exact matrix X is available. Therefore, for $\sigma > 0$, we carry out simulations only with $\nu = 5$ ($\kappa = 9$). We set $d = 2n$ for various choices of n. Tables 3 and 4 demonstrate that small kurtosis can be as much of a problem for recovering Z and X as noise. Indeed, errors for small κ do not decline as n and d grow as they do for larger κ and σ.

Conflict of interest: None declared.

References

Gao C., Ma Z., Zhang A. Y., & Zhou H. H. (2018). Community detection in degree-corrected block models. *Annals of Statistics*, *46*(5), 2153–2185. https://www.jstor.org/stable/26542860

Lei J., & Rinaldo A. (2015). Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, *43*(1), 215–237. https://doi.org/10.1214/14-AOS1274

The vote of thanks was passed by acclamation.

https://doi.org/10.1093/jrsssb/qkad031
Advance access publication 6 April 2023

Joshua Cape's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Joshua Cape

Department of Statistics, University of Wisconsin–Madison

Address for correspondence: Joshua Cape, 1250A Medical Sciences Center, 1300 University Avenue, Madison, WI 53706, USA. Email: jrcape@wisc.edu

I congratulate Professor Rohe and Dr Zeng on their illuminating paper. Their broad contributions will no doubt redouble contemporary research activity in multivariate analysis for years to come. Even the paper's appendices are full of valuable gems, not to be overlooked by readers.

Rohe and Zeng contextualize their findings vis-à-vis classical developments in psychometrics and statistics. This choice is apt, for the authors champion a newfound understanding of varimax factor rotations as inferential, not merely exploratory. In a secondary capacity, the paper further complements a burgeoning body of contemporary work dedicated to the entrywise study of eigenvector estimation, perturbations, and asymptotics, in network analysis (e.g., Abbe et al., 2020; Tang & Priebe, 2018) and in high-dimensional statistics (e.g., Cape et al., 2019b; Fan et al., 2018). Under certain data-generating conditions, these works characterize the behaviour of individual vectors and their coordinates in low-dimensional point cloud representations of data, in a much more precise manner than was considered in past decades, let alone during the time of Thurstone or Kaiser.

Rohe and Zeng emphasize two forms of sparsity in their treatment of stochastic blockmodel random graphs (SBMs) in the appendix: network sparsity, and latent variable sparsity. The contemporary literature has heretofore largely overlooked the latter and its implications, notably in papers dedicated to adjacency spectral embedding in the time since (Sussman et al., 2012).

For SBMs, Rohe and Zeng establish a high-probability uniform error bound which I summarize as $\max_{1\le i\le n} \|\widehat{Z} - ZP_n\|_{i,\ell_2} = o_{\mathbb{P}}(1)$ in the moderately sparse regime $n\rho_n \gtrsim (\log n)^c$. Here $\|\cdot\|_{i,\ell_2}$ denotes the ℓ_2 vector norm of the i-th row of a given matrix, while $n\rho_n$ reflects nodal expected degree. Consequently, perfect clustering (discrimination) is achieved using $\widehat{Z}$; more interestingly, the sparse latent variable matrix Z can be element-wise directly estimated uniformly well.

More can be said. The eigenvector analysis in Cape et al. (2019a) and Rubin-Delanchy et al. (2022) hints that asymptotic normality might hold for the varimax-based estimator at the scaling $\sqrt{n\rho_n}$. Indeed, it has recently been established in Cape (2022) that for certain SBM graphs, loosely speaking, the i-th row vector of $\sqrt{n\rho_n}(\widehat{Z} - ZP_n)$ is conditionally asymptotically multivariate Gaussian, with block-specific asymptotic covariance matrix.

Unanswered questions abound. Among them, it would be interesting to quantify the estimation performance of varimax rotations in dimension $\widehat{k} \neq k_{\text{true}}$ and when the coefficient matrix B is rank degenerate.

Conflict of interest: None declared.

References

Abbe E., Fan J., Wang K., & Zhong Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, *48*(3), 1452–1474. https://doi.org/10.1214/19-AOS1854

Cape J. (2022). On varimax asymptotics in network models and spectral methods for dimensionality reduction. Under review.

Cape J., Tang M., & Priebe C. E. (2019a). Signal-plus-noise matrix models: Eigenvector deviations and fluctuations. *Biometrika*, *106*(1), 243–250. https://doi.org/10.1093/biomet/asy070

Cape J., Tang M., & Priebe C. E. (2019b). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Annals of Statistics*, *47*(5), 2405–2439. https://doi.org/10.1214/18-AOS1752

Fan J., Wang W., & Zhong Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, *18*(207), 1–42.

Rubin-Delanchy P., Cape J., Tang M., & Priebe C. E. (2022). A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society, Series B*, *84*(4), 1446–1473. https://doi.org/10.1111/rssb.12509

Sussman D. L., Tang M., Fishkind D. E., & Priebe C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, *107*(499), 1119–1128. https://doi.org/10.1080/01621459.2012.699795

Tang M., & Priebe C. E. (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Annals of Statistics*, *46*(5), 2360–2415. https://doi.org/10.1214/17-AOS1623

https://doi.org/10.1093/jrsssb/qkad032
Advance access publication 5 April 2023

Kaizheng Wang's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Kaizheng Wang

Department of IEOR and Data Science Institute, Columbia University

Address for correspondence: Kaizheng Wang, Department of IEOR and Data Science Institute, Columbia University, 500 W 120th St Room 419, New York, NY 10027, USA. Email: kaizheng.wang@columbia.edu

I congratulate Drs. Rohe and Zeng for their elegant paper that advances our understanding of the classical Varimax algorithm. Varimax was originally proposed by Kaiser (1958) as an analytic criterion for determining interpretable factors. Despite its extensive applications and tremendous success, theoretical understanding is still lacking. Statisticians are further perplexed by negative results showing the nonidentifiability of Gaussian factor models. Toward resolving the dilemma, Rohe and Zeng proved that Varimax performed on the spectral embedding of data indeed recovers meaningful structures under a wide class of semi-parametric factor models. Below I briefly summarize their key insights and discuss new avenues of research.

The authors assumed that the observed data $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ satisfies $\mathbb{E}(\boldsymbol{A}|\boldsymbol{Z}, \boldsymbol{Y}) = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Y}^\top$, where $\boldsymbol{Z} \in \mathbb{R}^{n\times K}$, $\boldsymbol{Y} \in \mathbb{R}^{d\times K}$ are random latent factor matrices and $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ a deterministic matrix. The ith row $\boldsymbol{z}_i$ of $\boldsymbol{Z}$ consists of latent factors of the ith sample, e.g., community membership of an individual, topics of a document, etc. The authors showed that if elements of $\boldsymbol{z}_i$ are independent and *leptokurtic*, then Varimax consistently estimates them. Interestingly, leptokurticity can be guaranteed by sparsity (Theorem 4.1), which is ubiquitous in high-dimensional models.

The paper opens up exciting possibilities for further study. Recall that Varimax searches for a $K \times K$ orthonormal matrix that maximizes the sum of kurtoses of the K transformed factors. However, kurtosis maximization fails to recover factors if the leptokurticity assumption is violated. As a toy example, consider $K = d = 2$, $\boldsymbol{A} = \boldsymbol{Z}$ and let $\{\boldsymbol{z}_i\}_{i=1}^n$ be i.i.d. from the degenerate Gaussian mixture $\frac{1}{2}N(\boldsymbol{e}_1, \boldsymbol{e}_2\boldsymbol{e}_2^\top) + \frac{1}{2}N(-\boldsymbol{e}_1, \boldsymbol{e}_2\boldsymbol{e}_2^\top)$. The two canonical coordinates correspond to independent factors and the first one is platykurtic. Elementary calculation shows that as $n \to \infty$, Varimax rotates the data by 45 degrees and returns noninterpretable coordinates. So, how should one choose the objective function in general? Hyvarinen (1997) showed advantages of several objectives over the kurtosis. More research is needed for modern statistical models. Another question is whether one should compute an orthonormal matrix in one shot or K orthogonal directions one by one. The latter may have computational advantage while the former is more stable numerically. Moreover, they generally yield different solutions even as $n \to \infty$. Their statistical properties are worth investigating. Last but not least, the nonconvex objective functions call for studies of efficient algorithms with convergence guarantees.

Conflict of interest: None declared.

References

Hyvarinen A. (1997). One-unit contrast functions for independent component analysis: A statistical analysis. In *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*. IEEE.

Kaiser H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. https://doi.org/10.1007/BF02289233

https://doi.org/10.1093/jrsssb/qkad033
Advance access publication 5 April 2023

Rungang Han and Anru R. Zhangs contribution to the Discussion of 'Vintage factor analysis with varimax performs statistical inference' by Rohe & Zeng

Rungang Han and Anru R. Zhang

Duke University

Address for correspondence: Anru R. Zhang, 2424 Erwin Road, Durham, NC 27705, USA. Email: anru.stat@gmail.com

We wholeheartedly congratulate Drs. Rohe and Zeng for their insightful paper on vintage factor analysis with Varimax rotation. Varimax rotation is a basic scheme to simplify the expression of a particular subspace and is included in build-in standard packages *stats* in R and *PROC FACTOR* statement in SAS. Drs. Rohe and Zeng nicely show that the principal component analysis with Varimax rotation actually performs statistical inference for the explainable factors.

Drs. Rohe and Zeng suggested leptokurtosis as a key identifiability condition for Varimax rotation; the number of factors often increases as the data dimension and sample size grow. It is thus natural to ask whether Varimax works with vanishing leptokurtosis and/or a growing number of factors. This note discusses when Varimax recovers the subspace rotation in such high-dimensional regimes. As a first step, we assume the factor matrix $Z \in \mathbb{R}^{n\times k}$ includes a collection of i.i.d. centered random variables satisfying

$$\mathbb{E}Z_{ij} = \mathbb{E}Z_{ij}^3 = 0, \quad \mathbb{E}Z_{ij}^2 = 1, \quad \mathbb{E}Z_{ij}^4 = \kappa > 3. \tag{1}$$

We also assume Z_{ij}'s are sub-Gaussian such that $\mathbb{E}\exp(\lambda Z_{ij}) \le e^{c\lambda^2}$, $\forall \lambda \in \mathbb{R}$ for some constant $c > 0$. Let $\hat{Z} \in \mathbb{R}^{p\times r}$ be the observed factors generated as

$$\hat{Z} = ZR^*,$$

where R^* is an unknown k-dimensional orthogonal matrix that represents the rotation to be recovered. Due to vanishing mean of Z, we focus on the following centred Varimax:

$$\hat{R} = \underset{R\in\mathcal{O}(k)}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j=1}^{k}\left(\left(\hat{Z}R^\top\right)_{ij}\right)^4,$$

where $\mathcal{O}(k)$ is the set of k-by-k orthogonal matrices. Consider the following error metric:

$$\operatorname{dist}(R^*, \hat{R}) = \min_{P\in\mathcal{P}(k)} \frac{\|\hat{R} - PR^*\|_{\mathrm{F}}}{\|R^*\|_{\mathrm{F}}} = k^{-1/2} \min_{P\in\mathcal{P}(k)} \|\hat{R}R^{*\top} - P\|_{\mathrm{F}},$$

where $\mathcal{P}(k) = \mathcal{O}(k) \cap \{0, \pm 1\}^{k^2}$ is the set of orthogonal matrices that allow for column reordering and sign changes as defined in Equation (12). The following theorem characterizes the conditions of n, k, κ under which Varimax works or fails.

Theorem 1 [1]Let $\delta \in (0, 1/2]$ be any fixed value and $\kappa \le C_0$ for some universal constant $C_0 > 3$.

[1] See the online full version at `arXiv preprint arXiv:2205.10151` for a complete proof.

- If $n \gtrsim \max\{\frac{k\log n}{(\kappa-3)^2}, \frac{k^2\log^2 n}{\kappa-3}\}$, then

$$\limsup_{n\to\infty} \mathbb{P}(\mathrm{dist}(R^*, \hat{R}) \geq \delta) = 0;$$

- if $k < n \lesssim k^2$, then

$$\liminf_{n,k\to\infty} \mathbb{P}(\mathrm{dist}(R^*, \hat{R}) \geq \delta) = 1.$$

If the kurtosis $\kappa > 3$ is a constant, Theorem 1 suggests that a necessary and nearly sufficient condition for consistent rotation recovery is $n \gtrsim k^2$—this bound is tight up to $\log^2(n)$. Theorem 1 also shows when the leptokurtosis of factors is insignificant, i.e., $\kappa \to 3^+$, a sufficient sample size to ensure rotation recovery by Varimax is $\max\{k\log(n)/(\kappa-3)^2, k^2\log^2(n)/(\kappa-3)\}$, while it is unclear if this bound is sharp. It is of future interest to investigate the tight condition that guarantees the consistency of Varimax.

https://doi.org/10.1093/jrsssb/qkad034
Advance access publication 5 April 2023

Alexander Van Werde's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Alexander Van Werde

Eindhoven University of Technology

Address for correspondence: Alexander Van Werde, Department of Mathematics and Computer Science, Eindhoven University of Technology. Email: a.van.werde@tue.nl

I would like to point out that the author's results have potential applications in the theory of random graphs beyond those for block models which are explicitly remarked upon in the paper. In particular, the results could be used to resolve the nonidentifiability associated with random dot product graphs (Athreya et al., 2017, Remark 1).

