An odyssey to obtain *interpretable* spectral embeddings from graphs

We will find four monsters.
We will confront each one with multiple pieces of intuition.
Then, slay them with theorems.

Karl Rohe; IFDS summer school 2021

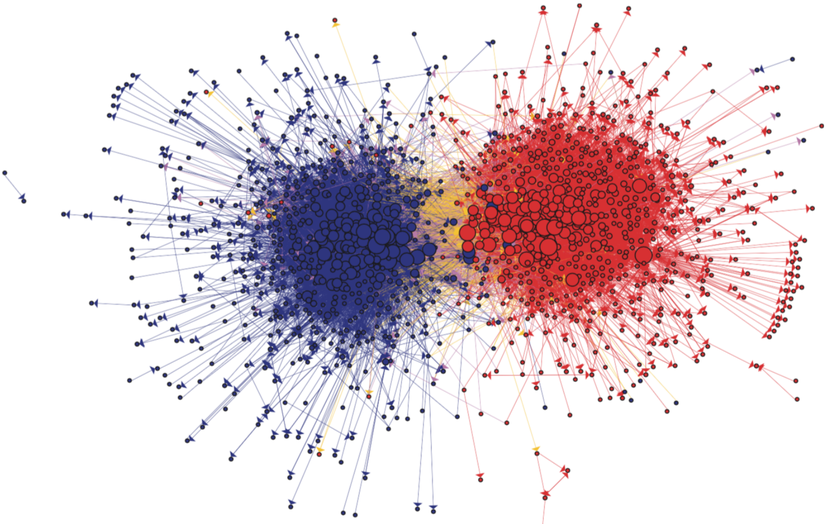
Please interrupt me

- It’s been awhile. We all have covid-isolation-brain. So, let’s practice.
- If you’ve read this sentence before I have read it aloud, please interrupt me before I get to this. ***Quick, say anything or “I don’t have covid-brain, that’s just you Karl”!***

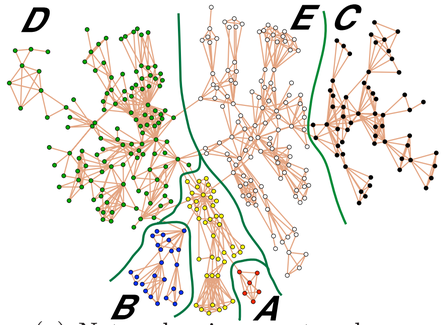
Poll the audience!

- How many PhD students? Postdocs? Other?
- Statistics / CS / EE / Math / Other?
- Who traveled the furthest?
- Who traveled the shortest?

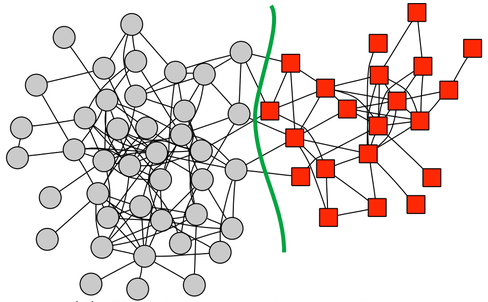
image source: http://ccn.ucla.edu/~jbrown/dti.html



Adamic, Glance. The political blogosphere and the 2004 US election: divided they blog. 36–43 (2005).



(g) Network science network . . .

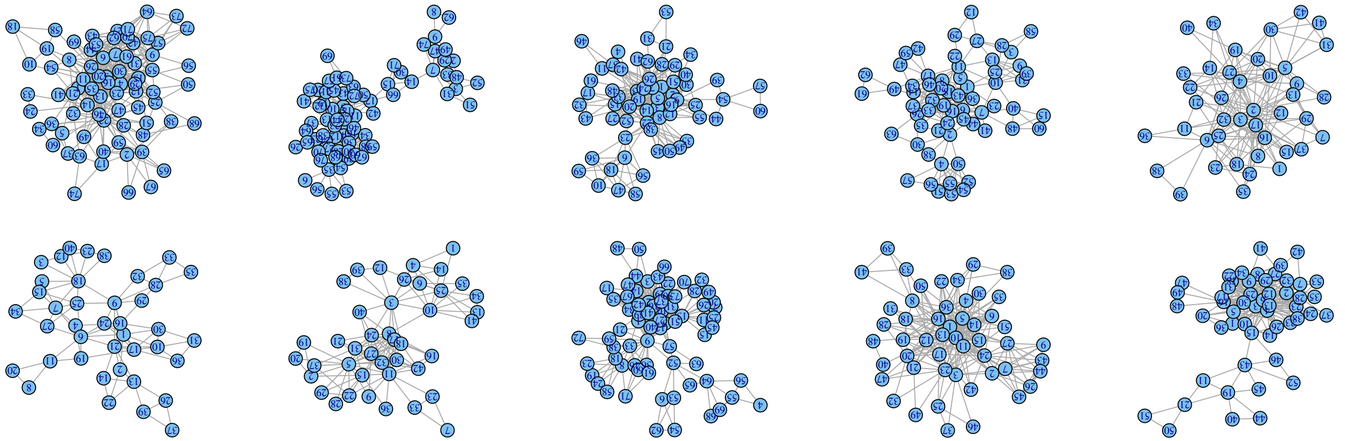


(c) Dolphins social network . . .

Leskovec et. al 2008

Graphs all over

- Brain graphs:
  DTI or correlation from fMRI
- Academic graphs:
  coauthorship or citations
- Animals observed together
- Online social network
  following/friends/retweets/comments/likes/…
- Webpages connected by hyperlinks
- “Bipartite graphs”
  which customers purchase which things or
  which documents contain which words (“LDA”)
- …



Graph Notation

- $V = \{1, \ldots, n\}$. The nodes in the people in the social network
- $A \in \mathbb{N}^{n \times n}$ where $A_{ij} > 0$ if i and j are friends. This is the adjacency matrix.

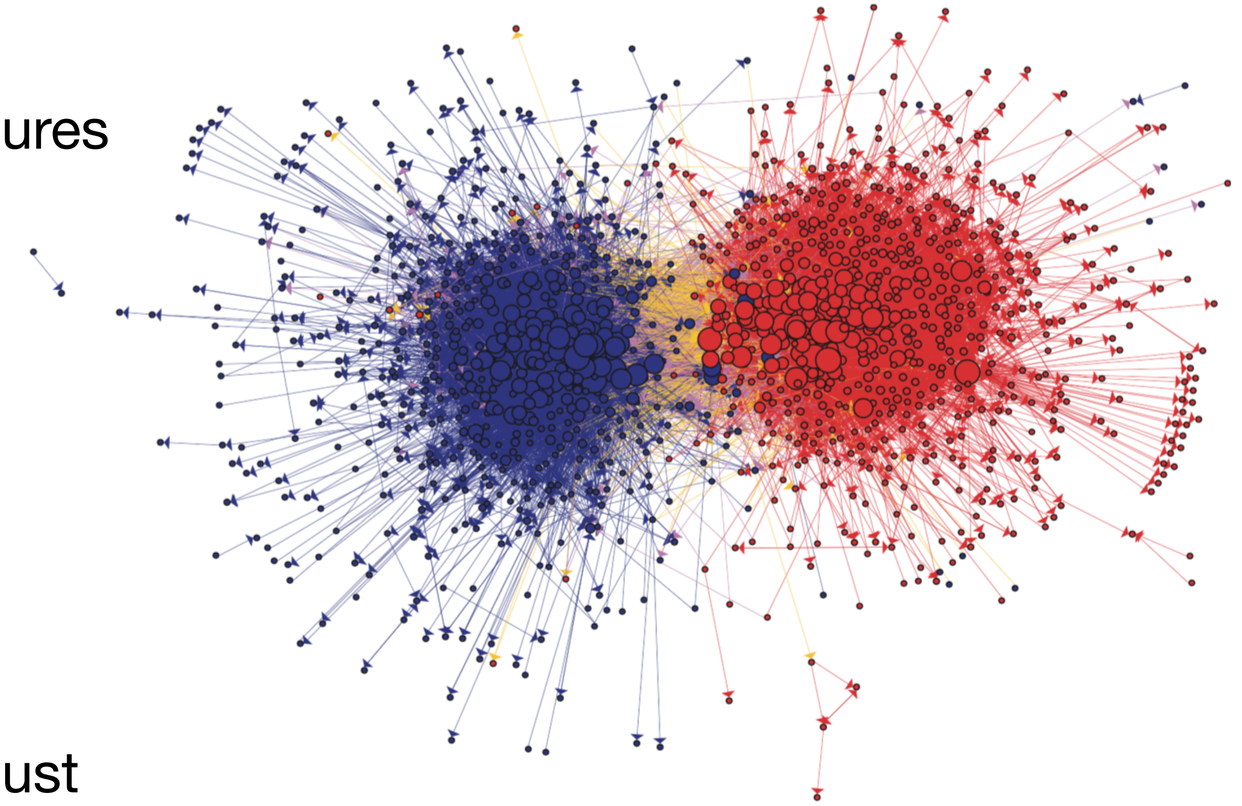
  $deg(i) = \sum_{\ell} A_{i\ell}$ is the "node degree"… the number of friends of node i.

- We don't infer A, as in graphical models. Throughout, A is data that we observe (e.g. friendships).

We want to “embed” the nodes/people in an interpretable way.

The embedding should (somehow) characterize the individuals.

- In this “political blogs” example, blogs link to each other.
- In the visualization, there are two latent features estimated from the data (democratic vs republican)
- From the graph, we give each node two numbers.
- We are interested in estimating WAY more dimensions than 2.
- In advance, we don’t know what the embeddings mean (e.g. D vs R)… they are just numbers computed from the graph.



Three motivating examples…

Time permitting, we will investigate each one.

- Example 1 is a graph on 22,000ish academic journals. If there are at least 100 citations from papers in journal i to papers in journal j, then $A_{ij} = 1$.
- Example 2 is a document-term bipartite graph from the abstracts of 144,136 academic papers; these papers all contain the phrase “factor analysis” in the title or abstract.
- Example 3 is a Twitter following graph on 120,000ish users sampled in the neighborhood of Hadley Wickham and Guido van Rossum on twitter.

Our approach centers on algorithms and science, guided by “statistical inference”

- We focus on algorithms that are easy to use.
    - quick to compute
    - no tuning parameters
    - no problems with local optima
- We validate the “easy to use” with simulation and data.
- We validate the “sense making” with heuristic understandings, statistical theory, and finally with applications.
- Our approach is “spectral” (i.e. uses eigenvectors / singular vectors).

The first monster:

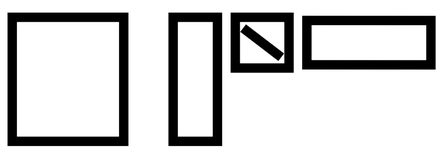
Why should eigenvectors of A tell us anything about a node embedding???

- We will confront this first monster with intuition from two places:
- First, from optimization… what are the eigenvectors solving?
- Second, from statistical models… if the graph is generated from a model with node embeddings, how do the eigenvectors “estimate” the model embeddings?
- (Then, we are going to confront a second monster and after that, a big theorem is going to slay both)

Gaining intuition from optimization…

Don't think about the $Ax = \lambda x$ definition of eigenvectors.
Instead, think about reconstruction error:

- For $U \in \mathbb{R}^{n \times k}$ with orthonormal columns and diagonal $\Lambda \in \mathbb{R}^{k \times k}$, minimize:
  $\|A - U\Lambda U^T\|_F^2 = \sum_{ij} \left(A_{ij} - U_{i\cdot}\Lambda U_{j\cdot}^T\right)^2$, where $U_{i\cdot} \in \mathbb{R}^k$ is the ith row of U



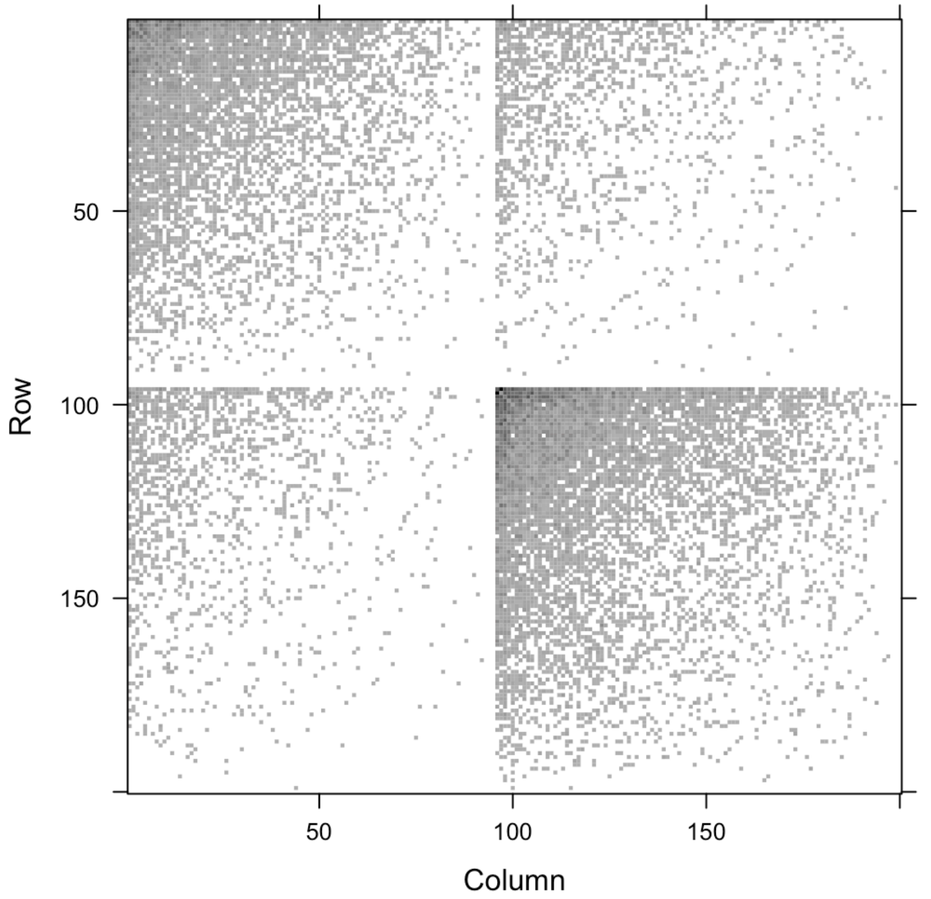
- The eigenvectors minimize this reconstruction error, subject to only getting k numbers $U_{i\cdot} \in \mathbb{R}^k$ for each node.
- The node embeddings $U_{i\cdot} \in \mathbb{R}^k$ are constructed so that $U_{i\cdot}\Lambda U_{j\cdot}^T \approx A_{ij}$

Gaining intuition from statistical inference…

The degree-corrected stochastic block model (DC-SBM) with 2 blocks.

- $\theta_1, \ldots, \theta_n \in \mathbb{R}_+$ control the node degrees; big θ_i, more friends.
- $z(1), \ldots, z(n) \in \{1,2\}$ assigns each node to a block (i.e. “community”) 1 or 2.
- $B \in \mathbb{R}^{2\times 2}$ describes the affinity between the two blocks.
- $\mathbb{E}A_{ij} = \theta_i \theta_j B_{z(i)z(j)}$
- I like Poisson distribution, but Bernoulli ok too.

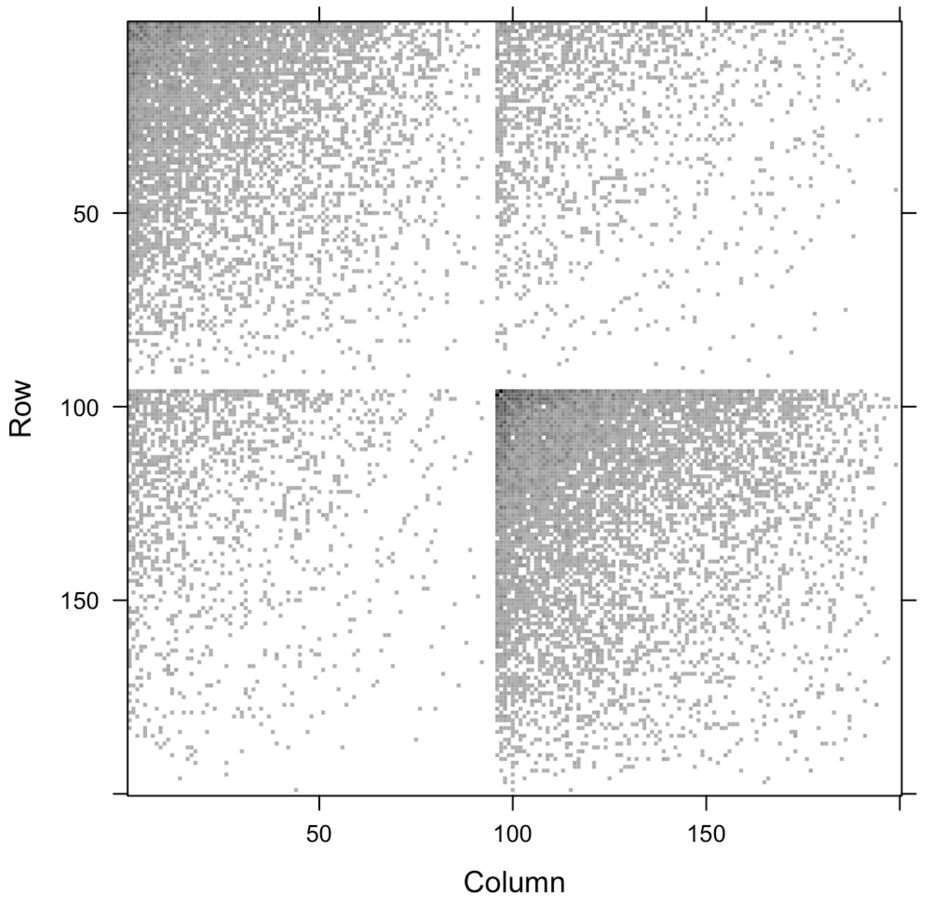
See code to generate this DC-SBM adjacency matrix:



What is the parametric-embedding for the DC-SBM?

- Define $Z \in \mathbb{R}_{+}^{n \times 2}$ with only one non-zero element in each row…
  - $Z_{i,z(i)} = \theta_i$
- Each row of Z embeds the rows.
- Moreover, $\mathscr{A} = \mathbb{E}A = ZBZ^T$ is linear algebra notation for the same model as before…
  - $\mathbb{E}A_{ij} = \theta_i \theta_j B_{z(i)z(j)} = Z_i B Z_j^T$
- We want to estimate Z.

See code to generate this DC-SBM adjacency matrix:

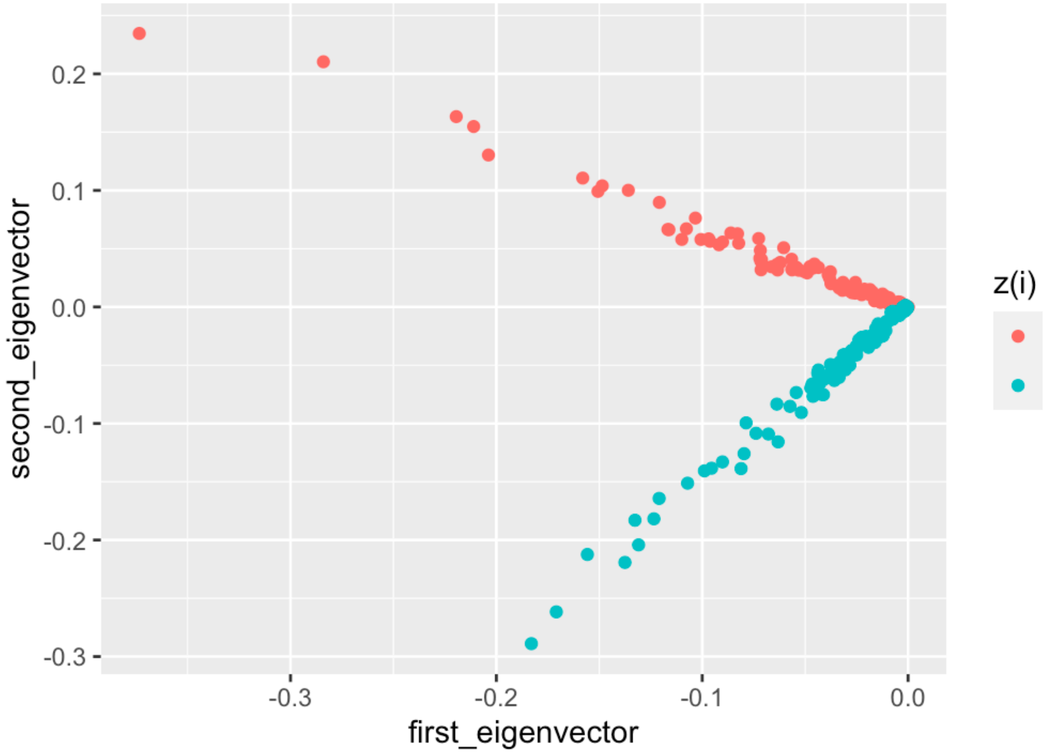


Eigenvectors estimate the parametric embedding (*kinda!!!*)

To see this…. Let's compute the eigendecomp of $\mathscr{A} = ZBZ^T$

- We want U, D as orthogonal and diagonal matrix with $\mathscr{A} = ZBZ^T = UDU^T$
- Take eigendecomp: $B = \Phi\Lambda\Phi^T$
- Define $U = Z\Phi; D = \Lambda$.
- Z has orthogonal columns.
  They are still orthogonal after orthogonal rotation Φ.
  Λ is diagonal. So, Λ is too.
- And they minimize reconstruction error:
  $UDU^T = (Z\Phi)\Lambda(Z\Phi)^T = Z\Phi\Lambda\Phi^T Z = ZBZ^T$.
- Indeed, sample eigenvectors are not sparse… they are a rotation Φ away from Z. See eigenvectors of A on right ->

Here are the top 2 eigenvectors of A, colored by block membership $z(i)$.
See code!

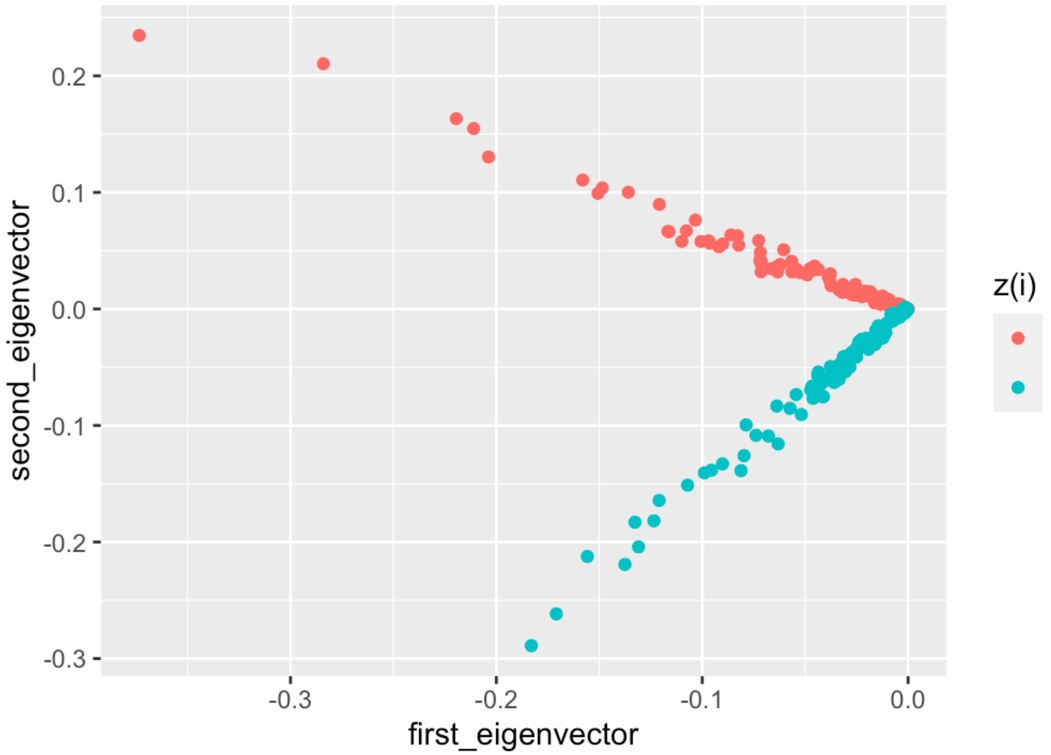


Fine print: to get unit length columns for U, normalize with diagonal matrix N of column norms of $Z\Phi$.
$U = Z\Phi N^{-1}; D = N\Lambda N$. This doesn't change orthogonality, diagonal structure, or reconstruction error.

