High-dimensional principal component analysis with heterogeneous missingness

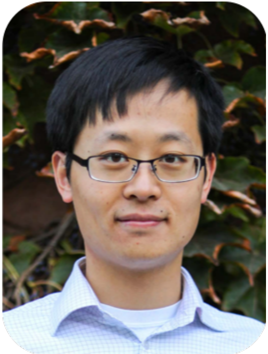
Tengyao Wang

University College London

Statistics Seminar, University of Kent

24 Oct 2019

Collaborators



Ziwei Zhu

Richard Samworth

Missing data

Missing data are ever more important in the Big Data era, because complete-case analysis is less feasible.

Consider a complete-case analysis with an $n \times d$ matrix, where each entry is observed independently with probability $p = 0.99$.

Missing data

Missing data are ever more important in the Big Data era, because complete-case analysis is less feasible.

Consider a complete-case analysis with an $n \times d$ matrix, where each entry is observed independently with probability $p = 0.99$.

- When $d = 5$, around 95% of observations are retained

Missing data

Missing data are ever more important in the Big Data era, because complete-case analysis is less feasible.

Consider a complete-case analysis with an $n \times d$ matrix, where each entry is observed independently with probability $p = 0.99$.

- When $d = 5$, around 95% of observations are retained
- When $d = 300$, only around 5% of observations are retained.

Missing data

Approaches to handle missing data:

- Imputation (Ford, 1983; Rubin, 2004)
- Factored likelihood (Anderson, 1957)
- Expectation-Maximisation (Dempster et al., 1977)

Missing data

Approaches to handle missing data:

- Imputation (Ford, 1983; Rubin, 2004)
- Factored likelihood (Anderson, 1957)
- Expectation-Maximisation (Dempster et al., 1977)

Recently, there has been increased emphasis on missing data in high-dimensional problems:

- Sparse regression (Loh and Wainwright, 2012; Belloni et al., 2017)
- Classification (Cai and Zhang, 2018b)
- Covariance and precision matrix estimation (Lounici, 2014; Loh and Tan, 2018)

High-dimensional PCA problem set-up

Suppose the (partially observed) matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ is of the form

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{V}_K^\top + \boldsymbol{Z},$$

where $\boldsymbol{V}_K \in \mathbb{R}^{d\times K}$ has orthonormal columns and $\boldsymbol{U}$ is a random $n \times K$ matrix (with $n > K$) having i.i.d. rows with mean zero.

High-dimensional PCA problem set-up

Suppose the (partially observed) matrix $\boldsymbol{Y} \in \mathbb{R}^{n \times d}$ is of the form

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{V}_K^\top + \boldsymbol{Z},$$

where $\boldsymbol{V}_K \in \mathbb{R}^{d \times K}$ has orthonormal columns and $\boldsymbol{U}$ is a random $n \times K$ matrix (with $n > K$) having i.i.d. rows with mean zero.

Let $\Omega_{ij} := \{Y_{ij}$ is observed$\}$ and $\boldsymbol{Y}_{\boldsymbol{\Omega}} := \boldsymbol{Y} \circ \boldsymbol{\Omega}$. We observe the pair $(\boldsymbol{Y}_{\boldsymbol{\Omega}}, \boldsymbol{\Omega})$ and wish to estimate $\mathrm{Col}(\boldsymbol{V}_K)$.

High-dimensional PCA problem set-up

Suppose the (partially observed) matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ is of the form

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{V}_K^\top + \boldsymbol{Z},$$

where $\boldsymbol{V}_K \in \mathbb{R}^{d\times K}$ has orthonormal columns and $\boldsymbol{U}$ is a random $n \times K$ matrix (with $n > K$) having i.i.d. rows with mean zero.

Let $\Omega_{ij} := \{Y_{ij} \text{ is observed}\}$ and $\boldsymbol{Y}_{\boldsymbol{\Omega}} := \boldsymbol{Y} \circ \boldsymbol{\Omega}$. We observe the pair $(\boldsymbol{Y}_{\boldsymbol{\Omega}}, \boldsymbol{\Omega})$ and wish to estimate $\mathrm{Col}(\boldsymbol{V}_K)$.

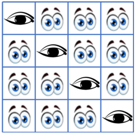
Performance of an estimator $\widehat{\boldsymbol{V}}_K$ measured by the loss function

$$L(\widehat{\boldsymbol{V}}_K, \boldsymbol{V}_K) := \|\sin\Theta(\widehat{\boldsymbol{V}}_K, \boldsymbol{V}_K)\|_{\mathrm{F}},$$

where $\Theta(\boldsymbol{U}, \boldsymbol{V})$ is the matrix of principal angles between $\mathrm{Col}(\boldsymbol{U})$ and $\mathrm{Col}(\boldsymbol{V})$.

