Estimating Eigenvectors of the Correlation Matrix

Nick Stiles

Dec 11, 2023

Context

- N random variables $X_1, \ldots, X_N$ with mean 0, variance 1.

Context

- N random variables $X_1, ..., X_N$ with mean 0, variance 1.
- Comovements of variables depend on eigenvectors/values of the covariance matrix $\mathbf{C}$ such that $\mathbf{C}_{ij} = \mathbb{E}[X_i X_j]$.

Context

- N random variables $X_1, ..., X_N$ with mean 0, variance 1.
- Comovements of variables depend on eigenvectors/values of the covariance matrix $\mathbf{C}$ such that $\mathbf{C}_{ij} = \mathbb{E}[X_i X_j]$.
- PCA: assume most of the covariance be described by $r << N$ linear combinations.
  - Given by eigenvectors of $\mathbf{C}$.

Context

- N random variables $X_1, ..., X_N$ with mean 0, variance 1.
- Comovements of variables depend on eigenvectors/values of the covariance matrix $\mathbf{C}$ such that $\mathbf{C}_{ij} = \mathbb{E}[X_i X_j]$.
- PCA: assume most of the covariance be described by $r << N$ linear combinations.
  - Given by eigenvectors of $\mathbf{C}$.
- Various applications: finance, genomics, engineering, medicine, physics, image analysis & more.

The Problem

Issue

We don't know C.

The Problem

Issue

We don't know C.

- Still can form empirical covariance matrix. Take T samples, form matrix $\mathbf{E}$ such that $\mathbf{E}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{jt}$ for observation matrix $\mathbf{X} \in \mathbb{R}^{N\times T}$.

The Problem

Issue

We don't know C.

- Still can form empirical covariance matrix. Take T samples, form matrix $\mathbf{E}$ such that $\mathbf{E}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{jt}$ for observation matrix $\mathbf{X} \in \mathbb{R}^{N\times T}$.
- The problem: how "close" is $\mathbf{E}$ to the true covariance matrix $\mathbf{C}$?
  - Analyze "closeness" by comparing eigenvalues, eigenvectors of $\mathbf{E}$ with those of $\mathbf{C}$.

Deterministic vs Random Approach

- Various approaches to solving problem.

Deterministic vs Random Approach

- Various approaches to solving problem.
- Deterministic: consider $\mathbf{C} + \mathbf{N}$ or $\sqrt{\mathbf{C}}\mathbf{N}\sqrt{\mathbf{C}}$ for deterministic & structured "perturbation" $\mathbf{N}$.
  - Pros: works with arbitrarily "fat-tailed" distributions, stronger bounds.
  - Cons: require stronger assumptions on structure of $\mathbf{C}, \mathbf{N}$, which may be unknown/unknowable.

Deterministic vs Random Approach

- Various approaches to solving problem.
- Deterministic: consider $\mathbf{C} + \mathbf{N}$ or $\sqrt{\mathbf{C}}\mathbf{N}\sqrt{\mathbf{C}}$ for deterministic & structured "perturbation" $\mathbf{N}$.
  - Pros: works with arbitrarily "fat-tailed" distributions, stronger bounds.
  - Cons: require stronger assumptions on structure of $\mathbf{C}, \mathbf{N}$, which may be unknown/unknowable.
- Random: compute expected "overlaps" of eigenvectors of $\mathbf{C}$ and $\sqrt{\mathbf{C}}\mathbf{N}\sqrt{\mathbf{C}}$ for random matrix $\mathbf{N}$, compute distribution of eigenvalues.

Fixed N,T or Asymptotic?

- Decision 2: nonasymptotic or asymptotic?

Fixed N,T or Asymptotic?

- Decision 2: nonasymptotic or asymptotic?
- Nonasymptotic gives results for specific T, N.
  - Con: using asymptotics gives stronger results, using machinery from free probability and RMT.

Fixed N,T or Asymptotic?

- Decision 2: nonasymptotic or asymptotic?
- Nonasymptotic gives results for specific T, N.
  - Con: using asymptotics gives stronger results, using machinery from free probability and RMT.
- Asmyptotic: what is going to infinity? Only T or both T and N?

Fixed N,T or Asymptotic?

- Decision 2: nonasymptotic or asymptotic?
- Nonasymptotic gives results for specific T, N.
  - Con: using asymptotics gives stronger results, using machinery from free probability and RMT.
- Asmyptotic: what is going to infinity? Only T or both T and N?
- If only $T \to \infty$, no need for RMT. SLLN says $\mathbf{E} \to \mathbf{C}$ almost surely.

Fixed N,T or Asymptotic?

- Decision 2: nonasymptotic or asymptotic?
- Nonasymptotic gives results for specific T, N.
  - Con: using asymptotics gives stronger results, using machinery from free probability and RMT.
- Asmyptotic: what is going to infinity? Only T or both T and N?
- If only $T \to \infty$, no need for RMT. SLLN says $\mathbf{E} \to \mathbf{C}$ almost surely.
- Assumption may be unjustified. When working with large datasets, N might be on the order of T.