The authors are most likely already aware of this connection: they have previously worked on generalized random dot product graphs in Rohe et al. (2018). Thus, I would like to invite the authors to share further insights concerning the connection between this work and random dot product graphs in their reply. Some concrete questions are as follows:

Q1: Does the assumption of leptokurtic entries for the latent positions appear to be satisfied in real-world graphs?

Q2: Are there other considerations which one should take into account when applying the theory of the current paper to random dot product graphs? For instance, are there situations where one should be careful?

For the sake of the reader unfamiliar with the topic, I will now sketch the connection between random dot product graphs and the current paper. A *random dot product graph* with latent positions $X_1, \ldots, X_n \in \mathbb{R}^k$ is a random graph whose adjacency matrix $A \in \mathbb{R}^{n \times n}$ has independent entries with $A_{ij} \sim \text{Bernoulli}(\langle X_i, X_j \rangle)$. Here, it should be assumed that the latent positions are such that $\langle X_i, X_j \rangle \in [0, 1]$ for all $i, j \in \{1, \ldots, n\}$. The key problem in the theory is to estimate the latent positions $X_1, \ldots, X_n$ given an observation of A. There is however an unidentifiability in the model: for any orthogonal transformation $R \in O(k)$ it holds that $\langle RX_i, RX_j \rangle = \langle X_i, X_j \rangle$. Thus, if no additional assumptions are made concerning the nature of the latent positions, then estimation is only possible up to an orthogonal transformation.

If, however, it is further assumed that the latent positions are generated as $X_i = BZ_i$ where $Z_1, \ldots, Z_n \in \mathbb{R}^k$ are i.i.d. random vectors with independent and leptokurtic components of variance one and $B \in \mathbb{R}^{k \times k}$ is a fixed matrix. Then, with $Z := [Z_1, \ldots, Z_n]^{\mathrm{T}}$ it holds that

$$\mathbb{E}[A \mid Z] = Z(B^{\mathrm{T}}B)Z^{\mathrm{T}}$$

which falls in the scope of the current paper. Hence, estimation of the Z_i and the kernel $B^{\mathrm{T}}B$ is possible up to permutations and sign flips. Viewing the Z_i as the true latent positions may lead to more interpretable representations of the graph since there is less ambiguity and more sparsity. The corresponding random graph model, allowing for a kernel $B^{\mathrm{T}}B$, is called a generalized random dot product graph.

References

Athreya A., Fishkind D. E., Tang M., Priebe C. E., Park Y., Vogelstein J. T., Levin K., Lyzinski V., Qin Y., & Sussman D. L. (2017). Statistical inference on random dot product graphs: A survey. *The Journal of Machine Learning Research*, *18*(1), 8393–8484.

Rohe K., Tao J., Han X., & Binkiewicz N. (2018). A note on quickly sampling a sparse matrix with low rank expectation. *The Journal of Machine Learning Research*, *19*(1), 3040–3052.

https://doi.org/10.1093/jrsssb/qkad035
Advance access publication 25 May 2023

Yinqiu He, Yuqi Gu and Zhilian Ying's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Yinqiu He[1], Yuqi Gu[2] and Zhiliang Ying[3]

[1]Department of Statistics, University of Wisconsin-Madison, Madison, USA
[2]Department of Statistics, Columbia University, New York, USA
[3]Columbia University, New York, USA

Address for correspondence: Yinqiu He, Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA.
Email: yinqiu.he@wisc.edu

We congratulate the authors on an impressive paper that demystifies the popular Varimax rotation. They showed that the Varimax rotation can identify true axes of latent factors under the

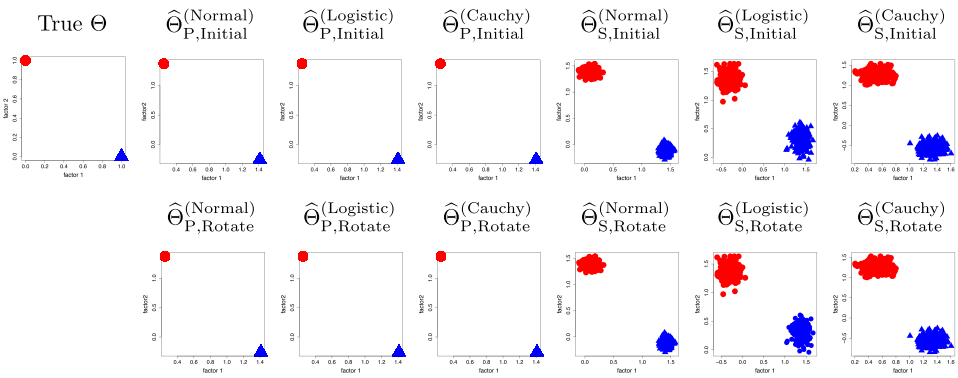

Figure 1. Setting (I): $N = J = 500$, $K = 2$. True factors $\Theta = (\theta_{ik})$ are binary and give orthogonal columns: $(\theta_{i1}, \theta_{i2})$ are i.i.d. Multinomial(0.5, 0.5). Data are not centred, as factors are orthogonal. The loading matrix Λ has i.i.d. entries following $U(-2, 2)$. Point i in red circle if $\theta_{i1} = 0$, and blue triangle otherwise. The first figure in the first row gives the scatterplot of true factors Θ, and the other figures plot estimated factors that are obtained following the subtitles above themselves.

leptokurtic condition in a low-rank semiparametric factor model

$$\mathrm{E}(A_{N\times J} \mid Z_{N\times K}, Y_{J\times K}) = Z_{N\times K} \times B_{K\times K} \times Y^{\top}_{J\times K}. \tag{1}$$

Different from (1), the exploratory Item Factor Analysis (IFA), widely used in social and behavioural sciences, adopts a *nonlinear* transformation of the low-rank structure:

$$\mathrm{E}(R_{N\times J} \mid \Theta_{N\times K}, \Lambda_{J\times K}) = F(\Theta_{N\times K} \times \Lambda^{\top}_{J\times K}), \tag{2}$$

where the binary data matrix R contains N individuals' responses to J items, Θ represents N individuals' K latent factors, Λ is the loading matrix, and F is a prespecified monotone increasing function. The relationship between the two models can be seen by letting $F(x) = x$, $B = \mathrm{I}_K$, $Z = \Theta$, and $Y = \Lambda$. However, in IFA, $F(\cdot)$ is usually taken from a distribution function, e.g., $F_{\text{Normal}}(x) = \int_{-\infty}^{x} e^{-t^2/2}/(2\pi)\mathrm{d}t$ and $F_{\text{Logistic}}(x) = e^x/(1+e^x)$; see Reckase (2009). For such a model, it is of interest to develop a similar Vintage Factor Analysis with Varimax.

For (2), Zhang et al. (2020) proposed a two-step approach: (1) use SVD to obtain a low-rank approximation to the data matrix, denoted as $\widehat{R}$; (2) use $F^{-1}(\widehat{R})$ as an estimate of $\Theta\Lambda^{\top}$ and apply SVD to obtain $\widehat{\Theta}$ and $\widehat{\Lambda}$. We add a third step that incorporates the Varimax rotation: (3) apply Varimax rotation to the initial estimate $\widehat{\Theta}$. Our preliminary studies show some promising signs. Specifically, we conduct simulations under four settings of Θ, given in Figures 1–4, respectively. In each setting, we consider three $F(\cdot)$ in (2): $F_{\text{Normal}}(x)$, $F_{\text{Logistic}}(x)$, and $F_{\text{Cauchy}}(x) = \pi^{-1}\arctan(x) + 1/2$. For $\mathrm{d} \in \{\text{Normal}, \text{Logistic}, \text{Cauchy}\}$, we generate a binary data matrix R_{d} with a population mean matrix $F_{\mathrm{d}}(\Theta\Lambda^{\top})$.

(i) Population (Columns 2–4 in Figures 1–4). We apply the two-step algorithm in Zhang et al. (2020) to $F_{\mathrm{d}}(\Theta\Lambda^{\top})$ and obtain initial estimated factors $\widehat{\Theta}^{(\mathrm{d})}_{\text{P,Initial}}$. We then apply Varimax to $\widehat{\Theta}^{(\mathrm{d})}_{\text{P,Initial}}$ and obtain rotated factors $\widehat{\Theta}^{(\mathrm{d})}_{\text{P,Rotate}}$. For $\mathrm{d} \in \{\text{Logistic}, \text{Cauchy}\}$, rotated factors $\widehat{\Theta}^{(\mathrm{d})}_{\text{P,Rotate}}$ are close to true Θ, whereas $\widehat{\Theta}^{(\mathrm{d})}_{\text{P, Initial}}$ with $\mathrm{d} \in \{\text{Normal}\}$ can substantially deviate from Θ. The results show that latent axes might be statistically identified given a nonlinear F in (2). However, the performance varies with the choice of F: F_{Logistic} and F_{Cauchy} (heavy tail) outperform F_{Normal} (light tail).

(ii) Sample (Columns 5–7 in Figures 1–4). Similarly, we first apply the two-step algorithm to binary matrix R_{d} to obtain $\widehat{\Theta}^{(\mathrm{d})}_{\text{S, Initial}}$, and then $\widehat{\Theta}^{(\mathrm{d})}_{\text{S,Rotate}}$ via Varimax rotation. In Figures 1–2, binary Θ renders finite number of clusters, and the estimated factors can recover the clusters. This could be

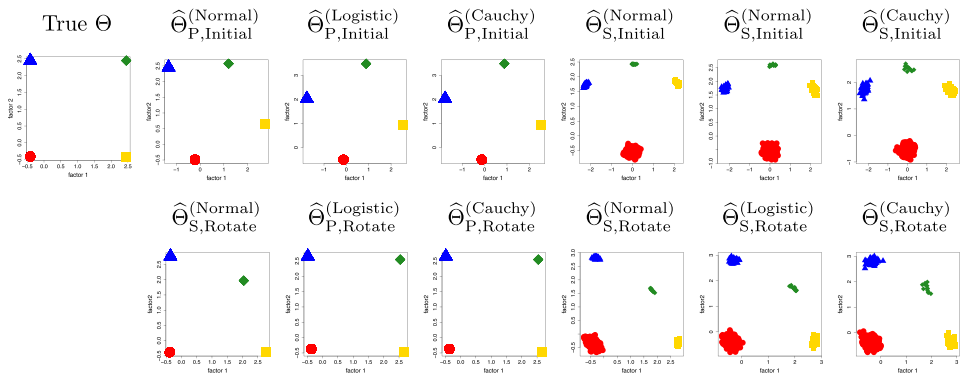

Figure 2. Setting (II): $N = J = 500$, $K = 2$. True factors $\Theta = (\theta_{ik})$ are binary and independent: for $k \in \{1, 2\}$, $\theta_{ik} = \text{norm}(\zeta_{ik})$, where ζ_{ik} are i.i.d. Bernoulli(1/7), and $\text{norm}(\zeta_{ik})$ represents normalizing ζ_{ik} by its population mean and standard deviation. Kurtosis$(\theta_{ik}) > 3$ for $k \in \{1, 2\}$. Data are centred, as true factors are mean zero. The loading matrix Λ has i.i.d. entries following $U(-2, 2)$. Let $\zeta_i = (\zeta_{i1}, \zeta_{i2})$. For $i = 1, \ldots, N$, point i is in red circle if $\zeta_i = (0, 0)$, in blue triangle if $\zeta_i = (0, 1)$, in yellow square if $\zeta_i = (1, 0)$, and in green diamond if $\zeta_i = (1, 1)$.

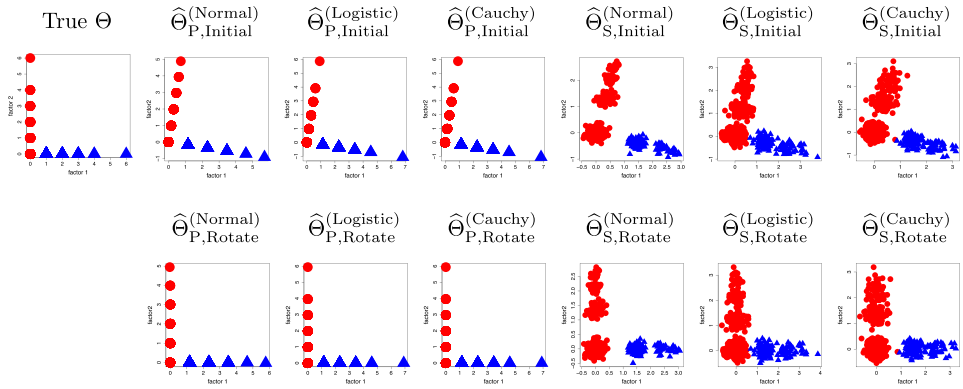

Figure 3. Setting (III): $N = J = 500$, $K = 2$. True factor $\Theta = (\theta_{ik})$ is nonbinary and give orthogonal columns: for $k \in \{1, 2\}$, $\theta_{ik} = \varrho_{ik} \times m_{ik}$, where ϱ_{ik} are i.i.d. Poisson(1), $(m_{i1}, m_{i2}) \sim$ Multinomial(0.5, 0.5), and Kurtosis$(\theta_{ik}) > 3$. Data are not centred, as factors are orthogonal. The loading matrix Λ has i.i.d. entries following $U(-2, 2)$. Point i is in red circle if $\theta_{i1} = 0$, and blue triangle otherwise.