Inverse-Probability Weighted estimator

Consider the p-*homogeneous setting*, where $\Omega_{ij} \overset{\mathrm{iid}}{\sim} \mathrm{Bern}(p)$. Then



$$\boldsymbol{P} := \mathbb{E}(\boldsymbol{\omega}_1 \boldsymbol{\omega}_1^\top) = p^2 \{\mathbf{1}_d \mathbf{1}_d^\top - (1 - p^{-1}) \boldsymbol{I}_d\}.$$

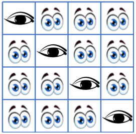
Its elementwise inverse is $\boldsymbol{W} := p^{-2}\{\mathbf{1}_d \mathbf{1}_d^\top - (1-p)\boldsymbol{I}_d\}$, and we can define the weighted sample covariance matrix

$$\boldsymbol{G} := \left(\frac{1}{n} \boldsymbol{Y}_{\boldsymbol{\Omega}}^\top \boldsymbol{Y}_{\boldsymbol{\Omega}}\right) \circ \boldsymbol{W}.$$

This ensures that $\mathbb{E}(\boldsymbol{G} \mid \boldsymbol{Y}) = n^{-1} \boldsymbol{Y}^\top \boldsymbol{Y}$.

Inverse-Probability Weighted estimator

Consider the p-*homogeneous setting*, where $\Omega_{ij} \overset{\text{iid}}{\sim} \mathrm{Bern}(p)$. Then



$$\boldsymbol{P} := \mathbb{E}(\boldsymbol{\omega}_1 \boldsymbol{\omega}_1^\top) = p^2 \{ \mathbf{1}_d \mathbf{1}_d^\top - (1 - p^{-1}) \boldsymbol{I}_d \}.$$

Its elementwise inverse is $\boldsymbol{W} := p^{-2} \{ \mathbf{1}_d \mathbf{1}_d^\top - (1 - p) \boldsymbol{I}_d \}$, and we can define the weighted sample covariance matrix

$$\boldsymbol{G} := \left(\frac{1}{n} \boldsymbol{Y}_{\boldsymbol{\Omega}}^\top \boldsymbol{Y}_{\boldsymbol{\Omega}} \right) \circ \boldsymbol{W}.$$

This ensures that $\mathbb{E}(\boldsymbol{G} \mid \boldsymbol{Y}) = n^{-1} \boldsymbol{Y}^\top \boldsymbol{Y}$. Define $\widehat{\boldsymbol{W}}$ by replacing p in $\boldsymbol{W}$ with $\widehat{p} := (nd)^{-1} \|\boldsymbol{\Omega}\|_1$, and set

$$\widehat{\boldsymbol{G}} := \left(\frac{1}{n} \boldsymbol{Y}_{\boldsymbol{\Omega}}^\top \boldsymbol{Y}_{\boldsymbol{\Omega}} \right) \circ \widehat{\boldsymbol{W}}.$$

The IPW estimator of $\boldsymbol{V}_K$ is given by the top K eigenvectors of $\widehat{\boldsymbol{G}}$, denoted $\widehat{\boldsymbol{V}}_K$ (Cai and Zhang, 2018a; Cho, Kim and Rohe, 2017).

Assumptions

For $r \in \mathbb{N}$ and a d-dimensional random vector $\boldsymbol{x}$, define its Orlicz norm

$$\|\boldsymbol{x}\|_{\psi_r} := \sup_{\boldsymbol{u} \in \mathcal{S}^{d-1}} \sup_{q \in \mathbb{N}} \frac{(\mathbb{E}|\boldsymbol{u}^\top \boldsymbol{x}|^q)^{1/q}}{q^{1/r}}$$

and a version that is invariant to invertible affine transformations:

$$\|\boldsymbol{x}\|_{\psi_r^*} := \sup_{\boldsymbol{u} \in \mathcal{S}^{d-1}} \frac{\|\boldsymbol{u}^\top(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})\|_{\psi_r}}{\mathrm{Var}^{1/2}(\boldsymbol{u}^\top \boldsymbol{x})}.$$

Assumptions

For $r \in \mathbb{N}$ and a d-dimensional random vector $\boldsymbol{x}$, define its Orlicz norm

$$\|\boldsymbol{x}\|_{\psi_r} := \sup_{\boldsymbol{u} \in \mathcal{S}^{d-1}} \sup_{q \in \mathbb{N}} \frac{(\mathbb{E}|\boldsymbol{u}^\top \boldsymbol{x}|^q)^{1/q}}{q^{1/r}}$$

and a version that is invariant to invertible affine transformations:

$$\|\boldsymbol{x}\|_{\psi_r^*} := \sup_{\boldsymbol{u} \in \mathcal{S}^{d-1}} \frac{\|\boldsymbol{u}^\top (\boldsymbol{x} - \mathbb{E}\boldsymbol{x})\|_{\psi_r}}{\mathrm{Var}^{1/2}(\boldsymbol{u}^\top \boldsymbol{x})}.$$

(A1) $\boldsymbol{U}$, $\boldsymbol{Z}$ and $\boldsymbol{\Omega}$ are independent;

(A2) $\|\boldsymbol{u}_1\|_{\psi_2^*} \leq \tau$;

(A3) $\boldsymbol{Z} = (z_{ij})_{i \in [n], j \in [d]}$ has i.i.d. entries with $\mathbb{E} z_{11} = 0$, $\mathrm{Var}\, z_{11} = 1$ and $\|z_{11}\|_{\psi_2^*} \leq \tau$;

(A4) $\|y_{1j}^2\|_{\psi_1} \leq M$ for all $j \in [d]$.

