Deterministic parallel analysis for selecting the number of factors

Edgar Dobriban

University of Pennsylvania

November 13, 2017

Overview

Factor analysis

Parallel analysis

A theory for parallel analysis

Improving parallel analysis

Summary

Overview

Factor analysis

Parallel analysis

A theory for parallel analysis

Improving parallel analysis

Summary

Factor analysis

- Fundamental statistical technique for unsupervised discovery of the factors driving variability in the data

Factor analysis

- Fundamental statistical technique for unsupervised discovery of the factors driving variability in the data
- Data: x_{ij}, j-th feature of i-th sample
  - e.g., education: p test scores of n students

Factor analysis

- Fundamental statistical technique for unsupervised discovery of the factors driving variability in the data
- Data: x_{ij}, j-th feature of i-th sample
  - e.g., education: p test scores of n students
- Factor analysis: Are there unobserved features that drive variability in the data?
  - factors: skills

Factor analysis

- Fundamental statistical technique for unsupervised discovery of the factors driving variability in the data
- Data: x_{ij}, j-th feature of i-th sample
  - e.g., education: p test scores of n students
- Factor analysis: Are there unobserved features that drive variability in the data?
  - factors: skills
- Very common problem in psychology, econometrics, biology etc.

Factor analysis

- Fundamental statistical technique for unsupervised discovery of the factors driving variability in the data
- Data: x_{ij}, j-th feature of i-th sample
  - e.g., education: p test scores of n students
- Factor analysis: Are there unobserved features that drive variability in the data?
  - factors: skills
- Very common problem in psychology, econometrics, biology etc.
- Dates back to Spearman's "general intelligence" (1904)

Factor analysis

- x_{ij} is a linear function of common factors η_{ik} and noise ε_{ij}:

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

Factor analysis

- x_{ij} is a linear function of common factors η_{ik} and noise ε_{ij}:

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

- Factor scores η_{ik} and the factor loadings λ_{jk} are not observed

Factor analysis

- x_{ij} is a linear function of common factors η_{ik} and noise ε_{ij}:

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

- Factor scores η_{ik} and the factor loadings λ_{jk} are not observed
- Education: η_{ik} is student i's level on the k-th skill, λ_{jk} is the relevance of the k-th skill to the j-th test.

Factor analysis

- In matrix form, $x_{ij} = \sum_{k=1}^{r} \eta_{ik}\lambda_{jk} + \varepsilon_{ij}$ is

$$X = \eta\Lambda^{\mathsf{T}} + \mathcal{E}.$$

  - $X = (x_1, \ldots, x_n)^{\mathsf{T}}$ is $n \times p$ data matrix; normalized st $\mathrm{Var}\left[x_{ij}\right] = 1$
  - η is the $n \times r$ matrix containing the factor values η_{ij}; normalized st $\mathrm{Var}\left[\eta_{ij}\right] = 1$
  - Λ is $p \times r$ factor loading matrix with entries λ_{jk}
  - $\mathcal{E}$ is $n \times p$ matrix containing the noise ε_{ij}

Factor analysis

- In matrix form, $x_{ij} = \sum_{k=1}^{r} \eta_{ik}\lambda_{jk} + \varepsilon_{ij}$ is

$$X = \eta\Lambda^{\mathsf{T}} + \mathcal{E}.$$

  - $X = (x_1, \ldots, x_n)^{\mathsf{T}}$ is $n \times p$ data matrix; normalized st $\mathrm{Var}\left[x_{ij}\right] = 1$
  - η is the $n \times r$ matrix containing the factor values η_{ij}; normalized st $\mathrm{Var}\left[\eta_{ij}\right] = 1$
  - Λ is $p \times r$ factor loading matrix with entries λ_{jk}
  - $\mathcal{E}$ is $n \times p$ matrix containing the noise ε_{ij}
- Statistical inference?

Inference is challenging[1]

- Not well-specified ($\eta \Lambda^{\mathsf{T}} = \eta M M^{-1} \Lambda^{\mathsf{T}}$). Need constraints.

[1] see e.g., Anderson, 2003, Intro Multivariate Analysis

Inference is challenging[1]

- Not well-specified ($\eta\Lambda^{\mathsf{T}} = \eta M M^{-1}\Lambda^{\mathsf{T}}$). Need constraints.
- MLE severely nonconvex. Unknown how to solve it. PCA often used.

[1] see e.g., Anderson, 2003, Intro Multivariate Analysis

Inference is challenging[1]

- Not well-specified ($\eta\Lambda^\mathsf{T} = \eta M M^{-1}\Lambda^\mathsf{T}$). Need constraints.
- MLE severely nonconvex. Unknown how to solve it. PCA often used.
- Inference in high dimensions requires delicate analysis:
  - Selecting number of factors
  - Estimating factors
  - Testing hypotheses

[1] see e.g., Anderson, 2003, Intro Multivariate Analysis

How to select the number of factors?

- Textbook by Brown (2014): "*the most crucial decision*" in FA. Affects every downstream step.
- Approaches:

How to select the number of factors?

- Textbook by Brown (2014): "*the most crucial decision*" in FA. Affects every downstream step.
- Approaches:
  1. Bartlett's likelihood ratio test (1950)

How to select the number of factors?

- Textbook by Brown (2014): "*the most crucial decision*" in FA. Affects every downstream step.
- Approaches:
  1. Bartlett's likelihood ratio test (1950)
  2. Kaiser's "eigenvalue larger than one" rule (1960)

How to select the number of factors?