Eigenvectors estimate the parametric embedding (*kinda!!!*)

- $U = Z\Phi$ is not Z …. ugh.
- This is the “kinda” in slide title.
- So, how to recover Z from $U = Z\Phi$??
- This is the second monster.

Here are the top 2 eigenvectors of A, colored by block membership $z(i)$.
See code!



Intuition for the first monster:

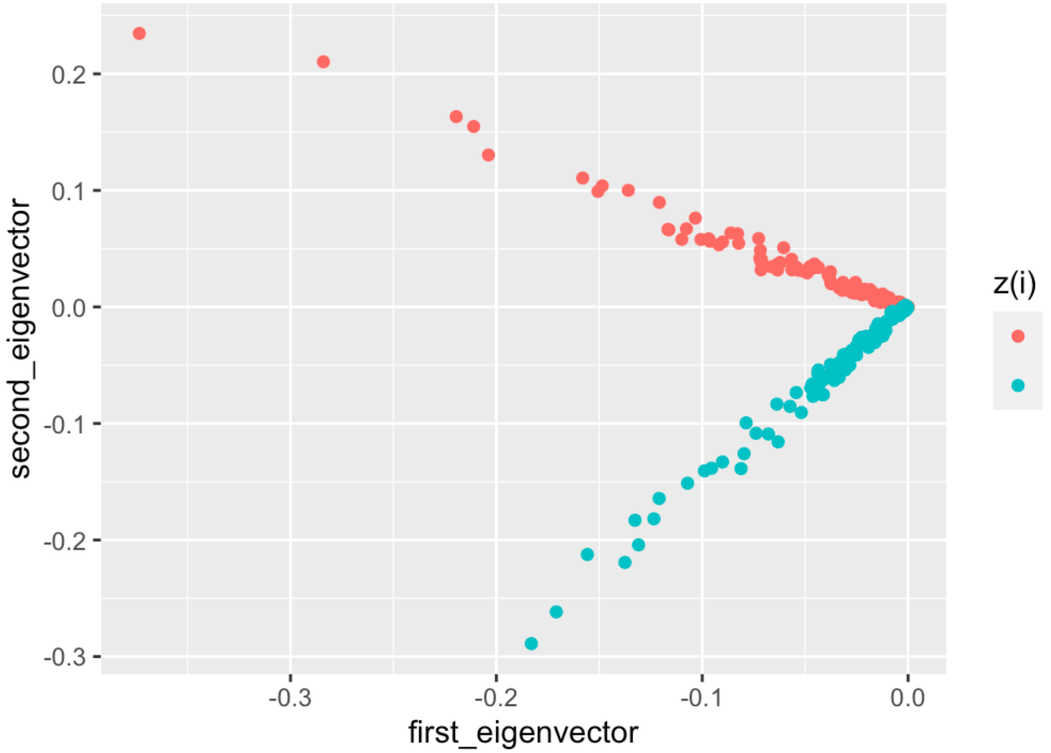
Why should eigenvectors of A tell us anything about a node embedding???

- **Optimization:**
  The eigenvectors provide the best $\mathbb{R}^k$ embedding in squared error loss:

$$\sum_{ij} \left(A_{ij} - U_{i\cdot} \Lambda U_{j\cdot}^T \right)^2$$

- **Inference:**
  If $\mathscr{A} = ZBZ^T$ and Z has orthogonal columns, then the eigenvectors of $\mathscr{A}$ are $U = Z\Phi$, for a small rotation matrix Φ.

In our path to recovering Z, we want to make sure to *mimic the structures of real data* as much as possible…



In our path to recovering Z, we want to make sure to *mimic the structures of real data* as much as possible…

Radial streaks in LL Thurstone's 1947 landmark textbook *Multiple Factor Analysis*.

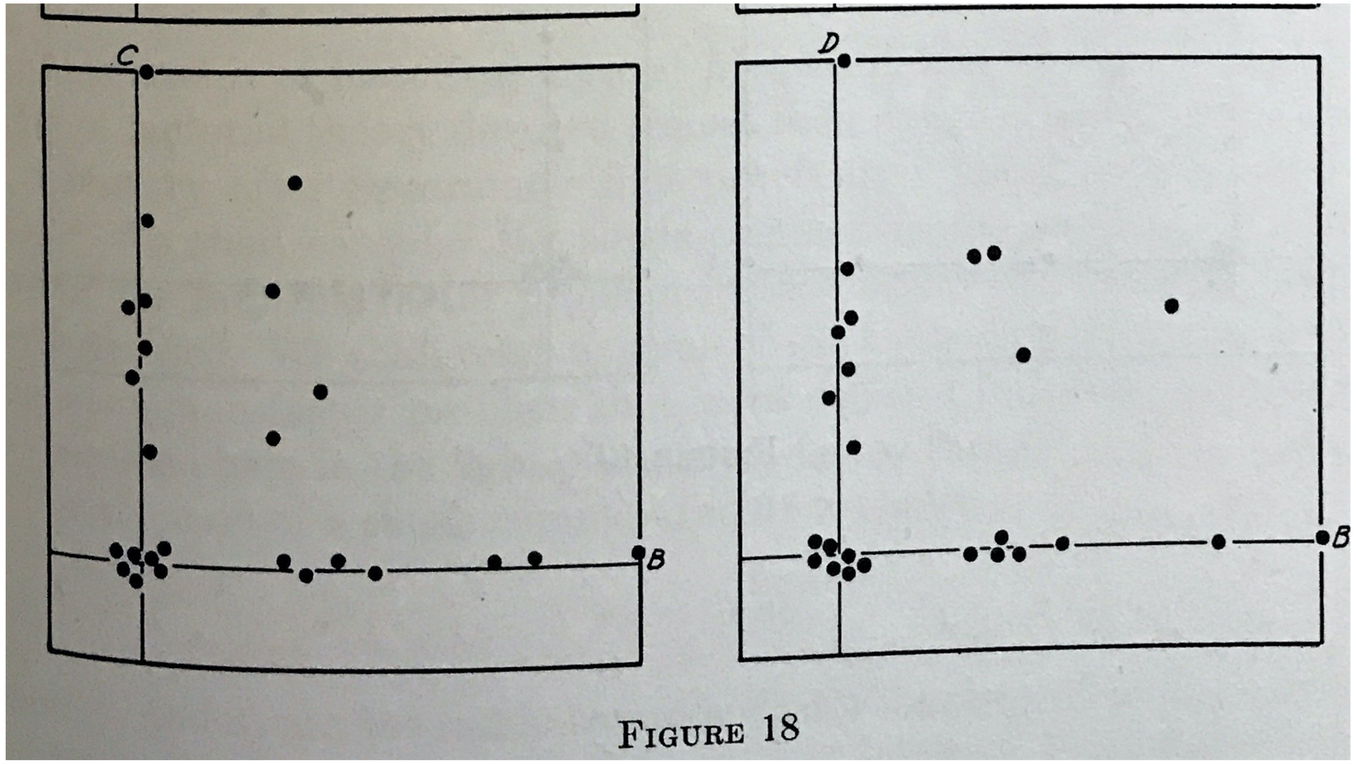
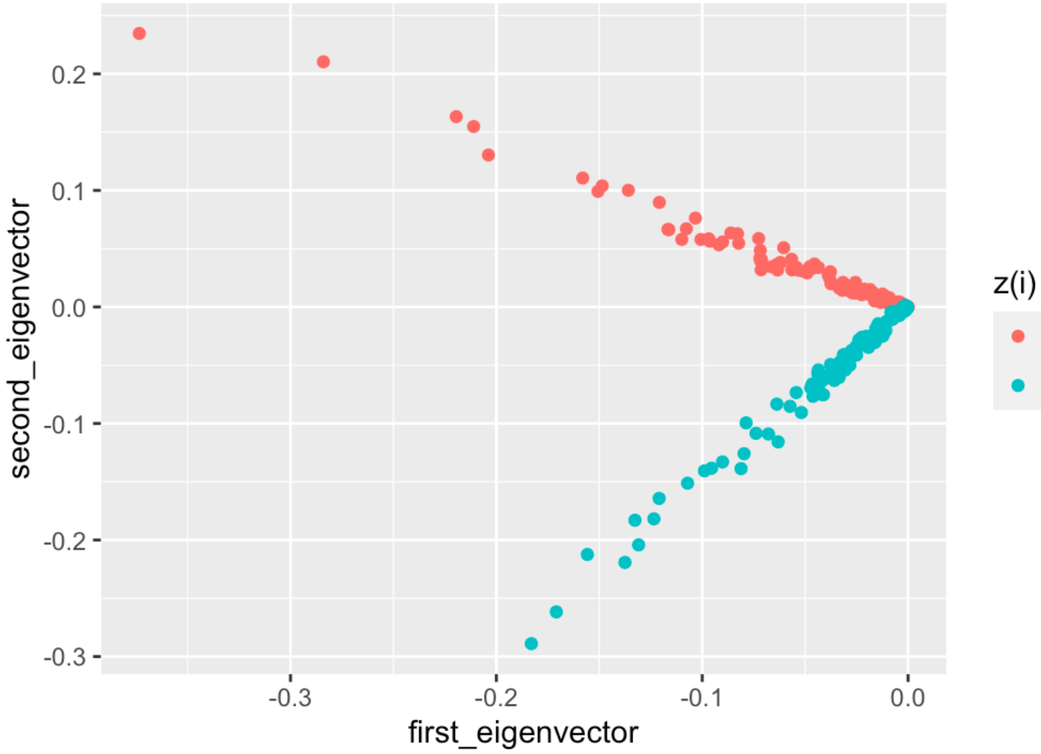
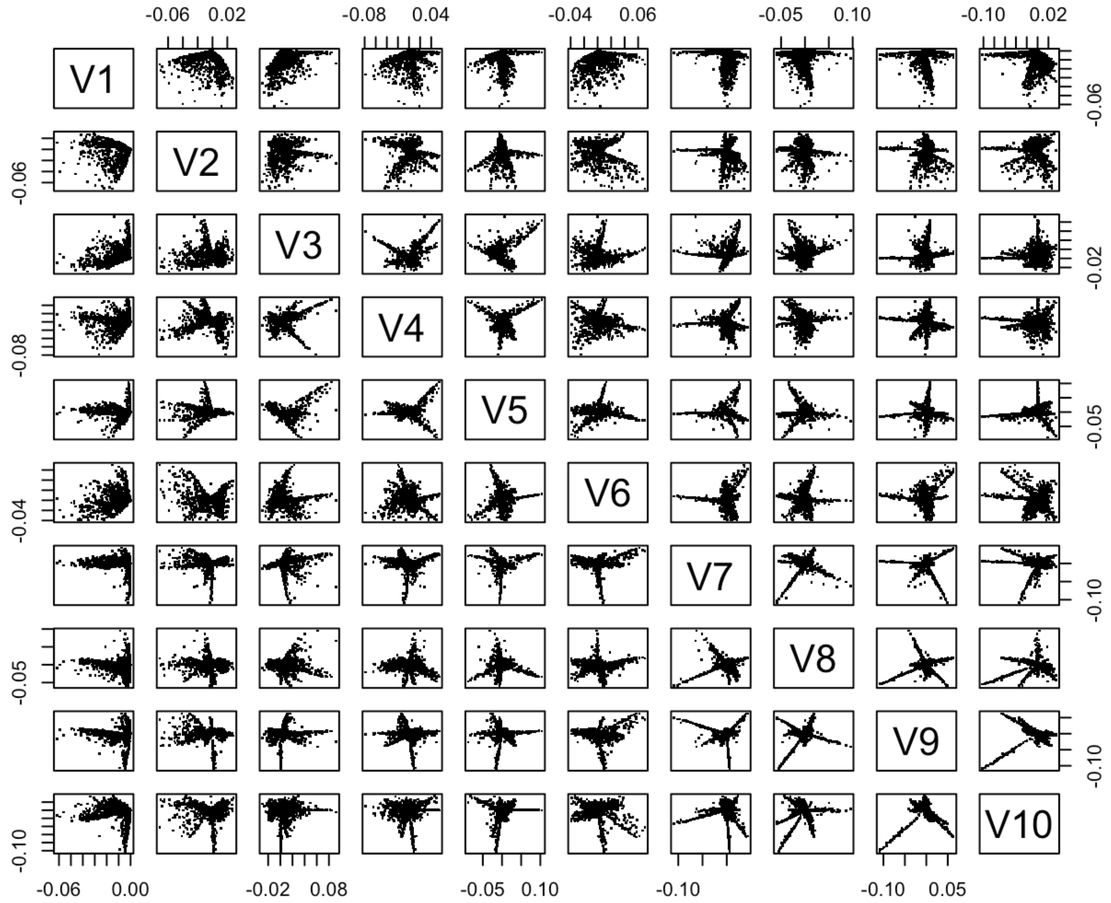


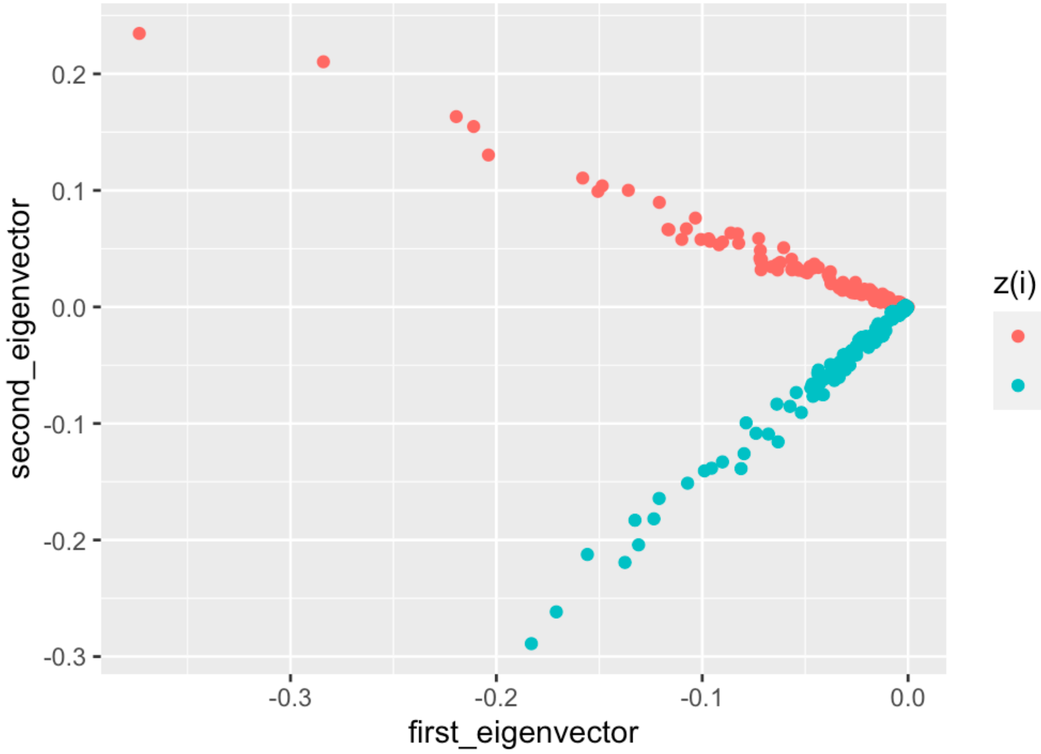
FIGURE 18



“Radial streaks” (LL Thurstone, 1947) appear in data!

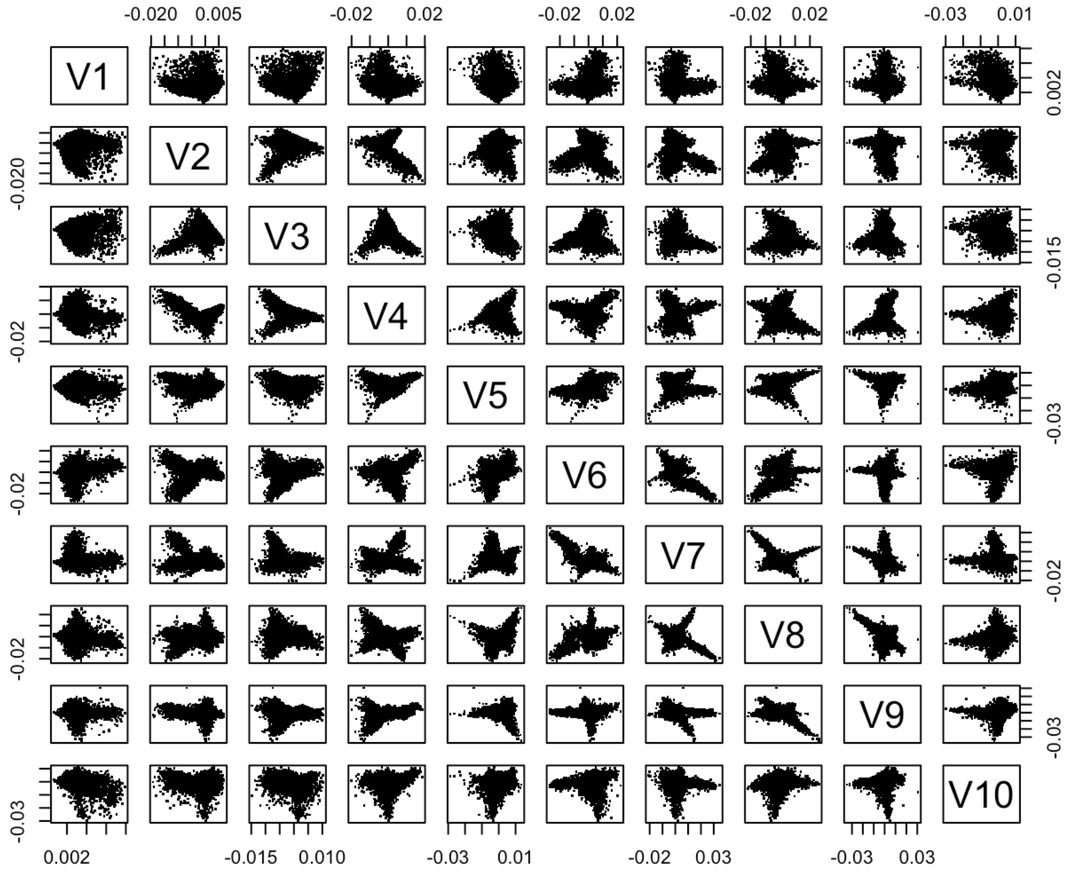
Journal-Journal citations; way more than 2 streaks!

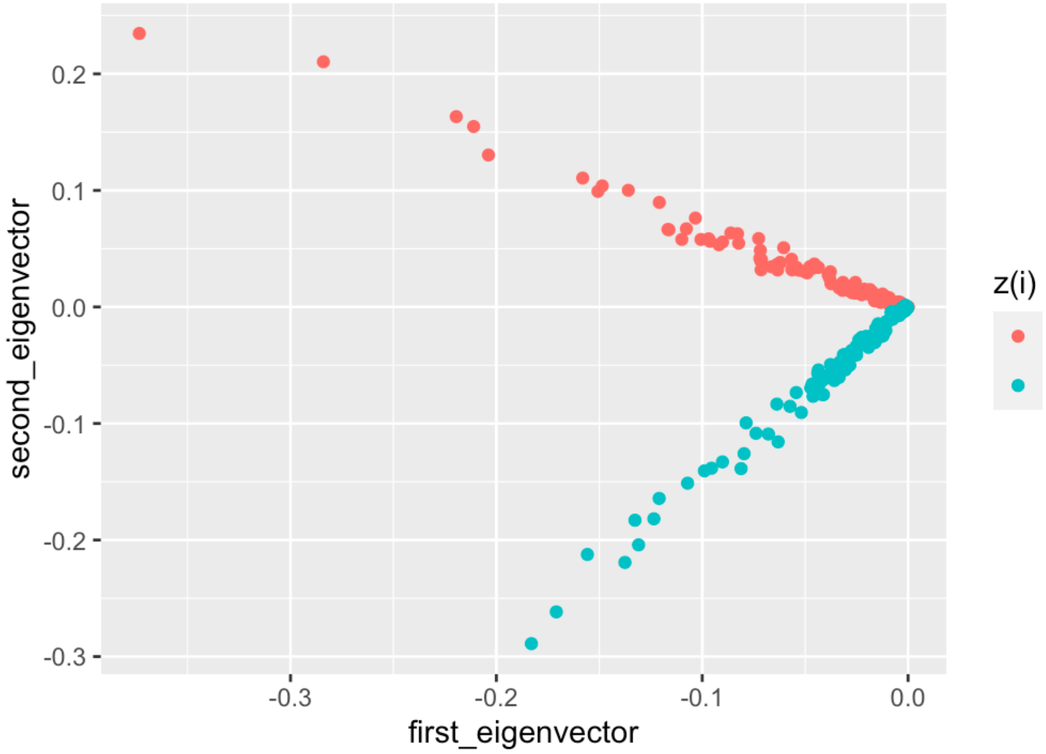




“Radial streaks” (LL Thurstone, 1947) appear in data!