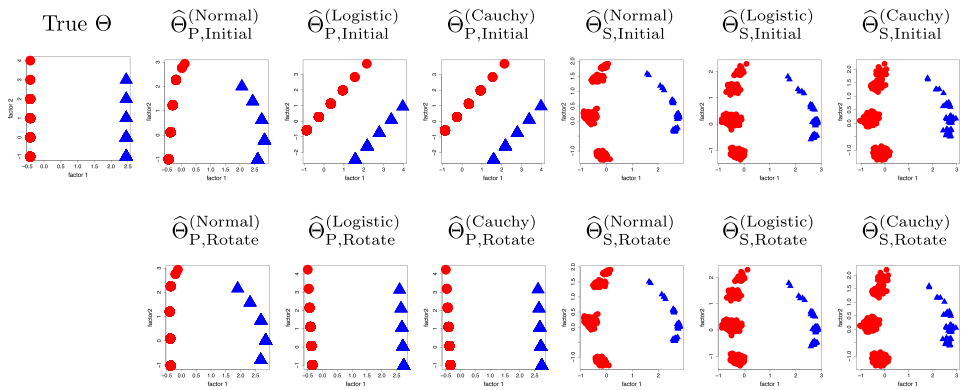

Figure 4. Setting (IV): $N = J = 500$, $K = 2$. True factors $\Theta = (\theta_{ik})$ are independent and include nonbinary: $\theta_{i1} \sim \text{norm}(\zeta_{i1})$, where ζ_{i1} are i.i.d. Bernoulli(1/7), and $\theta_{i2} \sim \rho_{i2} - 1$, where ρ_{i2} are i.i.d. Poisson(1). Data are centred, as true factors have zero mean. The loading matrix Λ has i.i.d. entries following $U(-2, 2)$. Point i is in red circle if $\zeta_{i1} = 0$, and blue triangle otherwise.

because in this case, (2) can be alternatively viewed as a Latent Class Model (Goodman, 1974), and $F_d(\Theta\Lambda^\top)$ always exhibits an exact low-rank structure even with a nonlinear F_d. This observation aligns with the authors' findings. However, in Figures 2–4, the sample estimate $\widehat{\Theta}^{(d)}_{S,Rotate}$ can deviate from true Θ more significantly than its population counterpart $\widehat{\Theta}^{(d)}_{P,Rotate}$ does. This may be due to the special signal-to-noise structure of a binary variable, i.e., its variance can always be computed from its mean, so the sampling noise can have a significant impact on the recovery accuracy. Nevertheless, comparing $\widehat{\Theta}^{(d)}_{S,Initial}$ and $\widehat{\Theta}^{(d)}_{S,Rotate}$, Varimax can indeed find one rotation aligned with the axes of the true latent factors. This suggests that similar Vintage Factor Analysis with varimax may also apply in the IFA model (2) with a nonlinear F and Θ that satisfy leptokurtic conditions.

Conflict of interest: None conflict of interest declared. The discussion does not contain real-world data. The codes of simulation that support the findings of this study are available from the author, Yinqiu He, upon reasonable request.

References

Goodman L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231. https://doi.org/10.1093/biomet/61.2.215

Reckase M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer.

Zhang H., Chen Y., & Li X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, *85*(2), 358–372. https://doi.org/10.1007/s11336-020-09704-7

The following contributions were received in writing after the meeting.

https://doi.org/10.1093/jrsssb/qkad036
Advance access publication 4 April 2023

Peter J. Bickel, Derek Bean, Aiyou Chen and Purnamrita Sarkar's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Peter J. Bickel[1], Derek Bean[2], Aiyou Chen[3] and Purnamrita Sarkar[4]

[1]Statistics, University of California, Berkeley
[2]Statistics, University of Wisconsin, Madison
[3]Statistician, Waymo LLC
[4]Statistics and Data Science, University of Texas, Austin

Address for correspondence: Peter J. Bickel, 411 Evans Hall, Berkeley, CA 94720-3860, USA.
Email: bickel@stat.berkeley.edu

We compliment the authors for their unification of old and new latent variable models. Their motivation seems to have come from Thurstone's Factor Analysis and the wish to explain the somewhat mysterious success of Kaiser's Varimax rule in that context. However, their major

applications and notation stem from overlapping network community detection and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Inference for these models can be unified in another way via a provably consistent algorithmic framework (Mao et al., 2018) because of the inherent similarities in the geometrical structure of eigenvectors of appropriate matrices. It would be interesting to see how the identifiability conditions tied to these models, which typically involve the existence of pure nodes or anchor words relate to those needed for Varimax.

We elaborate on a connection to Independent Component Analysis (ICA) (Hyvärinen, 2013; Hyvärinen et al., 2001), which they have noted, and make explicit a connection between Varimax and the FastICA algorithm of Hyvarinen (1999a). The factor analysis/ICA model we consider has n independent observations with the following structure:

$$A = BZ + E \tag{1}$$

where A and E are $d \times 1$, Z is a $K \times 1$ random vector independent of E, and B is $d \times K$, full rank and nonrandom. The elements of Z are independent with no Gaussian component, and have mean zero and unit variance. Also, $E \sim N(0, \Sigma)$, with unknown covariance matrix Σ. This noisy ICA model is a special case of the model of this paper. When $d = K$, the goal is to estimate a nonsingular C such that $CA = PZ + CE$, where P is some permutation matrix. If $E = 0$ and A is prewhitened, C (and B) are orthogonal yielding factors with Z as the vector of loadings. By a theorem of Comon (1994), B is identifiable up to a permutation as is Σ. FastICA (Hyvarinen, 1999a, 1999b), which estimates C by maximizing the sum of the squared empirical kurtoses of the coordinates of CA after prewhitening, yields consistent estimates if all the Z_j have nonzero kurtoses and $E = 0$ or Σ is known. The kurtosis assumption can be dropped (Chen & Bickel, 2005) using a different fast algorithm. The authors' main Theorem 6.1 requires 'leptokurtosis'. Is that necessary? Another latent variable model of interest is NonGaussian Components Analysis (Blanchard et al., 2006), where A is viewed as the sum of an unknown nonGaussian signal and independent Gaussian noise with the aim of estimating the linear space in which the signal lies. Is there a relation to the model of this paper?

Conflict of interest: None declared.

References

Blanchard G., Kawanabe M., Sugiyama M., Spokoiny V., & Müller K.-R. (2006). In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, *7*(9), 247–282.

Blei D. M., Ng A. Y., & Jordan M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Chen A., & Bickel P. J. (2005). Consistent independent component analysis and prewhitening. *IEEE Transactions on Signal Processing*, *53*(10), 3625–3632. https://doi.org/10.1109/TSP.2005.855098

Comon P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*(3), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9

Hyvarinen A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

Hyvarinen A. (1999b). Fast ica for noisy data using gaussian moments. In *1999 IEEE International Symposium on Circuits and Systems (ISCAS)* (Vol. 5, pp. 57–61).

Hyvärinen A. (2013). Independent component analysis: Recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984), 20110534. https://doi.org/10.1098/rsta.2011.0534

Hyvärinen A., Karhunen J., & Oja E. (2001). *Independent component analysis* (1st ed.). John Wiley & Sons.

Mao X., Sarkar P., & Chakrabarti D. (2018). Overlapping clustering models, and one (class) SVM to bind them all. In *Advances in neural information processing systems* (Vol. 31, pp. 2130–2140).

https://doi.org/10.1093/jrsssb/qkad037
Advance access publication 12 April 2023

Junhui Cai, Dan Yang, Linda Zhao and Wu Zhu's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Junhui Cai[1,*], Dan Yang[2], Linda Zhao[1] and Wu Zhu[3]

[1]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, USA
[2]Innovation and Information Management, The University of Hong Kong, Hong Kong, China
[3]Department of Finance, Tsinghua University, Beijing, China

*Present address: Department of Information Technology, Analytics, and Operations, University of Notre Dame, Notre Dame, IN, USA.
Address for correspondence: Junhui Cai, Department of Information Technology, Analytics, and Operations, University of Notre Dame, Notre Dame, IN, USA. Email: jcai2@nd.edu

We congratulate the authors on their excellent article (Rohe & Zeng, 2022). Factors and communities in networks are often hierarchically structured, as demonstrated in the academic bibliometrics example in the paper. In order to (1) identify factors/communities with hierarchical structure and (2) identify the important individuals/nodes within these factors/communities, we propose hierarchical vintage sparse PCA (Hvsp) to account for the hierarchical structure while taking advantages of vsp's capability of performing statistical inference.

Hvsp combines the idea of hierarchical clustering and vsp. Specifically, Hvsp follows a top-down hierarchical partitioning by recursively applying vsp with dimension $k = 2$ to split the nodes into two communities and eventually produce a binary tree. Compared with the existing hierarchical clustering methodologies in network analysis, Hvsp can explore the hierarchical structure but also inherits vsp's advantages in computation and interpretability. In addition, the rotated principal component provides a score of importance for each individual/node in its corresponding factor/community, analogous to the popular eigenvector centrality measure in network analysis. The detailed algorithm is described in Algorithm 1.

Algorithm 1. The hierarchical vintage sparse PCA (Hvsp) algorithm.

1. Apply vsp with dimension $k = 2$ to the network/community $\boldsymbol{A} \in \mathbb{R}^{n' \times n'}$ and obtain the factor $\hat{\boldsymbol{Y}} \in \mathbb{R}^{n' \times 2}$ at the corresponding level;
2. for each node $i = 1, \ldots, n'$:
   (a) Cluster label: assign cluster label as 0 if $|\hat{\boldsymbol{Y}}_{i,1}| \geq |\hat{\boldsymbol{Y}}_{i,2}|$ and as 1 otherwise;
   (b) Importance score: obtain the importance score as $\hat{\boldsymbol{Y}}_{i,1}$ if it was clustered as 0 in step 2 and otherwise as $\hat{\boldsymbol{Y}}_{i,2}$;
3. Repeat steps 1–2 for each community until the stopping rule is reached.

Remarks:

(a) The above algorithm is for the binary split but it can be extended to multiple split.

(b) For directed networks, one can instead use the other factor $\hat{\boldsymbol{Z}}$ obtained from vsp. The interpretations of using $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Z}}$ are different: $\hat{\boldsymbol{Y}}$ embeds the columns of $\boldsymbol{A}$ while $\hat{\boldsymbol{Z}}$ embeds the rows.

(c) Possible stopping rule includes Le & Levina (2015); Li et al. (2020); Chen et al. (2021); Jin et al. (2022).

(d) The package is available at https://jh-cai.com/Hvsp.

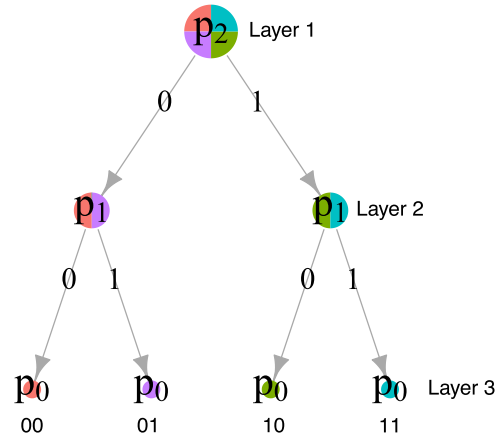

Figure 1. A four-cluster binary tree stochastic block models (BTSBM) (Li et al., 2020). The binary tree has three layers where Layer 1 includes all nodes, Layer 2 splits into two mega-communities {0, 1}, and Layer 3 further splits into four communities {00, 01, 10, 11}. Each colour corresponds to each community in Layer 3. In Layer 2, the mega-community {0} includes {00, 01} (red and purple) and the mega-community {1} includes {10, 11} (green and teal). Edges between nodes within the same community/mega-community are assumed to be independently Bernoulli with probability p_0, p_1, and p_2 depending on the layer. It is most natural to assume the communities are assortative $p_0 > p_1 > p_2$ so that the communities are more closely connected as the hierarchical tree goes deeper; or vice versa dis-assortative where $p_0 < p_1 < p_2$. In the toy example, we generate a balanced four-clustered BTSBM with 2,048 nodes where each mega-community at Layer 2 has 1,024 nodes and each community at Layer 3 has 512 nodes. We let $p_0 = 1$, $p_1 = 0.3$, and $p_2 = 0.09$ and scale accordingly so that the average degree of nodes is expected to be 50.