Upper bound

Theorem. Assume (A1)–(A4) and that $n, d \geq 2$, $dp \geq 1$. Let λ_j denote the jth largest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{u}}$. If $n \geq d \log^2 d \log^2 n/(\lambda_1 p + \log d)$, then

$$\mathbb{E}L(\widehat{\boldsymbol{V}}_K, \boldsymbol{V}_K) \lesssim_{M,\tau} \frac{1}{\lambda_K}\left(\frac{Kd(\lambda_1 p + \log d)\log^2 d}{np^2}\right)^{1/2}.$$

Upper bound

Theorem. Assume (A1)–(A4) and that $n, d \geq 2$, $dp \geq 1$. Let λ_j denote the jth largest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{u}}$. If $n \geq d \log^2 d \log^2 n/(\lambda_1 p + \log d)$, then

$$\mathbb{E}L(\widehat{\boldsymbol{V}}_K, \boldsymbol{V}_K) \lesssim_{M,\tau} \frac{1}{\lambda_K}\biggl(\frac{Kd(\lambda_1 p + \log d)\log^2 d}{np^2}\biggr)^{1/2}.$$

The sample size requirement is reasonable: with no missing data and when $\lambda_1 \gg 1$, the top eigenvector of the sample covariance matrix estimator is consistent if and only if $d/(n\lambda_1) \to 0$ (Shen et al., 2016).

Upper bound

Theorem. Assume (A1)–(A4) and that $n, d \geq 2$, $dp \geq 1$. Let λ_j denote the jth largest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{u}}$. If $n \geq d \log^2 d \log^2 n / (\lambda_1 p + \log d)$, then

$$\mathbb{E} L(\widehat{\boldsymbol{V}}_K, \boldsymbol{V}_K) \lesssim_{M,\tau} \frac{1}{\lambda_K} \left(\frac{K d (\lambda_1 p + \log d) \log^2 d}{np^2} \right)^{1/2}.$$

The sample size requirement is reasonable: with no missing data and when $\lambda_1 \gg 1$, the top eigenvector of the sample covariance matrix estimator is consistent if and only if $d/(n\lambda_1) \to 0$ (Shen et al., 2016).

The theorem reveals a phase transition depending on the relative magnitudes of $\lambda_1 p$ and $\log d$. In particular,

$$\mathbb{E} L(\widehat{\boldsymbol{V}}_K, \boldsymbol{V}_K) \lesssim_{M,\tau} \begin{cases} \frac{1}{\lambda_K} \left(\frac{K d \log^3 d}{np^2} \right)^{1/2} & \text{if } \lambda_1 p \lesssim \log d, \\ \frac{\lambda_1^{1/2}}{\lambda_K} \left(\frac{K d \log^2 d}{np} \right)^{1/2} & \text{if } \lambda_1 p \gtrsim \log d. \end{cases}$$

Minimax lower bound

Let $\mathcal{P}_{n,d}(\lambda_1, p)$ denote the class of distributions of pairs $(\boldsymbol{Y}_{\boldsymbol{\Omega}}, \boldsymbol{\Omega})$ satisfying (A1), (A2), (A3) with $K = 1$. Since we are now working with vectors instead of matrices, we write $\boldsymbol{v}$ in place of $\boldsymbol{V}_1$.

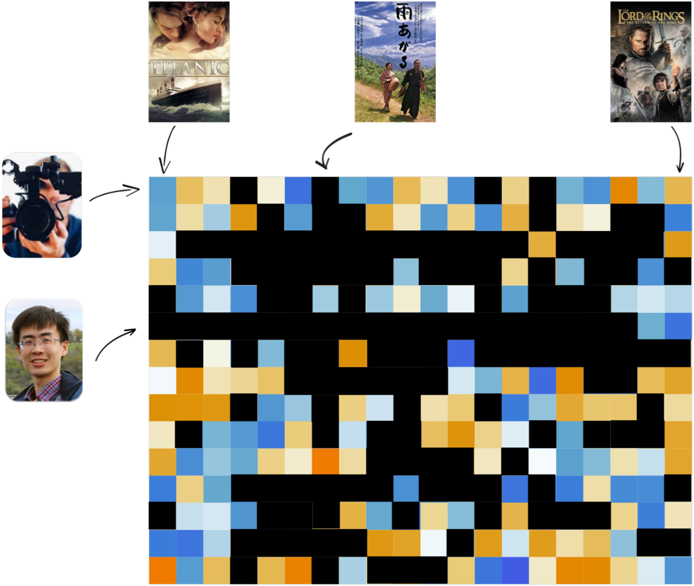
Theorem. There exists a universal constant $c > 0$ such that

$$\inf_{\widehat{\boldsymbol{v}}} \sup_{P \in \mathcal{P}_{n,d}(\lambda_1, p)} \mathbb{E}_P L(\widehat{\boldsymbol{v}}, \boldsymbol{v}) \geq c \min\biggl\{ \frac{1}{\lambda_1} \biggl(\frac{d(\lambda_1 p + 1)}{np^2} \biggr)^{1/2}, 1 \biggr\},$$

where the infimum is taken over all estimators $\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{v}}(\boldsymbol{Y}_{\boldsymbol{\Omega}}, \boldsymbol{\Omega})$ of $\boldsymbol{v}$.