Academic abstracts, represented as document-term graphs

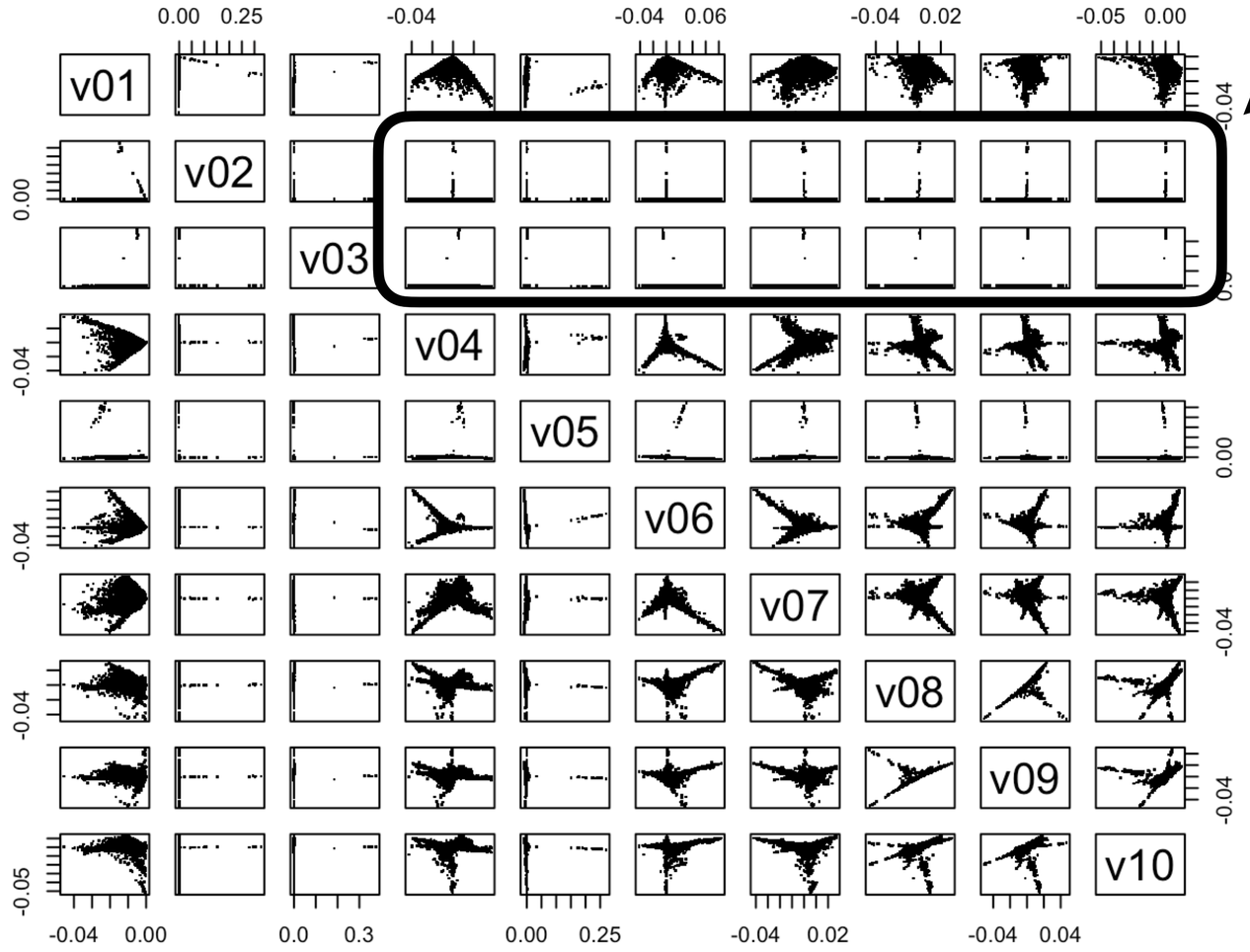


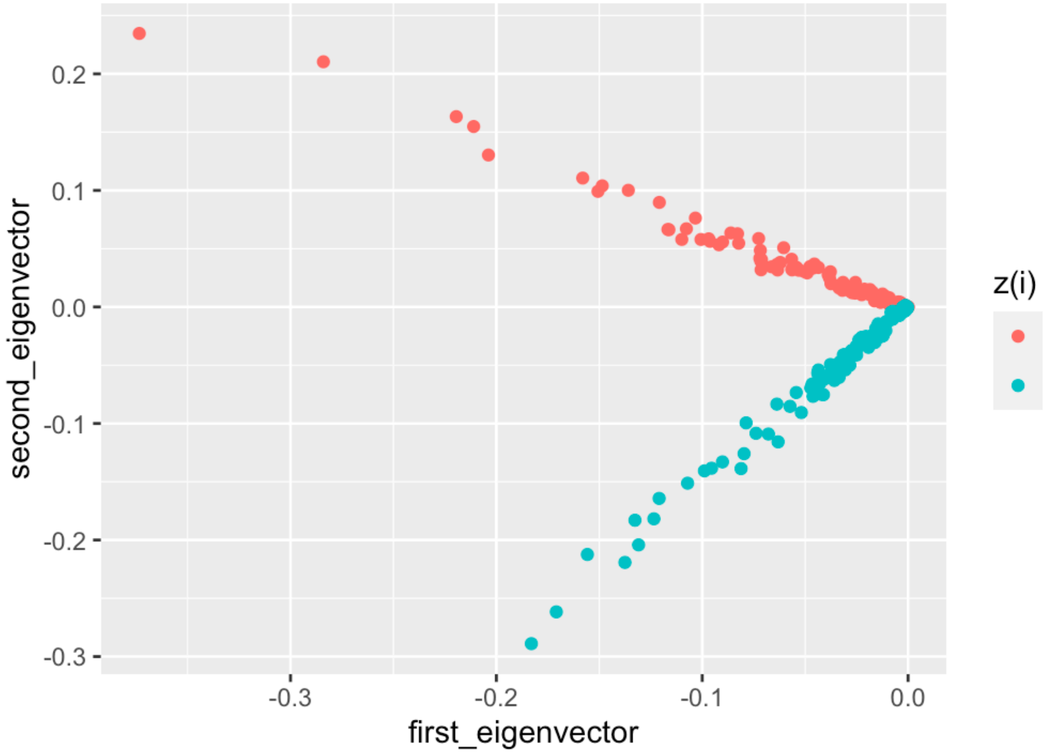


“Radial streaks” (LL Thurstone, 1947) appear in data!

A sample of the Twitter following graph

Localization! eeek!



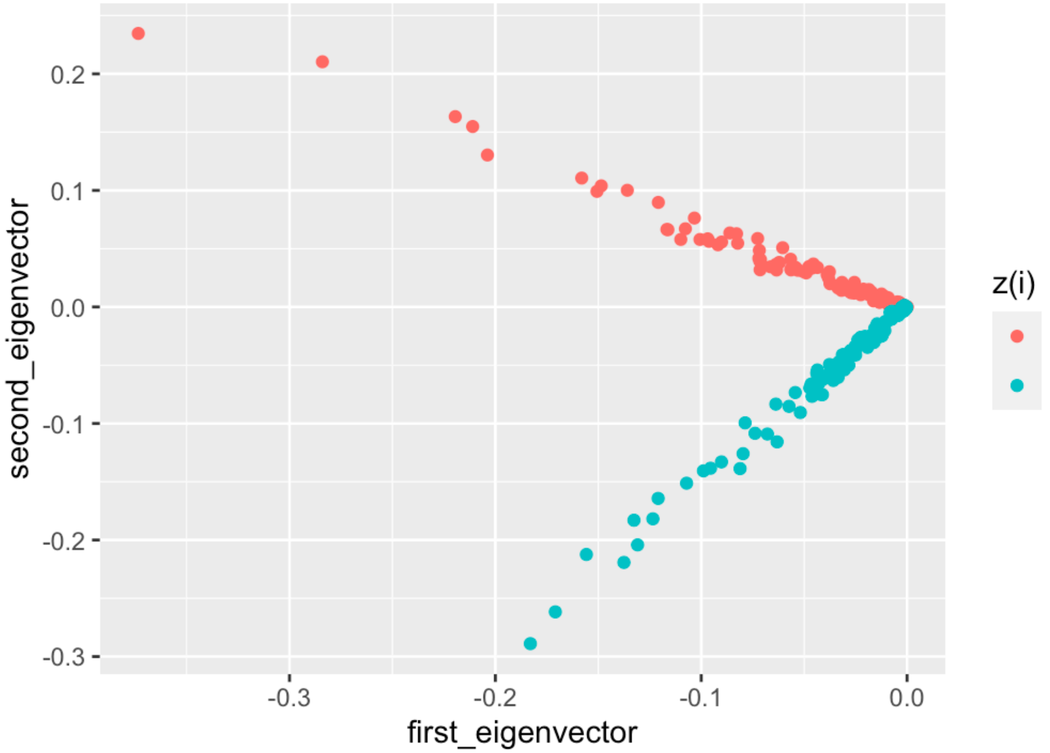


The second monster:

How do we recover Z from U ??

- We know that $U = Z\Phi$ and Φ comes from eigen of symmetric matrix. So, it is orthogonal: $U\Phi^T = Z\Phi\Phi^T = Z$
- If we can find Φ^T, then we can recover Z as $U\Phi^T$.
- "Orthogonal rotation" sounds fancy. Instead, think of Φ as re-drawing the (orthogonal) axes for the points in the rows of U.
- How would you re-draw the axes on the right?
- Note that Z is sparse and U is not!
- We are going to find a "sparse rotation" of U.

Here are the top 2 eigenvectors of A, colored by block membership $z(i)$.
See code!



Interlude poll:

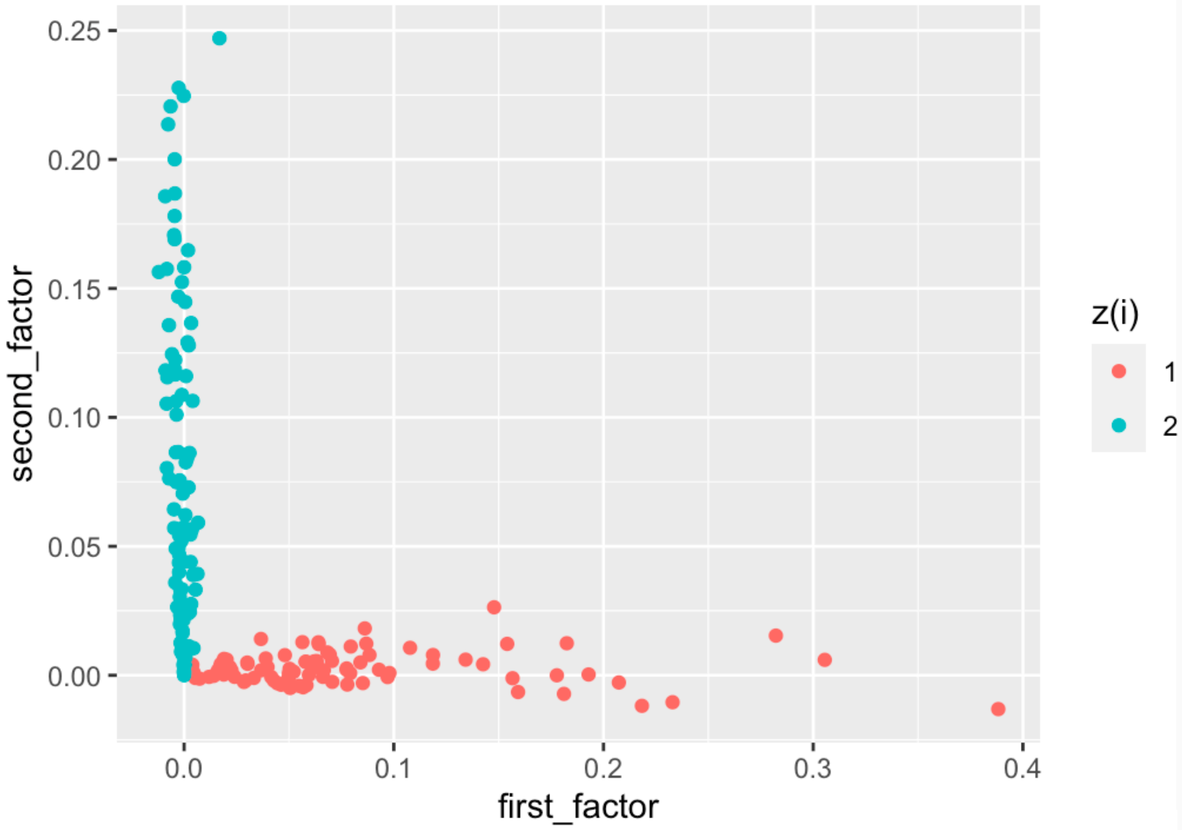
If you use k-means in a paper, would you give a citation?

The second monster:

To recover Z, take the "Varimax rotation" of U.

- Motivated by Thurstone, the Varimax rotation was proposed by Psychometrician Henry Kaiser in 1958.
- Already loaded into R. Widely used in factor analysis, often without citation. It's popularity there is akin to k-means to us. When you use k-means, do you cite any paper?
- This second monster is the key to making the embedding interpretable. Without this… we just have an embedding.

Here are the top 2 eigenvectors of A, after the Varimax rotation. See the code!

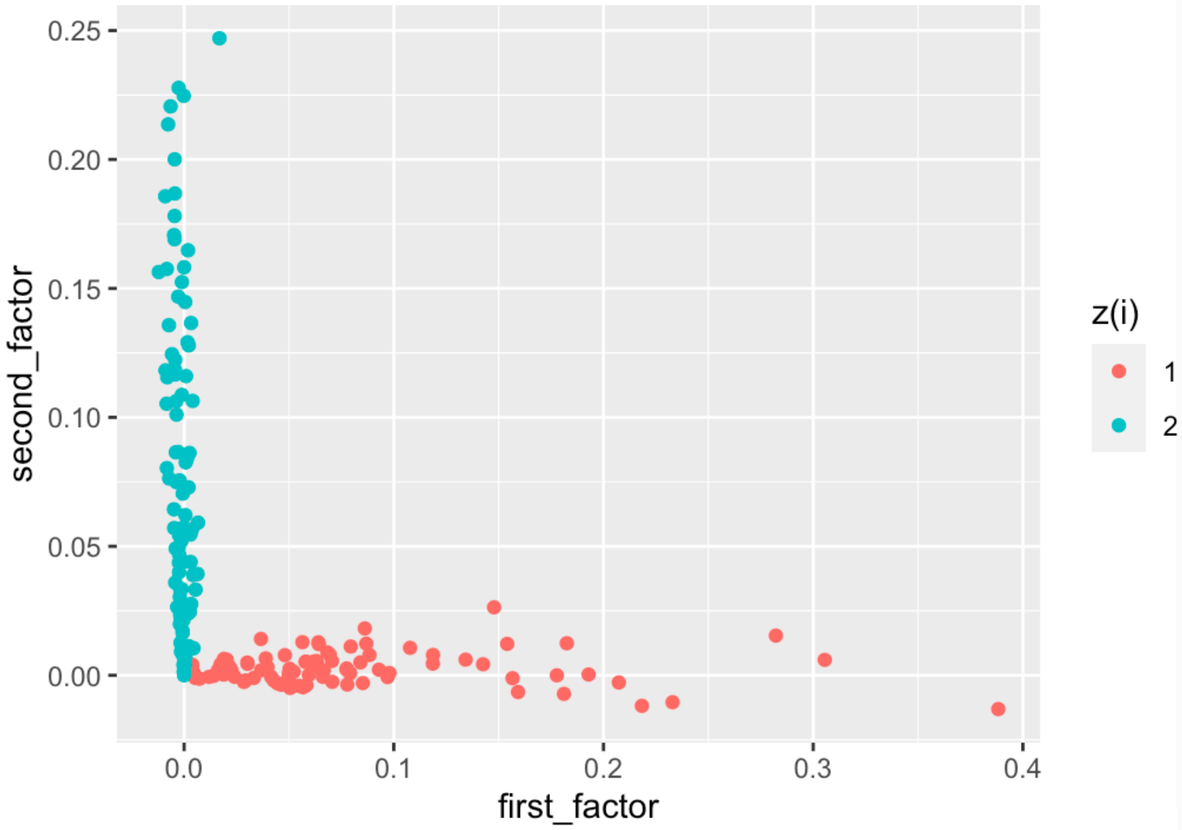


The second monster:

To recover Z, take the "Varimax rotation" of U.

- Varimax is hugely controversial. The only statistical theory for Varimax is from 1956 and it says that rotations are unidentifiable.
- Rohe, Zeng 2020 gives a positive theory for Varimax that I will describe.

Here are the top 2 eigenvectors of A, after the Varimax rotation. See the code!



The second monster:

To recover Z, take the "Varimax rotation" of U.

- The Varimax rotation solves

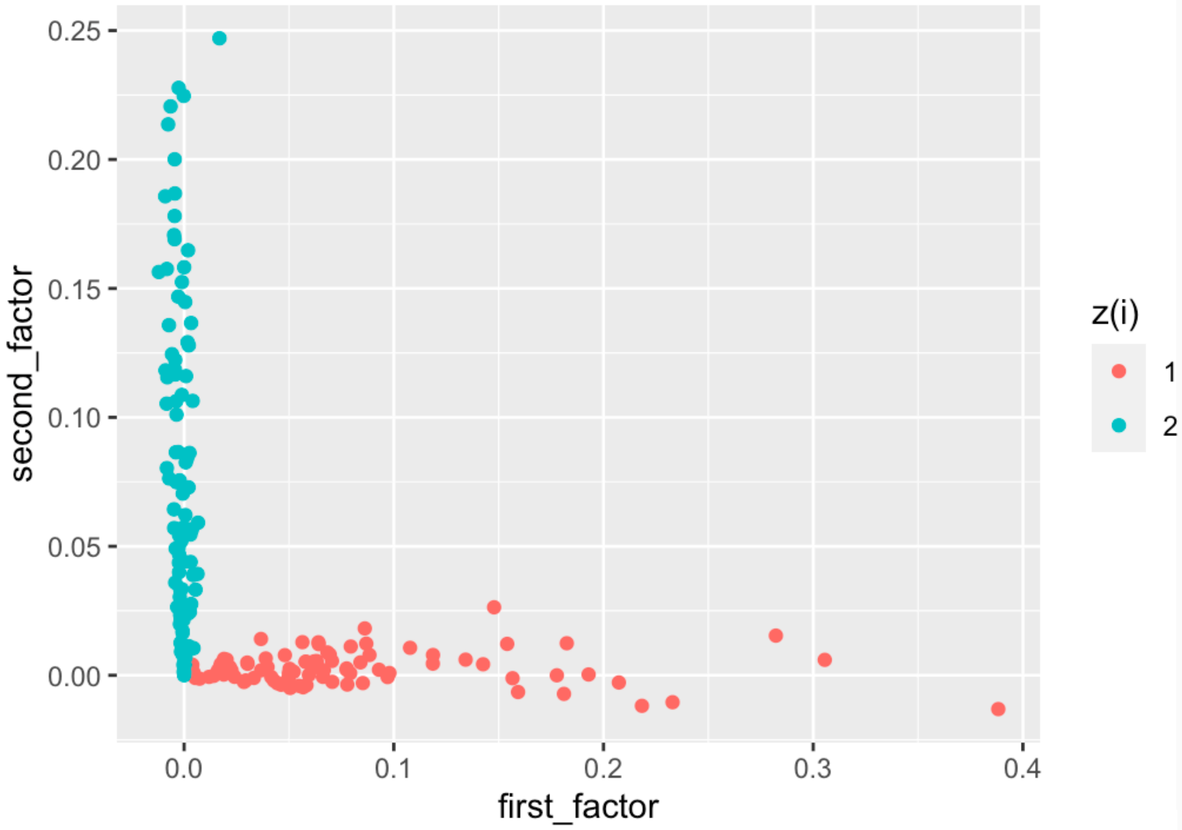
$$R^* = \min_{R;R^TR=I} \sum_{ij} (UR)_{ij}^4$$

- This figure gives the sample eigenvectors $\hat{U} \in \mathbb{R}^{n\times 2}$ computed from A, rotated with the Varimax rotation $\hat{R}$ computed from $\hat{U}$.

- $\hat{Z} = \hat{U}\hat{R}$

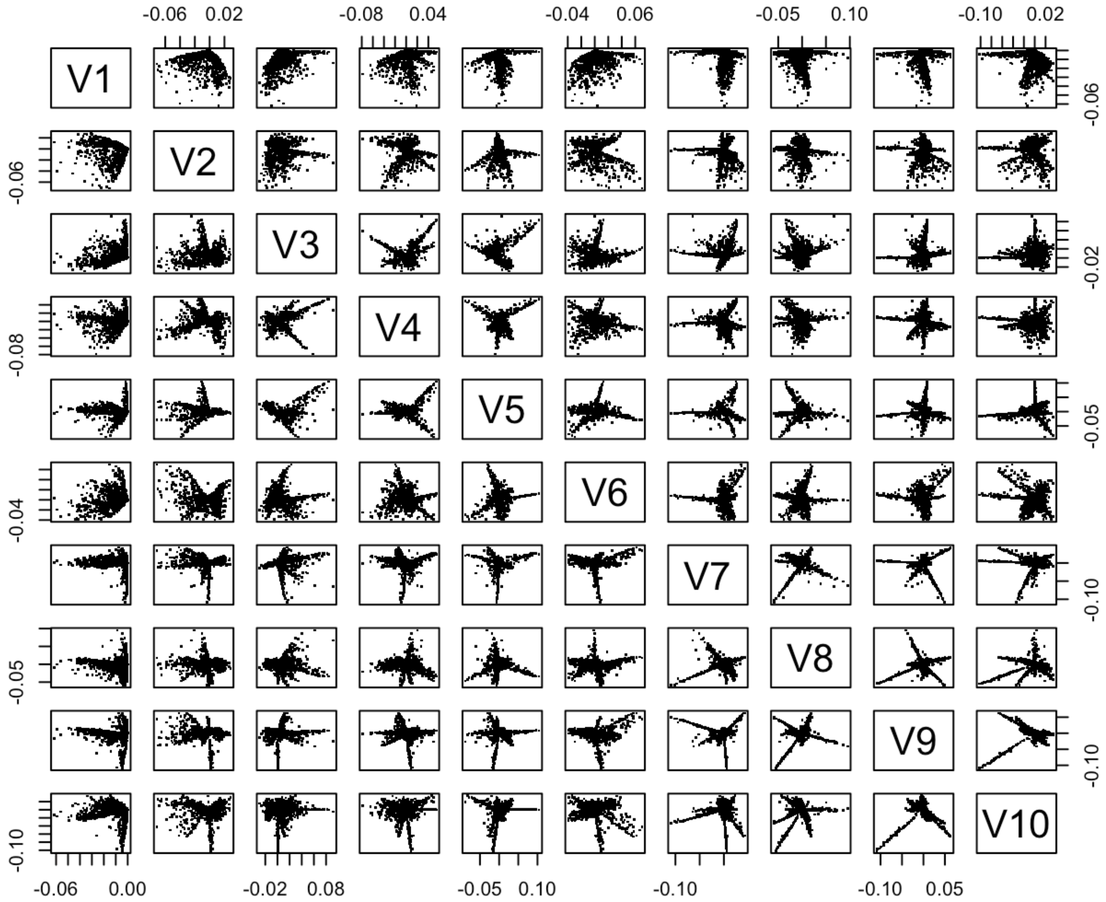
- The horizontal axis is the first column. The vertical axis is the second column.

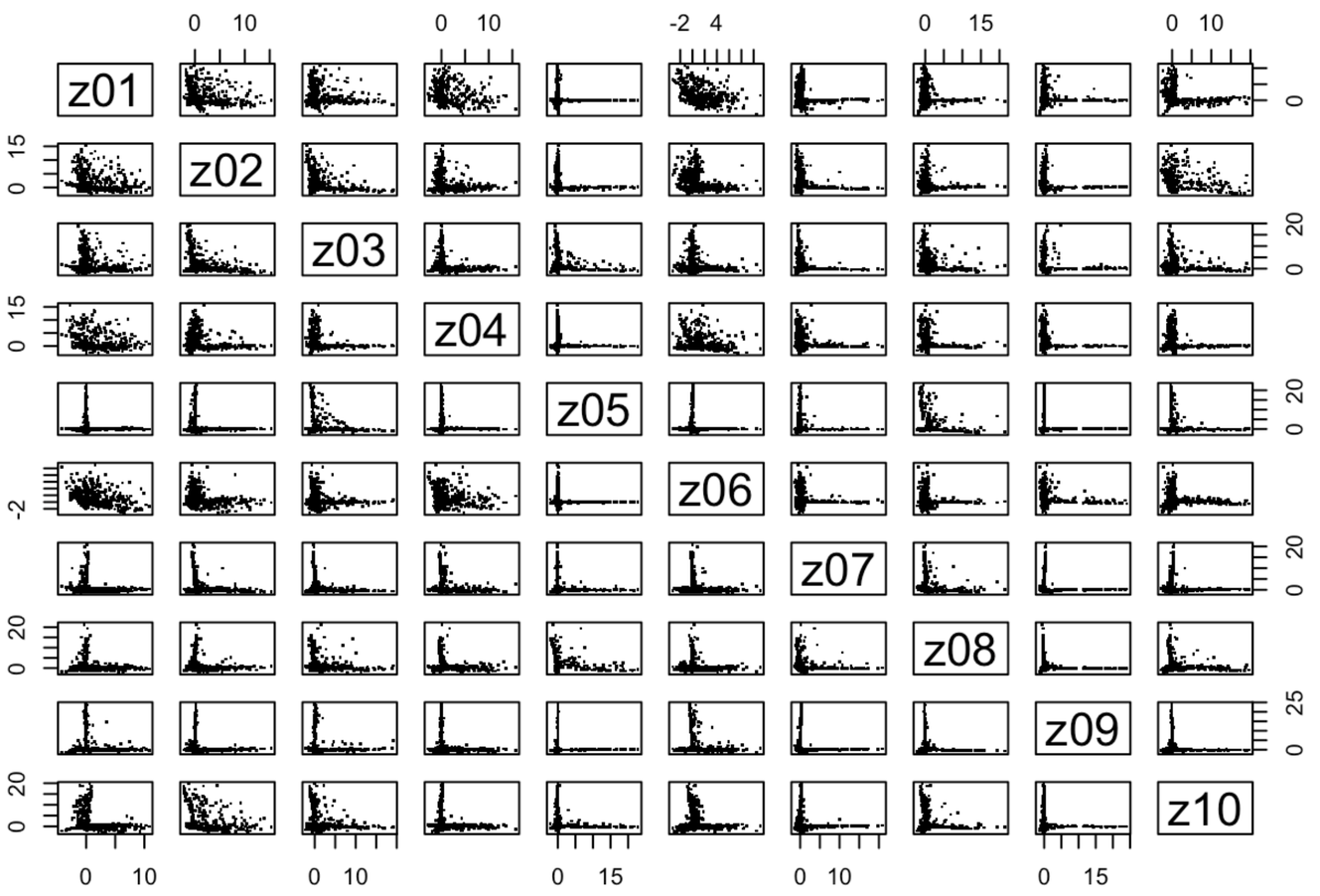
Here are the top 2 eigenvectors of A, after the Varimax rotation. See the code!