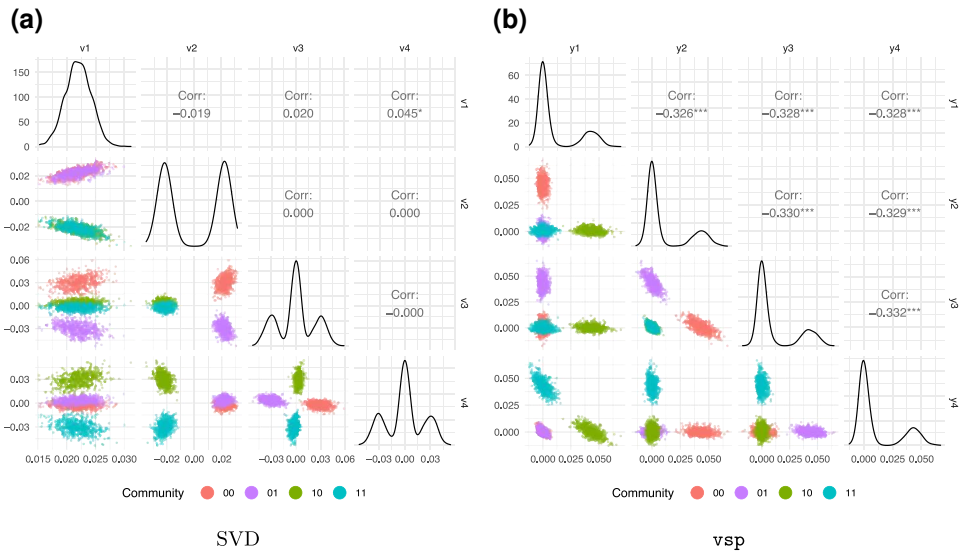

Figure 2. Scatter plot of pairs of principal components by SVD in figure (a) and pairs of varimax rotated components in figure (b). The colour corresponds to each community at Layer 3 in Figure 1. The radial streaks appear in figure (a) while the Varimax rotation aligns the streaks with the coordinate axes in figure (b), providing a sparse representation. However, neither provides a hierarchical structure.

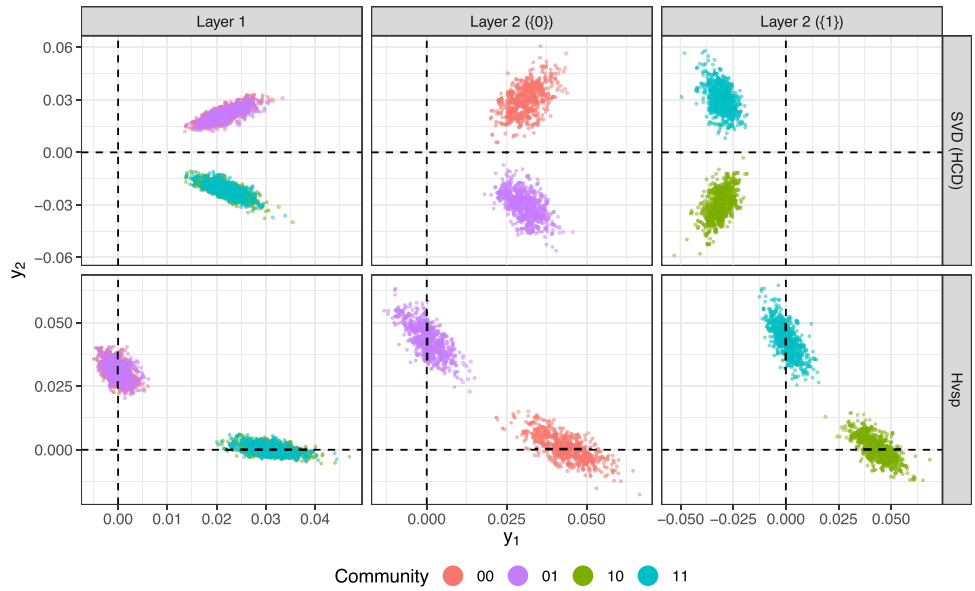

Figure 3. Scatter plot of pairs of principal components by SVD and pairs of Varimax rotated components. The rows correspond to hierarchical community detection (HCD-sign) (Li et al., 2020), which first performs SVD with dimension $k = 2$ and then assigns labels based on the sign of the second component, and the proposed Hvsp; while the columns correspond to the first split among all nodes (Layer 1) and the split of the mega-community {0} of {00, 01} and mega-community {1} of {10, 11} (Layer 2). The colour corresponds to each community at Layer 3 in Figure 1. HCD and Hvsp split the community layer by layer and reveal the hierarchical structure. In addition, Hvsp aligns the principal components to the coordinate axes so as to provide a sparse representation. The rotated components further provide a measure of the importance (importance score) of each node in each community as suggested by Rohe & Zeng (2022). Furthermore, the importance score can be provided by layers.

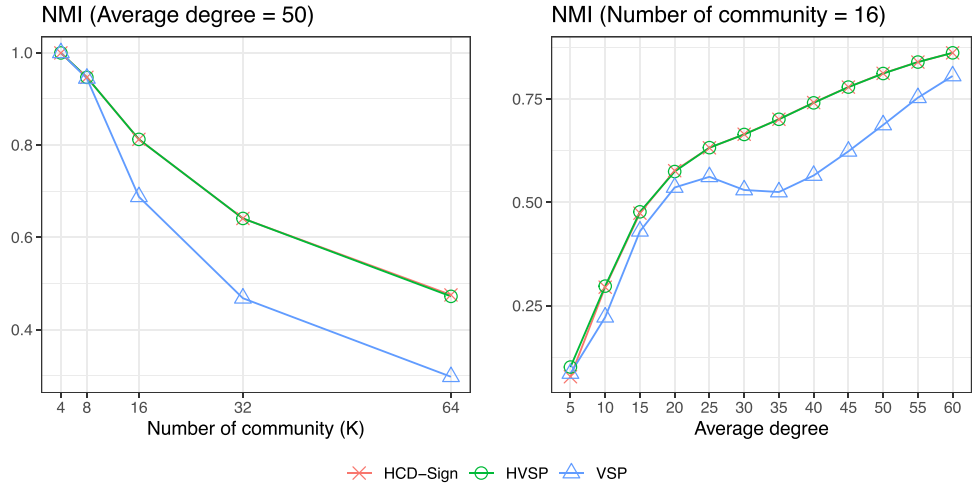

Figure 4. The normalized mutual information (NMI) (Yao, 2003) between the true and estimated labels obtained by HCD-sign, Hvsp, and vsp varying the number of communities and the average degree of nodes. The simulation setup follows Section 4.1 in Li et al. (2020). A larger NMI suggests better clustering performance. HCD-sign and Hvsp perform similarly while vsp falls behind. We compare the performance with more metrics at https://github.com/cccfran/Hvsp-paper.

Figure 5. The dendrogram of 11 communities of the three-core of the statistics citation network from 2003 to 2012 was obtained by Hvsp using edge cross-validation (ECV) as a stopping rule (Li et al., 2020). Research areas are manually labelled based on the research interests of the 10 statisticians with highest importance scores in Table 1, which are followed by the community size labelled in parentheses. The labelling can be made algorithmic such as using the 'best feature function' bff (Wang & Rohe, 2016). We provide the clustering result of using the nonbacktracking method (Le & Levina, 2015) as the stopping rule in https://github.com/cccfran/Hvsp-paper.

Table 1. The 15 statisticians with the highest importance scores in each community of the 2003–2012 citation network

Community (size)	Top 15 contributors
Bayesian methodology (66)	Alan E. Gelfand, David Dunson, Abel Rodriguez, Gary L. Rosner, Peter Muller, Steven N. MacEachern, Lawrence Carin, Mark F. J. Steel, Gareth Roberts, Ju-Hyun Park, Omiros Papaspiliopoulos, Yee Whye Teh, David M. Blei, Matthew J. Beal, Michael I Jordan
Bayesian theory (31)	Mike West, Hemant Ishwaran, J. Sunil Rao, Carlos M. Carvalho, James O. Berger, Helene Massam, James G. Scott, Chris Hans, Anirban Bhattacharya, Nicholas G. Polson, Adrian Dobra, Robert J. Kohn, Joseph E. Lucas, Frederick Wong, Christopher K. Carter
Design of experiments (40)	Boxin Tang, Randy R. Sitter, Derek Bingham, C. Devon Lin, Dennis K. J. Lin, David M. Steinberg, Neil A. Butler, Hongquan Xu, V. Roshan Joseph, Shan Ba, Ching-Shui Cheng, Peter Z G Qian, Frederick K. H. Phoa, Hegang H. Chen, John Stufken
Multivariate & dimension reduction (17)	Bing Li, R. Dennis Cook, Peng Zeng, Liqiang Ni, Francesca Chiaromonte, Yuexiao Dong, Robert E. Weiss, Zhishen Ye, Ronghua Luo, Xiangrong Yin, Shaoli Wang, Xin Chen, Louis Ferre, Tao Wang, Songqiao Wen
High-dimensional theory (54)	Alexandre B. Tsybakov, Marten H. Wegkamp, Iain M. Johnstone, Florentina Bunea, Vladimir Koltchinskii, Alexandre Belloni, Victor Chernozhukov, Karim Lounici, Yaacov Ritov, Bernard W. Silverman, Theofanis Sapatinas, Felix Abramovich, Emmanuel J. Candes, Olivier Bousquet, Peter L. Bartlett
Sampling & hypothesis testing (54)	Joseph P. Romano, Michael Wolf, Etienne Roquain, Gilles Blanchard, Sylvain Arlot, Larry Wasserman, Christopher Genovese, E. L. Lehmann, Chunming Zhang, Tao Yu, Luc Devroye, Nicolas Broutin, Louigi Addario-Berry, Isabella Verdinelli, M. Perone Paci_co
Multiple testing & inference (56)	John D Storey, T. Tony Cai, Yoav Benjamini, Jiashun Jin, Bradley Efron, David L Donoho, Jonathan E. Taylor, David Siegmund, Sanat K. Sarkar, Thorsten Dickhaus, Helmut Finner, Markus Roters, Wenge Guo, Daniel Yekutieli, Wenguang Sun
Functional data analysis (13)	Hans-Georg Muller, Jane-Ling Wang, Fang Yao, Peter Hall, R. Todd Ogden, Philip T. Reiss, David Ruppert, Je_rey S. Morris, Gerda Claeskens, Jianwei Chen, Bani Mallick, J. N. K. Rao, David Daniel Smith
Functional data & time series (60)	Lajos Horvath, Robertas Gabrys, Chong-Zhi Di, Ana-Maria Staicu, Siegfried Hormann, Piotr Kokoszka, Tailen Hsing, Kehui Chen, Ci-Ren Jiang, Pascal Sarda, Bitao Liu, Ciprian M Crainiceanu, Alois Kneip, Jeng-Min Chiou, Ulrich Stadtmuller
Non- & semiparametric methods (114)	Raymond J. Carroll, Xihong Lin, Naisyin Wang, Xuming He, Donglin Zeng, Enno Mammen, Guosheng Yin, Hua Liang, Joseph G. Ibrahim, Jing Qin, Zhezhen Jin, Arnab Maity, Kyusang Yu, Byeong U Park, Zhongyi Zhu
High-dimensional methodology (201)	Hui Zou, Jianqing Fan, Yi Lin, Peter Buhlmann, Trevor J. Hastie, Ming Yuan, Hao Helen Zhang, Jian Huang, Hansheng Wang, Ji Zhu, Cun-Hui Zhang, Runze Li, Heng Peng, Jinchi Lv, Shuangge Ma

To gauge the performance of Hvsp in community detection, we adopt the binary tree stochastic block models (BTSBM) that capture a binary tree community structure (Li et al., 2020).[1] We first use a toy example of a four-cluster balanced BTSBM (Figure 1) to provide insights and compare singular value decomposition (SVD) vs. vsp with dimension $k = 4$ and hierarchical community detection (HCD) (Li et al., 2020) vs. Hvsp in Figures 2 and 3. As expected, we observe the radial streaks from the pairs of principal components in Figure 2a, and the Varimax rotation aligns the streaks with the coordinate axes in Figure 2b. However, neither accounts for the hierarchical structure. On the other hand, HCD and Hvsp split the community layer by layer. In addition, Hvsp aligns the principal components to the coordinate axes, which provide a measure of the importance/centrality of each node in each community at different levels. We further compare the clustering performance using normalized mutual information (NMI) (Yao, 2003) of HCD, Hvsp, and vsp varying the number of communities and the average degree of nodes in Figure 4. HCD and Hvsp perform similarly while vsp falls behind.

Finally, we apply Hvsp to the three-core of the largest connected component of a statistics citation network (2003–2012) (Ji & Jin, 2016; Li et al., 2020). Figure 5 shows the hierarchical communities whose labels are based on the research interests of the ten statisticians with the highest scores within each community in Table 1. Hvsp clusters related communities together and the communities become more refined as the hierarchical tree goes deeper.

Conflicts of interest: none declared.

Data availability

We included a Github repo in our manuscript that provides all the codes and data.

References

Chen, F., Roch, S., Rohe, K., & Yu, S. (2021). Estimating graph dimension with cross-validated eigenvalues. arXiv preprint arXiv:2108.03336.

Ji, P., & Jin, J. (2016). Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, *10*(4), 1779–1812. https://doi.org/10.1214/15-AOAS896

Jin, J., Ke, Z. T., Luo, S., & Wang, M. (2022). Optimal estimation of the number of network communities. *Journal of the American Statistical Association*, 1–16. https://doi.org/10.1080/01621459.2022.2035736

Le, C. M., & Levina, E. (2015). Estimating the number of communities in networks by spectral methods. arXiv preprint arXiv:1507.00827.