Thus $\widehat{\boldsymbol{V}}_1$ achieves the minimax optimal rate of estimation up to a poly-logarithmic factor when M and τ are regarded as constants and $K = 1$.

General observation mechanisms



General observation mechanisms: example

Suppose that

$$\mathbb{P}\bigl\{\boldsymbol{\omega}_1 = (1,0,1,\ldots,1)^\top\bigr\} = \mathbb{P}\bigl\{\boldsymbol{\omega}_1 = (0,1,1,\ldots,1)^\top\bigr\} = 1/2.$$

Consider $\boldsymbol{\Sigma} = \boldsymbol{I}_d + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$, where $\boldsymbol{\alpha} = (2^{-1/2}, 2^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^d$, and $\boldsymbol{\Sigma}' = \boldsymbol{I}_d + \boldsymbol{\alpha}'(\boldsymbol{\alpha}')^\top$, where $\boldsymbol{\alpha}' = (2^{-1/2}, -2^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^d$.

Suppose that $\boldsymbol{y} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{y}' \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}')$. Then $(\boldsymbol{y} \circ \boldsymbol{\omega}, \boldsymbol{\omega})$ and $(\boldsymbol{y}' \circ \boldsymbol{\omega}, \boldsymbol{\omega})$ are identically distributed.

General observation mechanisms: example

Suppose that

$$\mathbb{P}\big\{\boldsymbol{\omega}_1 = (1,0,1,\ldots,1)^\top\big\} = \mathbb{P}\big\{\boldsymbol{\omega}_1 = (0,1,1,\ldots,1)^\top\big\} = 1/2.$$

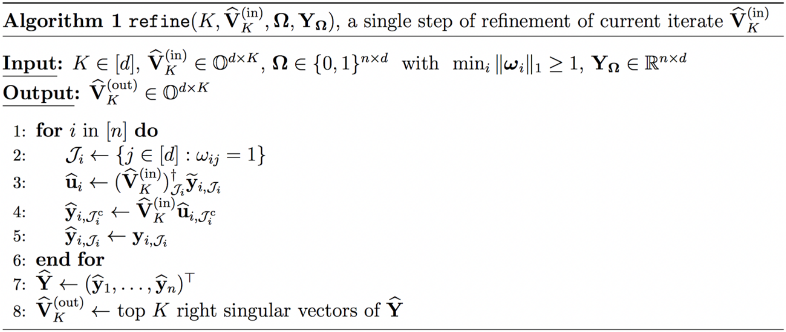
Consider $\boldsymbol{\Sigma} = \boldsymbol{I}_d + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top$, where $\boldsymbol{\alpha} = (2^{-1/2}, 2^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^d$, and $\boldsymbol{\Sigma}' = \boldsymbol{I}_d + \boldsymbol{\alpha}'(\boldsymbol{\alpha}')^\top$, where $\boldsymbol{\alpha}' = (2^{-1/2}, -2^{-1/2}, 0, \ldots, 0)^\top \in \mathbb{R}^d$.

Suppose that $\boldsymbol{y} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{y}' \sim N_d(\mathbf{0}, \boldsymbol{\Sigma}')$. Then $(\boldsymbol{y} \circ \boldsymbol{\omega}, \boldsymbol{\omega})$ and $(\boldsymbol{y}' \circ \boldsymbol{\omega}, \boldsymbol{\omega})$ are identically distributed.

But the respective leading eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ are $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, which are orthogonal!

`primePCA`: a single iteration of refinement

`primePCA` (projected refinement for imputation of missing entries in PCA) iteratively refines a warm initialiser. We write $\widetilde{\boldsymbol{y}}_i := \boldsymbol{y}_i \circ \boldsymbol{\omega}_i$.



Algorithm 1 `refine`$(K, \widehat{\mathbf{V}}_K^{(\mathrm{in})}, \mathbf{\Omega}, \mathbf{Y}_{\mathbf{\Omega}})$, a single step of refinement of current iterate $\widehat{\mathbf{V}}_K^{(\mathrm{in})}$

Input: $K \in [d]$, $\widehat{\mathbf{V}}_K^{(\mathrm{in})} \in \mathbb{O}^{d \times K}$, $\mathbf{\Omega} \in \{0,1\}^{n \times d}$ with $\min_i \|\boldsymbol{\omega}_i\|_1 \geq 1$, $\mathbf{Y}_{\mathbf{\Omega}} \in \mathbb{R}^{n \times d}$

