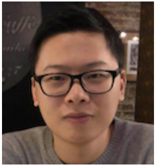
Understanding parallel analysis methods for rank selection in PCA

David Hong



Yue Sheng

Edgar Dobriban

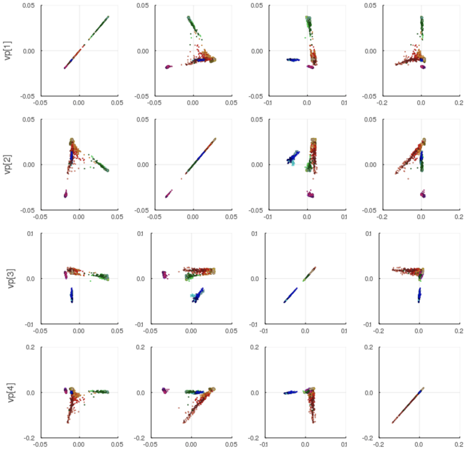
Wharton Statistics, University of Pennsylvania

Random Matrices and Complex Data Analysis Workshop
10 December 2019

This work was supported in part by NSF BIGDATA grant IIS 1837992 and NSF TRIPODS award 1934960.

An illustrative example: principal components for genetics

1000G genetics data: $n = 2318$ individuals, $p = 115019$ SNPs

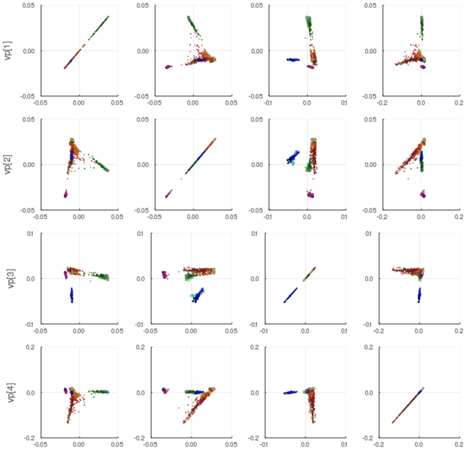


Rounak Dey

Xihong Lin

An illustrative example: principal components for genetics

1000G genetics data: $n = 2318$ individuals, $p = 115019$ SNPs



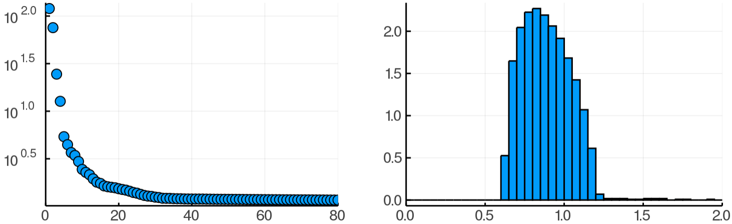
Rounak Dey

Xihong Lin

PC's can reveal population (and sub-population) structure, but how many are meaningful?

An illustrative example: principal components for genetics

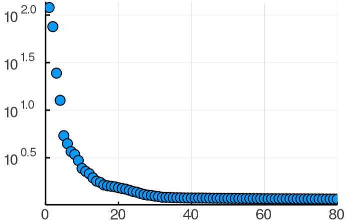
Often, we look at the scree plot and the spectrum:

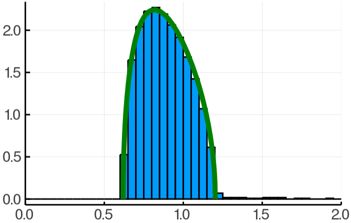


Question: how can we make principled selections and reason about them?

An illustrative example: principal components for genetics

Often, we look at the scree plot and the spectrum:





Question: how can we make principled selections and reason about them?

The spectrum looks like a spiked covariance model...

Rank selection for PCA

Rank selection is important – it affects every downstream step!

- too many: add noise to downstream analyses
- too few: lose signals that were in the data

Many excellent and practical methods:

- Likelihood ratio test (Bartlett 1950)
- Fixed threshold (Kaiser 1960)
- Scree plot (Cattell 1966)
- $4/\sqrt{3}$ (Gavish & Donoho 2014)
- bi-cross-validation (Owen & Wang 2016)
- ...

Today's talk: parallel analysis (Horn, 1965; Buja & Eyuboglu 1992)

Rank selection for PCA

Rank selection is important – it affects every downstream step!

- too many: add noise to downstream analyses
- too few: lose signals that were in the data

Many excellent and practical methods:

- Likelihood ratio test (Bartlett 1950)
- Fixed threshold (Kaiser 1960)
- Scree plot (Cattell 1966)
- $4/\sqrt{3}$ (Gavish & Donoho 2014)
- bi-cross-validation (Owen & Wang 2016)
- ...

Today's talk: parallel analysis (Horn, 1965; Buja & Eyuboglu 1992)

PA is a popular method with extensive empirical evidence, but limited theoretical understanding – exciting area for work!

Parallel analysis for rank selection

Parallel analysis is suggested in many reviews:

- Brown (2014): PA "is accurate in the vast majority of cases"
- Hayton et al. (2004): PA is "one of the most accurate factor retention methods" used in social science and management
- Costello and Osborne (2005): PA is "accurate and easy to use"
- Friedman et al. (2009): defaults to PA for rank selection

Parallel analysis for rank selection

Parallel analysis is suggested in many reviews:

- Brown (2014): PA "is accurate in the vast majority of cases"
- Hayton et al. (2004): PA is "one of the most accurate factor retention methods" used in social science and management
- Costello and Osborne (2005): PA is "accurate and easy to use"
- Friedman et al. (2009): defaults to PA for rank selection

Also gaining popularity in applied statistics (esp. biological sciences):

- Leek and Storey (2007)
- Leek and Storey (2008)
- Lin et al. (2016)
- Gerard and Stephens (2017)
- Zhou et al. (2017)
- ...

Parallel analysis for rank selection

Parallel analysis is suggested in many reviews:

- Brown (2014): PA "is accurate in the vast majority of cases"
- Hayton et al. (2004): PA is "one of the most accurate factor retention methods" used in social science and management
- Costello and Osborne (2005): PA is "accurate and easy to use"
- Friedman et al. (2009): defaults to PA for rank selection

Also gaining popularity in applied statistics (esp. biological sciences):

- Leek and Storey (2007)
- Leek and Storey (2008)
- Lin et al. (2016)
- Gerard and Stephens (2017)
- Zhou et al. (2017)
- ...