Varimax aligns “radial streaks” with the axes.

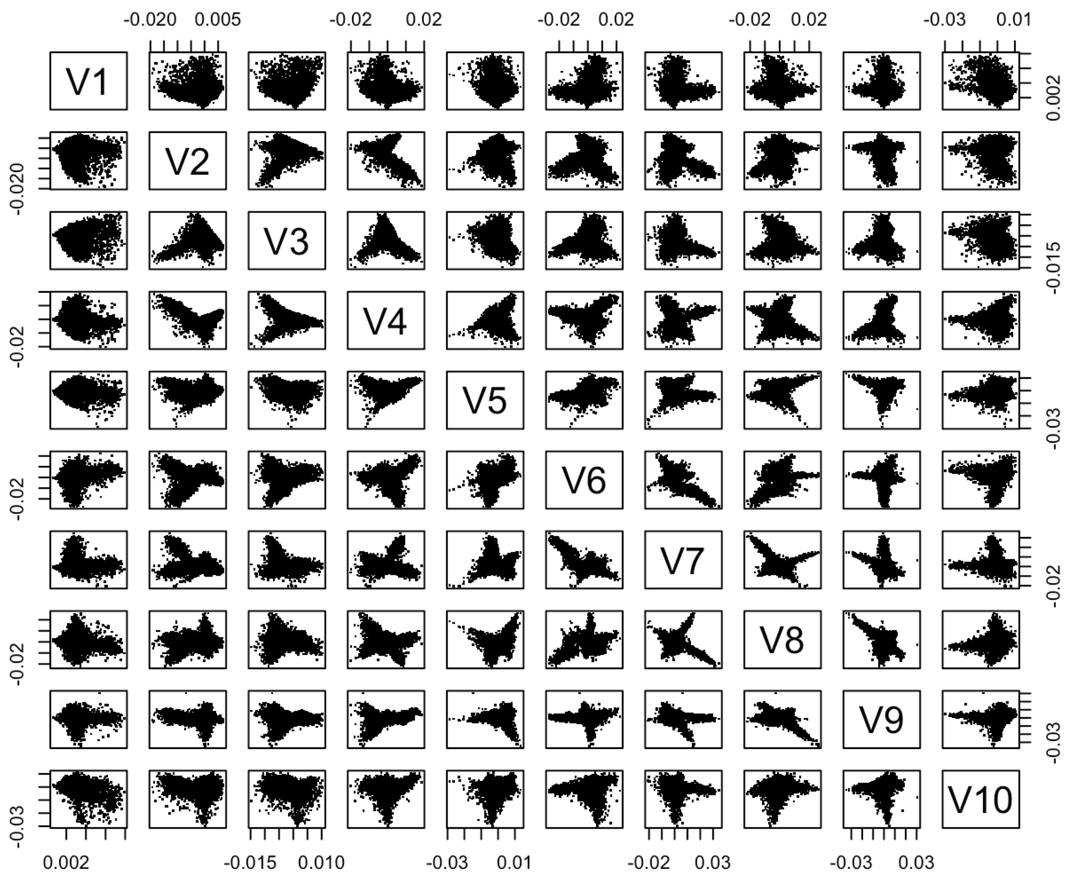
Journal citations

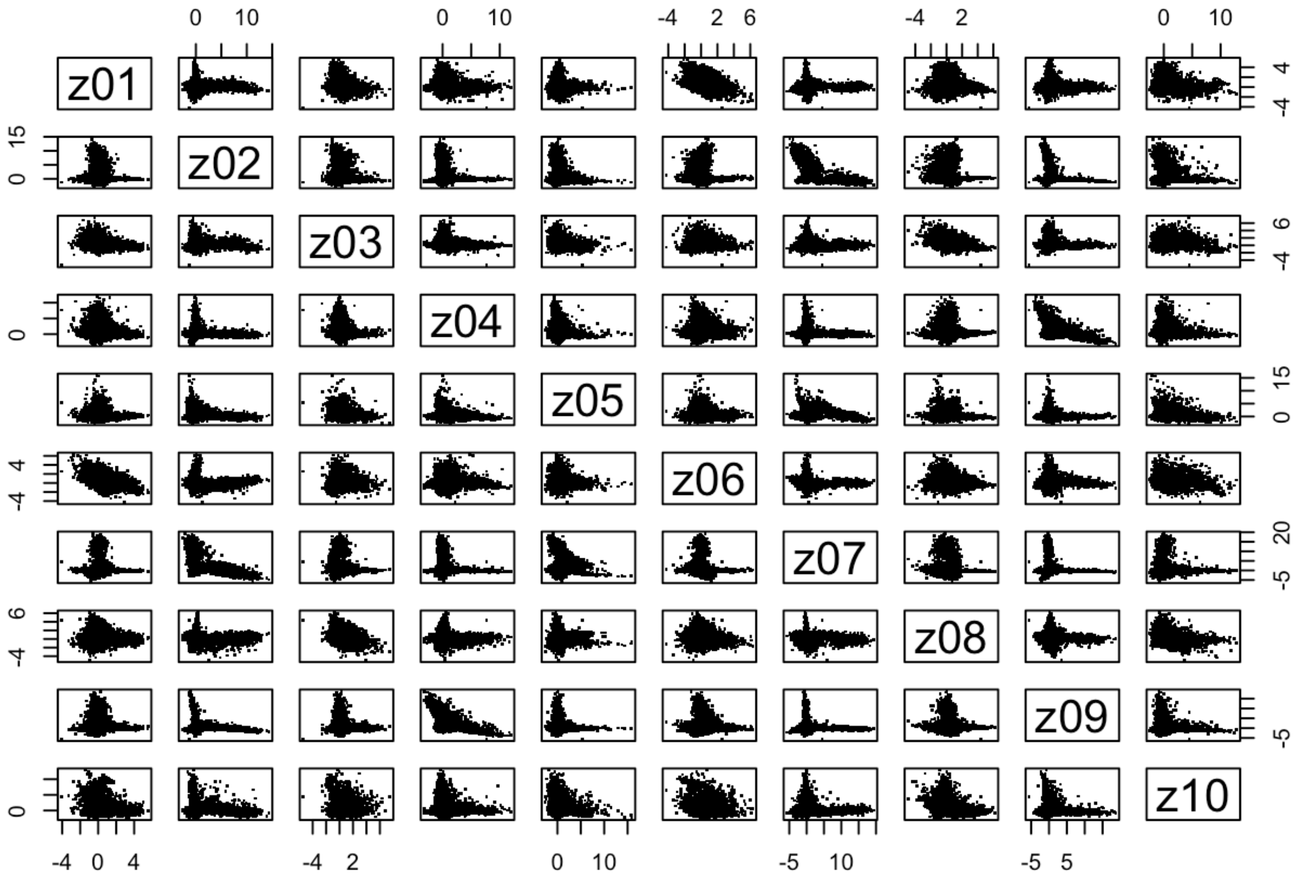




Varimax aligns “radial streaks” with the axes.

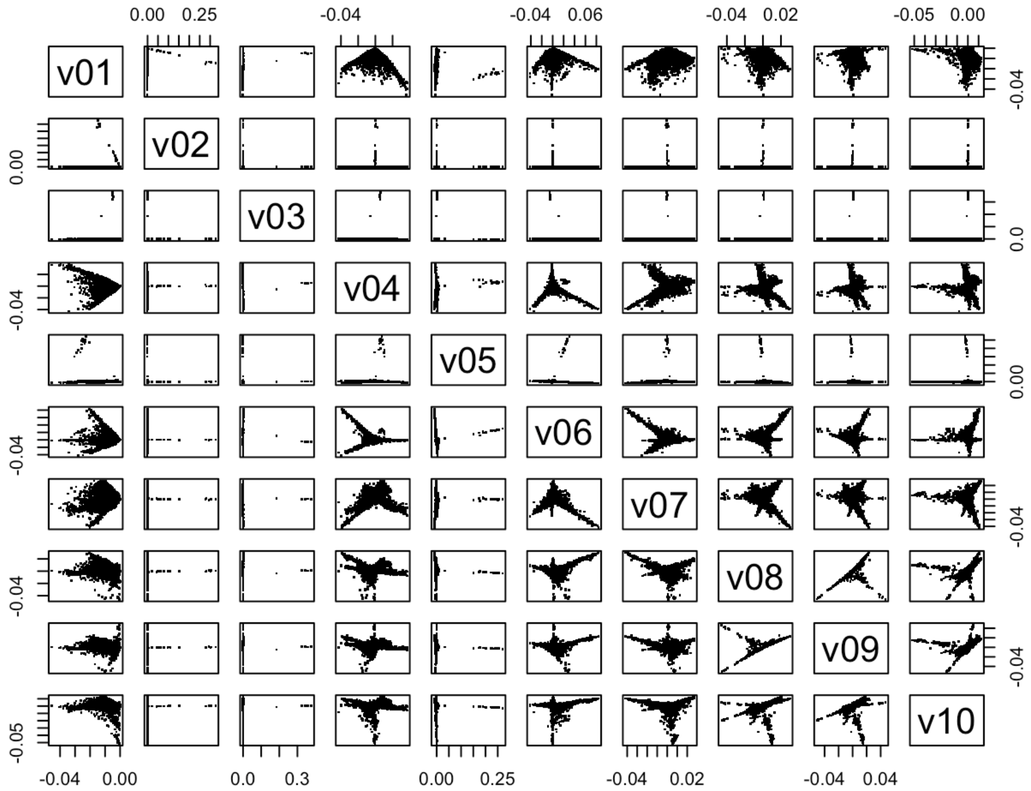
Academic abstracts, represented as document-term graphs

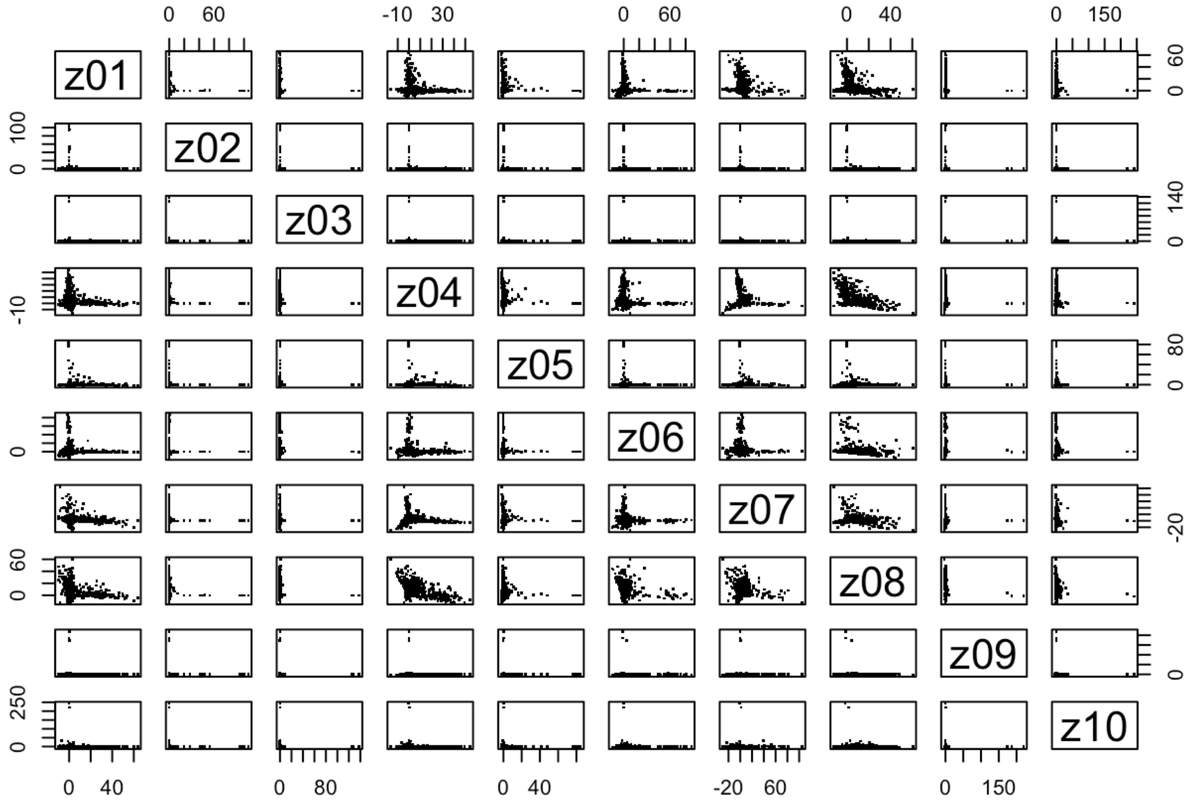




Varimax aligns “radial streaks” with the axes.

A sample of the Twitter following graph





We will give two pieces of intuition for why it works so well.

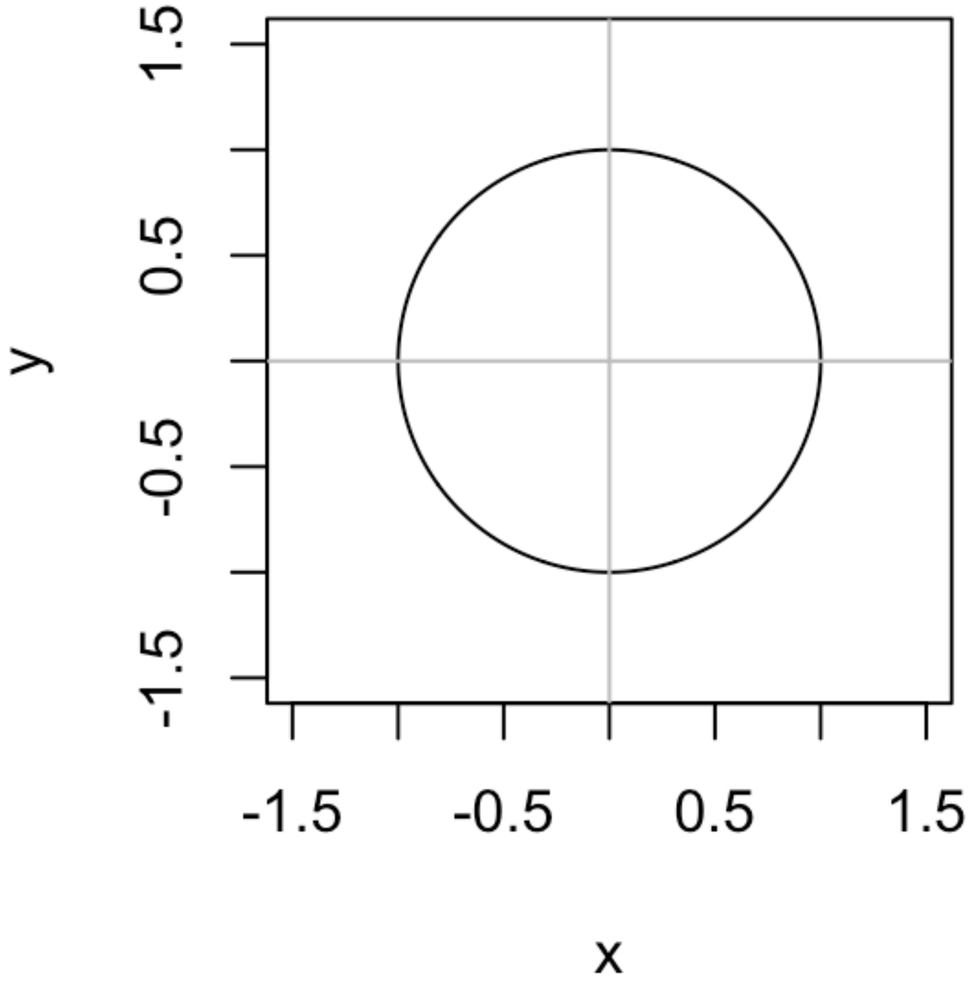
- We will investigate the optimization problem to give intuition for why Varimax finds a *sparse* rotation.
- The first heuristic is algebraic (9th grade algebra).
- The second heuristic is geometric (a pretty picture).
- Finally, we will loop back and give a big theorem to slay the first and second monsters.

Algebraic intuition for Varimax.

$$R^* = \min_{R;R^TR=I} \sum_{ij} (UR)_{ij}^4$$

Put a point anywhere on the circle to maximize the point's 4th moment.

- Suppose a single data point $(x, y) \in \mathbb{R}^2$
- Without loss of generality, presume it is on the unit sphere: $x^2 + y^2 = 1$
- You can put this point anywhere on the unit sphere. Maximize Varimax: $x^4 + y^4$.



Algebraic intuition for Varimax.

$$R^* = \min_{R;R^TR=I} \sum_{ij} (UR)_{ij}^4$$

To find the answer, complete a square

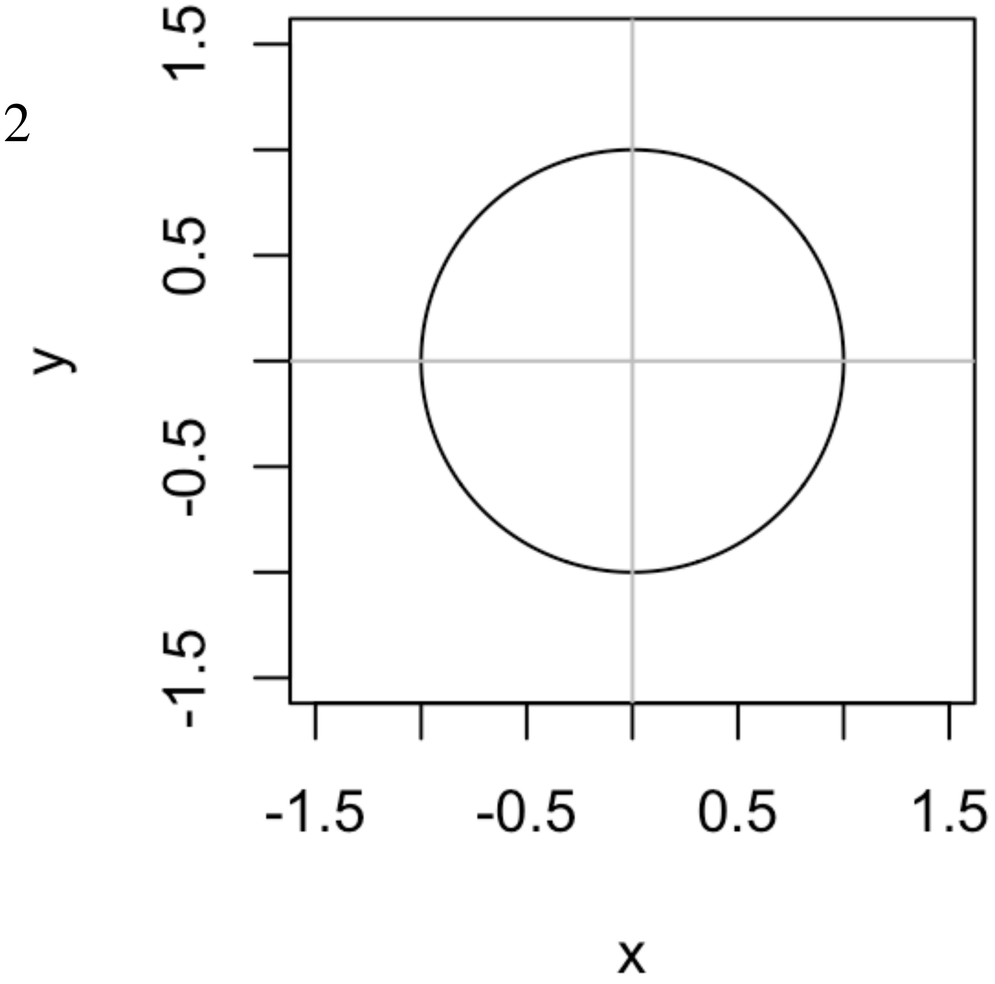
- Complete the square:
  $x^4 + y^4 = (x^2 + y^2)^2 - 2x^2y^2 = 1 - 2x^2y^2$
- Note: $2x^2y^2 \geq 0$
- So, you can maximize the objective by making $x^2y^2 = 0$.
- How to make $x^2y^2 = 0$?

Algebraic intuition for Varimax.

$$R^* = \min_{R;R^TR=I} \sum_{ij} (UR)_{ij}^4$$

To find the answer, complete a square

- Complete the square:
  $x^4 + y^4 = (x^2 + y^2)^2 - 2x^2y^2 = 1 - 2x^2y^2$
- Note: $2x^2y^2 \geq 0$
- So, you can maximize the objective by making $x^2y^2 = 0$.
- How to make $x^2y^2 = 0$?