Output: $\widehat{\mathbf{V}}_K^{(\mathrm{out})} \in \mathbb{O}^{d \times K}$

```
for $i$ in $[n]$ do
    $\mathcal{J}_i \leftarrow \{j \in [d] : \omega_{ij} = 1\}$
    $\widehat{\mathbf{u}}_i \leftarrow (\widehat{\mathbf{V}}_K^{(\mathrm{in})})^{\dagger}_{\mathcal{J}_i} \widetilde{\mathbf{y}}_{i,\mathcal{J}_i}$
    $\widehat{\mathbf{y}}_{i,\mathcal{J}_i^{\mathrm{c}}} \leftarrow \widehat{\mathbf{V}}_K^{(\mathrm{in})} \widehat{\mathbf{u}}_{i,\mathcal{J}_i^{\mathrm{c}}}$
    $\widehat{\mathbf{y}}_{i,\mathcal{J}_i} \leftarrow \mathbf{y}_{i,\mathcal{J}_i}$
end for
$\widehat{\mathbf{Y}} \leftarrow (\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_n)^\top$
$\widehat{\mathbf{V}}_K^{(\mathrm{out})} \leftarrow$ top $K$ right singular vectors of $\widehat{\mathbf{Y}}$
```

Two-to-infinity subspace distance

For $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{O}^{d \times K}$, let $\boldsymbol{W}_1 \boldsymbol{D}_{\boldsymbol{U},\boldsymbol{V}} \boldsymbol{W}_2^\top$ be an SVD of $\boldsymbol{V}^\top \boldsymbol{U}$ and let $\boldsymbol{W}_{\boldsymbol{U},\boldsymbol{V}} := \boldsymbol{W}_1 \boldsymbol{W}_2^\top$. Then $\boldsymbol{W}_{\boldsymbol{U},\boldsymbol{V}}$ solves the Procrustes problem in the sense that

$$\boldsymbol{W}_{\boldsymbol{U},\boldsymbol{V}} \in \operatorname*{arg\,min}_{\boldsymbol{W} \in \mathbb{O}^{K \times K}} \|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}\|_{\mathrm{F}}.$$

The two-to-infinity distance between $\mathrm{Col}(\boldsymbol{U})$ and $\mathrm{Col}(\boldsymbol{V})$ is then defined to be

$$\mathcal{T}(\boldsymbol{U}, \boldsymbol{V}) := \left\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{W}_{\boldsymbol{U},\boldsymbol{V}}\right\|_{2\to\infty},$$

where $\|\boldsymbol{A}\|_{2\to\infty} := \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_\infty$.

Contraction of each iteration

Proposition. Let $\widehat{\boldsymbol{V}}_K^{(\mathrm{out})} := \texttt{refine}\big(K, \widehat{\boldsymbol{V}}_K^{(\mathrm{in})}, \boldsymbol{\Omega}, \boldsymbol{Y}_{\boldsymbol{\Omega}}\big)$. We assume that $\min_{i\in[n]} \|\boldsymbol{\omega}_i\|_1 > K$, that $\min_{i\in[n]} \frac{d^{1/2}\sigma_K((\widehat{\boldsymbol{V}}_K^{(\mathrm{in})})_{\mathcal{J}_i})}{|\mathcal{J}_i|^{1/2}} \geq 1/\sigma_* > 0$, and write the SVD of $\boldsymbol{Y}$ as $\boldsymbol{L}\boldsymbol{\Gamma}\boldsymbol{R}^\top$.

Suppose that $\boldsymbol{Z} = \boldsymbol{0}$, and that both $\|\boldsymbol{L}\|_{2\to\infty} \leq \mu_1(K/n)^{1/2}$ and $\|\boldsymbol{R}\|_{2\to\infty} \leq \mu_2(K/d)^{1/2}$ hold for some $\mu_1, \mu_2 \geq 1$. Then there exist $c_1, C > 0$, depending only on μ_1, μ_2 and σ_*, such that whenever

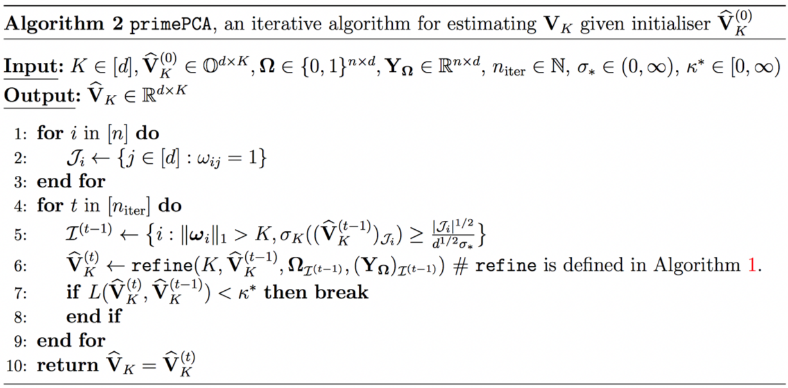
(i) $\mathcal{T}(\widehat{\boldsymbol{V}}_K^{(\mathrm{in})}, \boldsymbol{V}_K) \leq \frac{c_1\sigma_K(\boldsymbol{\Gamma})}{K^2\sigma_1(\boldsymbol{\Gamma})\sqrt{d}}$,

(ii) $\rho := \frac{CK^2\sigma_1(\boldsymbol{\Gamma})\|\boldsymbol{\Omega}^{\mathrm{c}}\|_{1\to1}}{\sigma_K(\boldsymbol{\Gamma})n} < 1$,

we have that

$$\mathcal{T}(\widehat{\boldsymbol{V}}_K^{(\mathrm{out})}, \boldsymbol{V}_K) \leq \rho\mathcal{T}(\widehat{\boldsymbol{V}}_K^{(\mathrm{in})}, \boldsymbol{V}_K).$$