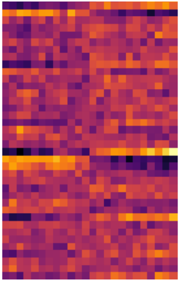
But there remains limited theoretical understanding:

PA is "at best a heuristic approach rather than a mathematically rigorous one" – Green et al. (2012)

Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

1. Generate X_π by randomly permuting the entries in each column

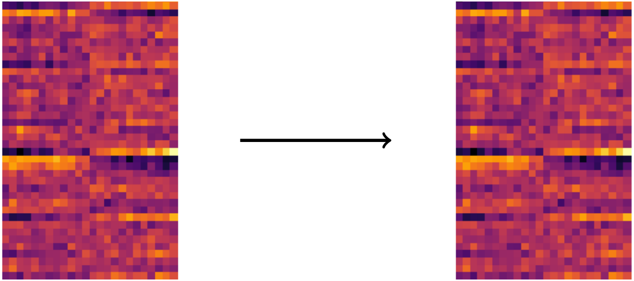


X

Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

1. Generate X_π by randomly permuting the entries in each column

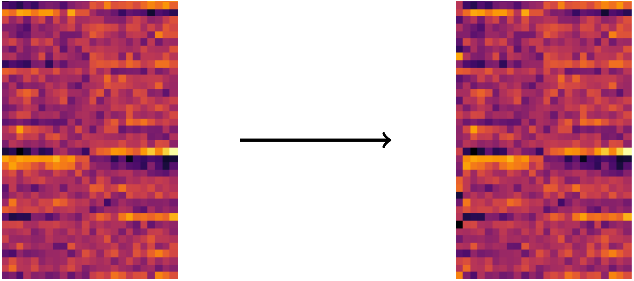


X $\qquad\qquad$ X_π

Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

1. Generate X_π by randomly permuting the entries in each column

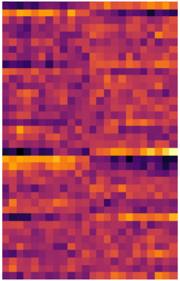


X $\qquad\qquad$ X_π

Parallel analysis for rank selection

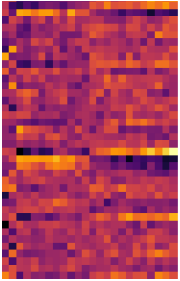
Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

1. Generate X_π by randomly permuting the entries in each column



X

$\longrightarrow$

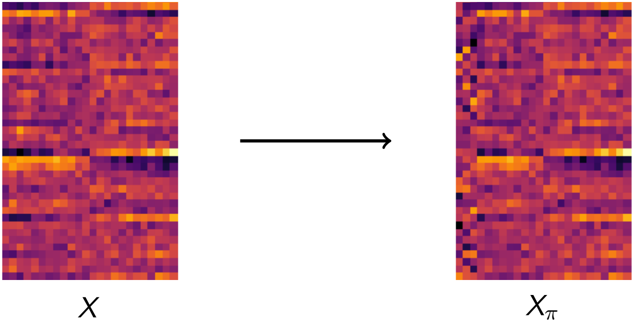


X_π

Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n\times p}$ and percentile $\alpha \in [0,1]$

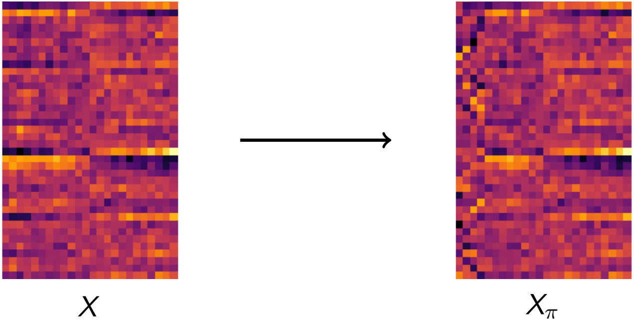
1. Generate X_π by randomly permuting the entries in each column



Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

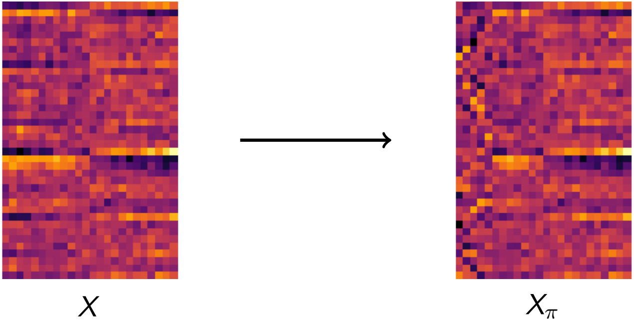
1. Generate X_π by randomly permuting the entries in each column



Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

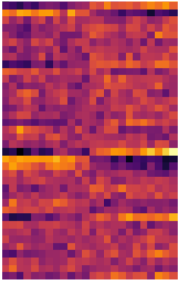
1. Generate X_π by randomly permuting the entries in each column



Parallel analysis for rank selection

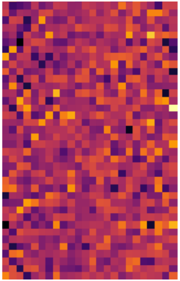
Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

1. Generate X_π by randomly permuting the entries in each column



X

$\longrightarrow$

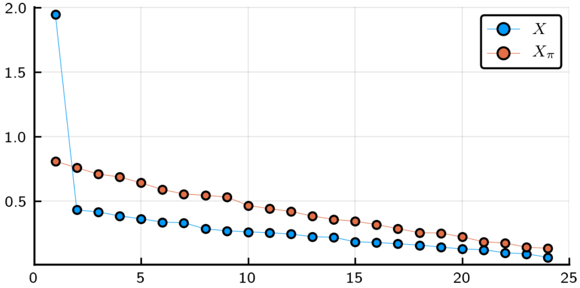


X_π

Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

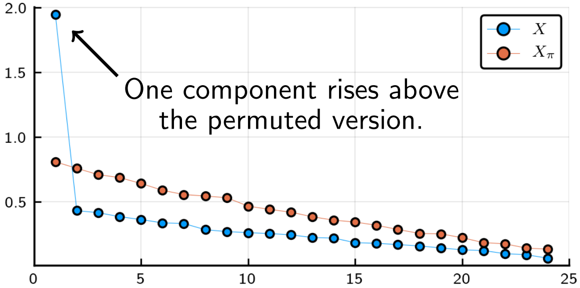
1. Generate X_π by randomly permuting the entries in each column
2. Repeat several times
3. Select the kth component if the kth singular value of X exceeds the α-percentile of the kth singular value of X_π



Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

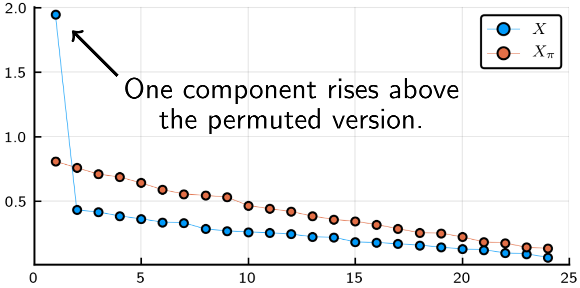
1. Generate X_π by randomly permuting the entries in each column
2. Repeat several times
3. Select the kth component if the kth singular value of X exceeds the α-percentile of the kth singular value of X_π



Parallel analysis for rank selection

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

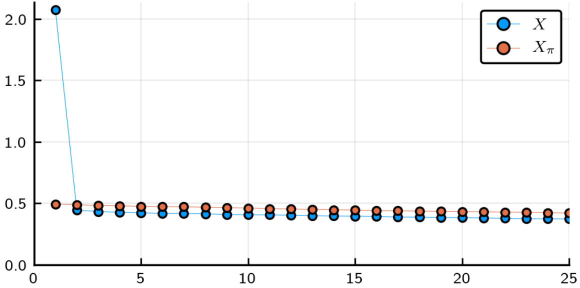
1. Generate X_π by randomly permuting the entries in each column
2. Repeat several times
3. Select the kth component if the kth singular value of X exceeds the α-percentile of the kth singular value of X_π



Idea: recover "null" by destroying correlations between features.