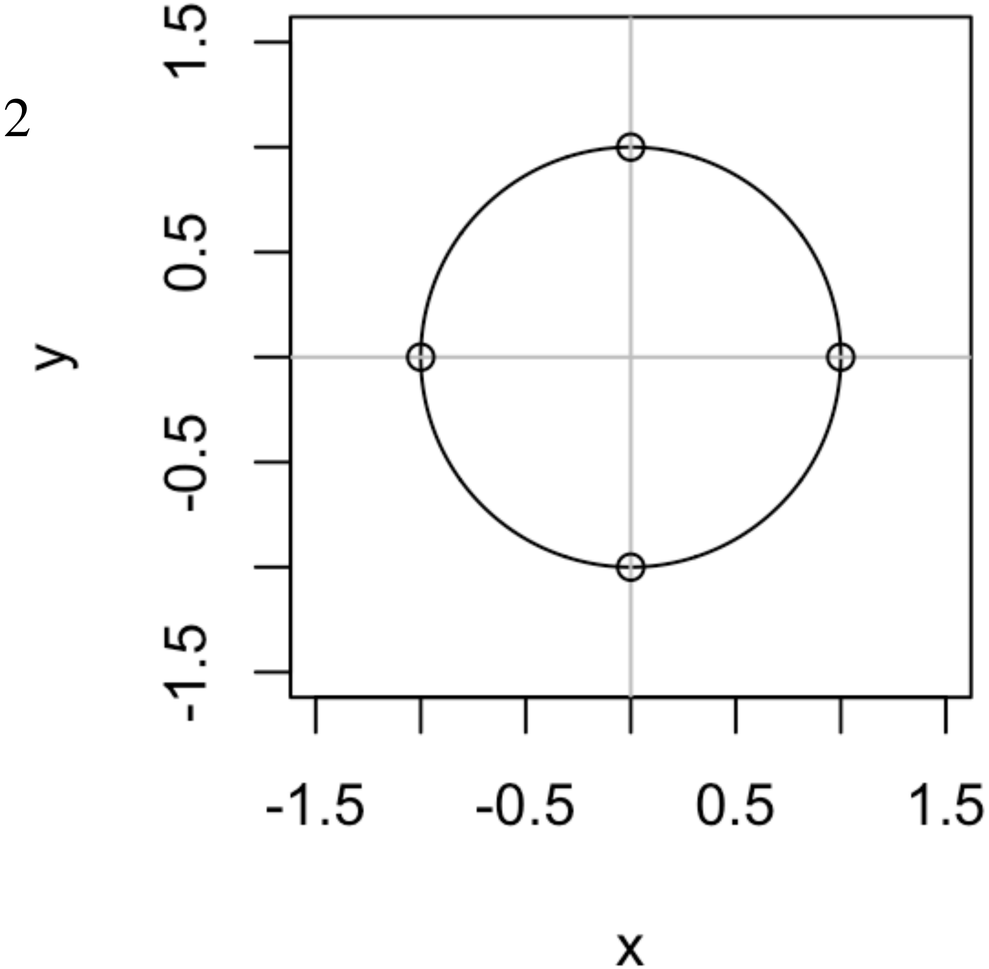


Algebraic intuition for Varimax.

$$R^* = \min_{R;R^TR=I} \sum_{ij} (UR)_{ij}^4$$

The maximum 4th moment points fall on the coordinate axes.

- Maximize
  $x^4 + y^4 = (x^2 + y^2)^2 - 2x^2y^2 = 1 - 2x^2y^2$
- By making (x, y) sparse… one coordinate equal to zero… $x^2y^2 = 0$
- Sparsity!



Algebraic intuition for Varimax.

$$R^* = \min_{R;R^TR=I} \sum_{ij} (UR)_{ij}^4$$

This intuition extends to higher dimensions

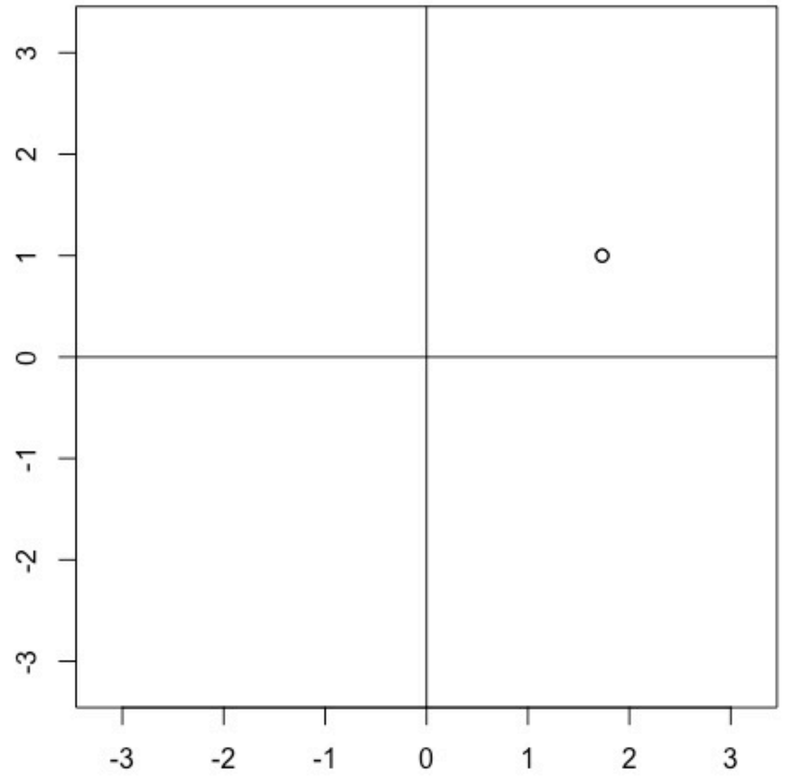
- Suppose $x \in \mathbb{R}^d$, on the unit sphere. Maximize:
- $$\sum_i x_i^4 = (\sum_i x_i^2)^2 - 2\sum_{i \neq j} x_i^2 x_j^2$$
$$= 1 - \text{[non-negative cross terms]}$$
- Maximize this by making cross terms all zero.
- How? Only one $x_i \neq 0$.
- Sparsity! Said another way… align the points and the axes.

Geometric intuition for Varimax

Before: Pick where to put the point (with axes fixed)

Now: Pick where to put the axes (with point fixed)

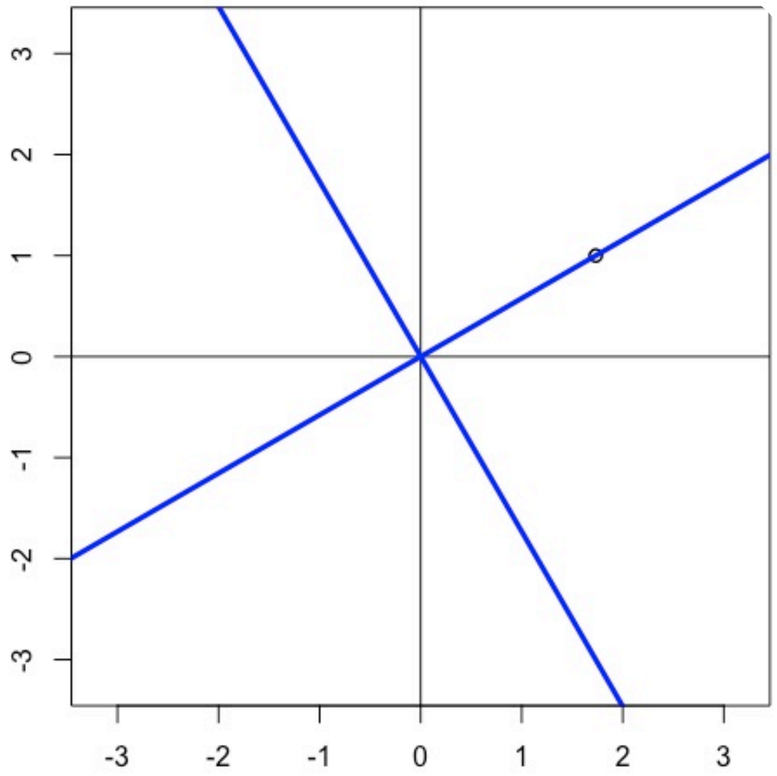
- Where do you draw the axes?



Geometric intuition for Varimax

Varimax aligns the points and axes

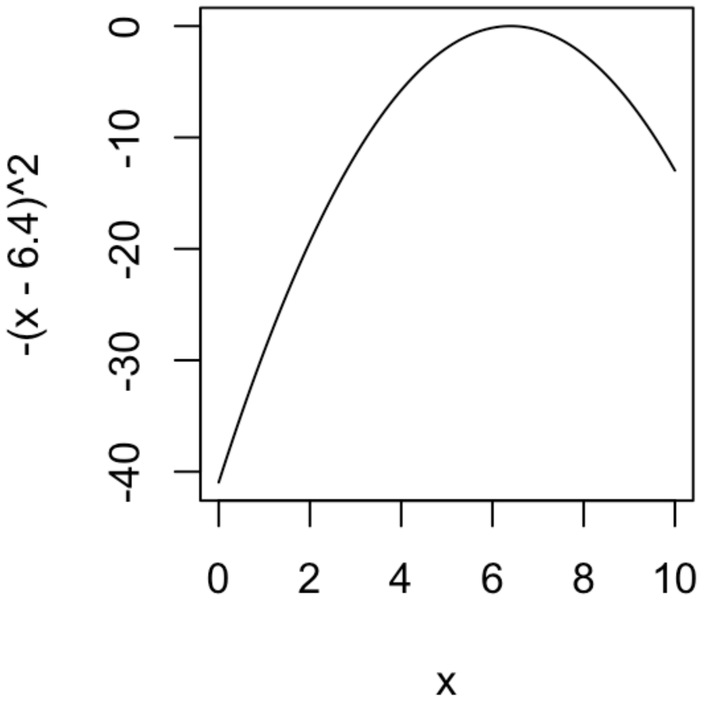
- For the same reason as before, aligning the axes with the point give Varimax solution.
- This isn't saying much yet.
- **But let's draw the objective function for Varimax. Hold on, it gets weird.**



To draw/“graph” the objective function for Varimax in 2d,

We need to understand {the argument} and {the value of the objective function with that argument}.

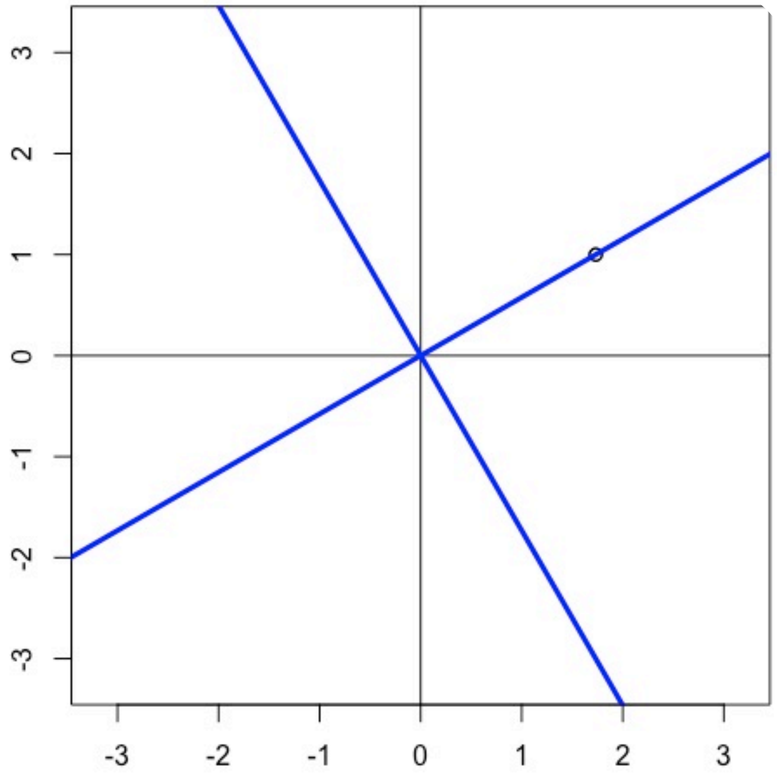
- By analogy, if you wanted to maximize
  $f(x) = -(x - 6.4)^2$
  you need to understand x and $f(x)$.
- This is comfortable. We “just plot it”
- How do we do that for Varimax??



Let's draw/"graph" the objective function for Varimax in 2d

In 2d: the axes / a rotation matrix / the argument to Varimax is parameterized by a single angle θ.

- Draw the first axis at angle θ from the right side of the horizontal axis.
- Then, make the second axis orthogonal to that.
- This is the rotation matrix $R_\theta = \begin{matrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{matrix}$
  Don't worry about this.
  I think rotation matrices make intuition tricky at first…



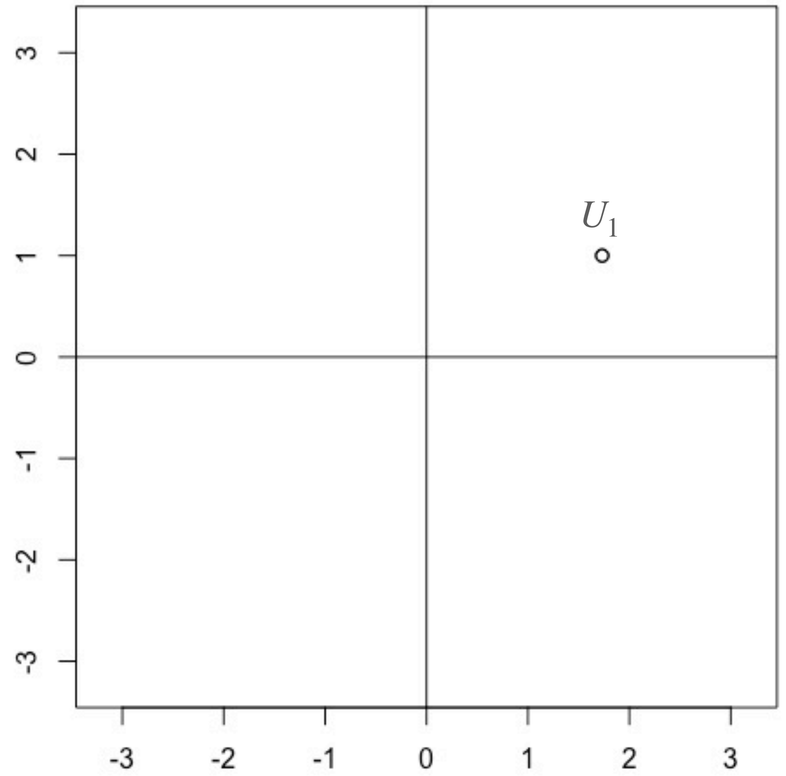
Let's draw/"graph" the objective function for Varimax in 2d

Let's plot the Varimax objective in polar coordinates!

- For each value θ, we can compute the Varimax objective for that rotation, with this single point.

$$f(\theta, U_1) = \sum_{j=1}^{d} (U_1 R_\theta)_j^4$$

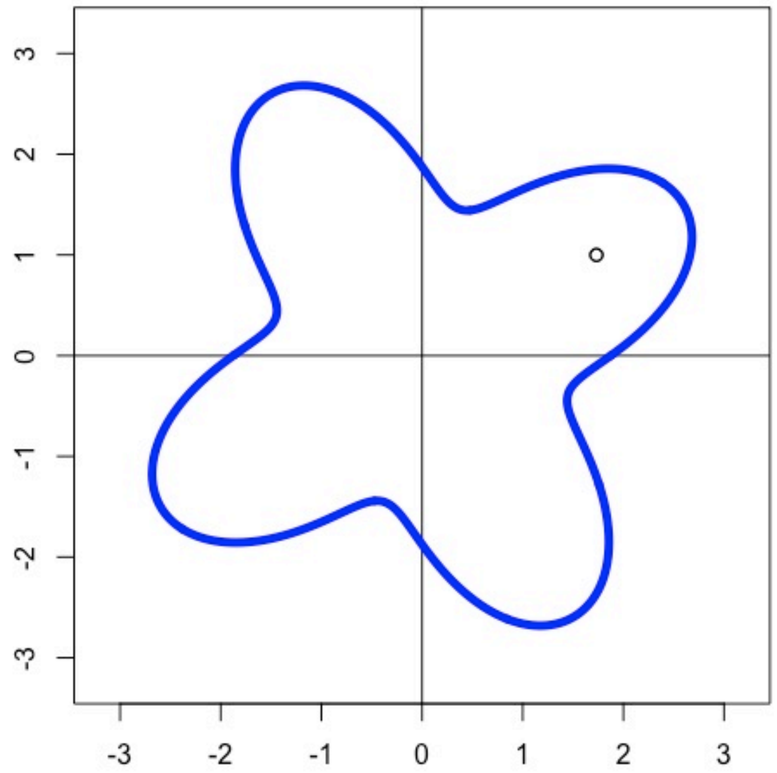
- Plot the objective $f(\theta, U_1)$ as the radius at θ.



Geometric intuition for Varimax

The Varimax objective in polar coordinates!

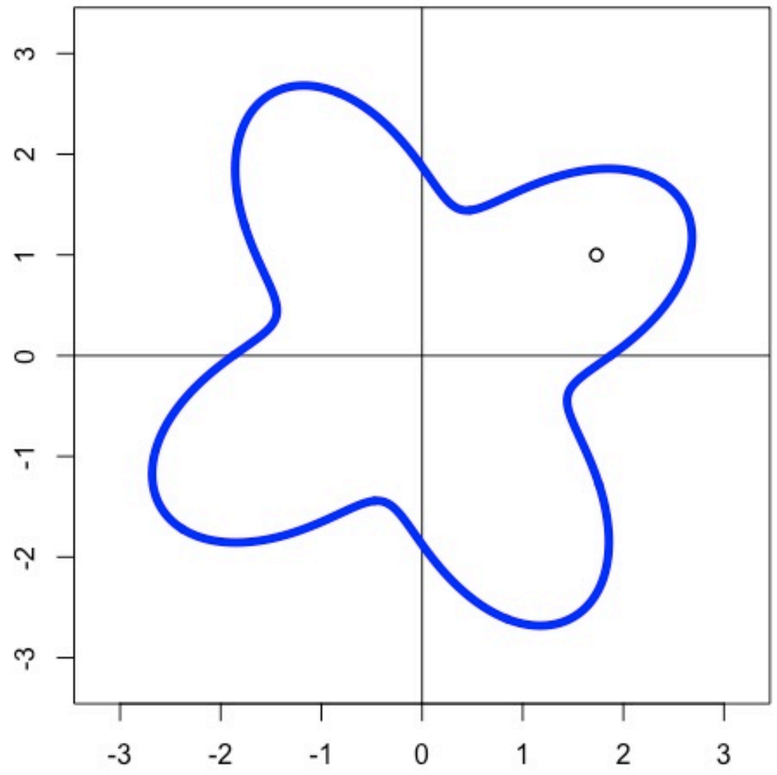
- Maximize the radius -> maximize Varimax.
- We call this a four petal flower,
  aligned with the data point.



Geometric intuition for Varimax

What about more than one point??

- This toy example has just one point.
- With more data points, align a 4-petal flower with each point.
- Sum up the flowers (to make the radius the sum of the radii)
- Then, maximize the sum of the flowers.



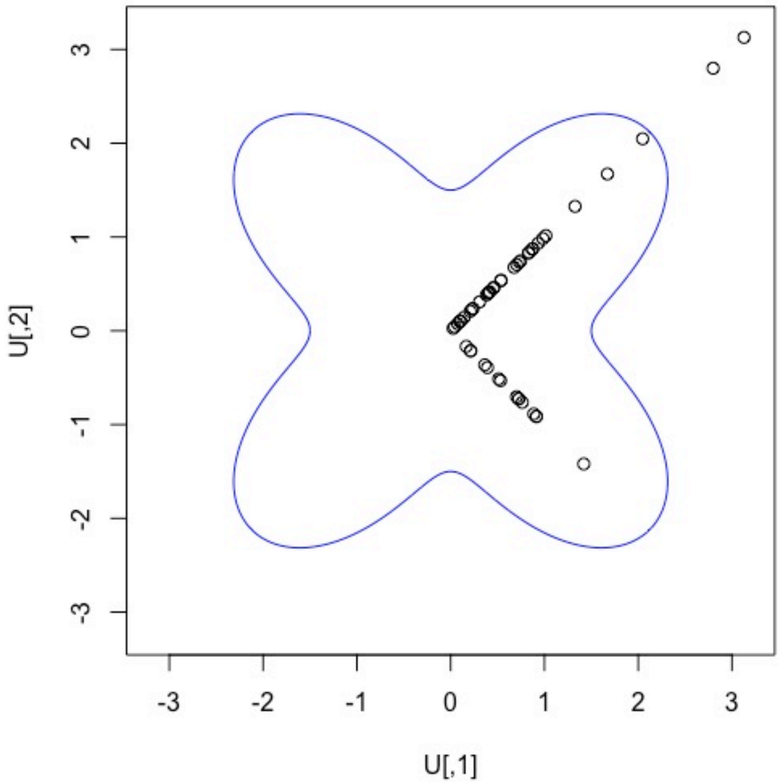
$$f(\theta) = \sum_{ij} (UR_\theta)_{ij}^4 = \sum_i \sum_j (U_i R_\theta)_j^4 = \sum_i f(\theta, U_i)$$

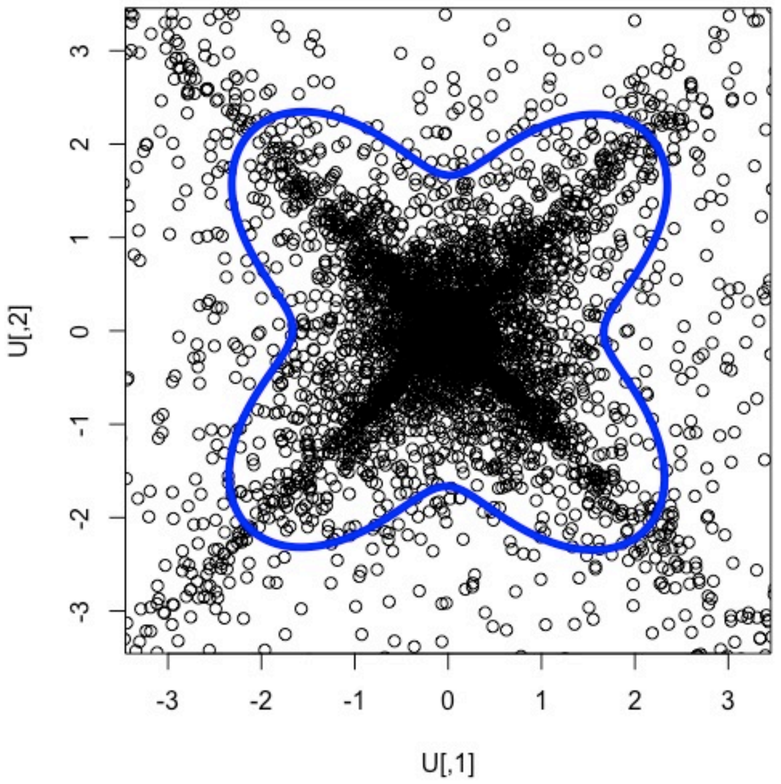
Full Varimax objective

Sum over each flower

Each point makes a flower. Sum up the flowers and find the max.

Notice how the “Varimax axes” align with the “radial streaks” in the data





$$f(\theta) = \sum_{ij} (UR_\theta)_{ij}^4 = \sum_i \sum_j (U_i R_\theta)_j^4 = \sum_i f(\theta, U_i)$$

Full Varimax objective

Sum over each flower

Intuition for the second monster:

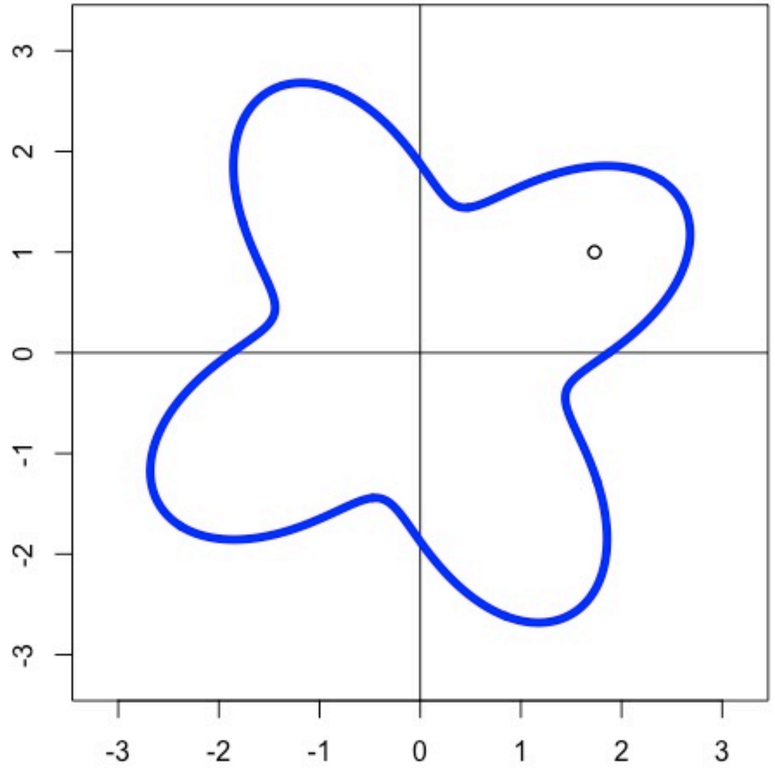
How do we recover Z from $U = Z\Phi$??

$$R^* = \min_{R;R^TR=I} \sum_{ij} (UR)_{ij}^4$$

- Notice that Z is sparse. We want to find axes in U that makes the points sparse. Why does Varimax do it?