`primePCA`



Algorithm 2 `primePCA`, an iterative algorithm for estimating $\mathbf{V}_K$ given initialiser $\widehat{\mathbf{V}}_K^{(0)}$

Input: $K \in [d], \widehat{\mathbf{V}}_K^{(0)} \in \mathbb{O}^{d \times K}, \mathbf{\Omega} \in \{0,1\}^{n \times d}, \mathbf{Y}_{\mathbf{\Omega}} \in \mathbb{R}^{n \times d}, n_{\text{iter}} \in \mathbb{N}, \sigma_* \in (0, \infty), \kappa^* \in [0, \infty)$

Output: $\widehat{\mathbf{V}}_K \in \mathbb{R}^{d \times K}$

```
for $i$ in $[n]$ do
    $\mathcal{J}_i \leftarrow \{j \in [d] : \omega_{ij} = 1\}$
end for
for $t$ in $[n_{\text{iter}}]$ do
    $\mathcal{I}^{(t-1)} \leftarrow \{i : \|\boldsymbol{\omega}_i\|_1 > K, \sigma_K((\widehat{\mathbf{V}}_K^{(t-1)})_{\mathcal{J}_i}) \geq \frac{|\mathcal{J}_i|^{1/2}}{d^{1/2}\sigma_*}\}$
    $\widehat{\mathbf{V}}_K^{(t)} \leftarrow \texttt{refine}(K, \widehat{\mathbf{V}}_K^{(t-1)}, \mathbf{\Omega}_{\mathcal{I}^{(t-1)}}, (\mathbf{Y}_{\mathbf{\Omega}})_{\mathcal{I}^{(t-1)}})$ # refine is defined in Algorithm 1.
    if $L(\widehat{\mathbf{V}}_K^{(t)}, \widehat{\mathbf{V}}_K^{(t-1)}) < \kappa^*$ then break
    end if
end for
return $\widehat{\mathbf{V}}_K = \widehat{\mathbf{V}}_K^{(t)}$
```

Geometric convergence in noiseless case

Theorem. For $t \in [n_{\text{iter}}]$, let $\widehat{\boldsymbol{V}}_K^{(t)}$ be the t^{th} iterate of Algorithm 2 with input $K, \widehat{\boldsymbol{V}}_K^{(0)}, \boldsymbol{\Omega} \in \{0,1\}^{n \times d}, \boldsymbol{Y}_{\boldsymbol{\Omega}} \in \mathbb{R}^{n \times d}, n_{\text{iter}} \in \mathbb{N}, \sigma_* \in (0, \infty)$ and $\kappa^* = 0$. Let

$$\mathcal{I} := \Big\{ i : \|\boldsymbol{\omega}_i\|_1 > K, \sigma_K\big((\boldsymbol{V}_K)_{\mathcal{J}_i}\big) \geq |\mathcal{J}_i|^{1/2} / \big(d^{1/2}\sigma_*\big) \Big\},$$

where $\mathcal{J}_i := \{j : \omega_{ij} = 1\}$. Let $\boldsymbol{Y}_{\mathcal{I}} = \boldsymbol{L}\boldsymbol{\Gamma}\boldsymbol{R}^\top$ be an SVD of $\boldsymbol{Y}_{\mathcal{I}}$. Suppose that both $\|\boldsymbol{L}\|_{2 \to \infty} \leq \mu_1 (K/|\mathcal{I}|)^{1/2}$ and $\|\boldsymbol{R}\|_{2 \to \infty} \leq \mu_2 (K/d)^{1/2}$. Let

$$\mathcal{Z} := \Big\{ \sigma_K\big((\boldsymbol{V}_K)_{\mathcal{J}_i}\big) d^{1/2} / |\mathcal{J}_i|^{1/2} : i \in [n], \|\boldsymbol{\omega}_i\|_1 > K \Big\},$$

and assume that $\epsilon := \min_{z \in \mathcal{Z}} |z - \sigma_*^{-1}| > 0$. Then there exist $c_1, C > 0$, depending only on μ_1, μ_2, σ_* and ϵ, such that whenever

$$\mathcal{T}(\widehat{\boldsymbol{V}}_K^{(0)}, \boldsymbol{V}_K) \leq \frac{c_1 \sigma_K(\boldsymbol{Y}_{\mathcal{I}})}{K^2 \sigma_1(\boldsymbol{Y}_{\mathcal{I}})\sqrt{d}} \quad \text{and} \quad \rho := \frac{C K^2 \sigma_1(\boldsymbol{Y}_{\mathcal{I}}) \|\boldsymbol{\Omega}_{\mathcal{I}}^{\mathrm{c}}\|_{1 \to 1}}{\sigma_K(\boldsymbol{Y}_{\mathcal{I}})|\mathcal{I}|} < 1,$$

we have $\mathcal{T}(\widehat{\boldsymbol{V}}_K^{(t)}, \boldsymbol{V}_K) \leq \rho^t \mathcal{T}(\widehat{\boldsymbol{V}}_K^{(0)}, \boldsymbol{V}_K)$ for every $t \in [n_{\text{iter}}]$.