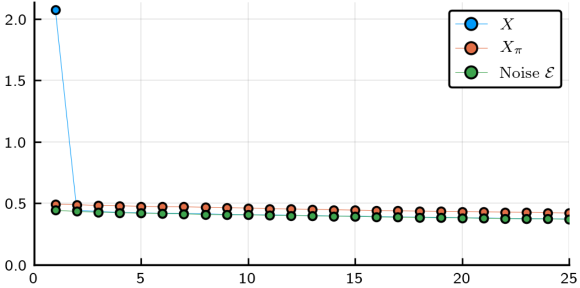
A quick sneak peak...

For a larger version of the same problem, i.e., bigger n, p:



A quick sneak peak...

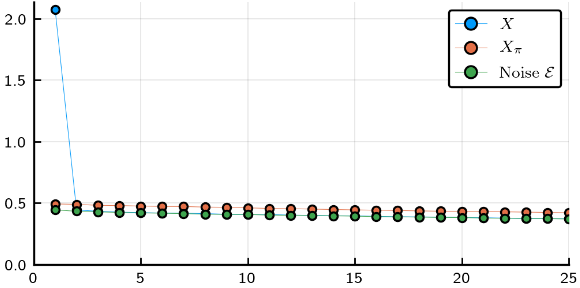
For a larger version of the same problem, i.e., bigger n, p:



Permutation provides a good estimate of the noise spectrum.

A quick sneak peak...

For a larger version of the same problem, i.e., bigger n, p:



Permutation provides a good estimate of the noise spectrum.
...let's begin characterizing this a bit!

Parallel analysis under factor models

Model: data is a linear combination of factors λ_{jk} with noise ε_{ij}

$$X_{ij} = \sum_{k=1}^{r} \eta_{ik}\lambda_{jk} + \varepsilon_{ij},$$

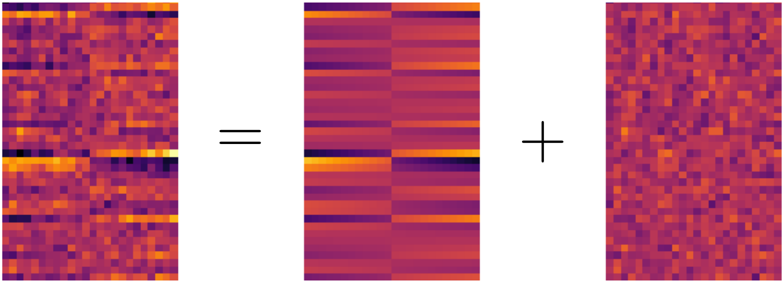
Parallel analysis under factor models

Model: data is a linear combination of factors λ_{jk} with noise ε_{ij}

$$X_{ij} = \sum_{k=1}^{r} \eta_{ik}\lambda_{jk} + \varepsilon_{ij},$$

i.e., low-rank signal + noise

$$X = \underbrace{\eta\Lambda^\top}_{S} + \mathcal{E} = S + \mathcal{E}.$$

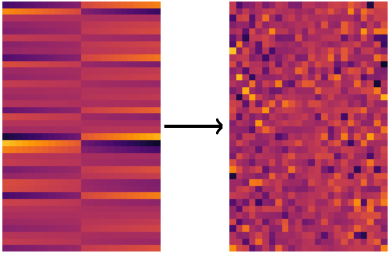


Parallel analysis under factor models

Key idea: permutation "destroys" the signal S but not the noise $\mathcal{E}$

Parallel analysis under factor models

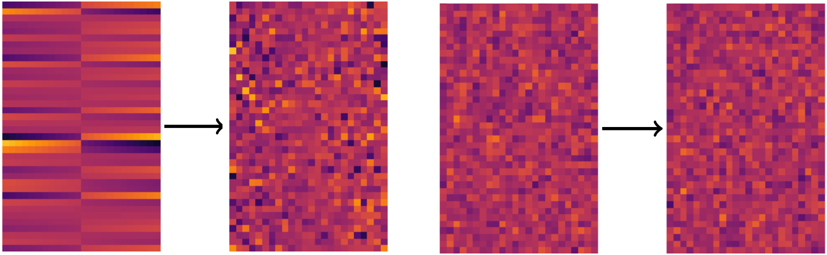
Key idea: permutation "destroys" the signal S but not the noise $\mathcal{E}$



S $\qquad$ $\|S_\pi\| \ll \|S\|$

Parallel analysis under factor models

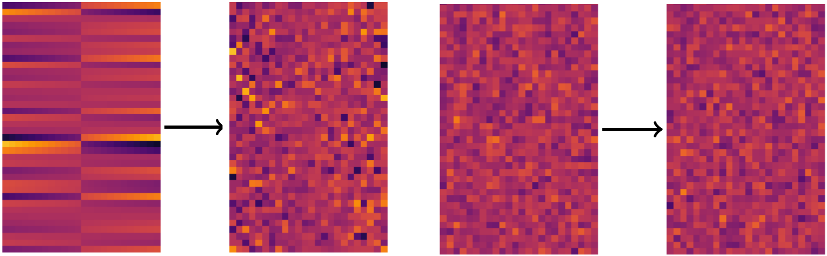
Key idea: permutation "destroys" the signal S but not the noise $\mathcal{E}$



S $\qquad \|S_\pi\| \ll \|S\|$ $\qquad \mathcal{E}$ $\qquad \mathcal{E}_\pi =_d \mathcal{E}$

Parallel analysis under factor models

Key idea: permutation "destroys" the signal S but not the noise $\mathcal{E}$



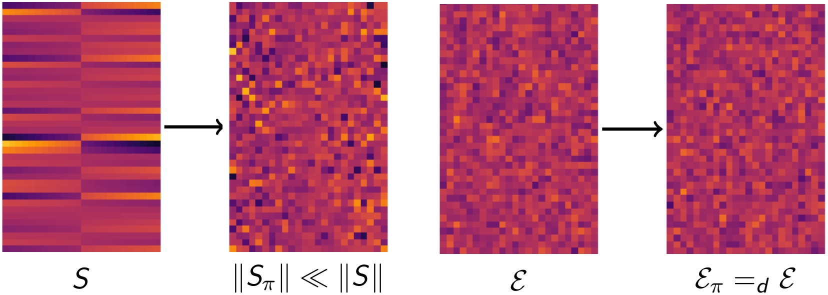
S $\qquad$ $\|S_\pi\| \ll \|S\|$ $\qquad$ $\mathcal{E}$ $\qquad$ $\mathcal{E}_\pi =_d \mathcal{E}$

Consequence: PA estimates noise spectrum (i.e., noise floor)

$$\sigma_k(X_\pi) = \sigma_k(S_\pi + \mathcal{E}_\pi) \approx \sigma_k(\mathcal{E}_\pi) =_d \sigma_k(\mathcal{E}_\pi).$$

Parallel analysis under factor models

Key idea: permutation "destroys" the signal S but not the noise $\mathcal{E}$



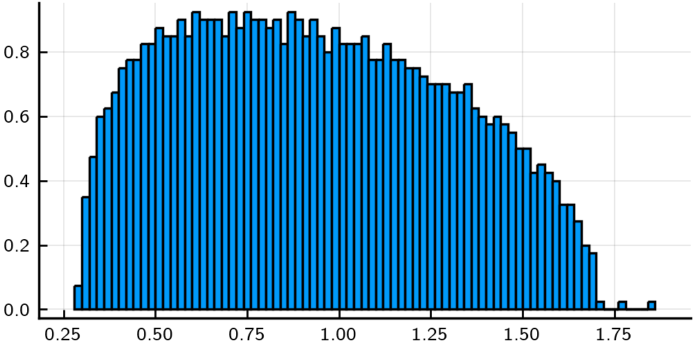
Consequence: PA estimates noise spectrum (i.e., noise floor)

$$\sigma_k(X_\pi) = \sigma_k(S_\pi + \mathcal{E}_\pi) \approx \sigma_k(\mathcal{E}_\pi) =_d \sigma_k(\mathcal{E}_\pi).$$

When does permutation successfully do this?

