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.

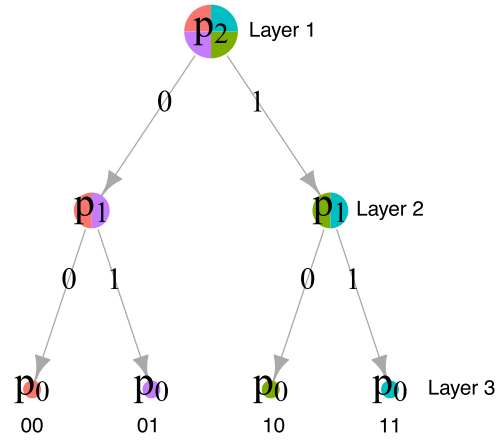

Figure 1. A four-cluster binary tree stochastic block models (BTSBM) (Li et al., 2020). The binary tree has three layers where Layer 1 includes all nodes, Layer 2 splits into two mega-communities {0, 1}, and Layer 3 further splits into four communities {00, 01, 10, 11}. Each colour corresponds to each 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.

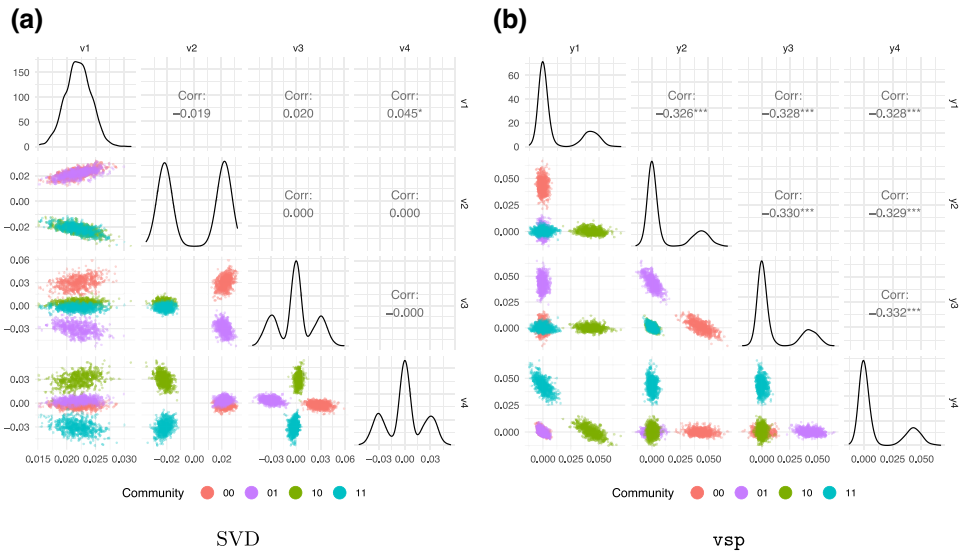

Figure 2. Scatter plot of pairs of principal components by SVD in figure (a) and pairs of varimax rotated components in figure (b). The colour corresponds to each community at Layer 3 in Figure 1. The radial streaks appear in figure (a) while the Varimax rotation aligns the streaks with the coordinate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

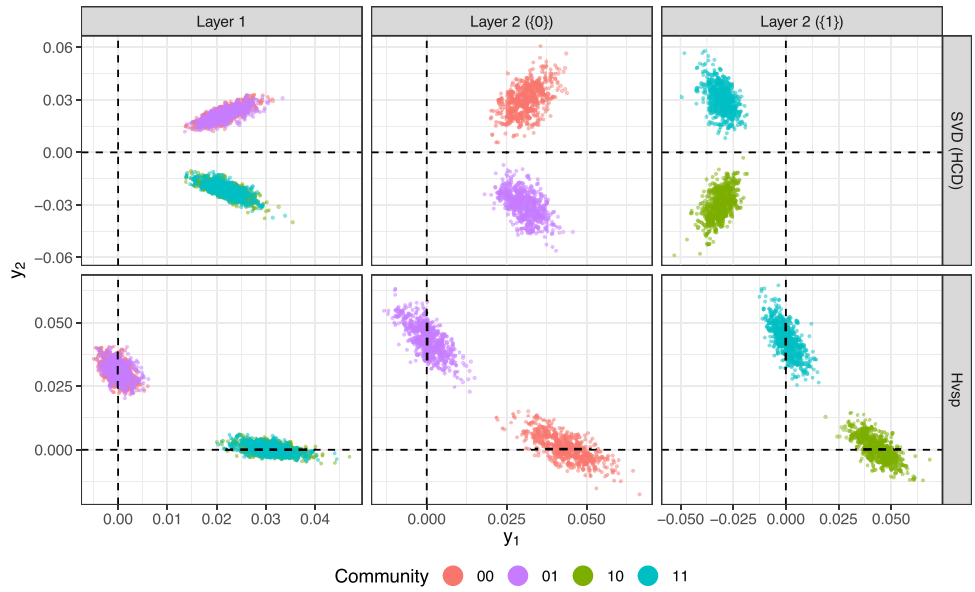

Figure 3. Scatter plot of pairs of principal components by SVD and pairs of Varimax rotated components. The rows correspond to hierarchical community detection (HCD-sign) (Li et al., 2020), which first performs SVD with dimension $k = 2$ and then assigns labels based on the sign of the second component, 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.

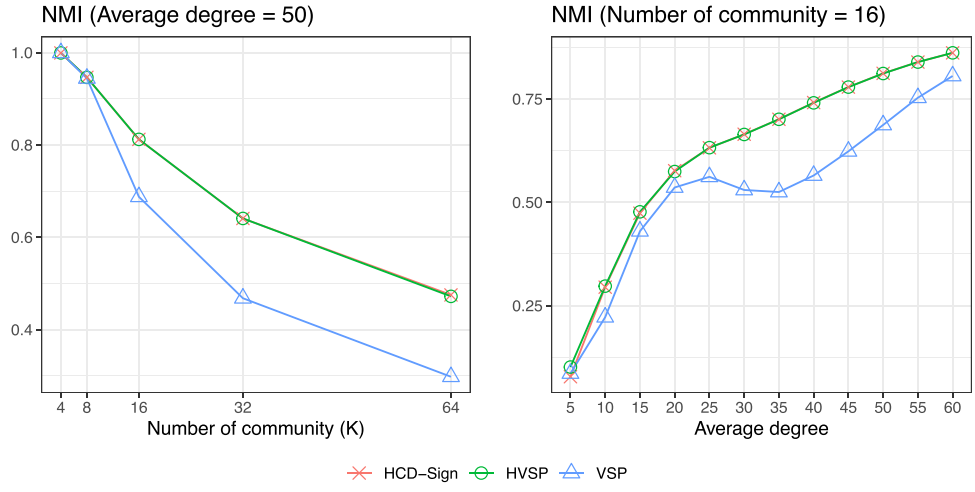

Figure 4. The normalized mutual information (NMI) (Yao, 2003) between the true and estimated labels obtained by HCD-sign, Hvsp, and vsp varying the number of communities and the average degree of nodes. The simulation setup follows Section 4.1 in Li et al. (2020). A larger NMI suggests better clustering 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.