Li, T., Lei, L., Bhattacharyya, S., Van den Berge, K., Sarkar, P., Bickel, P. J., & Levina, E. (2020). Hierarchical community detection by recursive partitioning. *Journal of the American Statistical Association*, *117*(538), 951–968. https://doi.org/10.1080/01621459.2020.1833888

Li, T., Levina, E., & Zhu, J. (2020). Network cross-validation by edge sampling. *Biometrika*, *107*(2), 257–276. https://doi.org/10.1093/biomet/asaa006

Rohe, K., & Zeng, M. (2022). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Association, Series B*.

Wang, S., & Rohe, K. (2016). Discussion of coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, *10*(4), 1820–1826. https://doi.org/10.1080/07350015.2022.2037432

Yao, Y. (2003). Information-theoretic measures for knowledge discovery and data mining. *Entropy measures, maximum entropy principle and emerging applications* (pp. 115–136). Springer.

https://doi.org/10.1093/jrsssb/qkad038
Advance access publication 5 April 2023

[1] Due to space limitations, we refer the readers to Li et al. (2020) for a detailed description of the BTSBM as well as the setup of the toy example and the following simulation study. The code for the toy example and the following simulation and citation network study is adapted from https://github.com/tianxili/HCD and is available at https://github.com/cccfran/Hvsp-paper where we provide more results on simulations and the citation network study in detail.

Christine P Chai's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Christine P. Chai

Microsoft Corporation, Redmond, WA, United States

Address for correspondence: Christine P. Chai, Microsoft Corporation, Redmond, WA, USA.
Email: cpchai21@gmail.com

Discussion

Rohe and Zeng (2020) made an advancement to show that Vintage Factor Analysis with Varimax can perform statistical inference, with a wide range of applications. I appreciate that the authors made the whole `vsp` (Vintage Sparse PCA) package[1] available in R for the general public.

I have several questions regarding the paper:

1. Why was Multiple Factor Analysis first developed by psychologists (Thurstone, 1935), instead of statisticians or computer scientists? Was there any historical context at that time?
2. Could the authors briefly describe the data cleaning process of the academic bibliometrics corpus? In the topic modelling results, the word 'biology' appears in both factors 2 and 7, which is a bit confusing. Some factors also include nontechnical words like 'on' and 'conference'. Finally, words of the same stem appear as separate words, such as 'chemistry' vs. 'chemical' and 'economics' vs. 'economic'.
3. The authors specified the compute resources for the code running time, indicating that Vintage Factor Analysis with Varimax can be implemented on a personal laptop within a reasonable time. Since the computation involves sparse matrices, are data structures like `scipy.sparse.csr_matrix`[2] used in the code to save memory space?
4. Are there particular types of applications that can benefit more from Vintage Sparse PCA, rather than just traditional PCA for dimensionality reduction? For instance, I investigated the relationship between the microbiome and environmental characteristics as a team (Beckman et al., 2015), and PCA projected the high-dimensional data to the first two principle components. Given the small n large p data, would Vintage Sparse PCA be more helpful than PCA?

References

Beckman E., Chai C., Lyu J., Mahserejian S., Tran H., Yavari S., Mitchell H., Calatroni A., & Kang E. L. (2015). Investigating the relationship between the microbiome and environmental characteristics. In *Twenty-first Mathematical and Statistical Modeling Workshop for Graduate Students* (pp. 89–112). North Carolina State University.

[1] https://rdrr.io/github/RoheLab/vsp/man/vsp.html
[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.csr_matrix.html

Rohe K., & Zeng M. (2020). 'Vintage factor analysis with Varimax performs statistical inference', arXiv, arXiv:2004.05387, preprint: not peer reviewed.
Thurstone L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press.

https://doi.org/10.1093/jrsssb/qkad039
Advance access publication 4 April 2023

Yunxiao Chen and Gongjun Xu's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Yunxiao Chen[1,*] **and Gongjun Xu**[2]

[1]London School of Economics and Political Science
[2]University of Michigan

Address for correspondence: Yunxiao Chen, Department of Statistics, London School of Economics and Political Science, Columbia House, Houghton Street, London WC2A 2AE, United Kingdom. Email: Y.Chen186@lse.ac.uk

We congratulate Rohe and Zeng for this insightful paper that elegantly connects psychometric methods and statistical and machine learning applications. We would like to mention several lines of related research. First, the problem is closely related to the latent variable selection problems (e.g., Y. Chen et al., 2015; Xu & Shang, 2018), where regularised estimation procedures are proposed to learn sparse loading structures. In fact, the vsp procedure can be viewed as the limiting case of a regularised estimation procedure, in the sense that $\hat{U}$ from Algorithm vsp is the limit of

$$\hat{U}^{\lambda} = \arg\min_{U} \left(\min_{D,V} \|\tilde{A} - UDV^T\|_F^2 - \lambda \left(\sum_{l=1}^{k} \frac{1}{n} \sum_{i=1}^{n} \left([U]_{il}^4 - \left(\frac{1}{n} \sum_{q=1}^{n} [U]_{ql}^2 \right)^2 \right) \right) \right), \quad \lambda > 0,$$

when λ goes to zero if the solution path is smooth, where U and V satisfy the same constraints as in singular value decomposition, while D is allowed to be non-diagonal (F. Chen & Rohe, 2020). Note that the regularisation is used to learn the sparse loading structure rather than to avoid overfitting, and thus, it does not require the tuning parameter to depend on the noise level. We agree that rotation is more convenient under many models, but the regularised estimation approach might be more general for some more complex latent variable models.

Second, various rotation methods have been proposed in the psychometric literature to find simple and scientifically meaningful factor loading structures. For example, consider the L_1 criterion that minimises the objective function

$$c(R, U) = \sum_{i=1}^{n} \sum_{l=1}^{k} |[UR]_{il}|.$$

This criterion is closely related to L_1 regularisation and ensures statistical consistency under suitable conditions (Jennrich, 2004, 2006; Liu et al., 2022). We run a small simulation study to compare the Varimax and L_1 rotations, where data are generated from the current factor model. Two settings are used to generate Z—one sparse setting where $P([Z]_{ij} = 0) = 0.5$ and one dense setting where $[Z]_{ij}$ follows a heavy-tail distribution. The L_1 rotation method only replaces $v(R, \hat{U})$ in step

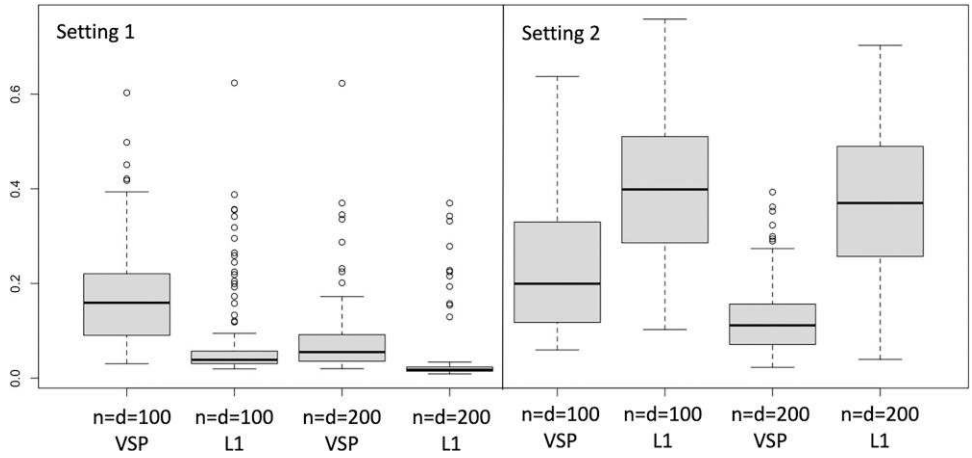

Figure 1. Box plots of mean squared errors $\|\hat{Z} - ZP_n\|_F^2/(nk)$ from 100 simulations. The four box plots correspond to the combinations of two settings ($k = 5$, $n = d = 100$, 200) and two rotation methods (Varimax and L_1). In the simulations, we generate $A = ZY^T + W$, where $[A]_{ij}$ and $[W]_{ij}$ are independent standard normal variables. Under Setting 1, $[Z]_{ij} = \sqrt{2}[C]_{ij}[S]_{ij}$, where $[C]_{ij}$ are independent standard normal random variables and $[S]_{ij}$ are independent Bernoulli random variables with success probability 0.5. Under Setting 2, $[Z]_{ij} = [T]_{ij}/\sqrt{5/3}$, where $[T]_{ij}$ follows a t distribution with five degrees of freedom. R code for the simulation can be found on https://stats.lse.ac.uk/cheny185/L1_rotation.R.

3 of the vsp algorithm with $c(R, \hat{U})$. The mean squared errors for the estimation of Z are given in Figure 1, where the L_1 rotation performs better under the sparse setting while the vsp outperforms under the dense setting.

Finally, another interesting extension of the current work is to non-linear factor models that assume $\mathbb{E}([A]_{ij} \mid Z, B, Y) = f([ZBY^T]_{ij})$, for some known smooth and strictly monotone non-linear function f (e.g., logistic function for binary data). Due to the non-linear transformation, the current vsp procedure does not directly apply. One solution is to first apply the universal singular value thresholding procedure (Chatterjee, 2015) to A to estimate $(f([ZBY^T]_{ij}))_{n\times d}$, which yields an estimate of ZBY^T through element-wise f^{-1} transformations; see Zhang et al. (2020) for more details and the related consistency theory. Then, one can learn Z by steps 2 and 3 of Algorithm vsp.

Conflict of interests: None declared.

Funding

G.X. is supported by National Science Foundation grants SES-1846747 and SES2150601.

Data availability

The data supporting the findings of this study are available within the article.

References

Chatterjee S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, *43*(1), 177–214. https://doi.org/10.1214/14-AOS1272

Chen F., & Rohe K. (2020). 'A new basis for sparse principal component analysis', arXiv, arXiv:2007.00596, preprint: not peer reviewed.

Chen Y., Liu J., Xu G., & Ying Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866. https://doi.org/10.1080/01621459.2014.934827

Jennrich R. I. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika*, *69*(2), 257–273. https://doi.org/10.1007/BF02295943

Jennrich R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, *71*(1), 173–191. https://doi.org/10.1007/s11336-003-1136-B

Liu X., Wallin G., Chen Y., & Moustaki I. (2022). 'Rotation to sparse loadings using L^p losses and related inference problems', arXiv, arXiv:2206.02263, preprint: not peer reviewed.
Xu G., & Shang Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*(523), 1284–1295. https://doi.org/10.1080/01621459.2017.1340889
Zhang H., Chen Y., & Li X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, *85*(2), 358–372. https://doi.org/10.1007/s11336-020-09704-7

https://doi.org/10.1093/jrsssb/qkad040
Advance access publication 4 April 2023

Kuldeep Kumar's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Kuldeep Kumar

Centre for Data Analytics, Bond University, Gold Coast, Australia

Address for correspondence: Kuldeep Kumar, Centre for Data Analytics, Bond University, Gold Coast, Queensland 4226, Australia. Email: kkumar@bond.edu.au

Data Analysts are quite often confused between principal component analysis (PCA) and factor analysis. Both are the same data reduction techniques that are immensely helpful in big data analysis in many other areas. I personally think factor analysis is just an offshoot of PCA and as the authors have mentioned PCA with Varimax is a vintage data analysis technique. So, should the title of the paper be Vintage PCA with Varimax performing statistical inference?

I am quite impressed with the systematic study of 144,136 papers done by the authors on 'How and where factor analysis is used'. No wonder it is one of the most widely used statistical techniques in many areas of social sciences, business, medicine, etc. I will just like to add that the 'vintage' book written by Professor Ian Jolliffe on PCA in 1986 and published by Springer has received more than 48,000 citations which is one of the highest citations of any statistical books. Professor Jolliffe has also written a series of well-cited seminal papers on PCA Jolliffe (2022), and his recent paper on his 'vintage' 50 years personal journey through time with PCA is worth reading in this context.

Conflicts of interest: None declared.

References

Jolliffe, I. T. (1986). *Principal component analysis* (290 p.). Springer-Verlag.
Jolliffe, I. T. (2022). A 50-year personal journey through time with principal component analysis. *Journal of Multivariate Analysis*, *188*, 104820. https://doi.org/10.1016/j.jmva.2021.104820

https://doi.org/10.1093/jrsssb/qkad041
Advance access publication 4 April 2023

Yang Liu's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Yang Liu

University of Maryland, College Park

Address for correspondence: Yang Liu, 3304R Benjamin Bldg, 3942 Campus Dr, College Park, MD 20742, USA.
Email: yliu87@umd.edu

I congratulate Rohe and Zeng for the fascinating piece of work. They discovered that sparse principle components analysis (PCA) with Varimax rotation (`vsp`) identifies and consistently recovers latent factors in a broad class of semi-parametric factor models. I have a number of comments to share regarding identification assumptions, asymptotic results, and interpretation of factors.