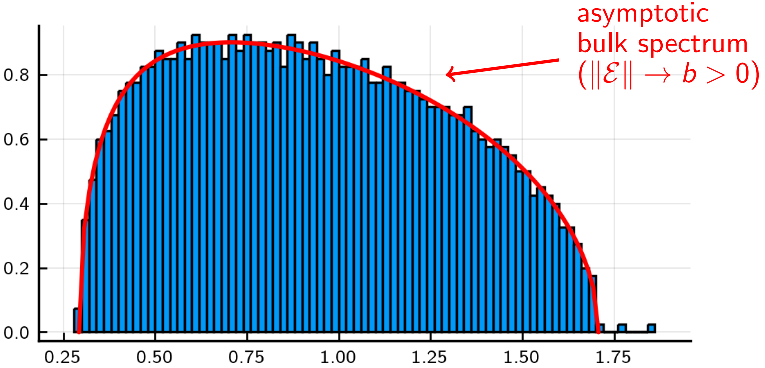
Important aside: small factors can fall below the noise

Example: Three factors, but only two rise above the phase transition.



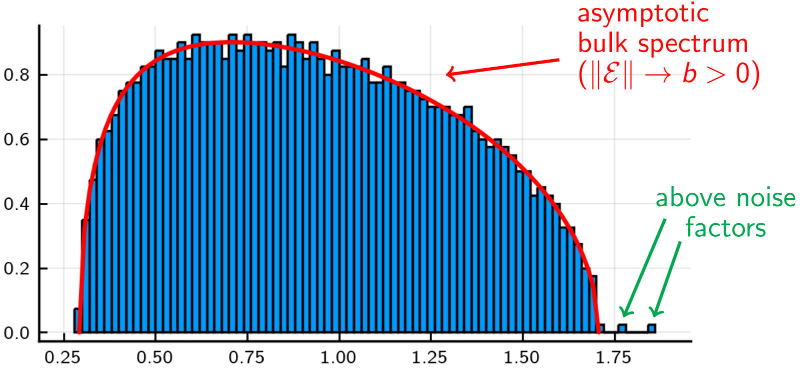
Important aside: small factors can fall below the noise

Example: Three factors, but only two rise above the phase transition.



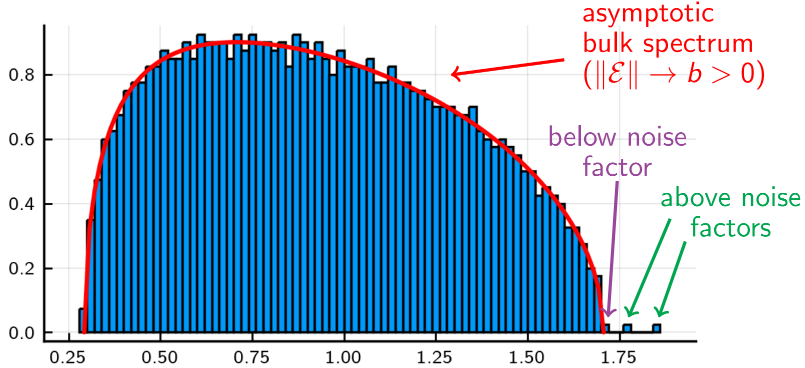
Important aside: small factors can fall below the noise

Example: Three factors, but only two rise above the phase transition.



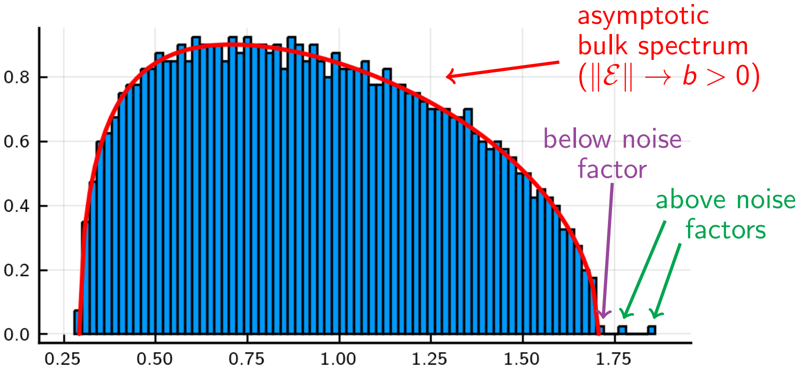
Important aside: small factors can fall below the noise

Example: Three factors, but only two rise above the phase transition.



Important aside: small factors can fall below the noise

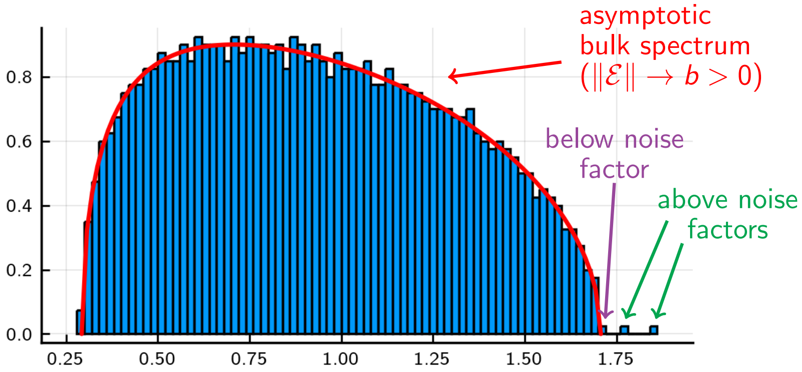
Example: Three factors, but only two rise above the phase transition.



Perceptible factor: singular value $\sigma_k > b + \delta$ a.s. for some $\delta > 0$
Imperceptible factors: singular value $\sigma_k < b - \delta$ a.s. for some $\delta > 0$

Important aside: small factors can fall below the noise

Example: Three factors, but only two rise above the phase transition.



Perceptible factor: singular value $\sigma_k > b + \delta$ a.s. for some $\delta > 0$
Imperceptible factors: singular value $\sigma_k < b - \delta$ a.s. for some $\delta > 0$

Question: when does parallel analysis identify <u>perceptible</u> factors?

Formalizing the intuition

Theorem. Suppose $X = S + \mathcal{E}$ with signal $S = \eta\Lambda^\top$ where

- $\eta = U\Psi^{1/2}$ for some Ψ where $U \in \mathbb{R}^{n\times r}$ has ind. stand. entries;
- $\Lambda\Psi^{1/2} = (f_1, \ldots, f_r)$ has bounded and delocalized columns, i.e., $\|f_k\|_2 \le Cn^{1/4-\delta/2}$ and $\|f_k\|_4/\|f_k\|_2 \to 0$;

and with noise $\mathcal{E} = Z\Phi^{1/2}$ where $\Phi = \mathrm{diag}(\phi)$ is diagonal,

- $Z \in \mathbb{R}^{n\times p}$ has ind. stand. entries with bounded fourth moment;
- entries of Z have bounded $(6+\Delta)$th moments;
- $p^{-1}\sum_j \delta_{\phi_j} \Rightarrow H$ and $\max_j \phi_j \to U(H)$ as $n, p \to \infty$ with $p/n \to \gamma > 0$.