Fixed N,T or Asymptotic?

- Decision 2: nonasymptotic or asymptotic?
- Nonasymptotic gives results for specific T, N.
  - Con: using asymptotics gives stronger results, using machinery from free probability and RMT.
- Asmyptotic: what is going to infinity? Only T or both T and N?
- If only $T \to \infty$, no need for RMT. SLLN says $\mathbf{E} \to \mathbf{C}$ almost surely.
- Assumption may be unjustified. When working with large datasets, N might be on the order of T.
- We will let $N, T \to \infty$ such that $N/T = \Theta(1)$.

Using Free Probability

- If the X_i are Gaussian, then the distribution of **E** given **C** is known for any N, T. Due to Wishart in a 1928 paper.

Using Free Probability

- If the X_i are Gaussian, then the distribution of $\mathbf{E}$ given $\mathbf{C}$ is known for any N, T. Due to Wishart in a 1928 paper.
- We want results for more general distributions.

Using Free Probability

- If the X_i are Gaussian, then the distribution of $\mathbf{E}$ given $\mathbf{C}$ is known for any N, T. Due to Wishart in a 1928 paper.
- We want results for more general distributions.

Key Idea

Use free probability, where we represent $\mathbf{E}$ as a **free product** of deterministic $\mathbf{C}$ with a Wishart matrix $\mathbf{W}$.

Using Free Probability

- If the X_i are Gaussian, then the distribution of $\mathbf{E}$ given $\mathbf{C}$ is known for any N, T. Due to Wishart in a 1928 paper.
- We want results for more general distributions.

Key Idea

Use free probability, where we represent $\mathbf{E}$ as a **free product** of deterministic $\mathbf{C}$ with a Wishart matrix $\mathbf{W}$.

- Recall: free product of $\mathbf{A}$ and $\mathbf{B}$ is

$$\sqrt{\mathbf{A}}\mathbf{Q}\mathbf{B}\mathbf{Q}^T\sqrt{\mathbf{A}},$$

where $\mathbf{Q}$ is a random orthogonal matrix (Haar measure).

Assumptions on distribution of the X_i

- Equivalently, $\sqrt{\mathbf{A}}\mathbf{B}\sqrt{\mathbf{A}}$ if $\mathbf{B}$ has orthogonally-invariant distribution. Satisfied if $\mathbf{B}$ is a Wishart matrix.

Assumptions on distribution of the X_i

- Equivalently, $\sqrt{\mathbf{A}}\mathbf{B}\sqrt{\mathbf{A}}$ if $\mathbf{B}$ has orthogonally-invariant distribution. Satisfied if $\mathbf{B}$ is a Wishart matrix.

- Recall: Wishart matrix is $\frac{1}{T}\mathbf{G}\mathbf{G}^T$ where $\mathbf{G}$ is a $N \times T$ matrix with i.i.d. $N(0,1)$ entries.

- Assume that $\mathbf{X} = \sqrt{\mathbf{C}}\mathbf{Y}$ for some $\mathbf{Y}$ such that $\mathbb{E}[\mathbf{Y}_{it}] = \mathbb{E}[\mathbf{Y}_{it}\mathbf{Y}_{jt}] = 0$, $\mathbb{E}[\mathbf{Y}_{it}^2] = 1$, and $\mathbb{E}[\mathbf{Y}_{it}^4]$ is bounded. Then can set $\mathbf{W} = \frac{1}{T}\mathbf{Y}\mathbf{Y}^T$, so that

$$\mathbf{E} = \frac{1}{T}\mathbf{X}\mathbf{X}^T = \frac{1}{T}\sqrt{\mathbf{C}}\mathbf{Y}\mathbf{Y}^T\sqrt{\mathbf{C}} = \sqrt{\mathbf{C}}\mathbf{W}\sqrt{\mathbf{C}}.$$

"Tails not too fat".

Assumptions on distribution of the X_i

- Equivalently, $\sqrt{\mathbf{A}}\mathbf{B}\sqrt{\mathbf{A}}$ if $\mathbf{B}$ has orthogonally-invariant distribution. Satisfied if $\mathbf{B}$ is a Wishart matrix.

- Recall: Wishart matrix is $\frac{1}{T}\mathbf{G}\mathbf{G}^T$ where $\mathbf{G}$ is a $N \times T$ matrix with i.i.d. $N(0,1)$ entries.

- Assume that $\mathbf{X} = \sqrt{\mathbf{C}}\mathbf{Y}$ for some $\mathbf{Y}$ such that $\mathbb{E}[\mathbf{Y}_{it}] = \mathbb{E}[\mathbf{Y}_{it}\mathbf{Y}_{jt}] = 0$, $\mathbb{E}[\mathbf{Y}_{it}^2] = 1$, and $\mathbb{E}[\mathbf{Y}_{it}^4]$ is bounded. Then can set $\mathbf{W} = \frac{1}{T}\mathbf{Y}\mathbf{Y}^T$, so that

$$\mathbf{E} = \frac{1}{T}\mathbf{X}\mathbf{X}^T = \frac{1}{T}\sqrt{\mathbf{C}}\mathbf{Y}\mathbf{Y}^T\sqrt{\mathbf{C}} = \sqrt{\mathbf{C}}\mathbf{W}\sqrt{\mathbf{C}}.$$

"Tails not too fat".