Identification Assumptions. The identification of the low-rank matrix $M = ZBY^\top$ (when treated as fixed) has been well understood in the literature of matrix completion (e.g., Chatterjee, 2015; Fan et al., 2019). The main contribution of the present paper is to establish that, not only M, but also the factorization of M can be uniquely identified when the factors are random and sparse/leptokurtic. The connection between sparseness and leptokurticity is an intriguing standalone result. However, the sparsity requirement in Theorem 4.1 is still restrictive in some applications when k is small: For example, a bifactor structure (Holzinger & Swineford, 1937) wherein each row of Z (or Y) has two non-zero entries does not satisfy the condition if $k < 12$. I am wondering if results similar to Theorem 5.1 (i.e., uniqueness of rotation) can be established under more general sparseness conditions.

Asymptotic Result. Seeing the close resemblance between `vsp` and matrix completion, it is natural to ask how different their asymptotic rates of convergence are. In matrix completion, the low-rank matrix M is considered fixed. I conjecture that an error bound for an identifiable Z may be obtained by combining

1. an error bound for the low-rank matrix M;
2. an error bound for the singular vectors U (up to a rotation matrix);
3. Proposition 5.2 that relates U to Z.

Such a result is not yet readily comparable to Theorem 6.1 as M is random in the present paper; however, it may be possible to conclude that the bound holds with a large probability given the distributional assumptions for Z and Y. Considering that `vsp` is computationally much cheaper, I do not expect that it outperforms matrix completion in accuracy.

Interpretation of Factors. My following comments are based on a more 'vintage' specification of factor analysis, in which Z is random and unstructured (i.e., scores) and $\tilde{Y} = YB^\top$ is fixed and sparse (i.e., loadings). First, factor analysis is not just about finding a low-dimensional approximation to the data—it also implies that the low-dimensional factors fully account for the dependencies in data (Fabrigar et al., 1999). Second, Thurstone's reference to 'scientifically meaningful category' cannot be simply translated to 'statistically identifiable factors'. It is sometimes meaningful to have two different explanations, corresponding to two differently rotated solutions, on the

same identifiable, low-dimensional structure. Here is a toy example deduced from the well-known Schmid-Leiman transform (Schmid & Leiman, 1957):

$$
\begin{aligned}
&Z_1 iid\sim\mathcal{N}(0,\Phi_1),\quad \Phi_1=\begin{pmatrix}1 & 1/4\\ 1/4 & 1\end{pmatrix},\quad \tilde{Y}_1=\begin{pmatrix}1 & 0\\ 1 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 1\\ 0 & 1\end{pmatrix},\\
&Z_2 iid\sim\mathcal{N}(0,\Phi_2),\quad \Phi_2=\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix},\quad \tilde{Y}_2=\begin{pmatrix}0.5 & \sqrt{3/4} & 0\\ 0.5 & \sqrt{3/4} & 0\\ 0.5 & \sqrt{3/4} & 0\\ 0.5 & 0 & \sqrt{3/4}\\ 0.5 & 0 & \sqrt{3/4}\\ 0.5 & 0 & \sqrt{3/4}\end{pmatrix},
\end{aligned}
\tag{1}
$$

in which Z_1 and Z_2 contain independent and identically distributed normal variates. It is obvious that $Z_1\tilde{Y}_1^\top \overset{d}{=} Z_2\tilde{Y}_2^\top$ because $\tilde{Y}_1\Phi_1\tilde{Y}_1^\top = \tilde{Y}_2\Phi_2\tilde{Y}_2^\top$. Both $\tilde{Y}_1$ and $\tilde{Y}_2$ have meaningful interpretations in psychological theory: The former represents correlated traits indicated by distinct observable responses, while the latter contains one overall and two specific traits that are mutually orthogonal.

Conflict of interest: None declared.

References

Chatterjee S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, *43*(1), 177–214. https://doi.org/10.1214/14-AOS1272

Fabrigar L. R., Wegener D. T., MacCallum R. C., & Strahan E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fan J., Gong W., & Zhu Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, *212*(1), 177–202. https://doi.org/10.1016/j.jeconom.2019.04.026

Holzinger K. J., & Swineford F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Schmid J., & Leiman J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61. https://doi.org/10.1007/BF02289209

https://doi.org/10.1093/jrsssb/qkad042
Advance access publication 5 April 2023

Xiaoyue Niu's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Xiaoyue Niu

Penn State University, University Park, USA

Address for correspondence: Xiaoyue Niu, Penn State University, University Park, USA. Email: xiaoyue@psu.edu

I want to congratulate Rohe and Zeng on solving a long-time puzzle (and fight between statisticians and practitioners) of 'why/whether the varimax rotation works'. Their inspiring results have some implications in the literature of latent space models for social networks (Hoff, 2021; Hoff et al., 2002). One of the appealing features of the latent space model is that one can extract the latent factors and analyse the unobserved features through the latent positions. However, it is a random orientation that we are interpreting due to the rotational invariance. It is less problematic if we are only interested in the relative locations. It becomes more of an issue when we are analysing dynamic networks or trying to compare networks across different subjects. In those situations, we need a 'true' orientation. A common approach is to pick a (random) orientation and perform Procrustes transformation, which could potentially diminish some of the structural changes and signals. Some more sophisticated effort towards solving the nonidentifiability issue includes Poworoznek et al. (2021), which also utilizes varimax. This rotational invariance happens because we assume normal factors, mostly due to computational convenience. If we are willing to move away from normality to something leptokurtic (such as a t-distribution), we would be able to have an 'identifiable latent space model'. In fact, the normal mixture that Handcock et al. (2007) used is leptokurtic so the factor version of their clustering model is identifiable and could potentially be extended to a dynamic model to study latent position changes without the trouble of matching the orientations.

References

Handcock M. S., Raftery A. E., & Tantrum J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *170*(2), 301–354. https://doi.org/10.1111/j.1467-985X.2007.00471.x

Hoff P. D. (2021). Additive and multiplicative effects network models. *Statistical Science*, *36*(1), 34–50. https://doi.org/10.1214/19-STS757

Hoff P. D., Raftery A. E., & Handcock M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*(460), 1090–1098. https://doi.org/10.1198/016214502388618906

Poworoznek E., Ferrari F., & Dunson D. (2021). Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching. arXiv:2107.13783.

https://doi.org/10.1093/jrsssb/qkad043
Advance access publication 5 April 2023

Florian Pargent, David Goretzko and Timo von Oertzen's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Florian Pargent[1], David Goretzko[2] and Timo von Oertzen[3]

[1]Department of Psychology, LMU Munich, Germany
[2]Methodology and Statistics, Utrecht University, The Netherlands
[3]Institute of Psychology, Germany and Max Planck Institute for Human Development, Center for Lifespan Psychology, University of the Bundeswehr Munich, Germany

Address for correspondence: Florian Pargent, Department of Psychology, LMU Munich, Leopoldstr. 13, 80802 Munich, Germany. Email: florian.pargent@psy.lmu.de

As psychologists, we appreciate Rohe & Zeng's (R&Z; Rohe & Zeng, 2022) new insights into 'vintage' principal component analysis with varimax rotation (PCA+VR). Theories of intelligence

and personality, perhaps psychology's contributions best known outside of our field, have been a direct product of PCA. PCA+VR is still widely used for developing and evaluating psychological tests and questionnaires, although the literature has fought against it in favour of more complex factor analytic techniques (Fokkema & Greiff, 2017).

In our opinion, abandoning the simpler PCA(+VR) is a mistake and R&Z refute a common argument by proving that PCA+VR *can* perform statistical inference in latent variable models: The factor indeterminacy problem which plagued VR since its invention only applies for the special case of normally distributed factors. For any other distribution, perfect factor indeterminacy does not apply, although identifiability might be weak. However, distributions producing sparse components fulfil a *sufficient* leptokurtic condition, which can be confirmed by simple diagnostics.

Because the results are complicated, we relate them to psychological applications. The examples in R&Z only deal with sparse binary network data, but in typical psychological applications, the A matrix consists of responses of n persons to d items which are either binary (e.g. intelligence tests), integer-valued (e.g. personality questionnaires), or continuous (e.g. digital sensors). Psychologists are often interested in whether (i) items can be structured in a simple way to represent a small number of meaningful components and (ii) those components can be interpreted as psychological constructs that describe interindividual differences. R&Z show that 'radial streaks' in the rotated loading matrix $\widehat{Y}$ suggest that item loadings are identified and can be estimated with PCA + VR from the data. Similarly, streaks in the component matrix $\widehat{Z}$ suggest that person scores can be estimated.

However, we question whether streaks are common in psychology with regard to both aspects. In our online materials (https://osf.io/5symf/), we analyse a data set (Stachl et al., 2020) containing both personality items ($n = 687$, $d = 300$) and smartphone sensing variables ($n = 624$, $d = 1821$). Streaks were found only in $\widehat{Y}$ but not in $\widehat{Z}$. It is also a cautionary example of how the imputation of missing values in combination with inappropriate data processing seemingly produce streaks in $\widehat{Z}$ that belong to uninterpretable components. Finally, we demonstrate R&Z's side result that the matrix $\widehat{Z}\widehat{B}$ from PCA+VR can estimate person scores simulated from oblique leptokurtic components.

In our opinion, the main usefulness of PCA + VR not necessarily stems from its ability to estimate latent variable models. PCA excels at providing meaningful descriptions in practical applications but R&Z's and our examples also show that there is rarely a single definite structure. Components are most useful when they predict other meaningful quantities, regardless of the assumed epistemological nature of psychological constructs (Yarkoni, 2020).

Conflicts of interest: None declared.

Data availability

All data and code are available at https://osf.io/5symf/.

References

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it. *European Journal of Psychological Assessment*, *33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460

Rohe, K., & Zeng, M. (2022). Vintage factor analysis with varimax performs statistical inference. Preprint Read Before the Royal Statistical Society on 11 May 2022. arXiv preprint arXiv:2004.05387, preprint: not peer reviewed.

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

Yarkoni, T. (2020). Implicit realism impedes progress in psychology: Comment on fried (2020). *Psychological Inquiry*, *31*(4), 326–333. https://doi.org/10.1080/1047840X.2020.1853478

https://doi.org/10.1093/jrsssb/qkad054
Advance access publication 23 May 2023

Mark Pilling's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Mark Pilling

Department of Public Health & Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge, UK

Address for correspondence: Mark Pilling, East Forvie Building, Robinson Way, Cambridge. CB2 0S. Email: mark.pilling@medschl.cam.ac.uk

Thanks to the authors for their very interesting, thoughtful, and thorough reconsideration of the topic of factor analysis.

We know to be a weakness of factor analysis that an issue is that analysts must chose the type of rotation (e.g., SPSS lists Varimax, Direct Oblimin, Quartimax, Equamax, & Promax) to use, and this choice tends to be subjectively based on how useful or interpretable the rotation was in the context of the study. So multiple rotations might be considered.

Given the author's assertion that rotations can be interpreted as performing statistical inference, could the authors additionally comment on whether this has consequences for multiple testing? i.e., Is there any danger that p-hacking could occur if multiple rotations are being considered? In some sense, searching for results that fit the preconceived ideas of the analyst.

Conflicts of interest: None declared.

https://doi.org/10.1093/jrsssb/qkad044
Advance access publication 5 April 2023

Konstantin Siroki and Korbinian Strimmer's contribution to the Discussion of 'Vintage factor analysis with varimax performs statistical inference' by Rohe and Zeng

Konstantin Siroki and Korbinian Strimmer

Department of Mathematics, University of Manchester, Alan Turing Building, Oxford Road, Manchester M13 9PL, UK

Address for correspondence: Korbinian Strimmer, Department of Mathematics, University of Manchester, Alan Turing Building, Oxford Road, Manchester M13 9PL, UK. Email: korbinian.strimmer@manchester.ac.uk

We congratulate Rohe and Zeng on their interesting paper and would like to share two comments and propose one question for their consideration.

First, it is impressive to see the large variety of models addressed by the 'vintage sparse PCA' (vsp) framework, ranging from traditional factor models to explainable and sparse PCA but also including topic models and stochastic block models. In our view, the demonstration of the broad applicability of the 'compress and rotate' strategy is a major contribution of this article. Unfortunately, much of this material is relegated to the Appendix of the article. This extends earlier discussions of the benefits of rotating and sparsifying principal components in more classical use cases, e.g. by Jolliffe (2002, Chapter 11) and by Trendafilov and Adachi (2015).

Second, the problem of identifiability of loadings and factors in factor analysis is mirrored in the problem of identifying optimal whitening transformations (Jendoubi & Strimmer, 2019; Kessy et al., 2018). There are infinitely many such transformations, and just like in factor analysis it is only the constraints on the covariance-based or correlation-based loadings that allow to distinguish among and select corresponding whitening methods (to include PCA-based whitening, ZCA whitening, and Cholesky whitening). Crucially, PCA-based whitening is itself already the result of a rotation, so vsp may be interpreted as two consecutive rotations, not just one, the first leading to optimal compression and the second to sparsify the loadings of the top-ranking factors.