Then PA selects *all perceptible* and *no imperceptible* factors with prob $\to 1$.

Formalizing the intuition

Theorem. Suppose $X = S + \mathcal{E}$ with signal $S = \eta\Lambda^\top$ where

- $\eta = U\Psi^{1/2}$ for some Ψ where $U \in \mathbb{R}^{n\times r}$ has ind. stand. entries;
- $\Lambda\Psi^{1/2} = (f_1, \dots, f_r)$ has bounded and delocalized columns, i.e., $\|f_k\|_2 \le Cn^{1/4-\delta/2}$ and $\|f_k\|_4/\|f_k\|_2 \to 0$;

and with noise $\mathcal{E} = Z\Phi^{1/2}$ where $\Phi = \mathrm{diag}(\phi)$ is diagonal,

- $Z \in \mathbb{R}^{n\times p}$ has ind. stand. entries with bounded fourth moment;
- entries of Z have bounded $(6+\Delta)$th moments;
- $p^{-1}\sum_j \delta_{\phi_j} \Rightarrow H$ and $\max_j \phi_j \to U(H)$ as $n, p \to \infty$ with $p/n \to \gamma > 0$.

Then PA selects *all perceptible* and *no imperceptible* factors with prob $\to 1$.

Key: Provide conditions so that

a) $\|N\| \to b > 0$, b) $N_\pi =_d N$, c) $\|S_\pi\| \to 0$.

Formalizing the intuition

Theorem. Suppose $X = S + \mathcal{E}$ with signal $S = \eta\Lambda^\top$ where

- $\eta = U\Psi^{1/2}$ for some Ψ where $U \in \mathbb{R}^{n\times r}$ has ind. stand. entries;
- $\Lambda\Psi^{1/2} = (f_1, \dots, f_r)$ has bounded and delocalized columns, i.e., $\|f_k\|_2 \le Cn^{1/4-\delta/2}$ and $\|f_k\|_4/\|f_k\|_2 \to 0$;

and with noise $\mathcal{E} = Z\Phi^{1/2}$ where $\Phi = \mathrm{diag}(\phi)$ is diagonal,

- $Z \in \mathbb{R}^{n\times p}$ has ind. stand. entries with bounded fourth moment;
- entries of Z have bounded $(6+\Delta)$th moments;
- $p^{-1}\sum_j \delta_{\phi_j} \Rightarrow H$ and $\max_j \phi_j \to U(H)$ as $n, p \to \infty$ with $p/n \to \gamma > 0$.

Then PA selects *all perceptible* and *no imperceptible* factors with prob $\to 1$.

Key: Provide conditions so that

a) $\|N\| \to b > 0$, b) $N_\pi =_d N$, c) $\|S_\pi\| \to 0$.

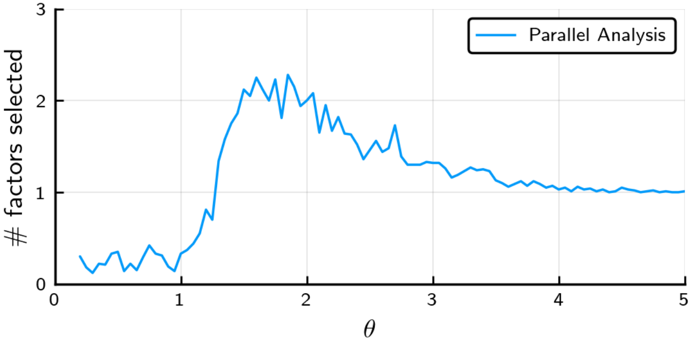
Involved deriving new moment bounds

Numerical experiment

Setup: $n = 500$ samples with $p = 300$ features, $r = 1$ latent factor.

$$X = \theta\sqrt{\gamma}\eta\Lambda^\top + \mathcal{E},$$

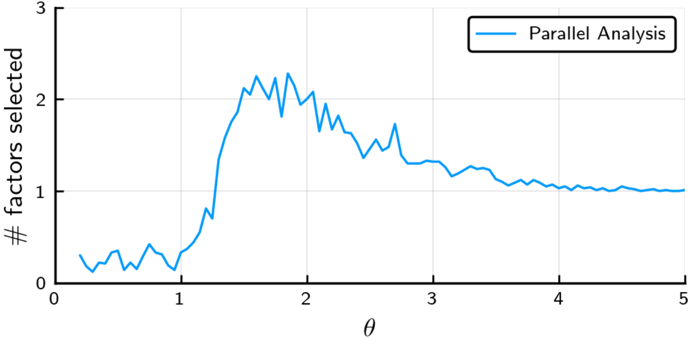
where $\eta \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, $\Lambda \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, and $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$.



Setup: $n = 500$ samples with $p = 300$ features, $r = 1$ latent factor.

$$X = \theta\sqrt{\gamma}\eta\Lambda^\top + \mathcal{E},$$

where $\eta \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, $\Lambda \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, and $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$.



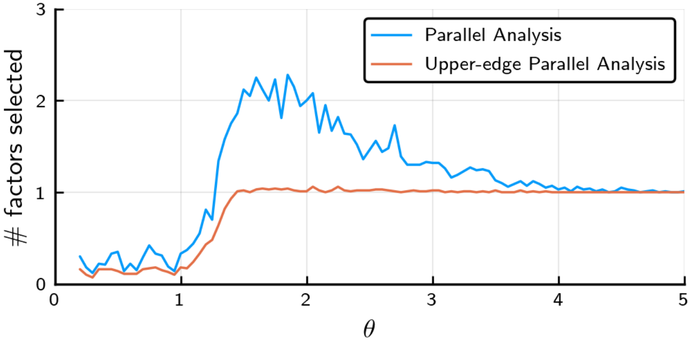
Comparing against $\sigma_1(X_\pi)$ can help combat overselection.

Numerical experiment

Setup: $n = 500$ samples with $p = 300$ features, $r = 1$ latent factor.

$$X = \theta\sqrt{\gamma}\eta\Lambda^\top + \mathcal{E},$$

where $\eta \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, $\Lambda \sim \mathrm{Unif}(\mathbb{S}^{p-1})$, and $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/n)$.



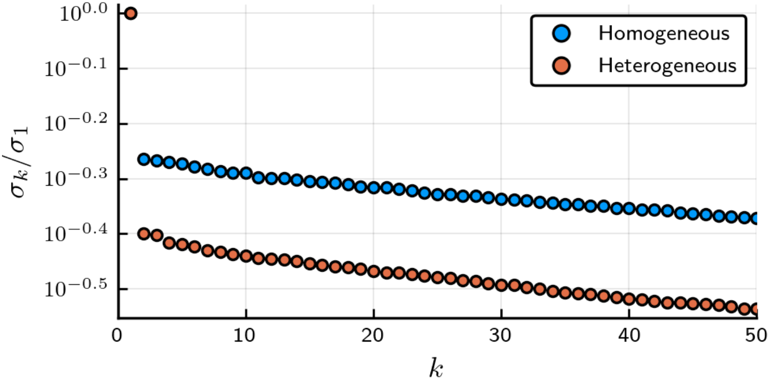
Comparing against $\sigma_1(X_\pi)$ can help combat overselection.