- Textbook by Brown (2014): "*the most crucial decision*" in FA. Affects every downstream step.
- Approaches:
  1. Bartlett's likelihood ratio test (1950)
  2. Kaiser's "eigenvalue larger than one" rule (1960)
  3. Scree plot (Cattell, 1966)

How to select the number of factors?

- Textbook by Brown (2014): "*the most crucial decision*" in FA. Affects every downstream step.
- Approaches:
    1. Bartlett's likelihood ratio test (1950)
    2. Kaiser's "eigenvalue larger than one" rule (1960)
    3. Scree plot (Cattell, 1966)
    4. Parallel analysis [PA] (Horn, 1965; Buja & Eyuboglu 1992)

How to select the number of factors?

- Textbook by Brown (2014): "*the most crucial decision*" in FA. Affects every downstream step.
- Approaches:
  1. Bartlett's likelihood ratio test (1950)
  2. Kaiser's "eigenvalue larger than one" rule (1960)
  3. Scree plot (Cattell, 1966)
  4. Parallel analysis [PA] (Horn, 1965; Buja & Eyuboglu 1992)
  5. Many others: Kritchman and Nadler (2008); Onatski (2012); Josse and Husson (2012); Gaskin and Happell (2014), etc

Overview

Factor analysis

Parallel analysis

A theory for parallel analysis

Improving parallel analysis

Summary

What is parallel analysis?

1. Generate X_π by permuting the entries in each column of X separately.

What is parallel analysis?

1. Generate X_π by permuting the entries in each column of X separately.
2. Repeat a few times.

What is parallel analysis?

1. Generate X_π by permuting the entries in each column of X separately.
2. Repeat a few times.
3. For $k = 1, 2, \ldots$
    - 3.1 Select the k-th factor if the k-th singular value of X is larger than 95%-th percentile of the k-th singular values of the permuted matrices.

What is parallel analysis?

1. Generate X_π by permuting the entries in each column of X separately.
2. Repeat a few times.
3. For $k = 1, 2, \ldots$
    3.1 Select the k-th factor if the k-th singular value of X is larger than 95%-th percentile of the k-th singular values of the permuted matrices.
    3.2 If this is not true, exit loop.

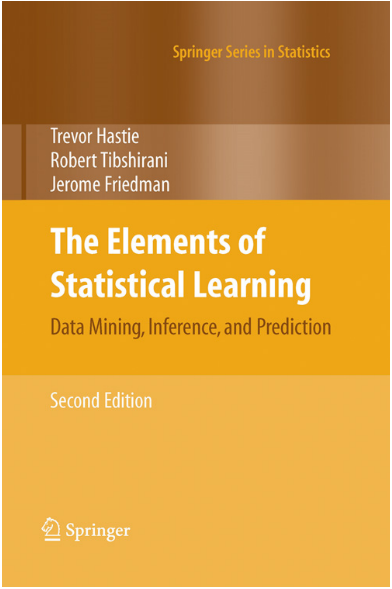
What is parallel analysis?

1. Generate X_π by permuting the entries in each column of X separately.
2. Repeat a few times.
3. For $k = 1, 2, \ldots$
    3.1 Select the k-th factor if the k-th singular value of X is larger than 95%-th percentile of the k-th singular values of the permuted matrices.
    3.2 If this is not true, exit loop.
4. Return $k - 1$

PA is recommended in many reviews on FA

1. Brown (2014): PA “*is accurate in the vast majority of cases*”
2. Hayton et al. (2004): evidence from social science and management that PA is “*one of the most accurate factor retention methods*”
3. Costello and Osborne (2005): PA is “*accurate and easy to use*”
4. Friedman et al. (2009) use it as the default method for selecting the number of PCs.

PA is used by leading applied statisticians



Springer Series in Statistics

Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer

PA is used by leading applied statisticians

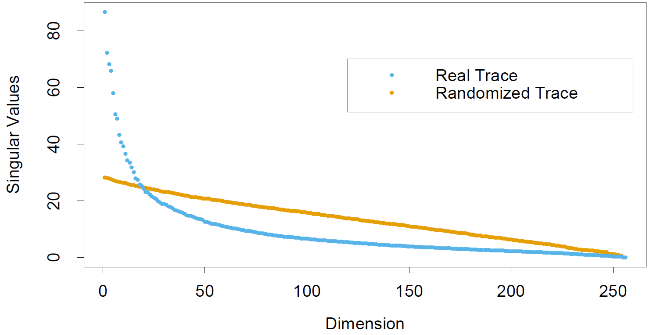
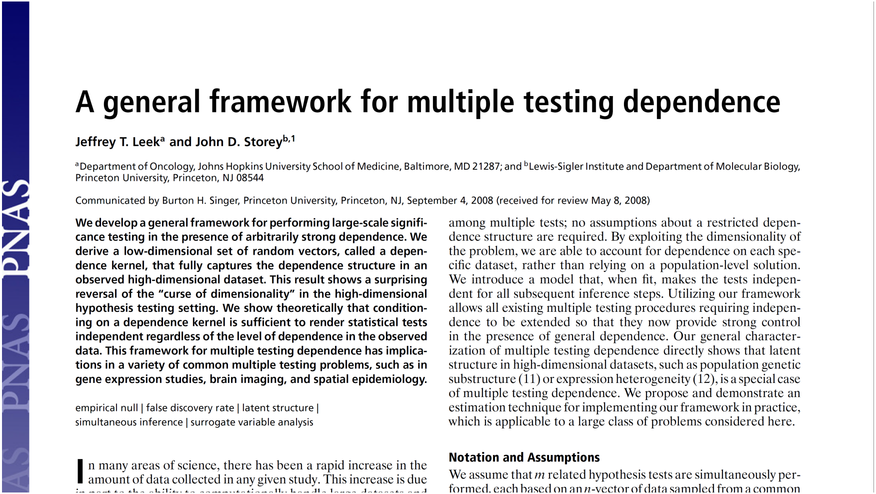