Key Fact

As $N, T \to \infty$, $\mathbf{W} = \frac{1}{T}\mathbf{Y}\mathbf{Y}^T$ is a Wishart matrix.

Distribution of the Eigenvalues

- Fact: as $N, T \to \infty$, distribution of eigenvalues of Wishart matrix $\to$ Marcenko-Pastur distribution.

Distribution of the Eigenvalues

- Fact: as $N, T \to \infty$, distribution of eigenvalues of Wishart matrix $\to$ Marcenko-Pastur distribution.
- Generally, distribution of eigenvalues of $\mathbf{E}$ converges, but limiting distribution different from that of $\mathbf{C}$.

Distribution of the Eigenvalues

- Fact: as $N, T \to \infty$, distribution of eigenvalues of Wishart matrix $\to$ Marcenko-Pastur distribution.
- Generally, distribution of eigenvalues of $\mathbf{E}$ converges, but limiting distribution different from that of $\mathbf{C}$.
- Implies existence of systematic deviations from eigenspectrum of $\mathbf{C}$, even in limit.

Distribution of the Eigenvalues

- Fact: as $N, T \to \infty$, distribution of eigenvalues of Wishart matrix $\to$ Marcenko-Pastur distribution.
- Generally, distribution of eigenvalues of $\mathbf{E}$ converges, but limiting distribution different from that of $\mathbf{C}$.
- Implies existence of systematic deviations from eigenspectrum of $\mathbf{C}$, even in limit.
- Recall: often assume $\mathbf{C}$ dominated by some $r << N$ factors, so assume top r eigenvalues are "outliers". Can be determined by separation from Marcenko-Pastur bulk.

Defining Overlap

- Define spectral decompositions as

$$\mathbf{E} = \sum_{i=1}^{N} \mu_i \mathbf{u}_i \mathbf{u}_i^T \qquad \mathbf{C} = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$$

where $\lambda_1 > ... > \lambda_N$, $\mu_1 > ... > \mu_N$.

Defining Overlap

- Define spectral decompositions as

$$\mathbf{E} = \sum_{i=1}^{N} \mu_i \mathbf{u}_i \mathbf{u}_i^T \qquad\qquad \mathbf{C} = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$$

  where $\lambda_1 > ... > \lambda_N$, $\mu_1 > ... > \mu_N$.

- Define overlap as

$$\Phi(\mu_i, \lambda_j) = N\mathbb{E}[(\mathbf{u}_i \mathbf{v}_j^T)^2].$$

  Squared to account for ambiguity in sign.

Result 1: Bulk Eigenvectors

Bulk Sample Eigenvectors

$\Phi(\mu_i, \lambda_j) = O(1)$ for $i < r$ and and $j = 1, ..., n$. Thus, sample eigenvectors in the bulk (i.e., $\mathbf{u}_j$ for $j > r$) are "delocalized" in the population basis.

- Sample eigenvectors corresponding to smaller eigenvalues contain little or no information about population eigenvectors.

Result 2: Outlier Eigenvectors

Outlier Sample Eigenvectors

Outlier sample eigenvectors $\mathbf{u}_i$ are distributed in a "cone" around the population eigenvector $\mathbf{v}_i$, but delocalized in all other directions.

- Can compute aperture of cone from $\mathbf{E}$.

Tools for Proofs

- Free probability tools: Resolvents, S-transforms, R-transforms, Stieltjes/Cauchy transform, inverse Cauchy/Blue transform, Marcenko-Pastur.

Tools for Proofs

- Free probability tools: Resolvents, S-transforms, R-transforms, Stieltjes/Cauchy transform, inverse Cauchy/Blue transform, Marcenko-Pastur.
- Complex analysis tools: Cauchy's integral formula, Residue theorem, Cauchy's integral theorem, Sokhotski–Plemelj theorem.

Tools for Proofs

- Free probability tools: Resolvents, S-transforms, R-transforms, Stieltjes/Cauchy transform, inverse Cauchy/Blue transform, Marcenko-Pastur.
- Complex analysis tools: Cauchy's integral formula, Residue theorem, Cauchy's integral theorem, Sokhotski–Plemelj theorem.
- Linear algebra tools: Schur complement formula. Forming "spikeless" covariance matrix by clipping eigenvalues greater than λ_{d+1}.

Takeaways

- May seem discouraging - most eigenvectors are informationless.

Takeaways

- May seem discouraging - most eigenvectors are informationless.
- On the bright side: know the unknowns.

Takeaways

- May seem discouraging - most eigenvectors are informationless.
- On the bright side: know the unknowns.
- Also, makes our estimation easier: "best we can do" in estimating eigenbasis of $\mathbf{C}$ is use that of $\mathbf{E}$.