What if the noise is not invariant under permutation?

Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.

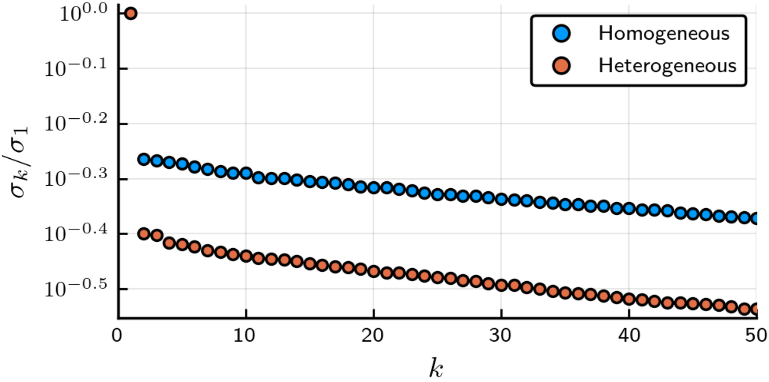
What if the noise is not invariant under permutation?

Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



What if the noise is not invariant under permutation?

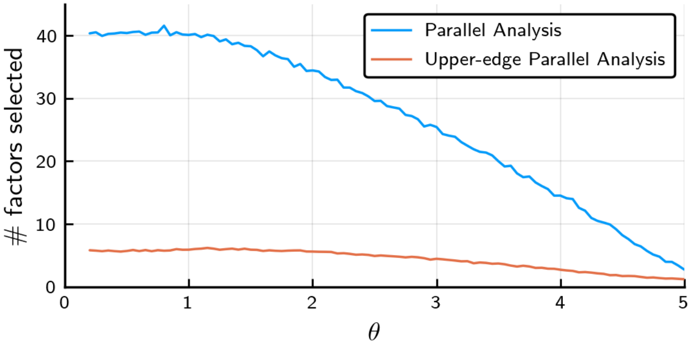
Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



This heterogeneous data is less noisy, should be easier!

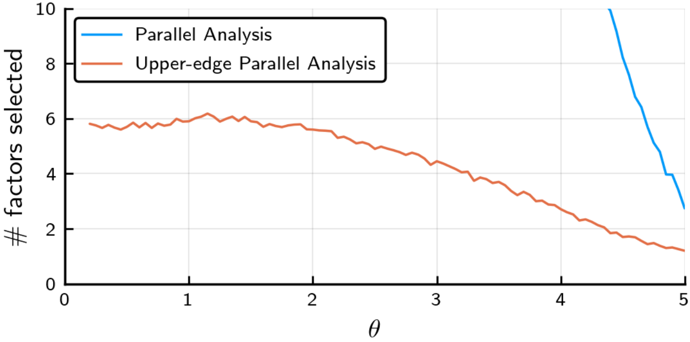
What if the noise is not invariant under permutation?

Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



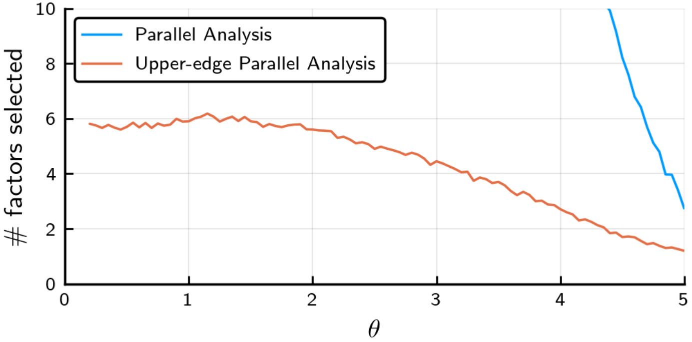
What if the noise is not invariant under permutation?

Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



What if the noise is not invariant under permutation?

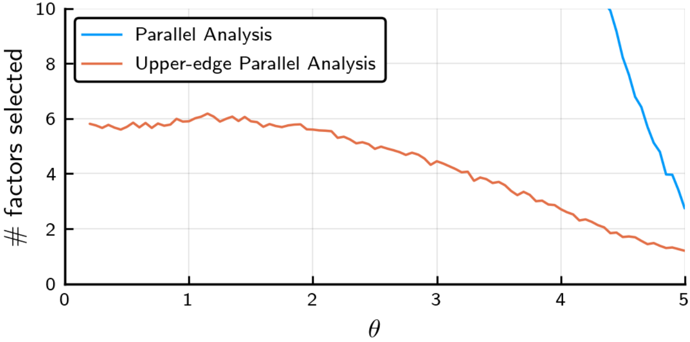
Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



But it performs <u>much</u> worse...

What if the noise is not invariant under permutation?

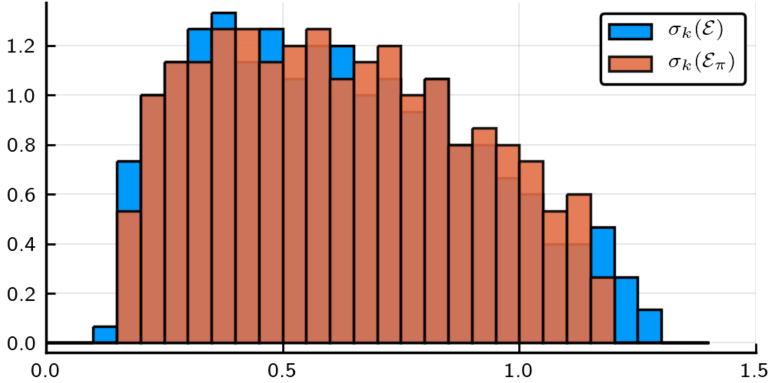
Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



But it performs <u>much</u> worse...what is happening?

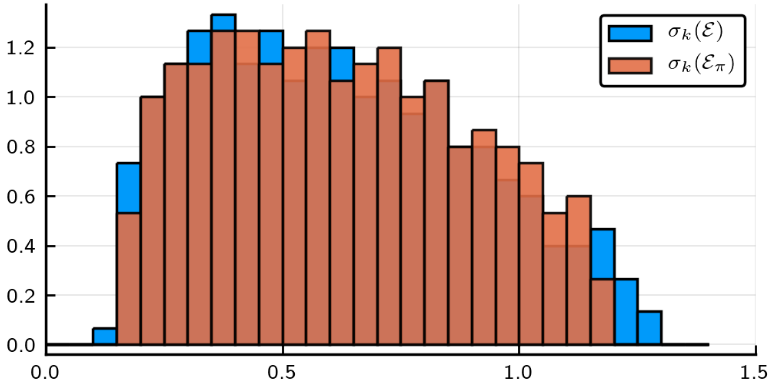
What if the noise is not invariant under permutation?

Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



What if the noise is not invariant under permutation?

Example: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.