FIGURE 14.24. *The* 256 *singular values for the digitized threes, compared to those for a randomized version of the data (each column of* **X** *was scrambled).*

PA is used by leading applied statisticians



A general framework for multiple testing dependence

Jeffrey T. Leek[a] and John D. Storey[b,1]

[a]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD 21287; and [b]Lewis-Sigler Institute and Department of Molecular Biology, Princeton University, Princeton, NJ 08544

Communicated by Burton H. Singer, Princeton University, Princeton, NJ, September 4, 2008 (received for review May 8, 2008)

We develop a general framework for performing large-scale significance testing in the presence of arbitrarily strong dependence. We derive a low-dimensional set of random vectors, called a dependence kernel, that fully captures the dependence structure in an observed high-dimensional dataset. This result shows a surprising reversal of the "curse of dimensionality" in the high-dimensional hypothesis testing setting. We show theoretically that conditioning on a dependence kernel is sufficient to render statistical tests independent regardless of the level of dependence in the observed data. This framework for multiple testing dependence has implications in a variety of common multiple testing problems, such as in gene expression studies, brain imaging, and spatial epidemiology.

empirical null | false discovery rate | latent structure | simultaneous inference | surrogate variable analysis

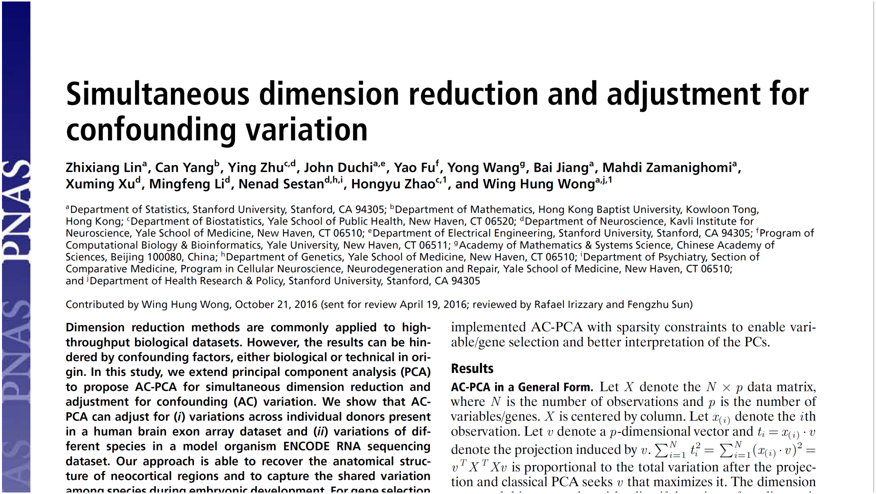
In many areas of science, there has been a rapid increase in the amount of data collected in any given study. This increase is due in part to the ability to computationally handle large datasets and

among multiple tests; no assumptions about a restricted dependence structure are required. By exploiting the dimensionality of the problem, we are able to account for dependence on each specific dataset, rather than relying on a population-level solution. We introduce a model that, when fit, makes the tests independent for all subsequent inference steps. Utilizing our framework allows all existing multiple testing procedures requiring independence to be extended so that they now provide strong control in the presence of general dependence. Our general characterization of multiple testing dependence directly shows that latent structure in high-dimensional datasets, such as population genetic substructure (11) or expression heterogeneity (12), is a special case of multiple testing dependence. We propose and demonstrate an estimation technique for implementing our framework in practice, which is applicable to a large class of problems considered here.

Notation and Assumptions

We assume that m related hypothesis tests are simultaneously performed, each based on an n-vector of data sampled from a common

PA is used by leading applied statisticians



Simultaneous dimension reduction and adjustment for confounding variation

Zhixiang Lin[a], Can Yang[b], Ying Zhu[c,d], John Duchi[a,e], Yao Fu[f], Yong Wang[g], Bai Jiang[a], Mahdi Zamanighomi[a], Xuming Xu[d], Mingfeng Li[d], Nenad Sestan[d,h,i], Hongyu Zhao[c,1], and Wing Hung Wong[a,j,1]

[a]Department of Statistics, Stanford University, Stanford, CA 94305; [b]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong; [c]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06520; [d]Department of Neuroscience, Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT 06510; [e]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [f]Program of Computational Biology & Bioinformatics, Yale University, New Haven, CT 06511; [g]Academy of Mathematics & Systems Science, Chinese Academy of Sciences, Beijing 100080, China; [h]Department of Genetics, Yale School of Medicine, New Haven, CT 06510; [i]Department of Psychiatry, Section of Comparative Medicine, Program in Cellular Neuroscience, Neurodegeneration and Repair, Yale School of Medicine, New Haven, CT 06510; and [j]Department of Health Research & Policy, Stanford University, Stanford, CA 94305

Contributed by Wing Hung Wong, October 21, 2016 (sent for review April 19, 2016; reviewed by Rafael Irizzary and Fengzhu Sun)

Dimension reduction methods are commonly applied to high-throughput biological datasets. However, the results can be hindered by confounding factors, either biological or technical in origin. In this study, we extend principal component analysis (PCA) to propose AC-PCA for simultaneous dimension reduction and adjustment for confounding (AC) variation. We show that AC-PCA can adjust for (*i*) variations across individual donors present in a human brain exon array dataset and (*ii*) variations of different species in a model organism ENCODE RNA sequencing dataset. Our approach is able to recover the anatomical structure of neocortical regions and to capture the shared variation among species during embryonic development. For gene selection

implemented AC-PCA with sparsity constraints to enable variable/gene selection and better interpretation of the PCs.