- **Algebraic:**
  Among points $(x, y) \in \mathbb{R}^2$ on the unit sphere,
  $x^4 + y^4$ is maximized when $x = 0$ or $y = 0$.
  $(x^4 + y^4 = (x^2 + y^2)^2 - 2x^2y^2 = 1 - 2x^2y^2)$

- **Geometric:**
  Inside Varimax, each point makes a flower. The pedals are aligned with that point. We sum up the flowers and find the biggest radius.



It's time for the theorem to slay the first two monsters.

Algorithm 1

- Input A, k:
- Compute $\hat{U} \in \mathbb{R}^{n \times k}$, leading k eigenvectors of A.
- Compute $\hat{R} \in \mathbb{R}^{k \times k}$, the Varimax rotation from $\hat{U}$
- Return $\hat{Z} = \hat{U}\hat{R}$

Theorem 1
(See Rohe, Zeng 2020 for rigorous statement)

Suppose that A contains independent, "sub-exponential" elements with $\mathbb{E}A = ZBZ^T$, each row of Z has one non-zero, non-zero elements in Z come from bounded distribution, elements of $B = \rho_n B_0$ can degrade asymptotically for sparse graphs, maximum expected degree Δ_n must grow faster than $\log^{11} n$, B_0 full rank, then *all rows* of $\hat{Z}$ converge:

$$\max_i \|\hat{Z}_i - Z_i\|_2 = O(\Delta_n^{-.24})$$

(mod permutation of columns and changing their sign)

Note: No parametric assumptions on elements of A or Z…

Can we remove the “hard clustering” assumption?

- Theorem 1 assumes “Only one non-zero element in each row of Z.”
  - This DC-SBM model is *hard clustering*; each column corresponds to a cluster.
- Suppose we want to relax this assumption to allow for “multiple memberships” or perhaps something more general. Is it still identifiable?
- Rotational invariance… how do we know we aren’t estimating ZR for some arbitrary orthogonal rotation R?

$$ZBZ^T = (ZR)(R^T BR)(ZR)^T$$

- Estimating a general Z is impossible… but

What about high dimensional regression?

- Suppose $Y = X\beta + \epsilon$ where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ with $p>>n$.
- While $\mathbb{E}Y = X\beta$, there is a huge space of $b \in \mathbb{R}^p$ with $b \neq \beta$ and $\mathbb{E}Y = Xb$
- So, is β identifiable??? Can we estimate it???

What about high dimensional regression?

- Suppose $Y = X\beta + \epsilon$ where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ with $p>>n$.
- While $\mathbb{E}Y = X\beta$, there is a huge space of $b \in \mathbb{R}^p$ with $b \neq \beta$ and $\mathbb{E}Y = Xb$
- So, is β identifiable??? Can we estimate it???
- A better question: What assumptions could we make?
- Sparsity resolves this invariance.
- The same happens with rotations…

Hard clustering assumption can be replaced with an *"old fashioned" sparsity assumption*

Algorithm 2

- Input A, k:
- $\tilde{A}_{ij} =$
  A_{ij} - rowmean(i) -colmean(j) + grandmean
- Compute $\hat{U} \in \mathbb{R}^{n \times k}$, leading k eigenvectors of $\tilde{A}$.
- Compute $\hat{R} \in \mathbb{R}^{k \times k}$, the Varimax rotation from $\hat{U}$
- Return $\hat{Z} = \hat{U}\hat{R} + \hat{\mu}_Z$
  (Details of $\hat{\mu}_Z$ in paper)

Definition

A mean zero, unit variance random variable X is *leptokurtic* if $\mathbb{E}X^4 > 3$.

Theorem (R, Zeng 2020)
X is leptokurtic if $5/6 < P(X = 0) < 1$

Theorem 2
(See R, Zeng 2020 for rigorous statement)

Same result as Theorem 1, except assumptions on Z are replaced. Instead, elements are
1) independent, unit variance
2) "sub-exponential"
3) *leptokurtic*

Most data is rectangular, not square & symmetric.

Use SVD instead of eigen. We call it *Vintage sparse PCA.*

- Input A, k:
- $\tilde{A}_{ij} =$
  A_{ij} - rowmean(i) -colmean(j) + grandmean
- Compute $\hat{U} \in \mathbb{R}^{n \times k}$, leading k
  left singular vectors of $\tilde{A}$.
- Compute $\hat{R} \in \mathbb{R}^{k \times k}$, the Varimax rotation from $\hat{U}$
- Return $\hat{Z} = \hat{U}\hat{R} + \hat{\mu}_Z$
  (Details of $\hat{\mu}_Z$ in paper)

Theorem 3
(See R, Zeng 2020 for rigorous statement)
Same result as Theorem 2, except $\mathbb{E}A = ZBY^T \in \mathbb{R}^{n \times d}$ and Y isn't too crazy (e.g. independent, sub-exponential)

Corollary
With slightly different normalization, this covers the Latent Dirichlet Allocation model in text analysis, both the document topic vectors $Z_i \in \mathbb{R}^k$ on the simplex and the word distributions by topic can be estimated.

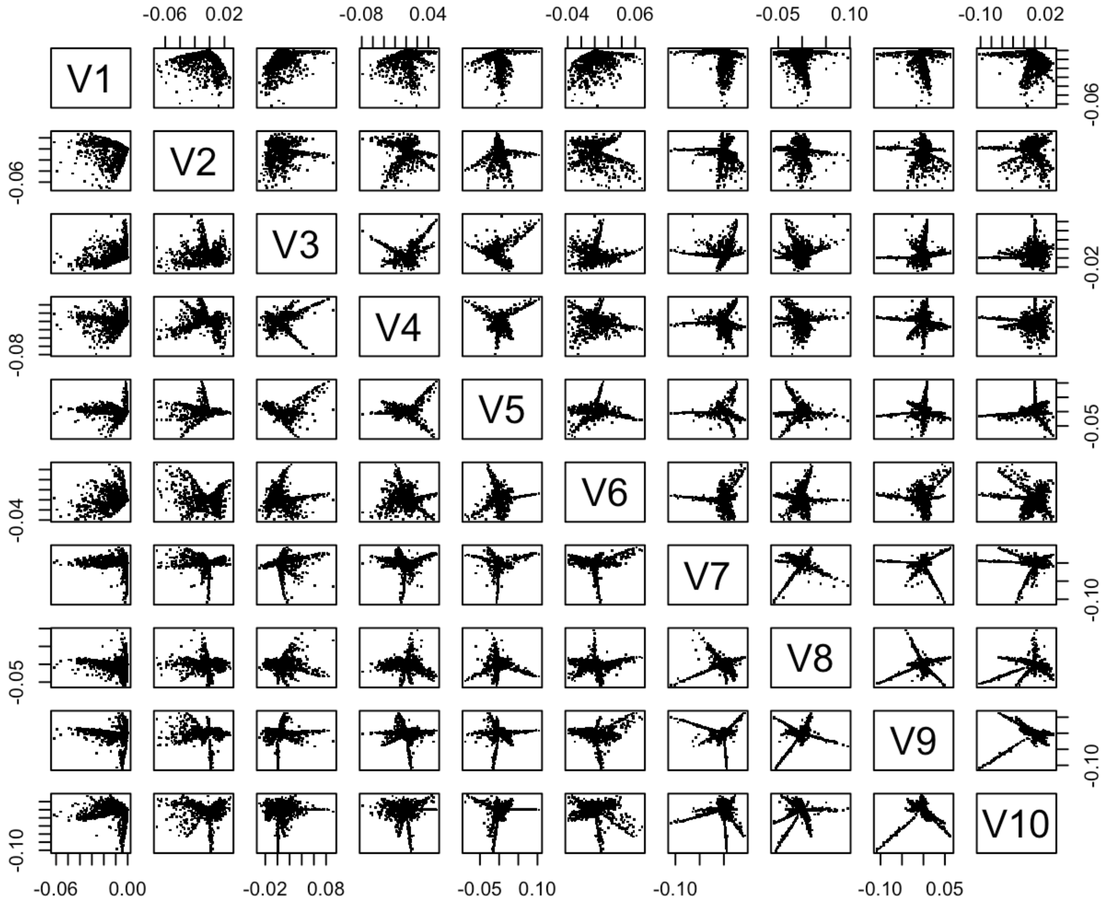
Monster 1: Why should eigen/svd give us an embedding?
Monster 2: How can we extract Z from eigenvectors?

- This theorem covers $\mathbb{E}A = ZBY^T$, a broad class of models.
  LDA, Stochastic Blockmodels, mixture of independent Gaussians,
  something weird you haven't thought of, etc.
  - Only distributional assumption on A: sub-exponential.
    Poisson, Gaussian, Bernoulli, Exponential, something weird, a mix… all ok.
  - Only distributional assumption on Z: leptokurtic.
    Dirichlet, Exponential, Bernoulli, spike&slab, something weird, a mix… all ok.
- We call this result "semi-parametric" because of the linearity in the expression ZBY^T.
- "Radial streaks" and sparsity are a key diagnostic to ensure identifiability.

No time to rest. Two more monsters.

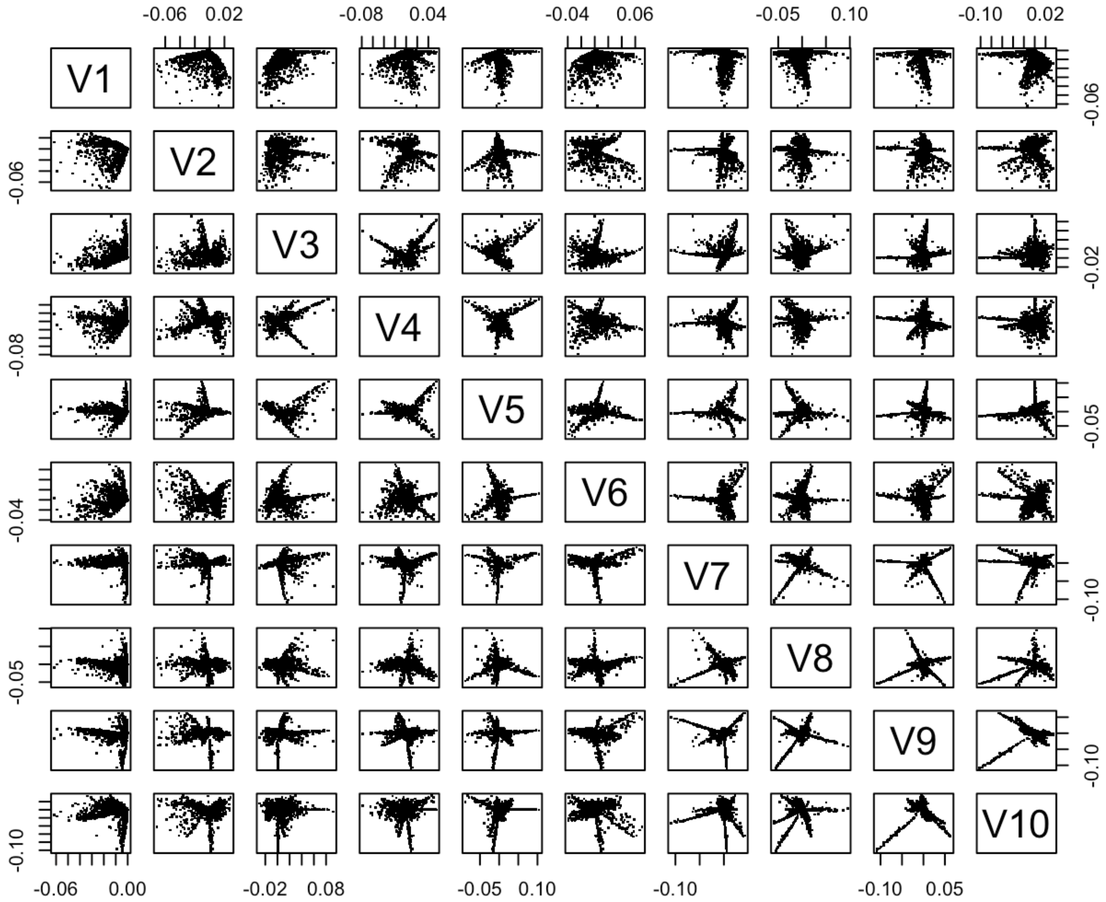
What I showed you before was a fairy tale.

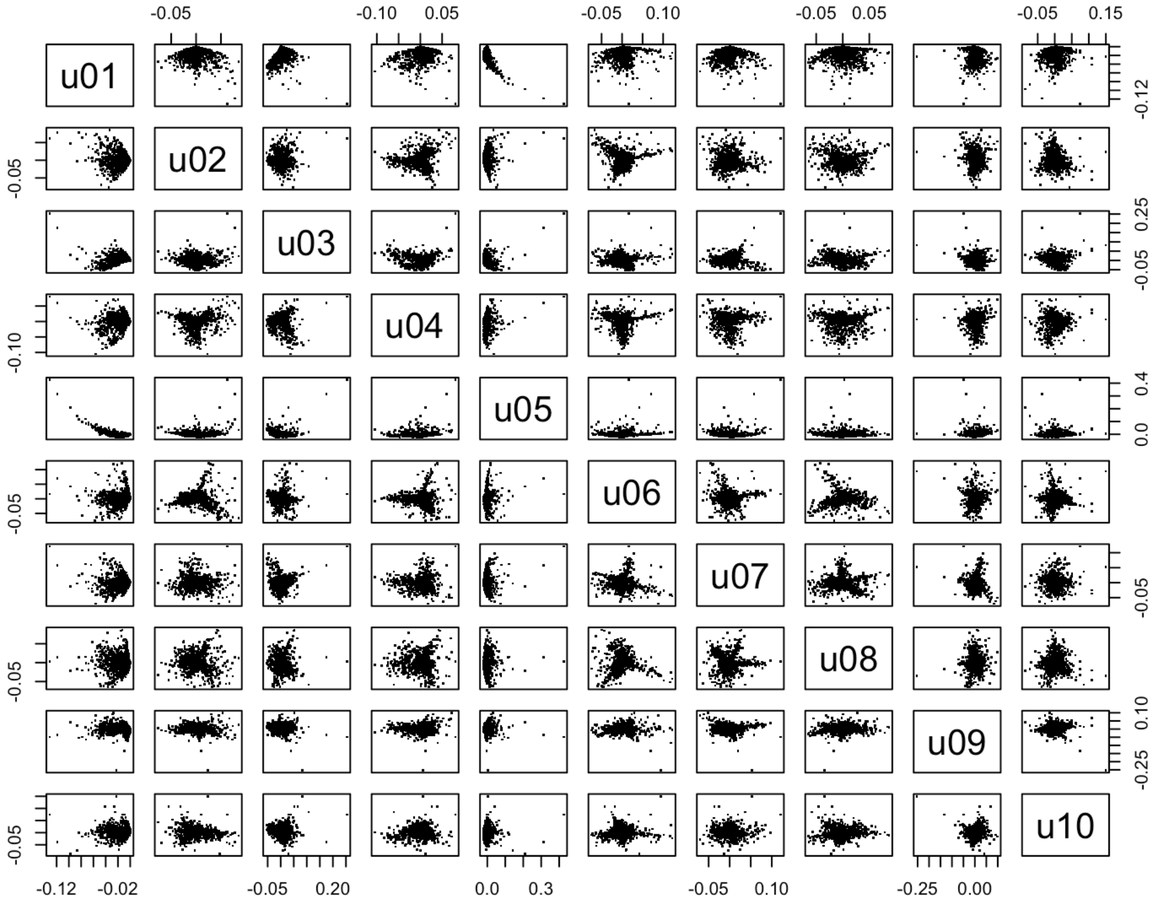
Journal citations



What I showed you before was a fairy tale.

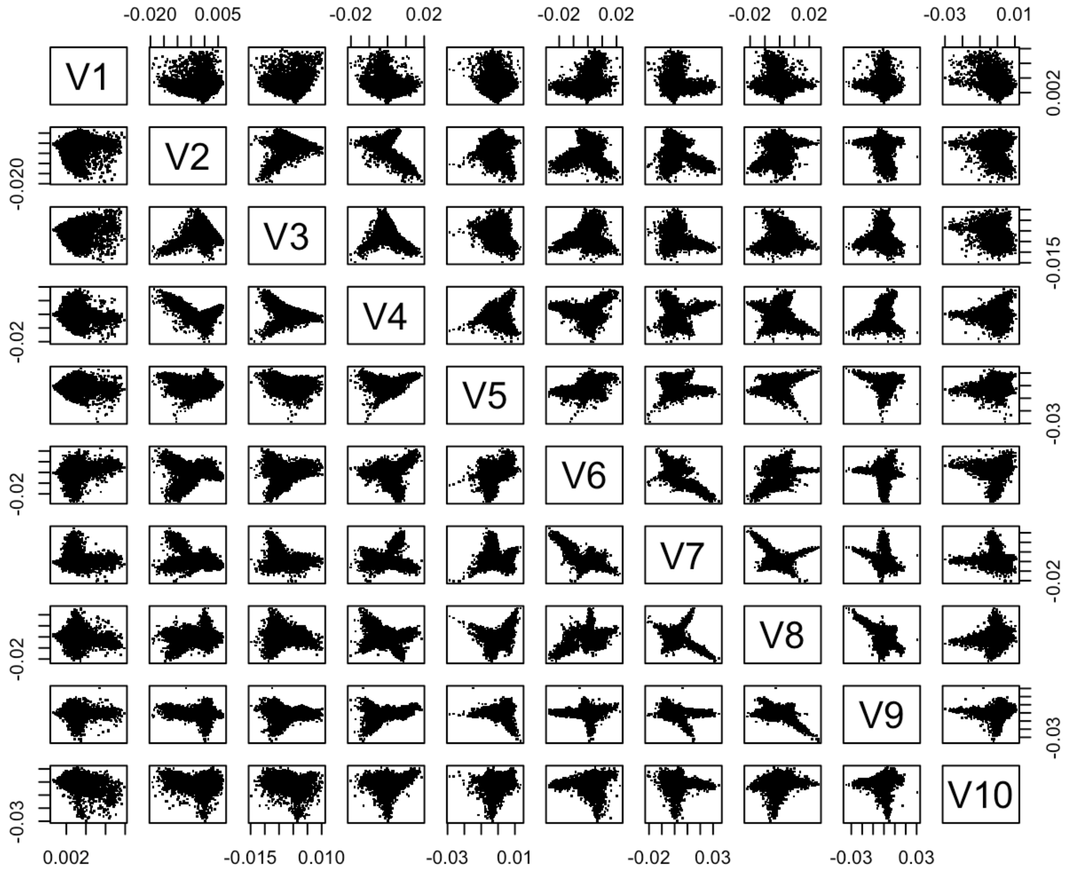
Journal citations





What I showed you before was a fairy tale.

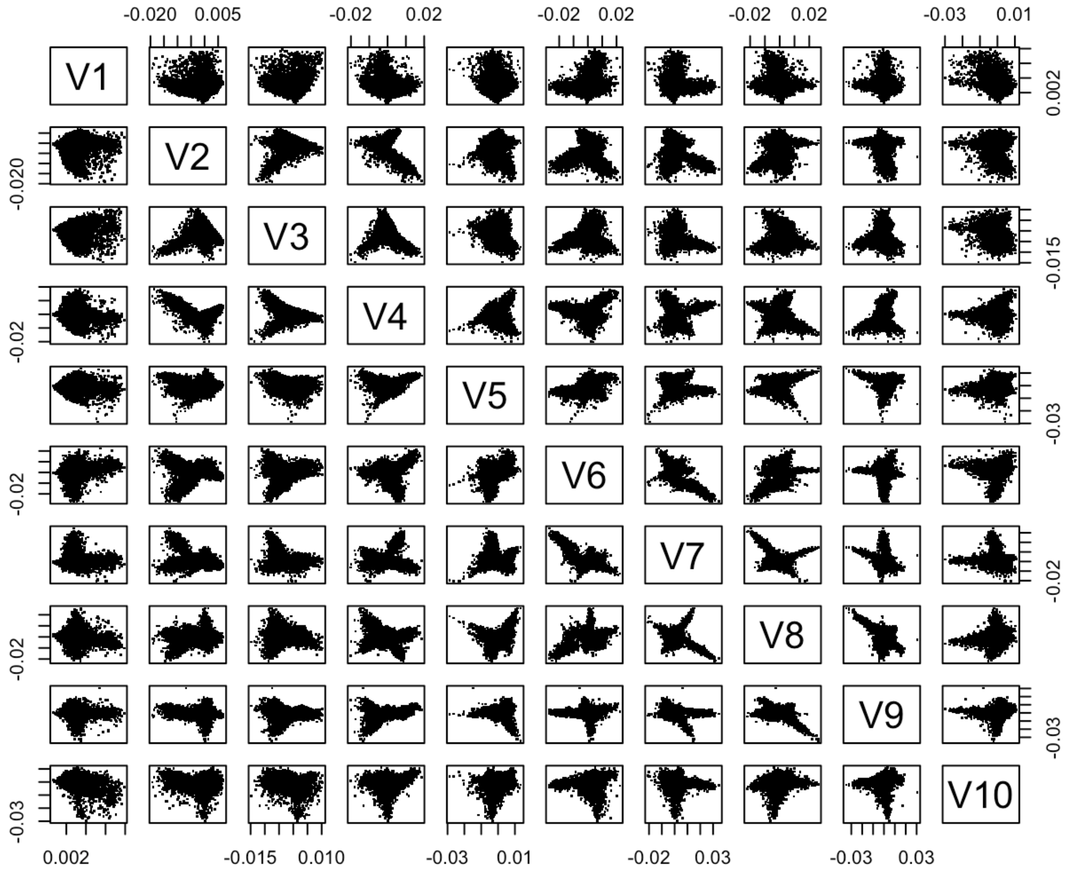
Academic abstracts, represented as document-term graphs

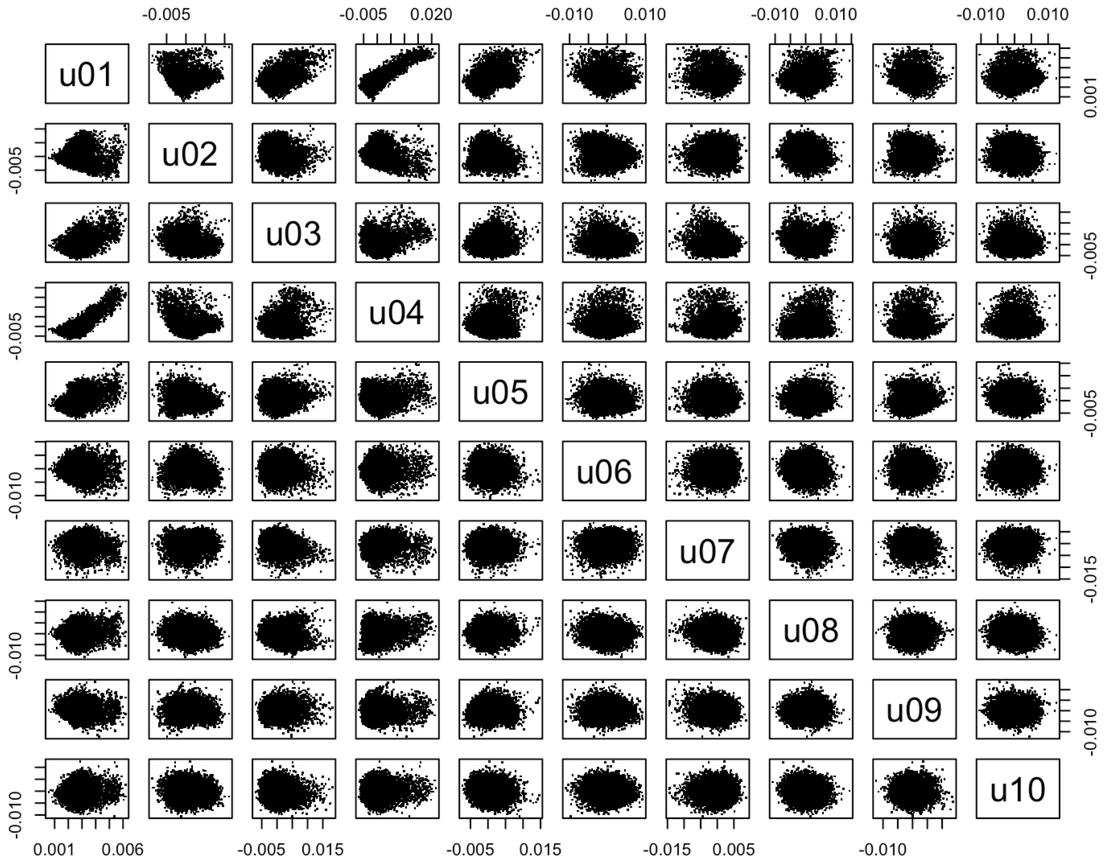


Get ready for this one.