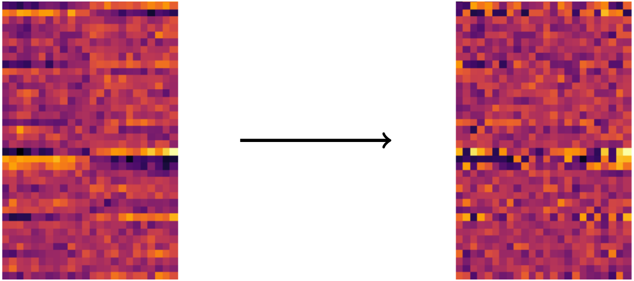


Permutation shrinks the noise spectrum, leading to overselection.

Idea: Replace permutation with signflips $\rightarrow$ Signflip PA

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

1. Generate X_π by randomly sign-flipping all entries

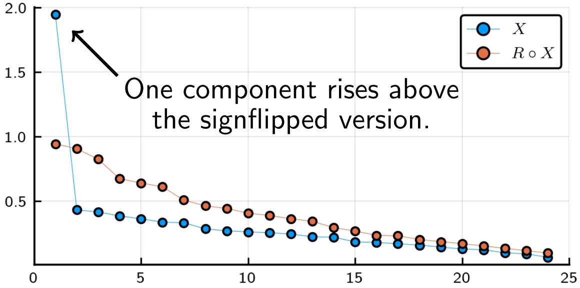


X $\qquad$ $R \circ X$

Idea: Replace permutation with signflips → Signflip PA

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

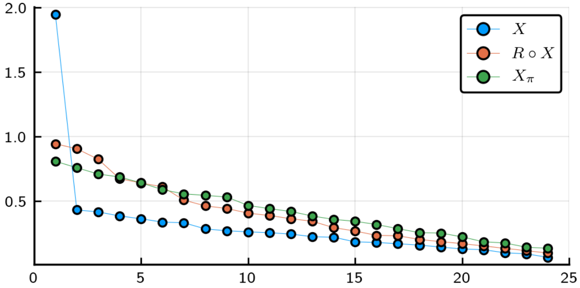
1. Generate X_π by randomly sign-flipping all entries
2. Repeat several times
3. Select the kth component if the kth singular value of X exceeds the α-percentile of the kth singular value of X_π



Idea: Replace permutation with signflips → Signflip PA

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

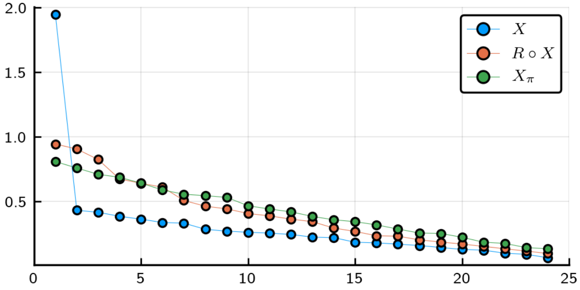
1. Generate X_π by randomly sign-flipping all entries
2. Repeat several times
3. Select the kth component if the kth singular value of X exceeds the α-percentile of the kth singular value of X_π



Idea: Replace permutation with signflips → Signflip PA

Given: data matrix $X \in \mathbb{R}^{n \times p}$ and percentile $\alpha \in [0, 1]$

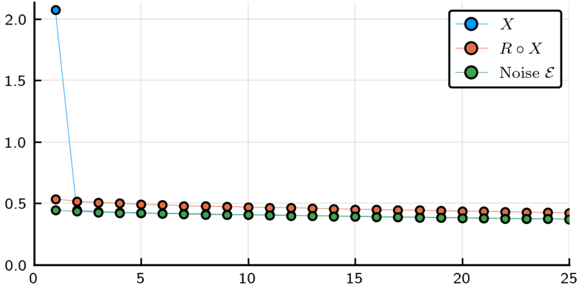
1. Generate X_π by randomly sign-flipping all entries
2. Repeat several times
3. Select the kth component if the kth singular value of X exceeds the α-percentile of the kth singular value of X_π



Sign-flipping also recovers the "null" by destroying correlations.

Idea: Replace permutation with signflips → Signflip PA

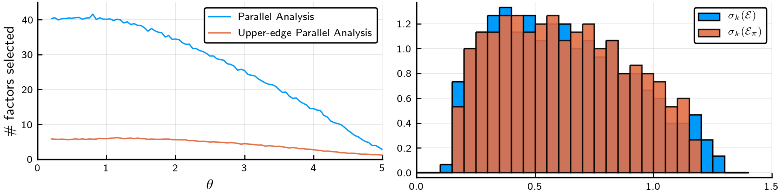
For a larger version of the same problem, i.e., bigger n, p:



Signflip PA also provides a good estimate of the noise spectrum.

Revisit: PA for the heterogeneous example

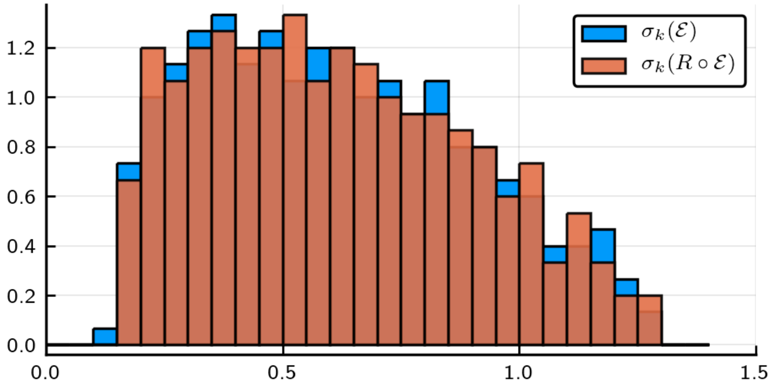
Recall: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



Permutation shrinks the noise spectrum, leading to overselection.

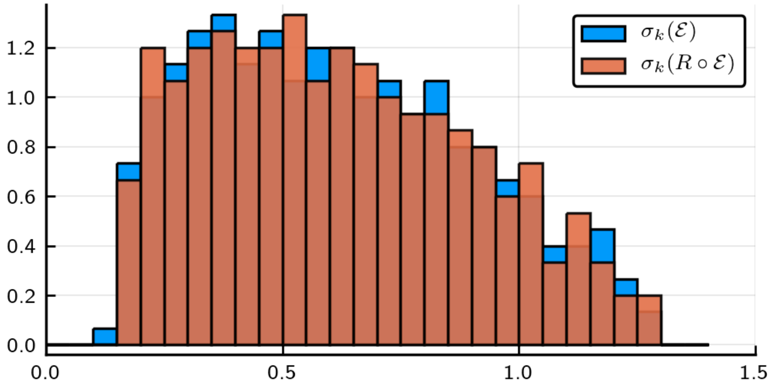
Revisit: Signflip PA for the heterogeneous example

Recall: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



Revisit: Signflip PA for the heterogeneous example

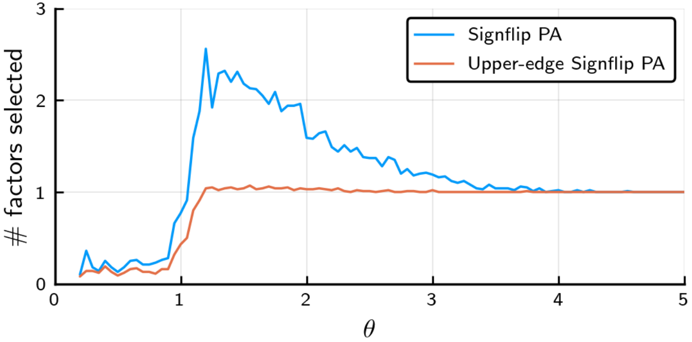
Recall: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



Signflips preserve the noise spectrum (in distribution).