Results

AC-PCA in a General Form. Let X denote the $N \times p$ data matrix, where N is the number of observations and p is the number of variables/genes. X is centered by column. Let $x_{(i)}$ denote the ith observation. Let v denote a p-dimensional vector and $t_i = x_{(i)} \cdot v$ denote the projection induced by v. $\sum_{i=1}^{N} t_i^2 = \sum_{i=1}^{N} (x_{(i)} \cdot v)^2 = v^T X^T X v$ is proportional to the total variation after the projection and classical PCA seeks v that maximizes it. The dimension

PA is used by leading applied statisticians

ME] 23 May 2017

Unifying and Generalizing Methods for Removing Unwanted Variation Based on Negative Controls

David Gerard[1] and Matthew Stephens[1,2]
Departments of Human Genetics[1] and Statistics[2],
University of Chicago, Chicago, IL, 60637, USA

May 23, 2017

Abstract

Unwanted variation, including hidden confounding, is a well-known problem in many fields, particularly large-scale gene expression studies. Recent proposals to use control genes — genes assumed to be unassociated with the covariates of interest — have led to new methods to deal with this problem. Going by the moniker **R**emoving **U**nwanted **V**ariation (RUV), there are many versions — RUV1, RUV2, RUV4, RUVinv, RUVrinv, RUVfun. In this paper, we introduce a general framework, RUV*, that both unites and generalizes these approaches. This unifying framework helps clarify connections between existing methods. In particular we provide conditions under which RUV2 and RUV4 are equivalent. The

What is known about PA?

- Extensive empirical evidence that it “works”

What is known about PA?

- Extensive empirical evidence that it “works”
- No theory or formal understanding; mysterious?

Overview

Factor analysis

Parallel analysis

A theory for parallel analysis

Improving parallel analysis

Summary

A theory for parallel analysis

- A theory for PA, using random matrices
- Clarifies what it does
- Leads to improvements (joint work with Art Owen)

What does PA do?

- signal-plus-noise $X = S + N$ (e.g., $X = \eta\Lambda^{\mathsf{T}} + \mathcal{E}$)

What does PA do?

- signal-plus-noise $X = S + N$ (e.g., $X = \eta \Lambda^{\mathsf{T}} + \mathcal{E}$)
- permute each column independently: $X_\pi = S_\pi + N_\pi$

What happens when you permute the low-rank signal?

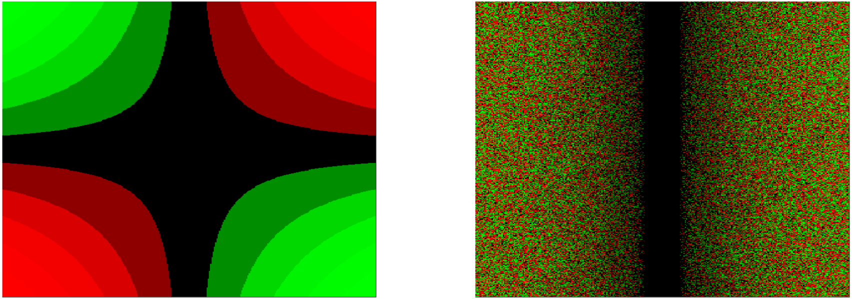


Figure : Heatmap of rank one S (left) and S_π (right).

What happens when you permute the low-rank signal?

- original signal $S = uv^\mathsf{T} = [v_1 u, \ldots, v_p u]$ of rank one

What happens when you permute the low-rank signal?

- original signal $S = uv^\mathsf{T} = [v_1 u, \ldots, v_p u]$ of rank one
- permuted signal $S_\pi = [v_1 \pi_1(u), \ldots, v_p \pi_p(u)]$

What happens when you permute the low-rank signal?

- original signal $S = uv^{\mathsf{T}} = [v_1 u, \ldots, v_p u]$ of rank one
- permuted signal $S_\pi = [v_1 \pi_1(u), \ldots, v_p \pi_p(u)]$
- if v is “spread out”, S_π typically has full rank.

What happens when you permute the low-rank signal?

- original signal $S = uv^{\mathsf{T}} = [v_1 u, \ldots, v_p u]$ of rank one
- permuted signal $S_\pi = [v_1 \pi_1(u), \ldots, v_p \pi_p(u)]$
- if v is "spread out", S_π typically has full rank.
- each column has effectively independent entries with variance v_j^2/n

What happens when you permute the low-rank signal?

- original signal $S = uv^\top = [v_1 u, \ldots, v_p u]$ of rank one
- permuted signal $S_\pi = [v_1 \pi_1(u), \ldots, v_p \pi_p(u)]$
- if v is “spread out”, S_π typically has full rank.
- each column has effectively independent entries with variance v_j^2/n
- total energy preserved: $|S_\pi|_{Fr} = |S|_{Fr}$

What happens when you permute the low-rank signal?

- original signal $S = uv^{\mathsf{T}} = [v_1 u, \ldots, v_p u]$ of rank one
- permuted signal $S_\pi = [v_1 \pi_1(u), \ldots, v_p \pi_p(u)]$
- if v is "spread out", S_π typically has full rank.
- each column has effectively independent entries with variance v_j^2/n
- total energy preserved: $|S_\pi|_{Fr} = |S|_{Fr}$
- Conclusion:
  - signal becomes small noise: $|S_\pi|_{op} \ll |S|_{op}$

What happens when you permute the noise?

- original noise $N = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$
- in the classical factor model, the n samples are iid

What happens when you permute the noise?