Initialisation

Consider the following modified weighted sample covariance matrix

$$\widetilde{\boldsymbol{G}} := \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{y}}_i \widetilde{\boldsymbol{y}}_i^\top \circ \widetilde{\boldsymbol{W}},$$

where for any $j, k \in [d]$,

$$\widetilde{\boldsymbol{W}}_{jk} := \begin{cases} \frac{n}{\sum_{i=1}^{n} \omega_{ij}\omega_{ik}} & \text{if } \sum_{i=1}^{n} \omega_{ij}\omega_{ik} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We take as our initial estimator of $\boldsymbol{V}_K$ the matrix of top K eigenvectors of $\widetilde{\boldsymbol{G}}$, denoted $\widetilde{\boldsymbol{V}}_K$.

Performance of initialiser

Proposition. Assume the same conditions as in the previous theorem. Then there exists a universal constant $C > 0$ such that for any $\xi > 1$, if

$$\lambda_K > C\biggl\{\biggl(\frac{M\tau^2 R\|\widetilde{\boldsymbol{W}}\|_{1\to 1}\xi\log d}{n}\biggr)^{1/2} + \frac{M\|\widetilde{\boldsymbol{W}}\|_{\mathrm{op}}\xi\log^2 d}{n}\biggr\},$$

then with $\mathbb{P}^{\boldsymbol{\Omega}}$-probability at least $1-(2K+4)d^{-(\xi-1)}$, we have

$$L(\widetilde{\boldsymbol{V}}_K, \boldsymbol{V}_K) \leq \frac{2^{9/2}e\tau\mu}{\lambda_K}\biggl(\frac{KMR}{d}\biggr)^{1/2}\biggl(\frac{\xi^{1/2}\|\widetilde{\boldsymbol{W}}\|_1^{1/2}\log^{1/2} d}{n^{1/2}} + \frac{\xi\|\widetilde{\boldsymbol{W}}\|_{\mathrm{F}}\log d}{n}\biggr).$$

Performance of initialiser

Proposition. Assume the same conditions as in the previous theorem. Then there exists a universal constant $C > 0$ such that for any $\xi > 1$, if

$$\lambda_K > C\biggl\{\biggl(\frac{M\tau^2 R\|\widetilde{\boldsymbol{W}}\|_{1\to 1}\xi \log d}{n}\biggr)^{1/2} + \frac{M\|\widetilde{\boldsymbol{W}}\|_{\mathrm{op}}\xi \log^2 d}{n}\biggr\},$$

then with $\mathbb{P}^{\boldsymbol{\Omega}}$-probability at least $1 - (2K+4)d^{-(\xi-1)}$, we have

$$L(\widetilde{\boldsymbol{V}}_K, \boldsymbol{V}_K) \leq \frac{2^{9/2}e\tau\mu}{\lambda_K}\biggl(\frac{KMR}{d}\biggr)^{1/2}\biggl(\frac{\xi^{1/2}\|\widetilde{\boldsymbol{W}}\|_1^{1/2}\log^{1/2} d}{n^{1/2}} + \frac{\xi\|\widetilde{\boldsymbol{W}}\|_{\mathrm{F}}\log d}{n}\biggr).$$

N.B. The bound depends on $\widetilde{\boldsymbol{W}}$ only through the *entrywise* ℓ_1 and ℓ_2 norms of the whole matrix.

Simulations: Noiseless case

Fix $n = 2000$, $d = 500$, $K = 2$ and $\boldsymbol{u}_i \sim N_d(0, \boldsymbol{\Sigma}_u)$ where $\boldsymbol{\Sigma}_{\boldsymbol{u}} = 100\boldsymbol{I}_2$. Set

$$\boldsymbol{V}_K = \sqrt{\frac{1}{500}}\begin{pmatrix} \mathbf{1}_{250} & \mathbf{1}_{250} \\ \mathbf{1}_{250} & -\mathbf{1}_{250} \end{pmatrix} \in \mathbb{R}^{500\times 2}.$$

Simulations: Noiseless case

Fix $n = 2000$, $d = 500$, $K = 2$ and $\boldsymbol{u}_i \sim N_d(0, \boldsymbol{\Sigma}_u)$ where $\boldsymbol{\Sigma}_{\boldsymbol{u}} = 100\boldsymbol{I}_2$. Set

$$\boldsymbol{V}_K = \sqrt{\frac{1}{500}} \begin{pmatrix} \mathbf{1}_{250} & \mathbf{1}_{250} \\ \mathbf{1}_{250} & -\mathbf{1}_{250} \end{pmatrix} \in \mathbb{R}^{500\times 2}.$$

(H1) Homogeneous: $\mathbb{P}(\omega_{ij} = 1) = 0.05$ for all $i \in [n], j \in [d]$;

(H2) Mildly heterogeneous: $\mathbb{P}(\omega_{ij} = 1) = P_i Q_j$ for $i \in [n], j \in [d]$, where $P_1, \ldots, P_n \overset{\text{iid}}{\sim} U[0, 0.2]$ and $Q_1, \ldots, Q_d \overset{\text{iid}}{\sim} U[0.05, 0.95]$ independently;

(H3) Highly heterogeneous columns: $\mathbb{P}(\omega_{ij} = 1) = 0.19$ for $i \in [n]$ and all odd $j \in [d]$ and $\mathbb{P}(\omega_{ij} = 1) = 0.01$ for $i \in [n]$ and all even $j \in [d]$.