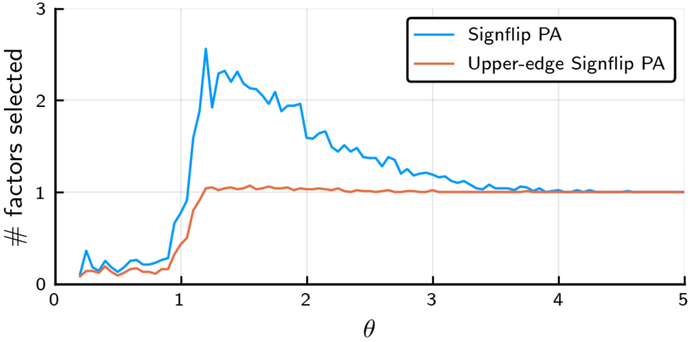
Revisit: Signflip PA for the heterogeneous example

Recall: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



Revisit: Signflip PA for the heterogeneous example

Recall: $\varepsilon_{ij} \overset{ind}{\sim} \mathcal{N}(0, \omega_i^2/n)$, 90% have $\omega_i^2 = 0.4$, 10% have $\omega_i^2 = 1$.



Preserving the noise distribution with signflips addresses the overselection of permutation.

Application to single cell RNA sequencing

Work with: Thomas Zhang, George Linderman, Yuval Kluger (Yale)

Question: how to select rank for single-cell RNA sequencing data?

Challenge: data does not (readily) fit our signal + noise setups.

Application to single cell RNA sequencing

Work with: Thomas Zhang, George Linderman, Yuval Kluger (Yale)

Question: how to select rank for single-cell RNA sequencing data?

Challenge: data does not (readily) fit our signal + noise setups.

Model: n samples are drawn independently from a multinomial

$$x_i \stackrel{ind}{\sim} \text{Multinomial}(s_i, k_i),$$

where $S = (s_1, \ldots, s_n)^\top$ is row-stochastic and low-rank.

Application to single cell RNA sequencing

Work with: Thomas Zhang, George Linderman, Yuval Kluger (Yale)

Question: how to select rank for single-cell RNA sequencing data?

Challenge: data does not (readily) fit our signal + noise setups.

Model: n samples are drawn independently from a multinomial

$$x_i \overset{ind}{\sim} \text{Multinomial}(s_i, k_i),$$

where $S = (s_1, \ldots, s_n)^\top$ is row-stochastic and low-rank.

Writing it in a signal + noise form

$$X = S + (X - S) = S + N,$$

where $N = X - S$ is centered (since $\mathbb{E}X = S$), but has dep. entries.

Application to single cell RNA sequencing

Work with: Thomas Zhang, George Linderman, Yuval Kluger (Yale)

Question: how to select rank for single-cell RNA sequencing data?

Challenge: data does not (readily) fit our signal + noise setups.

Model: n samples are drawn independently from a multinomial

$$x_i \stackrel{ind}{\sim} \text{Multinomial}(s_i, k_i),$$

where $S = (s_1, \ldots, s_n)^\top$ is row-stochastic and low-rank.

Writing it in a signal + noise form

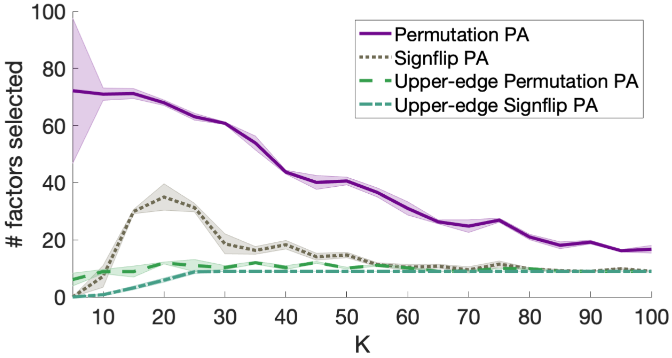
$$X = S + (X - S) = S + N,$$

where $N = X - S$ is centered (since $\mathbb{E}X = S$), but has dep. entries.

Ongoing work: how do our insights about PA apply here?

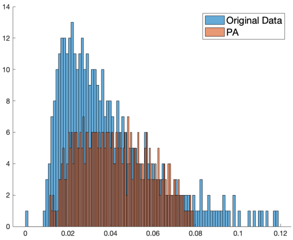
Application to single cell RNA sequencing

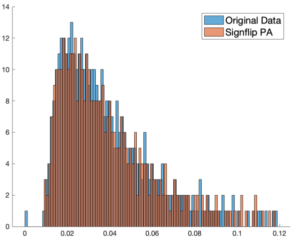
Prelim experiment: rank-10 S matrix, diverse total count rates, ...



Application to single cell RNA sequencing

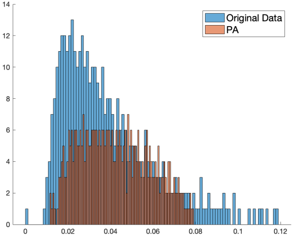
Prelim experiment: rank-10 S matrix, diverse total count rates, …

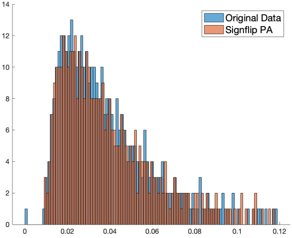




Application to single cell RNA sequencing

Prelim experiment: rank-10 S matrix, diverse total count rates, ...

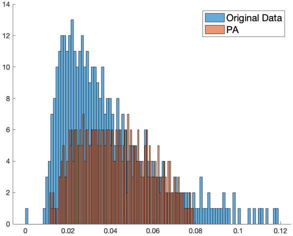


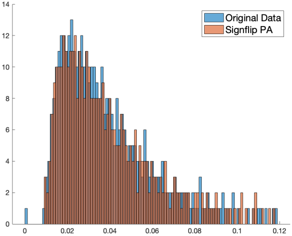


Permutations seem to shrink the noise spectrum sometimes and signflips seem to preserve them...

Application to single cell RNA sequencing

Prelim experiment: rank-10 S matrix, diverse total count rates, ...





Permutations seem to shrink the noise spectrum sometimes and signflips seem to preserve them...

Ongoing: theoretical analysis/characterization – how to deal with the dependence among noise entries?

Conclusions

Today:

- explaination for how parallel analysis works using insights/tools from *random matrix theory*
- some theoretical guarantees/characterization for parallel analysis
- signflip variant to handle alternative noise models
- preliminary work on applications to scRNAseq

Ongoing:

- characterization/analysis of signflip parallel analysis
- characterization of behavior under multinomial models
- application of similar ideas to other models?
- more evaluation in real data

Conclusions

Today:

- explaination for how parallel analysis works using insights/tools from *random matrix theory*
- some theoretical guarantees/characterization for parallel analysis
- signflip variant to handle alternative noise models
- preliminary work on applications to scRNAseq

Ongoing:

- characterization/analysis of signflip parallel analysis
- characterization of behavior under multinomial models
- application of similar ideas to other models?
- more evaluation in real data

Thanks!