- original noise $N = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$
- in the classical factor model, the n samples are iid
- the noise in each coordinate is independent

What happens when you permute the noise?

- original noise $N = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$
- in the classical factor model, the n samples are iid
- the noise in each coordinate is independent
- therefore: noise is permutation-invariant in distribution $N_\pi =_d N$

What does PA do?

- signal-plus-noise $X = S + N$ (e.g., $X = \eta\Lambda^\mathsf{T} + \mathcal{E}$)
- permute each column independently: $X_\pi = S_\pi + N_\pi$
  - signal becomes small noise: $|S_\pi|_{op} \ll |S|_{op}$
  - noise invariance: $N_\pi =_d N$

What does PA do?

- signal-plus-noise $X = S + N$ (e.g., $X = \eta\Lambda^\mathsf{T} + \mathcal{E}$)
- permute each column independently: $X_\pi = S_\pi + N_\pi$
  - signal becomes small noise: $|S_\pi|_{op} \ll |S|_{op}$
  - noise invariance: $N_\pi =_d N$
- Recall PA selects first factor if $\sigma_1(X) > |X_\pi|_{op}$

What does PA do?

- signal-plus-noise $X = S + N$ (e.g., $X = \eta \Lambda^{\mathsf{T}} + \mathcal{E}$)
- permute each column independently: $X_\pi = S_\pi + N_\pi$
  - signal becomes small noise: $|S_\pi|_{op} \ll |S|_{op}$
  - noise invariance: $N_\pi =_d N$
- Recall PA selects first factor if $\sigma_1(X) > |X_\pi|_{op}$
- PA estimates noise operator norm:

$$|X_\pi|_{op} = |S_\pi + N_\pi|_{op} \approx |N_\pi|_{op} =_d |N|_{op}$$

What does PA do?

- signal-plus-noise $X = S + N$ (e.g., $X = \eta\Lambda^\mathsf{T} + \mathcal{E}$)
- permute each column independently: $X_\pi = S_\pi + N_\pi$
  - signal becomes small noise: $|S_\pi|_{op} \ll |S|_{op}$
  - noise invariance: $N_\pi =_d N$
- Recall PA selects first factor if $\sigma_1(X) > |X_\pi|_{op}$
- PA estimates noise operator norm:

$$|X_\pi|_{op} = |S_\pi + N_\pi|_{op} \approx |N_\pi|_{op} =_d |N|_{op}$$

- PA selects factors above noise op norm: $\sigma_k(X) > |N|_{op}$

Formalizing the intuition

- Factor model $x_i = \Lambda\eta_i + \varepsilon_i$,
- Matrix form $X = \eta\Lambda^{\mathsf{T}} + \mathcal{E}$

Formalizing the intuition

- Factor model $x_i = \Lambda\eta_i + \varepsilon_i$,
- Matrix form $X = \eta\Lambda^{\mathsf{T}} + \mathcal{E}$
- Asymptotic setting, $n, p \to \infty$ (will specify later how)

Formalizing the intuition

- Factor model $x_i = \Lambda\eta_i + \varepsilon_i$,
- Matrix form $X = \eta\Lambda^\mathsf{T} + \mathcal{E}$
- Asymptotic setting, $n, p \to \infty$ (will specify later how)
- Define *the size of the noise* $b > 0$ st (after normalizing)

  $$|\mathcal{E}|_{op} \to b$$

  almost surely (a.s.), or in probability.

Formalizing the intuition

- Define *perceptible factors* as those indices k for which

  $$\sigma_k(X) > b + \varepsilon$$

  a.s or in probability, for some $\varepsilon > 0$.

Formalizing the intuition

- Define *perceptible factors* as those indices k for which

  $$\sigma_k(X) > b + \varepsilon$$

  a.s or in probability, for some $\varepsilon > 0$.
- Similarly, define *imperceptible factors* as those indices k for which

  $$\sigma_k(X) < b - \varepsilon$$

  for some $\varepsilon > 0$.

Formalizing the intuition

- Define *perceptible factors* as those indices k for which

  $$\sigma_k(X) > b + \varepsilon$$

  a.s or in probability, for some $\varepsilon > 0$.
- Similarly, define *imperceptible factors* as those indices k for which

  $$\sigma_k(X) < b - \varepsilon$$

  for some $\varepsilon > 0$.
- Closely related to “above/below the phase transition” in spiked models.

Parallel analysis selects the perceptible factors

Theorem

n iid samples from the p-dimensional factor model $x_i = \Lambda\eta_i + \varepsilon_i$. *Assume:*

Parallel analysis selects the perceptible factors

Theorem

n iid samples from the p-dimensional factor model $x_i = \Lambda\eta_i + \varepsilon_i$. Assume:

1. **Factors**: *$\eta_i = \Psi^{1/2}U_i$, where U_i have r independent standardized entries.*

Parallel analysis selects the perceptible factors

Theorem

n iid samples from the p-dimensional factor model $x_i = \Lambda\eta_i + \varepsilon_i$. Assume:

1. **Factors***: $\eta_i = \Psi^{1/2} U_i$, where U_i have r independent standardized entries.*
2. **Idiosyncratic terms***: $\varepsilon_i = \Phi^{1/2} Z_i$, where $\Phi^{1/2}$ is a diagonal matrix, and Z_i have p indepedent standardized entries.*

Parallel analysis selects the perceptible factors

Theorem

n iid samples from the p-dimensional factor model $x_i = \Lambda\eta_i + \varepsilon_i$. Assume:

1. **Factors**: *$\eta_i = \Psi^{1/2} U_i$, where U_i have r independent standardized entries.*
2. **Idiosyncratic terms**: *$\varepsilon_i = \Phi^{1/2} Z_i$, where $\Phi^{1/2}$ is a diagonal matrix, and Z_i have p indepedent standardized entries.*
3. **Asymptotics**: *$n, p \to \infty$ st one of the following holds:*

    3.1 *$p/n \to \gamma > 0$, while $p^{-1} \sum_j \delta_{\Phi_j} \Rightarrow H$, and $\max \Phi_j \to U(H)$.*

Parallel analysis selects the perceptible factors

Theorem

n iid samples from the p-dimensional factor model $x_i = \Lambda\eta_i + \varepsilon_i$. Assume:

1. **Factors**: *$\eta_i = \Psi^{1/2} U_i$, where U_i have r independent standardized entries.*
2. **Idiosyncratic terms**: *$\varepsilon_i = \Phi^{1/2} Z_i$, where $\Phi^{1/2}$ is a diagonal matrix, and Z_i have p indepedent standardized entries.*
3. **Asymptotics**: *$n, p \to \infty$ st one of the following holds:*
   - 3.1 *$p/n \to \gamma > 0$, while $p^{-1} \sum_j \delta_{\Phi_j} \Rightarrow H$, and $\max \Phi_j \to U(H)$.*
   - 3.2 *$p/n \to \infty$, while the entries $\Phi_j \leq C \operatorname{tr}[\Phi]/p$ for all j.*

Parallel analysis selects the perceptible factors

Theorem

n iid samples from the p-dimensional factor model $x_i = \Lambda\eta_i + \varepsilon_i$. Assume:

1. **Factors**: *$\eta_i = \Psi^{1/2} U_i$, where U_i have r independent standardized entries.*
2. **Idiosyncratic terms**: *$\varepsilon_i = \Phi^{1/2} Z_i$, where $\Phi^{1/2}$ is a diagonal matrix, and Z_i have p indepedent standardized entries.*
3. **Asymptotics**: *$n, p \to \infty$ st one of the following holds:*
   - 3.1 *$p/n \to \gamma > 0$, while $p^{-1}\sum_j \delta_{\Phi_j} \Rightarrow H$, and $\max \Phi_j \to U(H)$.*
   - 3.2 *$p/n \to \infty$, while the entries $\Phi_j \le C\,\mathrm{tr}[\Phi]/p$ for all j.*
4. **Factor loadings**: *Let $\Lambda\Psi^{1/2} = [b_1, \ldots, b_r]$. Then b_k are delocalized, $|b_k|_4/|b_k|_2 \to 0$.*

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

Comments

- Shows that PA “works” when
    - dimension p is relatively large
    - the factors are “delocalized”; load on more than just a few variables
    - the strength of factors is comparable

Comments

- Shows that PA "works" when
  - dimension p is relatively large
  - the factors are "delocalized"; load on more than just a few variables
  - the strength of factors is comparable
- Proof: new bounds on operator norms of permutation random matrices
- Use moment method

Overview

Factor analysis

Parallel analysis

A theory for parallel analysis

Improving parallel analysis

Summary

Limitations of parallel analysis

- Requires random permutations

Limitations of parallel analysis

- Requires random permutations
    - *Randomness* may lead to superfluous variability
    - *Extra work* in computing SVDs (20 permutations - 20x work)

Limitations of parallel analysis

- Requires random permutations
    - *Randomness* may lead to superfluous variability
    - *Extra work* in computing SVDs (20 permutations - 20x work)
- Does not work well with both *strong and weak* factors

Limitations of parallel analysis

- Requires random permutations
    - *Randomness* may lead to superfluous variability
    - *Extra work* in computing SVDs (20 permutations - 20x work)
- Does not work well with both *strong and weak* factors
    - The noise generated by strong factors “shadows” the weak ones

Our contribution

- Develop a method that addresses these two limitations.

Do we need randomness?

- We have seen that PA selects factors above the noise operator norm $b > 0$ st

$$|\mathcal{E}|_{op} \to b$$

Do we need randomness?

- We have seen that PA selects factors above the noise operator norm $b > 0$ st
$$|\mathcal{E}|_{op} \to b$$
- Permutations are a randomized estimator

Do we need randomness?

- We have seen that PA selects factors above the noise operator norm $b > 0$ st

$$|\mathcal{E}|_{op} \to b$$

- Permutations are a randomized estimator
- Can we estimate this deterministically?

Do we need randomness?

- We have seen that PA selects factors above the noise operator norm $b > 0$ st

$$|\mathcal{E}|_{op} \to b$$

- Permutations are a randomized estimator
- Can we estimate this deterministically?
- Yes! This is the upper edge of the MP distribution, well understood

What is the Marchenko-Pastur distribution?

- Under our conditions $\varepsilon_i = \Phi^{1/2} Z_i$, the noise entries are independent with variances Φ_j

What is the Marchenko-Pastur distribution?

- Under our conditions $\varepsilon_i = \Phi^{1/2} Z_i$, the noise entries are independent with variances Φ_j
- Asymptotics: $n, p \to \infty$ st $p/n \to \gamma > 0$
- The variance distribution $H_p = p^{-1} \sum_j \delta_{\Phi_j} \Rightarrow H$

What is the Marchenko-Pastur distribution?

- Under our conditions $\varepsilon_i = \Phi^{1/2} Z_i$, the noise entries are independent with variances Φ_j
- Asymptotics: $n, p \to \infty$ st $p/n \to \gamma > 0$
- The variance distribution $H_p = p^{-1} \sum_j \delta_{\Phi_j} \Rightarrow H$
- Let λ_p be the eigenvalues of $n^{-1} \mathcal{E}^\top \mathcal{E}$

What is the Marchenko-Pastur distribution?