What I showed you before was a fairy tale.

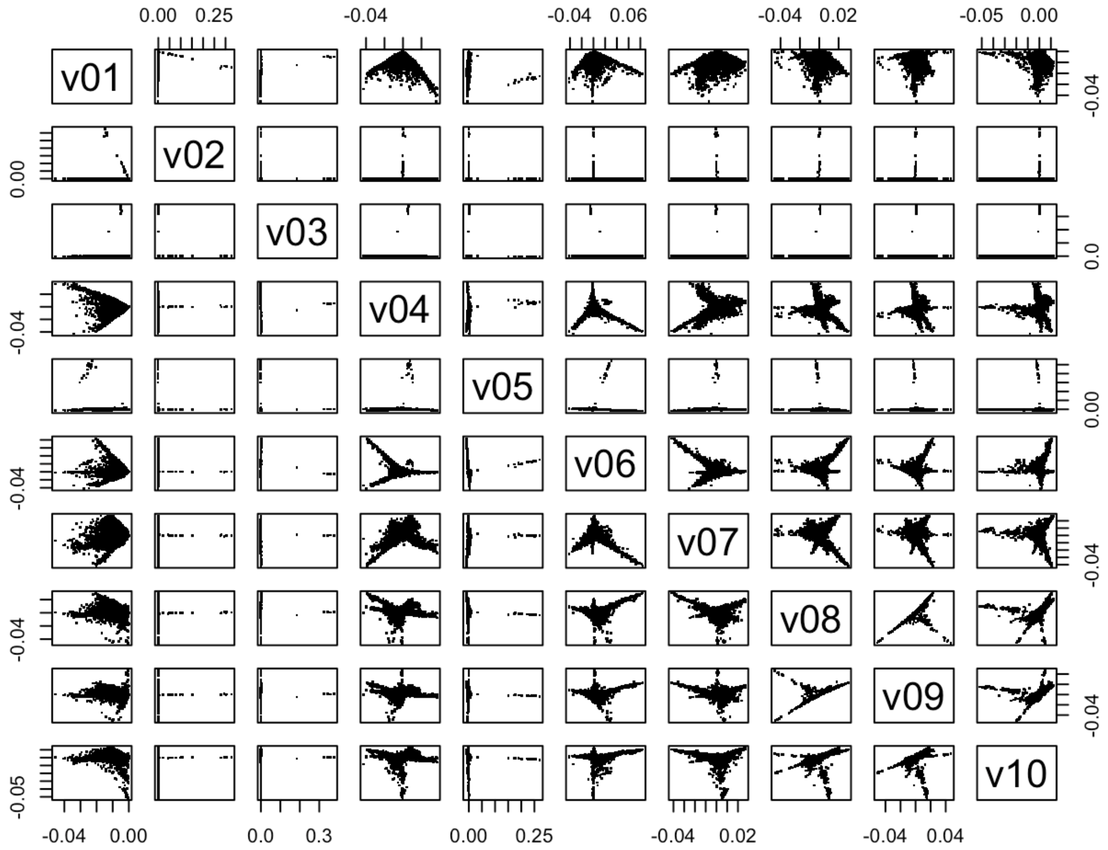
Academic abstracts, represented as document-term graphs





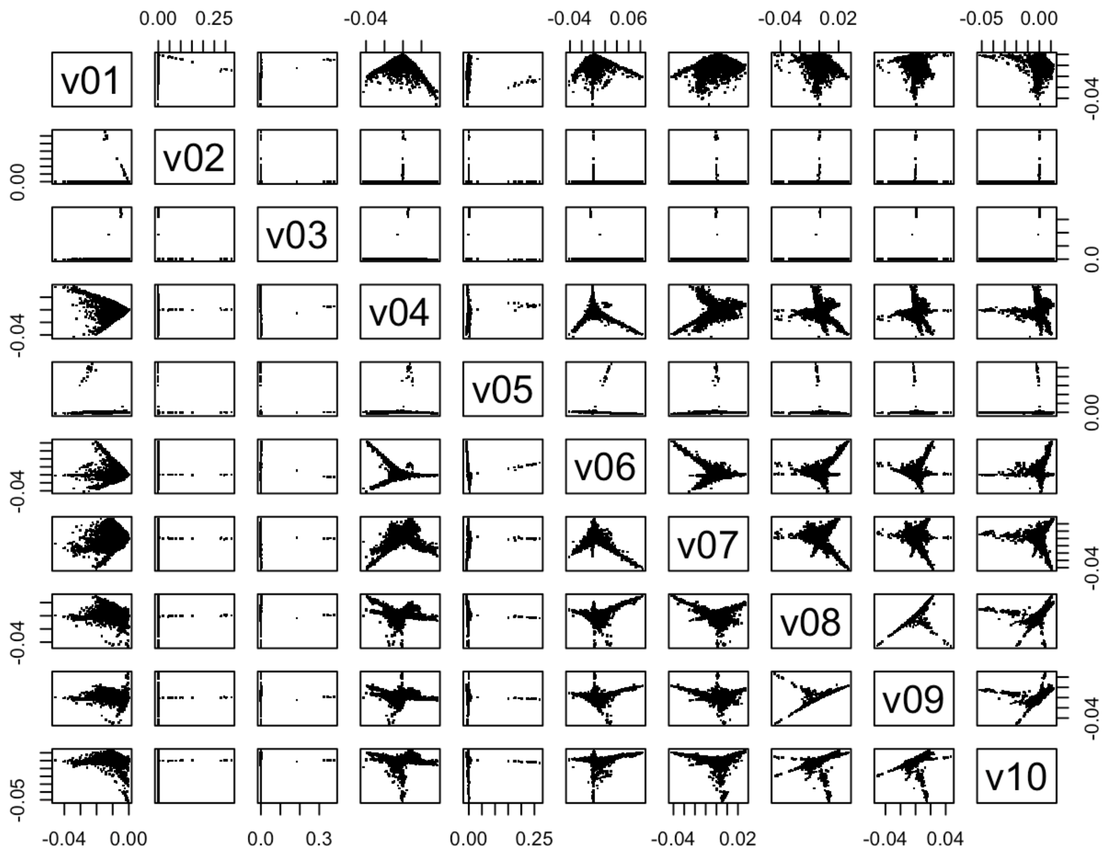
What I showed you before was a fairy tale.

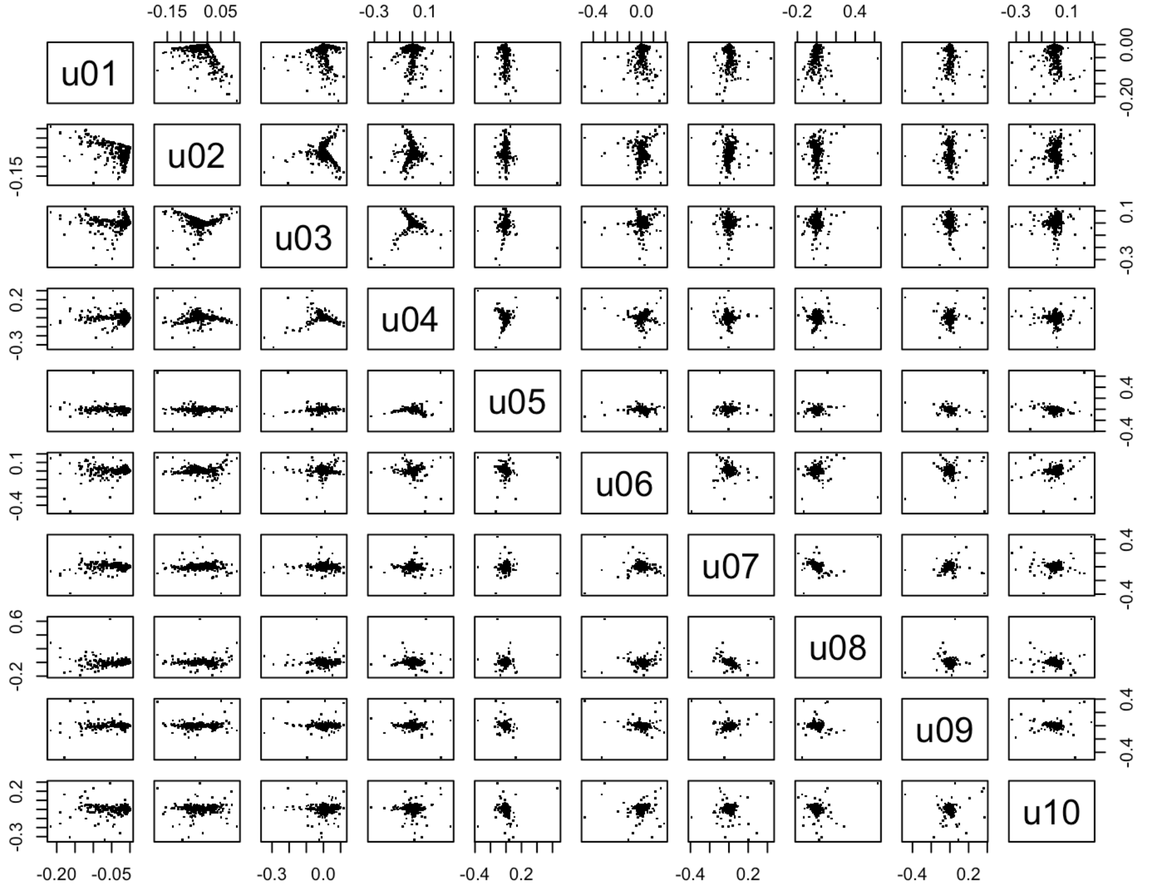
A sample of the Twitter following graph



What I showed you before was a fairy tale.

A sample of the Twitter following graph





The third monster:

The eigenvectors of A are not optimal… (and actually pretty bad)

- There exists a sparse two block Stochastic Blockmodels for which
  (1) the leading eigenvectors of A do not reveal Z
  (2) alternative algorithms do (e.g. belief propagation)

- This is not just a technical issue. There are fundamental problem with A.

There is a simple trick that is magic.

And a big theorem to show it provides for “optimal estimation”

- We construct a new matrix…
- “normalize” and “regularize” A

This new matrix has better concentration properties

The proof uses the Grothendieck-Pietsch factorization.

CONCENTRATION AND REGULARIZATION OF RANDOM GRAPHS

CAN M. LE, ELIZAVETA LEVINA AND ROMAN VERSHYNIN

ABSTRACT. This paper studies how close random graphs are typically to their expectations. We interpret this question through the concentration of the adjacency and Laplacian matrices in the spectral norm. We study inhomogeneous Erdös-Rényi random graphs on n vertices, where edges form independently and possibly with different probabilities p_{ij}. Sparse random graphs whose expected degrees are $o(\log n)$ fail to concentrate; the obstruction is caused by vertices with abnormally high and low degrees. We show that concentration can be restored if we regularize the degrees of such vertices, and one can do this in various ways. As an example, let us reweight or remove enough edges to make all degrees bounded above by $O(d)$ where $d = \max np_{ij}$. Then we show that the resulting adjacency matrix A' concentrates with the optimal rate: $\|A' - \mathbb{E}\, A\| = O(\sqrt{d})$. Similarly, if we make all degrees bounded below by d by adding weight d/n to all edges, then the resulting Laplacian concentrates with the optimal rate: $\|\mathcal{L}(A') - \mathcal{L}(\mathbb{E}\, A')\| = O(1/\sqrt{d})$. Our approach is based on Grothendieck-Pietsch factorization, using which we construct a new decomposition of random graphs. We illustrate the concentration results with an application to the community detection problem in the analysis of networks.

This thing is magic. So, cast the spell!

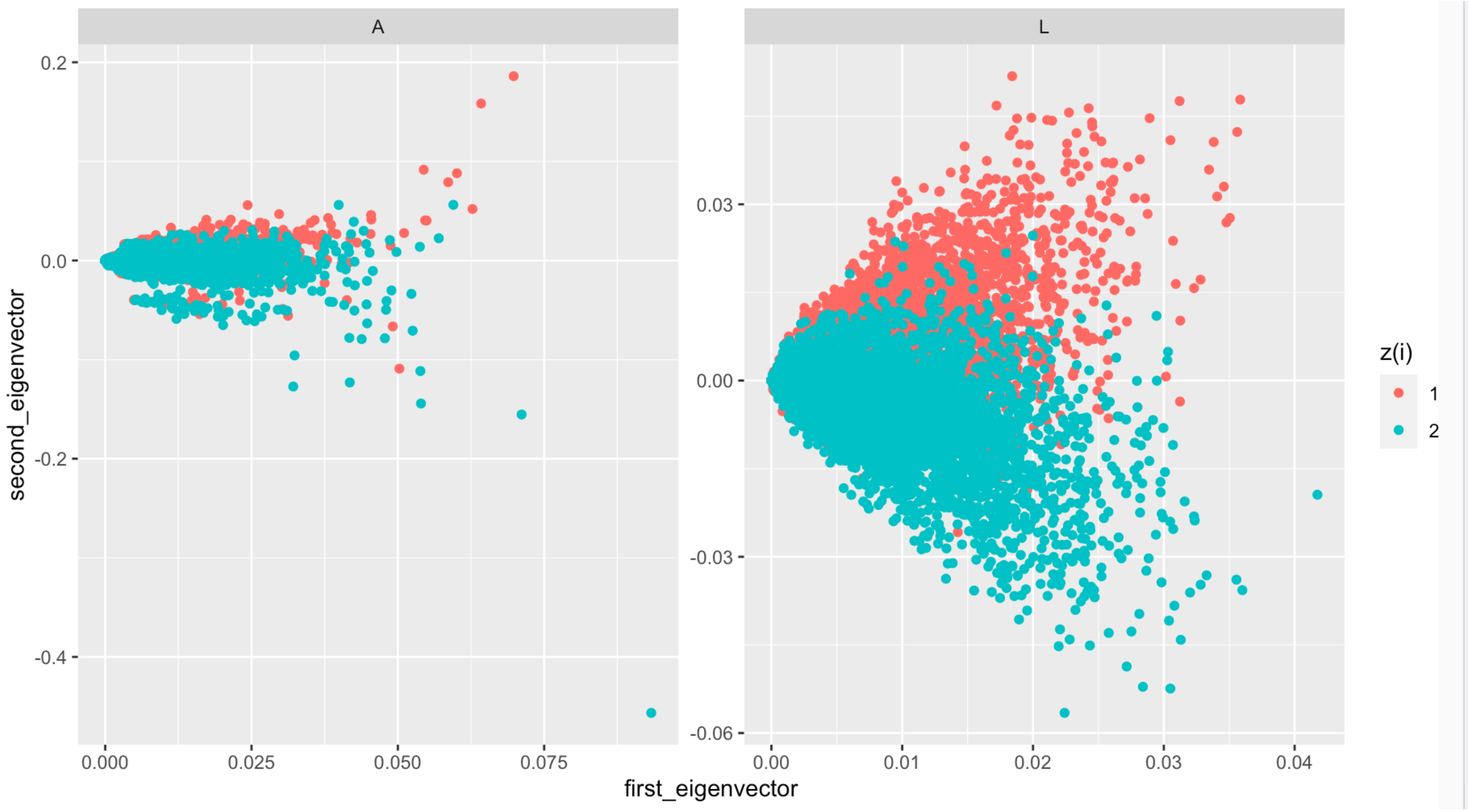
Normalize and regularize the adjacency matrix:

- Set the regularization parameter: $\tau = n^{-1} \sum_i deg(i)$
- $D \in \mathbb{R}^{n \times n}$ is a diagonal degree matrix (with inflated/regularized degrees): $D_{ii} = deg(i) + \tau$
- Normalized and regularized version: $L = D^{-1/2} A D^{-1/2}$

$$L_{ij} = \frac{A_{ij}}{\sqrt{(deg(i) + \tau)(deg(j) + \tau)}}$$

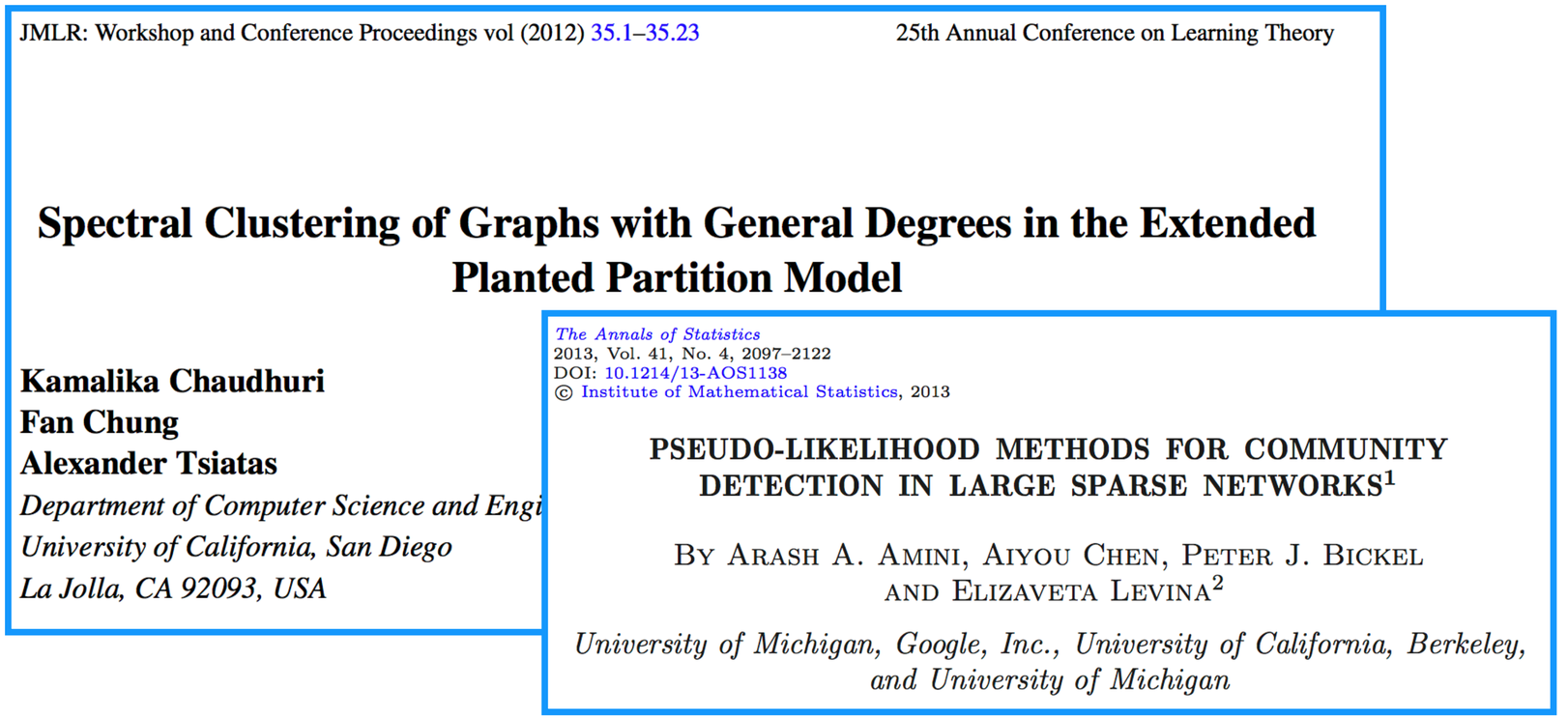
A on the left. L on the right.

Both generated from the same simulated data (2 block DC-SBM).



If you don’t normalize and regularize, then you will not be impressed with spectral results.

- I know this from personal experience!
- The idea was first proposed in two different papers (after I had already become disillusioned!)
- Please use it!



JMLR: Workshop and Conference Proceedings vol (2012) 35.1–35.23 25th Annual Conference on Learning Theory

Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model

Kamalika Chaudhuri
Fan Chung
Alexander Tsiatas
Department of Computer Science and Engi
University of California, San Diego
La Jolla, CA 92093, USA

The Annals of Statistics
2013, Vol. 41, No. 4, 2097–2122
DOI: 10.1214/13-AOS1138

PSEUDO-LIKELIHOOD METHODS FOR COMMUNITY DETECTION IN LARGE SPARSE NETWORKS[1]

BY ARASH A. AMINI, AIYOU CHEN, PETER J. BICKEL AND ELIZAVETA LEVINA[2]

University of Michigan, Google, Inc., University of California, Berkeley, and University of Michigan

Not much intuition for monster three.

It’s magic. Cast the spell!

This is a great spot for questions!

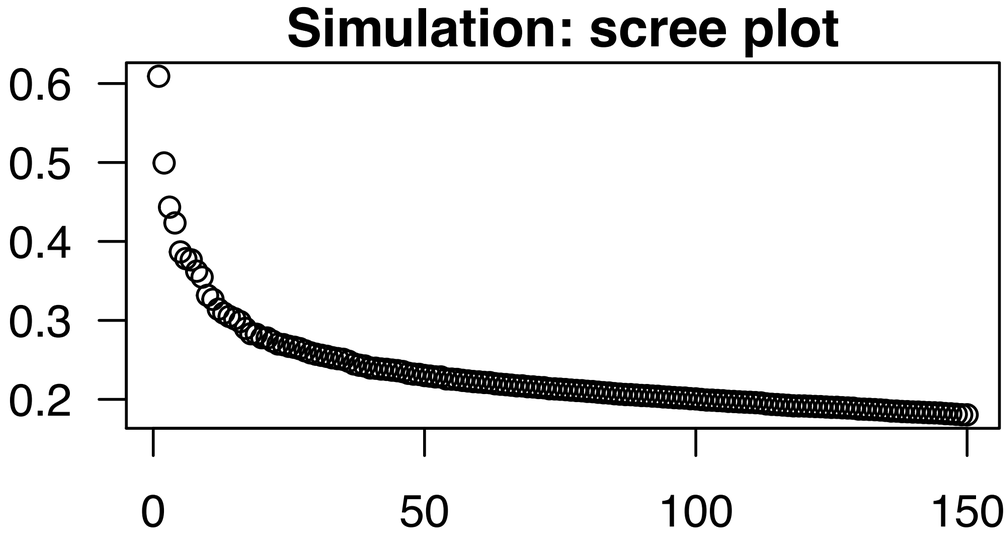
We’ve now arrived at the final boss.