This leads to our main question. It is shown by Rohe and Zeng that under the leptokurtic distributional assumption the varimax rotation is able to identify the original underlying factors. Now, varimax that lies at the heart of vsp is part of the larger orthomax family (Harman, 1976) which comprises not only varimax, but quartimax, biquartimax, equamax, and parsimax, and in fact many other types of rotations. In turn, the orthomax family is itself a special case of even richer families (e.g. Browne, 2001) that also include oblique transformations. A result by Bernaards and Jennrich (2003) suggests that all members of the orthomax family allow to recover simple structure, not just varimax. The key argument in Bernaards and Jennrich (2003) is essentially the same as the motivation put forward by Rohe and Zeng about varimax and sparsity. Hence, we ask whether the leptokurtic condition is strict enough and sufficient to single out varimax or whether perhaps the whole orthomax family may be compatible with it?

Conflict of interest: None declared.

References

Bernaards C. A., & Jennrich R. I. (2003). Orthomax rotation and perfect simple structure. *Psychometrika*, *68*(4), 585–588. https://doi.org/10.1007/BF02295613

Browne M. W. (2001). An overview over analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Harman H. H. (1976). *Modern factor analysis*. 3rd ed., University of Chicago Press.

Jendoubi T., & Strimmer K. (2019). A whitening approach to probabilistic canonical correlation analysis for omics data integration. *BMC Bioinformatics*, *20*(1), 15. https://doi.org/10.1186/s12859-018-2572-9

Jolliffe I. T. (2002). *Principal component analysis*. 2nd ed., Springer.

Kessy A., Lewin A., & Strimmer K. (2018). Optimal whitening and decorrelation. *The American Statistician*, *72*(4), 309–314. https://doi.org/10.1080/00031305.2016.1277159

Trendafilov N. T., & Adachi K. (2015). Sparse versus simple structure loadings. *Psychometrika*, *80*(3), 776–790. https://doi.org/10.1007/s11336-014-9416-y

https://doi.org/10.1093/jrsssb/qkad055
Advance access publication 23 May 2023

Tyler J. VanderWeele's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Tyler J. VanderWeele

John L. Loeb and Frances Lehman Loeb Professor of Epidemiology, Harvard University, Cambridge, MA, USA

Address for correspondence: Tyler J. VanderWeele, 677 Huntington Avenue, Boston MA 02115, USA. Email: tvanderw@hsph.harvard.edu

Rohe and Zeng (2022) have provided important results concerning properties of 'vintage factor analysis' that assist with interpretability, especially for non-Gaussian and leptokurtic data. I would like to focus my remarks on one particular comment, also present in Thurstone (1935) that, one aims "to ensure that each factor.... corresponds to a 'scientifically meaningful category.'" This is an important goal, but I would argue that (i) questions of the meaning of factors always extend beyond empirical analyses; (ii) the very notion of a 'scientifically meaningful category' is ambiguous; and (iii) that the neglect of the prior two points has led to inappropriate interpretation of factors in practice. On the first point, the interpretation of factors is often controversial and always requires some external knowledge. Rohe and Zeng's interpretation of their analyses, while sensible, did require the use of external features (words in journal-title; Section 3.1.1) and then still requires a reader who is aware that the cluster of title words forms a coherent whole and how to interpret that whole. Meaning is always external, and in some sense prior, to the empirical analysis (Mauran, 1996). On the second point, what is a 'scientifically meaningful category'? The notion of a 'discipline,' supposedly corresponding to the factors in their analysis, is a meaningful concept, but does it correspond to a 'scientific category'? There are various ways specific disciplines might be precisely defined, with varying levels of scope, but this then is again a conceptual, rather than an empirical, question. Moreover, the non-hierarchical nature of the categories as the number of factors increases (Section 3.1.3), indicates that while these are potentially interpretable categories, they are not fixed scientific realities. Perhaps more problematic is when factors are thought to correspond to univariate latent variables that actually represent real entities that are causally efficacious. Often it is presumed that if we had knowledge of the entity's quantity, the indicators would be irrelevant. This assumption, however, is so strong that it has empirically testable implications even though we never observe the supposed latent variable; the assumption will often be false, and, moreover, even if such entities did exist, the factor model is still consistent with multiple causal structures (VanderWeele & Vansteelandt, 2022). Consequently, concerning the third point, the neglect of questions of meaning and different causal structures being consistent with factor models routinely results in the practice of erroneously concluding that we uncover real entities by factor analysis and that the meaning of a construct can be established by empirical analysis. Vintage factor analysis is a powerful tool for uncovering patterns of association, but we must still go through the difficult interpretative work of trying to assess why the patterns of correlation are present, and what they mean (VanderWeele, 2022; VanderWeele & Batty, 2022).

Conflicts of interest: None declared.

References

Maraun, M. D. (1996). Meaning and mythology in the factor analysis model. *Multivariate Behavioral Research*, *31*(4), 603–616. https://doi.org/10.1207/s15327906mbr3104_13

Rohe, K., & Zeng, M. (2022). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society, Series B*.
Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press.
VanderWeele, T. J. (2022). Constructed measures and causal inference: Towards a new model of measurement for psychosocial constructs. *Epidemiology*, *33*(1), 141–151. https://doi.org/10.1097/EDE.0000000000001434
VanderWeele, T. J., & Batty, C. J. K. (2023). On the dimensional indeterminacy of one-wave factor analysis under causal effects. *Journal of Causal Inference*, in press. https://arxiv.org/abs/2001.10352
VanderWeele, T. J., & Vansteelandt, S. (2022). A statistical test to reject the structural interpretation of a latent factor model. *Journal of the Royal Statistical Society, Series B*, *84*, 2032–2054.

https://doi.org/10.1093/jrsssb/qkad045
Advance access publication 6 April 2023

Tao Wang's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Tao Wang

Department of Economics, University of Victoria, Victoria, BC V8W 2Y2, Canada

Address for correspondence: Tao Wang, Department of Economics, University of Victoria, Victoria, BC V8W 2Y2, Canada. Email: taow@uvic.ca

I congratulate the authors on their excellent paper studying the fundamental mechanism of PCA with the Varimax rotation. The paper thoughtfully establishes theorems to demonstrate that the Varimax rotation can supply an unified estimating technique for a wide range of semiparametric factor models. My discussion will be primarily centred on clarification and potential extensions.

The paper illustrates that if PCA is performed on a matrix with independent elements, PCA with the Varimax rotation can be utilized to fit the semiparametric model under certain assumptions. Is the proposed technique applicable in the case of Quartimax or Equamax rotation? In comparison to oblique rotation, the interpretability of the resulting components from orthogonal rotation is not always satisfied (may oversimplify the data). Although the paper reveals that the developed technique can handle correlated factors, to broaden the scope of the paper's application, I wonder if the built theorems apply to oblique rotations such as Oblimin and Promax. Also, the paper employs the scree plot (better if using parallel analysis scree plot), which is somewhat subjective, to argue the number of PCs to extract in order to obtain the most parsimonious factor structure. According to the paper, 'there is not a single correct answer for the choice of k'. Provided that the primary purpose of the paper is to lay the theoretical groundwork for a widely used model, I am intrigued whether the authors can deliver a more theoretically justified method for selecting which factors to retain.

In addition, the paper demonstrates that the procedure as a whole is effective on condition that the principal component matrix entries are reasonably nonnormal. Practically, we may assume that the original dataset is normally distributed to ensure that the PCs are independent and the results are more robust. In this case, I am concerned with the implications of the nonnormal condition on factor analysis using a normally distributed dataset. Given the leptokurtic condition on the elements of

Z that must be satisfied for Varimax rotation to function, the validity of the multivariate normality assumption regarding the distribution of observable variables or latent constructs appears to be compromised. Aside from that, the variation around the mean for symmetric distributions is a useful measure of dispersion. Nonetheless, it can fail when dealing with skewed or asymmetric distributions; see (Tran et al., 2019). Is the developed method capable of accommodating asymmetric distributions or distributions lacking moments (or with undefined or infinite variance)? There has been substantial research towards the robustification of PCA in the field of robust statistics; see (Candès et al., 2011). Will the Varimax rotation work if we consider a more robust PCA method or are interested in capturing the tail behaviour of the data such as quantile estimation?

Conflict of interest: None declared.

Data availability

Data sharing not applicable–no new data generated.

References

Candès E. J., Li X., Ma Y., & Wright J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*(11), 1–37. https://doi.org/10.1145/1970392.1970395

Tran N. M., Burdejová P., Ospienko M., & Härdle W. K. (2019). Principal component analysis in an asymmetric norm. *Journal of Multivariate Analysis*, *171*, 1–21. https://doi.org/10.1016/j.jmva.2018.10.004

https://doi.org/10.1093/jrsssb/qkad046
Advance access publication 5 April 2023

Ying Zhou and Xinyi Zhang's contribution to the Discussion of 'Vintage Factor Analysis with Varimax Performs Statistical Inference' by Rohe & Zeng

Ying Zhou and Xinyi Zhang

Department of Statistical Sciences, University of Toronto, Toronto, ON M5G 1Z5, Canada

Address for correspondence: Ying Zhou, Department of Statistical Sciences, University of Toronto, 700 University Ave., Toronto, ON M5G 1Z5, Canada. Email: yingx.zhou@mail.utoronto.ca

We congratulate Rohe and Zeng for their inspiring work on providing theoretical support for Varimax rotation in factor analysis. A key contribution of the article is the demonstration that if the factors in the semi-parametric factor model follow heavy-tailed distribution, then performing principal components analysis (PCA) on the observed data matrix along with Varimax rotation applied to the principal components does produce interpretable explanatory variables. We have the following comments and questions:

(a) The article mainly discussed the semi-parametric model, which seems to exclude the classical factor model in the form $A^T = LF + \varepsilon$, where observation matrix $A \in \mathbb{R}^{n\times d}$, loading matrix $L \in \mathbb{R}^{d\times k}$, factor matrix $F \in \mathbb{R}^{k\times n}$, error term matrix $\varepsilon \in \mathbb{R}^{d\times n}$. One question is whether there

is a similar result applicable to the classical factor model. If not, what are the obstacles? And what properties of semi-parametric factor model facilitate the identification? The main theorem in the article concerns the factor matrix F (A in their notation), while in many applications, people are interested in the loading matrix L. It appears there is no loading matrix in semi-parametric factor model. Perhaps BY^T in Definition 1 is somewhat related to loading matrix. This lead to the question that, if there is a parallel result for Y.

(b) As one of the modern factor models, the Independent Components Analysis is an unsupervised learning algorithm and can be applied for feature extraction. Would it be possible to integrate class information with the Varimax rotation for extracting features that belong to well-separated classes? It would be interesting to see if the Vintage Factor Analysis can be used in a supervised fashion.

(c) If we understand it correctly, derivation of the population results for PCA with latent variable models and Varimax uses $\widehat{\Sigma}_Z^{-1/2}$ to show how U can be recovered from Z. In theory, dimension d can be of the same order as n. However, in this case, the sample covariance matrix $\widehat{\Sigma}_Z$ may not be invertible and an alternative estimation of $\Sigma_Z^{-1/2}$ is needed. Can similar results be established? The factors Z are allowed to be correlated, will the corresponding theory be a direct generalisation from the independent setting?

Conflicts of interest: None declared.

The authors replied later, in writing, as follows.

https://doi.org/10.1093/jrsssb/qkad056
Advance access publication 2 June 2023

Karl Rohe and Muzhe Zeng's reply to the Discussion of 'Vintage factor analysis with varimax performs statistical inference'

Karl Rohe and Muzhe Zeng

Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA

Address for correspondence: Karl Rohe, Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA. Email: karl.rohe@wisc.edu

Statistics as a field has been tremendously successful in creating and propagating the quantitative theories, techniques, and tools to do quantitative science. Because of our success, our community is continually fragmented into methodological subdisciplines within other fields[1] , each a certain type of universe in which methodologies are continuously evolving (almost independently) in parallel. This phenomenon has happened slowly over time and as a result, we think it is time to reappraise the role of Statisticians with a capital S (i.e. in Statistics departments). In this vein, the history of Varimax provides a parable.

Statisticians often perceive of our field as producing methodology by deriving it from our foundational theories (e.g. Maximum Likelihood, Bayesian, etc.). And then, other fields consume our

[1] psychometrics, signal processing, econometrics, epidemiology, demography, chemeometrics, actuarial sciences, machine learning, bioinformatics, etc.

methodologies. One issue with this perception is that the methodological subdisciplines are producing their own methodologies; what about the Statisticians with a capital S? Sometimes our foundational theories help them. Sometimes our theories do not help them. Sometimes their theories go beyond our own and then we call them Statisticians (with a capital S). Regardless, these are all skilled craftspeople and we fool ourselves when we pretend that Statisticians are the all-knowing producers.

The history of factor analysis is an antidote to our illusions of grandeur. In 1935, Thurstone (a psychologist) was inspired by the idea of multiple types of intelligence and set out to create a way to measure them. Inspired by this and without an 'electronic' computer, he whittled a blunt tool, a way of iteratively plotting data and cleverly picking a tiny bit of a rotation and then iterating again. This was not derived from our theories. Moreover, this process was 'subjective' at every iteration. In 1956, Anderson and Rubin wrote a Gaussian theory that discounted factor rotations; their theory appears in every textbook on Multidimensional Statistics. But already in 1953, Carroll introduced an optimisation problem to 'objectively' pick a rotation using 4th-order moments. Then, Varimax came in 1958. Despite the protesting Statisticians, psychologists used factor rotations and conveyed them to future generations because factor rotations solved a problem. They did not need a 'theoretical foundation' and they should not need one now. We hope that the 'theoretical foundation' we have started to provide is that it might convince researchers in other disciplines to try using factor rotations on their problems.