- Under our conditions $\varepsilon_i = \Phi^{1/2} Z_i$, the noise entries are independent with variances Φ_j
- Asymptotics: $n, p \to \infty$ st $p/n \to \gamma > 0$
- The variance distribution $H_p = p^{-1} \sum_j \delta_{\Phi_j} \Rightarrow H$
- Let λ_p be the eigenvalues of $n^{-1} \mathcal{E}^\mathsf{T} \mathcal{E}$
- The Marchenko-Pastur distribution is the (weak) limit of their distribution function $F_p \Rightarrow F_{\gamma,H}$

What is the Marchenko-Pastur distribution?

- Under our conditions $\varepsilon_i = \Phi^{1/2} Z_i$, the noise entries are independent with variances Φ_j
- Asymptotics: $n, p \to \infty$ st $p/n \to \gamma > 0$
- The variance distribution $H_p = p^{-1} \sum_j \delta_{\Phi_j} \Rightarrow H$
- Let λ_p be the eigenvalues of $n^{-1} \mathcal{E}^\top \mathcal{E}$
- The Marchenko-Pastur distribution is the (weak) limit of their distribution function $F_p \Rightarrow F_{\gamma,H}$
- Under conditions, $\max \lambda_i \to U(F_{\gamma,H})$, where $U(F) = \operatorname{supp}\sup(F)$.

How to estimate the upper edge?

- To estimate upper edge $U(F_{\gamma,H})$, enough to estimate variance distribution H

How to estimate the upper edge?

- To estimate upper edge $U(F_{\gamma,H})$, enough to estimate variance distribution H
- The Spectrode algorithm (Dobriban, 2015) computes the MP distribution $H \to F_{\gamma,H}$

How to estimate the upper edge?

- To estimate upper edge $U(F_{\gamma,H})$, enough to estimate variance distribution H
- The Spectrode algorithm (Dobriban, 2015) computes the MP distribution $H \to F_{\gamma,H}$
- To estimate H, use plug-in: $\hat{H}_p = n^{-1} \operatorname{diag}(X^\top X)$

How to estimate the upper edge?

- To estimate upper edge $U(F_{\gamma,H})$, enough to estimate variance distribution H
- The Spectrode algorithm (Dobriban, 2015) computes the MP distribution $H \to F_{\gamma,H}$
- To estimate H, use plug-in: $\hat{H}_p = n^{-1}\,\mathrm{diag}(X^\top X)$
- Accurate if empirical variances are not too affected by factors: In $X = \eta\Lambda^\top + Z\Phi^{1/2}$, factors are delocalized

How to deal with strong and weak factors?

- A strong factor is $\eta\lambda^{\mathsf{T}}$ for large $|\lambda|_2$.

How to deal with strong and weak factors?

- A strong factor is $\eta\lambda^{\mathsf{T}}$ for large $|\lambda|_2$.
- Transformed into large noise by permutation. Noise level is overestimated, weaker factors are “shadowed”

How to deal with strong and weak factors?

- A strong factor is $\eta\lambda^{\mathsf{T}}$ for large $|\lambda|_2$.
- Transformed into large noise by permutation. Noise level is overestimated, weaker factors are “shadowed”
- How to remove the strong factors?

How to deal with strong and weak factors?

- A strong factor is $\eta\lambda^\mathsf{T}$ for large $|\lambda|_2$.
- Transformed into large noise by permutation. Noise level is overestimated, weaker factors are "shadowed"
- How to remove the strong factors?
- Deflate:
  - If select top factor, set $X \leftarrow X - \sigma_1 u_1 v_1^\mathsf{T}$

How to deal with strong and weak factors?

- A strong factor is $\eta\lambda^\mathsf{T}$ for large $|\lambda|_2$.
- Transformed into large noise by permutation. Noise level is overestimated, weaker factors are "shadowed"
- How to remove the strong factors?
- Deflate:
  - If select top factor, set $X \leftarrow X - \sigma_1 u_1 v_1^\mathsf{T}$
- Works if strong factors are well estimated by empirical PCs

An algorithm: DDPA

Algorithm 1 DDPA: Deflated Deterministic Parallel Analysis

```
1: input: Data $X \in \mathbb{R}^{n \times p}$, centered, containing $p$ features of $n$ samples
2: Initialize: $k \leftarrow 0$.
3: Compute variance distribution: $\hat{H}_p \leftarrow \mathrm{diag}(n^{-1} X^\mathsf{T} X)$.
4: if $\sigma_1(n^{-1/2}X) > (1+\epsilon_p)\mathcal{U}(F_{\gamma_p,\hat{H}_p})^{1/2}$, [by default $\epsilon_p = 0$] then
5:     $k \leftarrow k+1$
6:     $X \leftarrow X - \sigma_1 u_1 v_1^\mathsf{T}$ (from the SVD of $X$)
7:     Return to step 3.
8: return: Selected number of factors $k$.
```

DDPA selects perceptible factors

Theorem

Let $x_i = \Lambda\eta_i + \varepsilon_i$, $i = 1, \ldots, n$. Assume, for some $\epsilon, \delta > 0$

1. **Factors***: $\eta_i = \Psi^{1/2} U_i$, where U_i have r independent standardized entries with bdd moment $4 + \delta$.*
2. **Idiosyncratic terms***: $\varepsilon_i = \Phi^{1/2} Z_i$, where $\Phi^{1/2}$ is a diagonal matrix, and Z_i have p indepedent standardized entries bdd moment $8 + \epsilon$.*
3. **Asymptotics***: $n, p \to \infty$, $p/n \to \gamma > 0$, $\mathrm{ESD}(\Phi) \Rightarrow H$ and $\max_{1 \le j \le p} \Phi_j \to \mathcal{U}(H)$.*