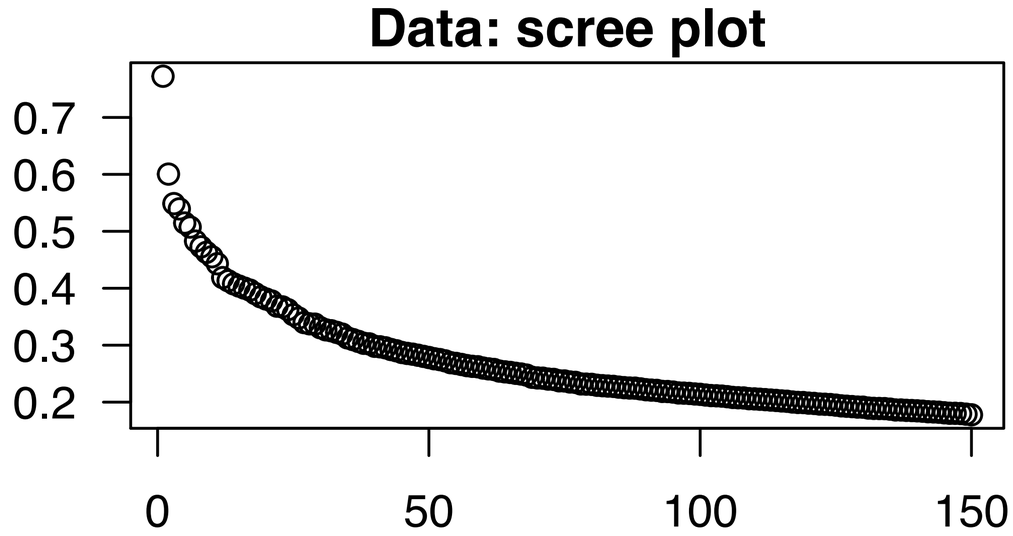
Monster 4: How to pick k, the number of eigenvectors?

Monster 4: How to pick k, the number of eigenvectors?

This is not a new problem.

- A classical solution is to look at the eigenvalues…

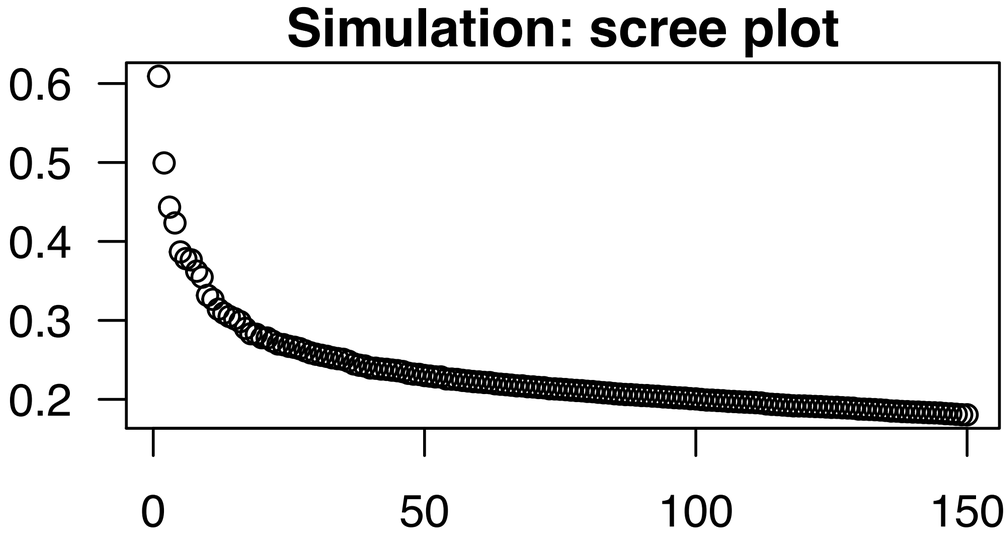


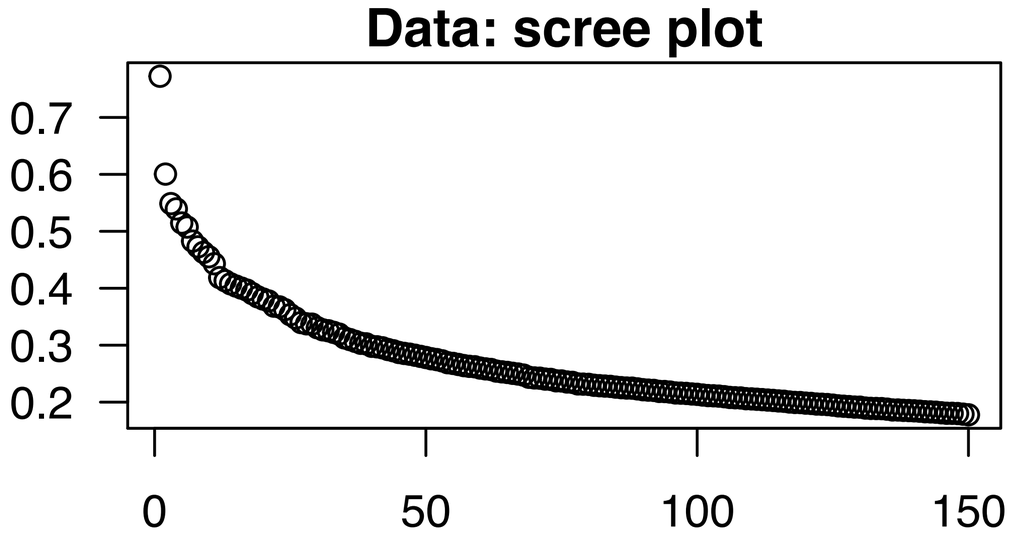


- What would you pick on left? On right?

Don’t mind the eigen-gap!

- On left, there are 128 dimensions in the simulation model.
  About 50 eigenvectors estimate some signal (correlate with Z)





- On right, the journal-journal graph has at least 100 dimensions
  (we will interpreted them)

The screeplot does not work because the eigenvalues are biased.

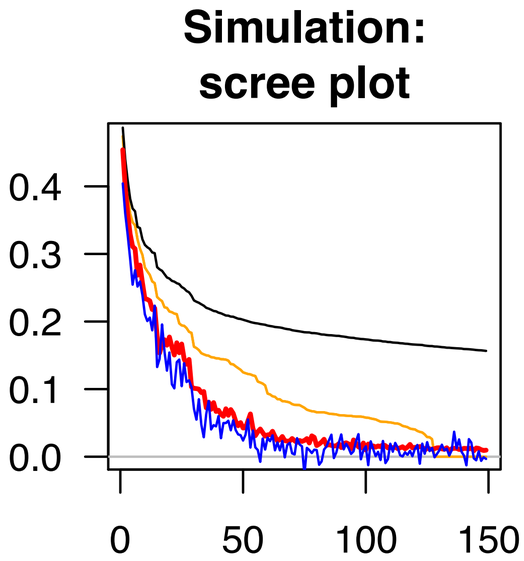
- $\hat{x} = \arg\max x^T A x$ (subject to some constraints)
- $\hat{\lambda} = \hat{x}^T A \hat{x}$
- Claim: we should be estimating $\lambda(\hat{x}) = \hat{x}^T P \hat{x}$, where $P = \mathbb{E}A$.
  This tells you how much signal $\hat{x}$ reveals about $P = ZBZ^T$.
- Old approach "double dips" on A to find $\hat{x}$ *and* $\hat{\lambda}$
- This makes $\hat{\lambda}$ a biased estimate of $\lambda(\hat{x})$.

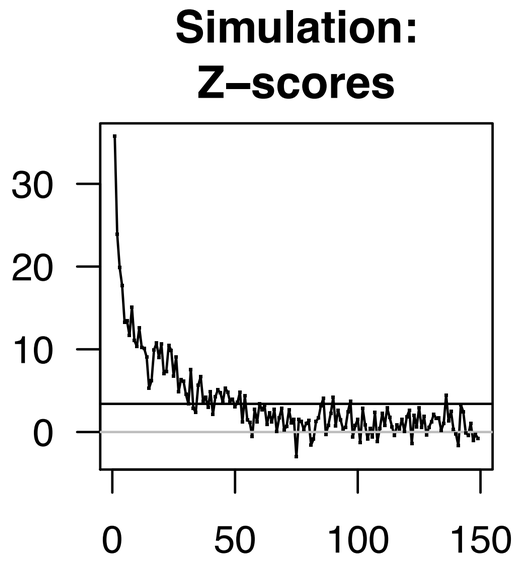
This is joint work with Fan Chen, Shuqi Yu, and Sebastien Roch

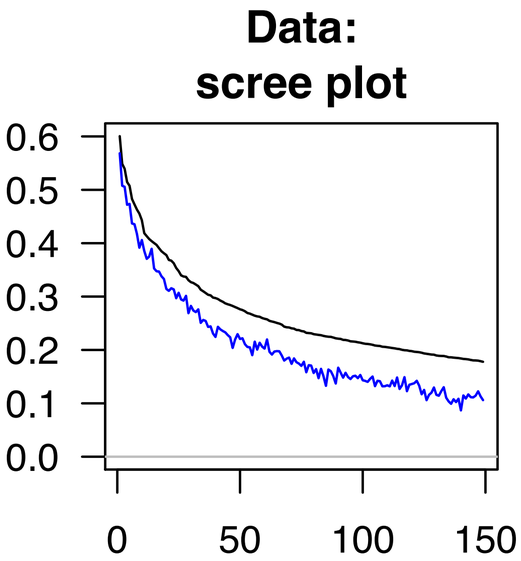
Solution: use “cross-validated eigenvalues”

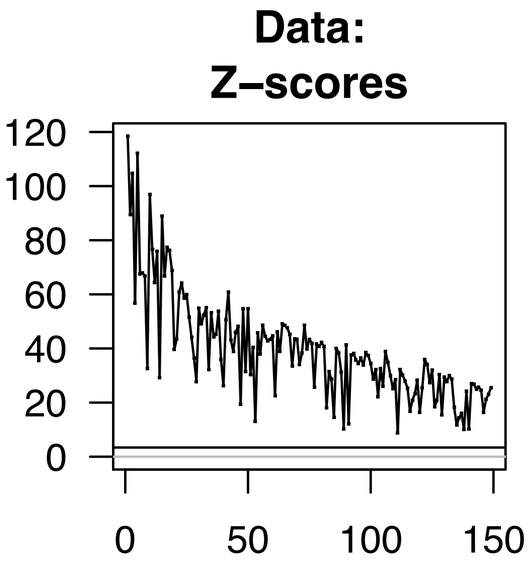
Blue line: CV-eigenvalues.

They come with Z-scores (p-values)!









This is joint work with Fan Chen, Shuqi Yu, and Sebastien Roch

Three key pieces to make CV-eigenvalues.

First piece: Make a “fitting graph” and a “test graph”

Second piece: The new graphs have the same k

Third piece: cross-validated eigenvalue has a CLT!

First piece: Make a “fitting graph” and a “test graph”

- Suppose $A_{ij} \sim Poisson(P_{ij})$. For sparse graphs, this is a common approximation.
- Define
  $\tilde{A}_{ij} \sim Binomial(A_{ij}, 1-\epsilon)$ and $[\tilde{A}_{test}]_{ij} = A_{ij} - \tilde{A}_{ij}$
- Wild fact: $\tilde{A}$ and $\tilde{A}_{test}$ are independent! Are you serious?? Yes, I’m serious.
- If A contains Bernoulli elements the simulation results are very good because the graphs are *negatively* dependent (test becomes conservative)

Second piece: All three graphs have the same k

- k is the number of non-zero eigenvalues of $\mathbb{E}A = P$.
- $\mathbb{E}\tilde{A} = \mathbb{E}\mathbb{E}(\tilde{A} \,|\, A) = \mathbb{E}(1-\epsilon)A = (1-\epsilon)P$
- Similarly, $\mathbb{E}\tilde{A}_{test} = \epsilon P$
- Same eigenvectors and the same k! …
- If $Px = \lambda x$, then x is also an eigenvector of $\mathbb{E}\tilde{A} = (1-\epsilon)P$ and $\mathbb{E}\tilde{A}_{test} = \epsilon P$ with eigenvalue $(1-\epsilon)\lambda$ and $\epsilon\lambda$ respectively.

$$\tilde{A}_{ij} \sim Binomial(A_{ij}, 1-\epsilon) \quad \text{and} \quad [\tilde{A}_{test}]_{ij} = A_{ij} - \tilde{A}_{ij}$$

Third piece: the cross-validated eigenvalue has a CLT!

- Let $\tilde{x}$ be an eigenvector from $\tilde{A}$ (or $\tilde{L}$).
- Define the cross validated eigenvalue: $\hat{\lambda}(\tilde{x}) = \tilde{x}^T \tilde{A}_{test} \tilde{x}$
- Because the graphs are independent, it is an unbiased estimator of $\lambda(\tilde{x}) = \tilde{x}^T P \tilde{x}$:

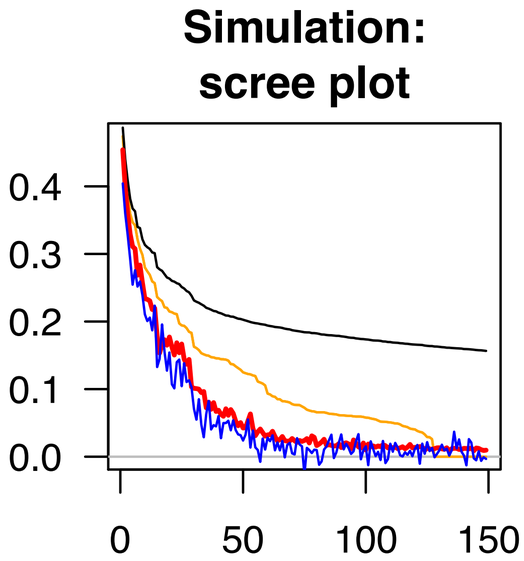
$$\mathbb{E}(\hat{\lambda}(\tilde{x}) \,|\, \tilde{A}) = \tilde{x}^T (\epsilon P) \tilde{x} = \epsilon \lambda(\tilde{x})$$

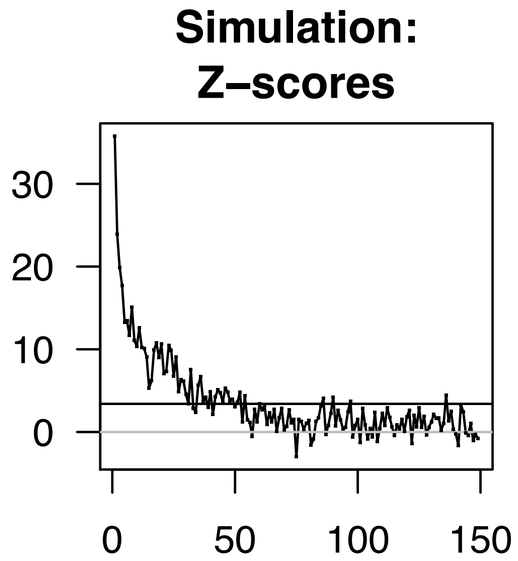
- Also, $\hat{\lambda}(\tilde{x})$ is the sum of weighted independent variables in $\tilde{A}_{test}$, hence a CLT.
- Test the null: $H_0 : \lambda(\tilde{x}) = \tilde{x}^T P \tilde{x} = 0$ with test statistic $\hat{\lambda}(\tilde{x})$

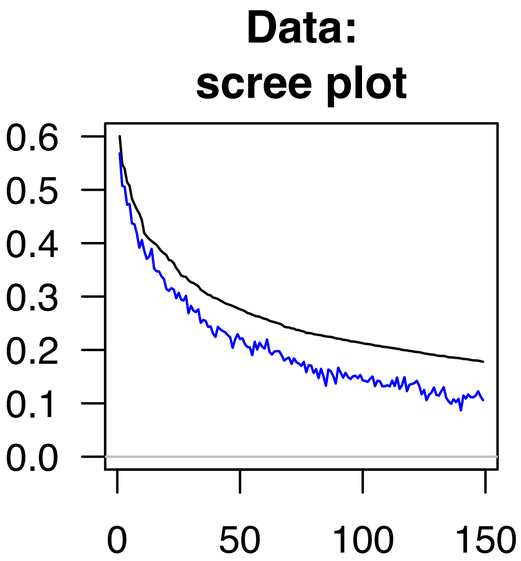
Advantages of this approach

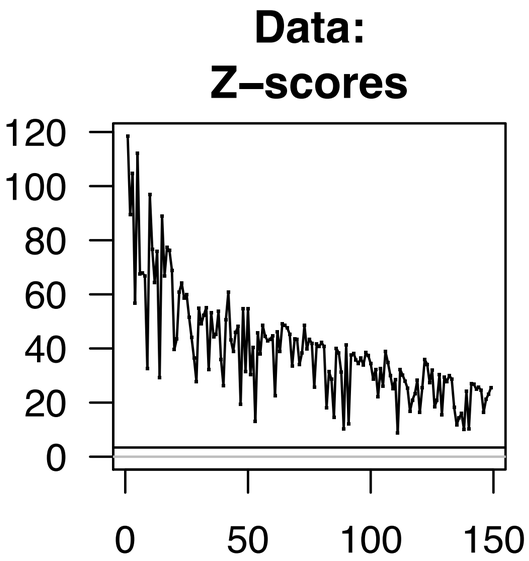
- Fast to compute!
- Does not require that we can estimate all “true k” eigenvectors.
- You get a p-value from a very simple CLT.
- Insensitive to degree distribution; compares favorably in time and performance to alternative approaches
- We have a theorem that it is consistent (under an asymptotic setting where you can estimate all k eigenvectors)

In the journal-journal graph, the first 150 dimensions are all highly statistically significant (all Z-scores > 8)









The final boss is not quite dead (yet?)

- Downside: You compute the embedding with only $(1-\epsilon) \approx 95\,\%$ of the data.

There are four steps to interpretable embeddings.

1. Normalize and Regularize the adjacency matrix A, to make L.
2. Compute a bunch of (eigenvector, eigenvalue) pairs of L.
3. Estimate k, the number of eigenvectors that are revealing signal.
4. Varimax rotate the leading k eigenvectors.

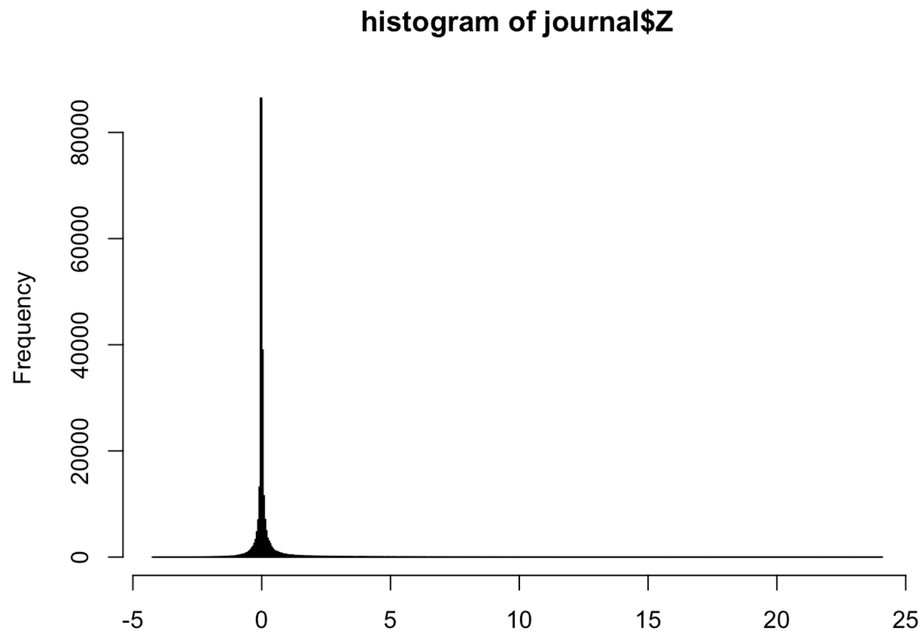
Finally, you need to interpret each dimension
(e.g. give it a name)

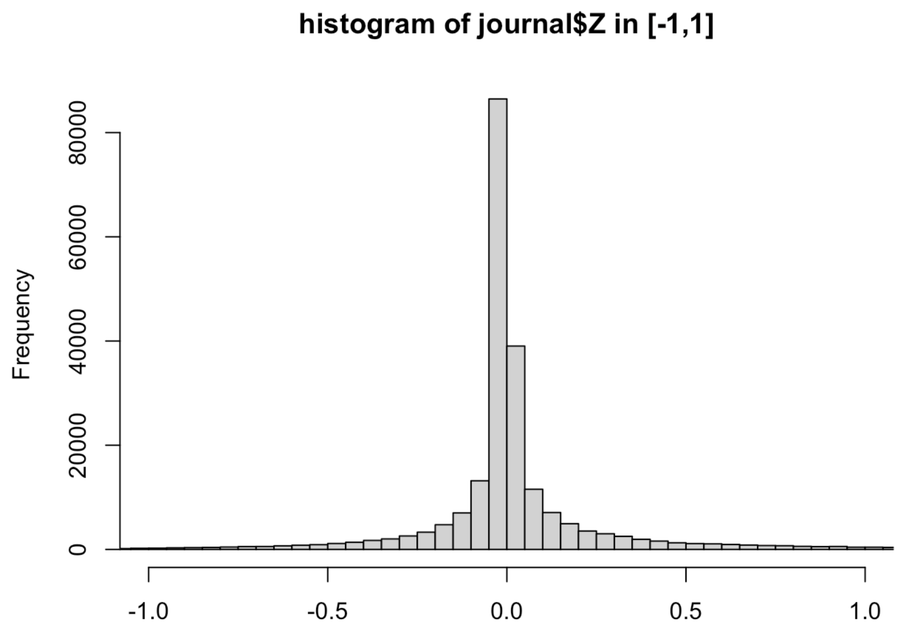
This is a great spot for questions!

How to interpret an embedding?

First, ignore negative loadings.

- Always define factors so that they skew positive.
- Empirical observation: when k is appropriately large, the negative "loadings" become less and less meaningful.
  "Embeddings put information in negative loadings when k is too small"
- Even if k is too small, negative loadings often are mixtures of different categories, while the positive loadings have a "coherent meaning"
- Be brave. Ignore negative loadings!





How to interpret an embedding?

One column of Z at a time…

- Each element of that column corresponds to a single node in the graph.
- List the nodes in the graph that have the biggest loading values *in that column*?
  This works if you have a way to understand the nodes
  (e.g. journal names, words, twitter handles have meaning!)
- "Contextualize" with high dimensional features
  (e.g. journal titles as bag-of-words)
- Use the best feature function (bff) with those features…

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

Top 7 words in journal titles for each factor (bff on $k = 10$)

Each journal is embedded in $\mathbb{R}^k$. So, each journal gets $k = 10$ values. Journals with large 1st value are likely to have "medicine" in the title.

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

Empirical observation:

Increasing k gives a "more refined interpretation"

- At $k = 10$,
  AOS, JASA, and JRSS-B mix between factors 7 (ieee) and 9 (econ)
  AOP loads primarily on 8 (math).

- At $k = 50$,
  they all combine into a probability and statistics factor:

- At $k = 100$:
  one statistics factor
  one probability factor.

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.

First bff word on journal titles with $k = 100$

Each column in $k = 100$ aligns with a discipline (or sub-discipline)

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

We've made it to the end.

- *First monster: Why does spectral stuff reveal anything?*
  $\mathbb{E}A$ is low rank & low rank factors determined by the embedding.

- *Second monster: How do we get Z out of the spectral stuff?*
  Varimax! It finds sparse axes for the embedding.

- *Third monster: A is not optimal.*
  Cast the spell! Normalize and regularize A to get L

- *Fourth monster: Picking k*
  For each eigenvector $\tilde{x}$, test $H_0 : \tilde{x}^T \mathbb{E}A\tilde{x} = 0$ by constructing a test graph and using CLT.