This parable of factor analysis is extreme because Statisticians have opposed it for nearly 90 years. Our fundamental claim is more general. Successful methodologies, the ones that spread, will have a consilience of a *product-market-fit*, 'statistical theory' (of models, theorems, and algorithms), and *practical-know-how*. In the successful diffusion of a methodology, Statistics as a discipline has an essential role. Ideally, we would be in a position to develop methodologies with product-market-fit, but often we are not. Our primary strength is that we can (1) provide a 'statistical theory', and then (2) leverage our central position in the academic network of methodological subdisciplines to convey this methodology and statistical theory.

Before (1) developing a theory and (2) propagating it, there is a step (0). We believe Statisticians need to get better at step zero for our field to continue flourishing. The zero-problem is this: which methodologies should we support? Where should we direct our attention? Our journals are (unfortunately) filled with methodologies that lack a product-market-fit.

There is a direct path to ensuring the methodology is fit: we can learn from others. This alternative path leverages *and reinforces* our central role in the network of quantitative subdisciplines. When we learn popular techniques from others, they already have product-market-fit and likely already have practical-know-how. Our job is to give it a model and a theory (broadly interpreted) and to make it into something that other researchers might enjoy. Maybe this also inspires new algorithms and new estimators. Maybe it does not. By developing this framework, we enable other fields to learn. Importantly, this is not an entirely new way of doing Statistics; its the way it is already happening.

In this process, methodologies are not derived, but rather methodologies evolve. We Statisticians can play a fundamental role in methodologies evolving and reaching consilience, but we should stop assuming that product-market-fit is easy. Instead, we should recognise the more realistic role that we play in the evolution of quantitative methodologies and leverage the subdisciplines that are simultaneously and endlessly refining numerous different methodologies and testing their product-market-fit. Varimax is a parable for this point.

1 Thank you

Thank you to the Discussion Meetings Committee for hosting this discussion. Thank you Dr. C. Grazian for chairing the meeting. Thank you Professors Hoff and Pensky for leading the discussion. Thank you to the discussants for your thoughtful and inspiring comments. It was an honour to take part in this venerable tradition.

We are excited for the interest in Varimax. Many of the discussants highlight additional areas that need more exploration. We agree. There is so much more work to be done. We will use the rest of this rejoinder to find common themes for future directions and emphasise some ways that we might learn from the subdisciplines to pursue these directions. We aim bring together some key

threads in this discussion; as such, many excellent and interesting questions and comments from the discussants are not captured here. In places where we express scepticism, we do not say this to discourage any research in these directions, but instead to identify the key challenges that we suspect will need to be addressed with various approaches.

2 Interpreting factors

2.1 Rotations

The basis for Principal Components Analysis (PCA) comes from the singular value decomposition (SVD), which provides the optimal low-rank approximation for a data matrix in a least-squares sense. We show that Varimax can pick new axes for this subspace that are reasonable under a broad class of models. However, Varimax is certainly not the only way to choose the axes (i.e. rotate the PCs).

Kaizheng Wang gives a counter example where Varimax fails because the natural (sparse) factor coordinates are platykurtic. Similarly, Bickel, Bean, Chen, and Sarkar highlight that our theorem requires Leptokurtosis, a specific kind of non-Gaussianity. Some Independent Components Analysis (ICA) objectives can recover natural coordinates in other or perhaps more general conditions. In our limited experience with social networks, there is always degree heterogeneity and this heterogeneity always manifests with radial streaks in the pairs plot (e.g. Figure 2 in the discussed paper). These radial streaks are a key diagnostic for leptokurtic factors. In such situations, in our experience, other rotation techniques tend to create local optima that are not empirically beneficial.

Bickel, Bean, Chen, and Sarkar mention an assumption from the literature on Non-Negative Matrix Factorisation about 'pure nodes or anchor words' (Donoho & Stodden, 2003). Under this identifying assumption, for each column in Z, there must be a row of Z that only loads in this column. While the existence of such pure nodes will likely (approximately) exist under independent leptokurtic factors in fixed dimension k, finding them might be tricky. Moreover, aesthetically, it seems not as satisfying as an objective function (like Varimax) that smoothly incorporates all data. In fact, it was precisely the assumption of 'pure nodes' that led us to search for smoother alternatives; we then happened upon Varimax.

It is tempting to consider new types of rotations and we do not wish to discourage that. In addition, we hope that we can also direct our attention to unifying the disjoint literatures on rotations. The literature on ICA has developed theory and algorithms that are comparatively underdeveloped in the literature on factor analysis. At the same time, the deeper theoretical understandings in the literature on Independent Component Analysis rests primarily on the heuristic of non-gaussianity, which seems comparatively underdeveloped to the heuristic notions of sparsity, radial streaks, and simple structure in the literature on factor analysis.

To re-emphasise a point of the paper, Varimax has survived for close to 70 years despite strong pressure against it. We suspect that this survival is linked to its empirical fitness. What is it about Varimax that make researchers tend to prefer it to other rotations? Perhaps 'leptokurtosis' is a better assumption? Alternatively, perhaps Varimax is better behaved algorithmically. Perhaps, it has better statistical performance in the presence of noise? Perhaps it is the 'Kaiser normalisation' that we do not account for in our paper? It is possible that we will discover more beneficial rotations for certain settings and as theoreticians, we are all drawn to that prospect. It is also important to further understand the successes that have existed over the past 90 years and ensure our theory captures those success as fully as possible. This point aligns with the first paragraphs of this rejoinder.

Going forward, a key aspect of this theory could be a more unifying lens to understand rotations and their relationships. In this vein, Siroki and Strimmer ask a brilliant question that unifies the spirit of ICA with the algorithms of factor analysis: '[is the] leptokurtic condition strict enough and sufficient to single out Varimax or whether perhaps the whole orthomax family may be compatible with it?' We do not know, but we are curious and hope that someone will find out! For a path forward, it will likely be helpful that Chu and Trendafilov (1998) give the first- and second-order conditions for Orthomax. We hope for more inquiries that bring together the richness of understanding that we have from 90 years of factor rotations and 30 years of ICA.

2.2 Sense making

> Vintage factor analysis is a powerful tool for uncovering patterns of association, but we must still go through the difficult interpretative work of trying to assess why the patterns of correlation are present, and what they mean.
> -Tyler VanderWeele

VanderWeele argues that the meaning of factors extends beyond anything inside the factor analysis; it is a necessary post-hoc step to give them meaning. We agree. He continues that Thurstone's notion of a 'Scientifically Meaningful Category' is ambiguous. We strongly agree. Speaking to our analysis of scientific publishing, he writes 'There are various ways specific disciplines might be precisely defined, with varying levels of scope, but this then is again a conceptual, rather than an empirical, question'. Here, we partially disagree. Whenever we seek to cut nature at the seams (e.g. perform clustering or factor analysis), it is important to know *the purpose of cutting*; identifying our purpose is clearly conceptual. That said, once we understand our purpose, we see no reason to exclude empirical evidence from such tailoring.

Yang Liu discusses the classical framing of 'sense making' with rotations. In particular, multiple different rotations can all make sense and can have different types of sparsity. Liu presents an interesting counter example from *The development of hierarchical factor solutions* (Schmid & Leiman, 1957), where the second solution is rank deficient, but has a convenient *hierarchical* interpretation. This suggests a connection to hierarchical clustering.

2.3 Interpreting with hierarchies

Liu was not the only discussant to highlight a connection to hierarchies. Cai, Yang, Zhao, and Zhu propose using Varimax within previous approaches to hierarchical clustering. Moreover, VanderWeele correctly critiques our analysis and the semi-parametric factor model for lacking hierarchical structure; 'the non-hierarchical nature of the categories as the number of factors increases (Section 3.1.3), indicates that while these are potentially interpretable categories, they are not fixed scientific realities'. Here, there is room for the theory and methods of hierarchies and factor analysis to hybridise and grow.

Unbeknownst to most Statisticians, a great deal of statistical theory for hierarchical clustering has been developed in the literature on phylogenetic tree reconstruction. In that literature, 'hierarchical clustering' is not simply an exploratory technique. Rather, it deeply informed by the theories and models of evolution. This backdrop has provided an environment with strong selection pressures for the fittest hierarchical clustering methodologies. We should all seek to incorporate what they have learned, identify the parts that likely generalise to other fields, and hybridise it with techniques from other subdisciplines.

In a forthcoming manuscript, Sijia Fang and co-author Rohe study one such relationship in hierarchical modelling of social networks. They propose the $\mathbb{T}$-Stochastic Graph model; `vsp` can be used to identify part of the hierarchy $\mathbb{T}$. We hope to see much more inquiry into the relationship between factoring and hierarchies and much more inquiry into hierarchical clustering.

3 Other models

Marianna Pensky's simulations suggest that in certain settings, more fully (and still correctly) specified model fitting techniques can out-perform `vsp`. We imagine `vsp` as one step in a fitting pipeline. Even in situations where one does not use `vsp` in the final estimates, it can (1) provide the opportunity to first diagnose what structure the latent space appears to have and then, potentially and (2) provide a statistically consistent initialisation for a more refined fitting procedure. In particular, we hope that this insight could be built into the choice of latent space prior for Bayesian approaches as Xiaoyue Niu discusses. This choice of prior can be diagnosed by inspecting the pairs plots (i.e. Figure 2 in the discussed paper).

What other models can we better understand with rotations? We are optimistic that Statisticians will become more involved in making sense of the black box Large Language Models (LLMs) such as the GPTs; within these models, there are 'low-dimensional' embeddings with $K > 700$. These essentially have $K > 700$ factors, where the individual dimensions have not yet been interpreted.

These dimensions were not computed with SVD, but rather with Stochastic Gradient Descent and a far more complicated objective functions. But the same questions remain. What do these dimensions mean? Do they display radial streaks (we hypothesise that they do)? If so, perhaps one component of that sense making might be factor rotations and hierarchies.

4 Localisation; regularisation, normalisation

Alexander Van Werde asks 'Are there situations where one should be careful [applying the theory of the current paper]?' In our experience, the biggest obstacle to successfully using PCA and Varimax in empirical applications is the problem of localisation, something that is drastically under discussed in any related literature.

Localisation can be diagnosed. Look at the pairs plots of your Varimax factors (i.e. Figure 2 in the discussed paper). Do you see any axes that do not have a streak, but instead look like a one to three outliers? If so, then we say that the component has localised. It might have identified a very important data point/measurement. Or, it could be an artefact of noise. Either way, it is irritating to dedicate an entire dimension to identifying a few measurements. Sometimes, all of your dimensions will look this way and you might need to dramatically increase k to find meaningful factors. In classical factor analysis, there are 'Heywood cases' and in neural networks there is 'neural collapse'; in an heuristic sense, we think of these three things as similar types of failures of three distinct techniques.

Chen and Xu combine `vsp` into one objective function. However, there is an extreme tension between PCA and Varimax that is not discussed in the paper. PCA can localise, with extreme outliers in one element (or perhaps a few elements). Such localised dimensions have a massive kurtosis. So, both PCA and Varimax will enjoy finding these dimensions. By combining PCA and Varimax into a single objective, we worry that it would amplify these failures.

Le et al. (2017) provide the first improved bounds for spectral convergence with regularisation. Ke and Wang (2022) use the phrase 'pre-PCA normalisation' and discuss the benefits of normalisation and regularisation. Zhang and Rohe (2018) illustrate the types of patterns in random graphs that generate localisation and how regularisation addresses this issue. More work is needed in these directions. In particular, there is relatively little methodological and applied work that could communicate most directly with newcomers to PCA with high dimensional data.

5 Theory

Han and Zhang provide a large step forward in our understanding of the Varimax solution to U when the dimension K increases. There is room to explore how it behaves with $\hat{U}$. In a related vein, we suspect that the bound in our main theorem can and will be improved. This is not something that we are currently pursuing. Joshua Cape hints at a result on the asymptotic normality of the rows of $\hat{Z}$. We hope that others will join them in exploring these direction.

References

Chu M. T., & Trendafilov N. T. (1998). Orthomax rotation problem. a differential equation approach. *Behaviormetrika*, *25*(1), 13–23. https://doi.org/10.2333/bhmk.25.13

Donoho D., & Stodden V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? *Advances in Neural Information Processing Systems*, *16*.

Ke Z. T., & Wang J. (2022). 'Optimal network membership estimation under severe degree heterogeneity', arXiv, arXiv:2204.12087, preprint: not peer reviewed.

Le C. M., Levina E., & Vershynin R. (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms*, *51*(3), 538–561. https://doi.org/10.1002/rsa.20713

Schmid J., & Leiman J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61. https://doi.org/10.1007/BF02289209

Zhang Y., & Rohe K. (2018). Understanding regularized spectral clustering via graph conductance. *Advances in Neural Information Processing Systems*, *31*.

https://doi.org/10.1093/jrsssb/qkad074
Advance access publication 24 August 2023