DDPA selects perceptible factors

Theorem

Let $x_i = \Lambda\eta_i + \varepsilon_i$, $i = 1, \ldots, n$. *Assume, for some* $\epsilon, \delta > 0$

1. **Factors**: $\eta_i = \Psi^{1/2} U_i$, *where* U_i *have* r *independent standardized entries with bdd moment* $4 + \delta$.
2. **Idiosyncratic terms**: $\varepsilon_i = \Phi^{1/2} Z_i$, *where* $\Phi^{1/2}$ *is a diagonal matrix, and* Z_i *have* p *indepedent standardized entries bdd moment* $8 + \epsilon$.
3. **Asymptotics**: $n, p \to \infty$, $p/n \to \gamma > 0$, $\mathrm{ESD}(\Phi) \Rightarrow H$ *and* $\max_{1 \le j \le p} \Phi_j \to \mathcal{U}(H)$.
4. **Factor loadings**: *are delocalized:* $\|d_\ell\|_\infty \to 0$. *Also delocalized wrt* Φ*:*

$$\frac{x^{\mathsf{T}}(\Phi - zI_p)^{-1} d_\ell - m_{\gamma,H}(z) \cdot x^{\mathsf{T}} d_\ell}{\|d_\ell\|} \to 0$$

uniformly for $\|x\| \le 1$, $\ell = 1, \ldots, r$, *and* $z \in \mathbb{C}$ *with* $\mathrm{Im}(z) > 0$ *fixed.*

Then wp $\to 1$, *DDPA* selects all perceptible factors, *and* no imperceptible factors.

How does it work?

- DDPA works well in simulations, but selects too many factors on empirical data

How does it work?

- DDPA works well in simulations, but selects too many factors on empirical data
- We think it is because it does not take estimation accuracy into account

How does it work?

- DDPA works well in simulations, but selects too many factors on empirical data
- We think it is because it does not take estimation accuracy into account
- To fix this, we raise threshold, generalizing Perry (2009); Gavish and Donoho (2014)
- Call this method DDPA+

How does it work now? HGDP example

- Human Genome Diversity Project (HGDP) dataset (e.g., Li et al., 2008). Goal was “to evaluate the diversity in the patterns of genetic variation across the globe.”

How does it work now? HGDP example

- Human Genome Diversity Project (HGDP) dataset (e.g., Li et al., 2008). Goal was “to evaluate the diversity in the patterns of genetic variation across the globe.”
- 51 populations from Africa, Europe, Asia, Oceania and the Americas.

How does it work now? HGDP example

- Human Genome Diversity Project (HGDP) dataset (e.g., Li et al., 2008). Goal was "to evaluate the diversity in the patterns of genetic variation across the globe."
- 51 populations from Africa, Europe, Asia, Oceania and the Americas.
- $n = 1043$ samples, $p = 9730$ SNPs on chromosome 22. $n \times p$ data matrix X, where $X_{ij} \in \{0, 1, 2\}$ is the number of copies of the minor allele of SNP j in the genome of individual i.

HGDP example

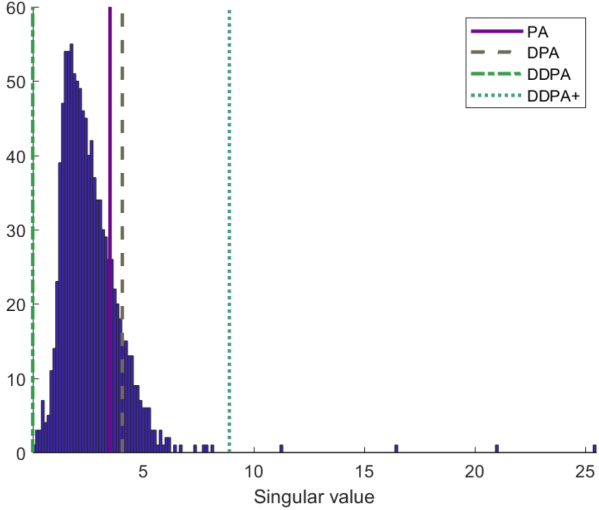


Figure : Singular value histogram of HGDP data, and the thresholds where factor selection stops. PA: 212, DPA: 122, DDPA: 1042, DDPA+: 4.

Overview

Factor analysis

Parallel analysis

A theory for parallel analysis

Improving parallel analysis

Summary

Summary

- Theory for PA
- Deterministic Deflated PA (DDPA+):
  - fast, derandomized, adapts to signal strength
- References:
  - E. Dobriban. Factor selection by permutation. arxiv.
  - E. Dobriban, A.B. Owen. Deterministic parallel analysis. arxiv.
- Talk slides: github.com/dobriban/Talks (can get there from my webpage)

T. A. Brown. *Confirmatory factor analysis for applied research*. The Guilford Press, New York, 2nd edition, 2014.

A. B. Costello and J. W. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research & evaluation*, 10(7):1–9, 2005.

E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer series in statistics, 2009.

C. J. Gaskin and B. Happell. On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International journal of nursing studies*, 51(3):511–521, 2014.

M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

J. C. Hayton, D. G. Allen, and V. Scarpello. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, 7 (2):191–205, 2004.

J. Josse and F. Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6):1869–1879, 2012.

S. Kritchman and B. Nadler. Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems*, 94(1):19–32, 2008.

J. Z. Li, D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104, 2008.

A. Onatski. Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258, 2012.

P. O. Perry. *Cross-validation for unsupervised learning*. PhD thesis, Stanford University, 2009.