(H4) Highly heterogeneous rows: $\mathbb{P}(\omega_{ij} = 1) = 0.18$ for $j \in [d]$ and all odd $i \in [n]$ and $\mathbb{P}(\omega_{ij} = 1) = 0.02$ for $j \in [d]$ and all even $i \in [n]$.

Simulations: Noiseless case

Fix $n = 2000$, $d = 500$, $K = 2$ and $\boldsymbol{u}_i \sim N_d(0, \boldsymbol{\Sigma}_u)$ where $\boldsymbol{\Sigma}_{\boldsymbol{u}} = 100\boldsymbol{I}_2$. Set

$$\boldsymbol{V}_K = \sqrt{\frac{1}{500}} \begin{pmatrix} \mathbf{1}_{250} & \mathbf{1}_{250} \\ \mathbf{1}_{250} & -\mathbf{1}_{250} \end{pmatrix} \in \mathbb{R}^{500 \times 2}.$$

(H1) Homogeneous: $\mathbb{P}(\omega_{ij} = 1) = 0.05$ for all $i \in [n], j \in [d]$;

(H2) Mildly heterogeneous: $\mathbb{P}(\omega_{ij} = 1) = P_i Q_j$ for $i \in [n], j \in [d]$, where $P_1, \ldots, P_n \overset{\text{iid}}{\sim} U[0, 0.2]$ and $Q_1, \ldots, Q_d \overset{\text{iid}}{\sim} U[0.05, 0.95]$ independently;

(H3) Highly heterogeneous columns: $\mathbb{P}(\omega_{ij} = 1) = 0.19$ for $i \in [n]$ and all odd $j \in [d]$ and $\mathbb{P}(\omega_{ij} = 1) = 0.01$ for $i \in [n]$ and all even $j \in [d]$.

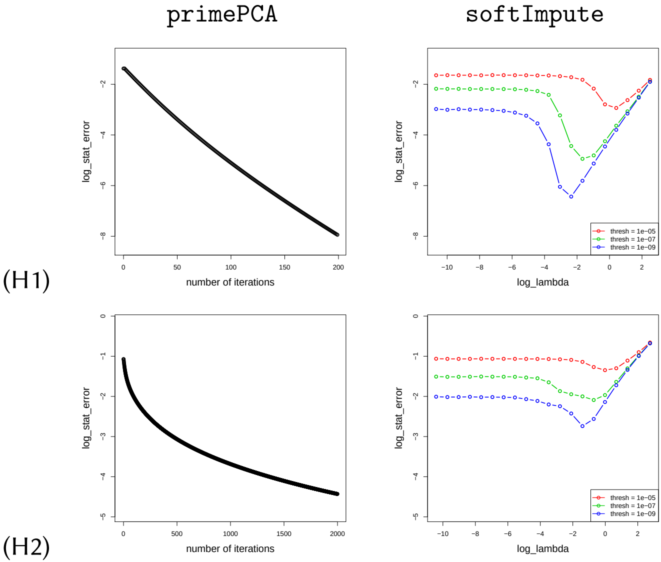
(H4) Highly heterogeneous rows: $\mathbb{P}(\omega_{ij} = 1) = 0.18$ for $j \in [d]$ and all odd $i \in [n]$ and $\mathbb{P}(\omega_{ij} = 1) = 0.02$ for $j \in [d]$ and all even $i \in [n]$.

Compare with `softImpute`: fix $\lambda > 0$ and take the top K eigenvectors of

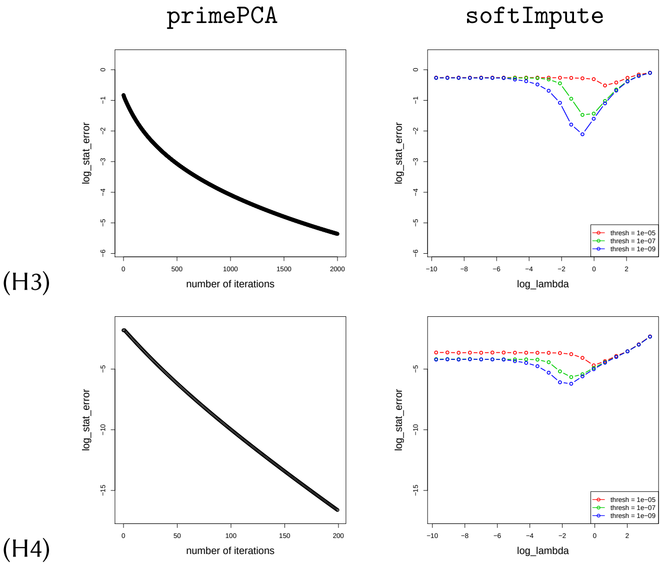
$$\widehat{\boldsymbol{Y}}^{\text{soft}} := \mathop{\arg\min}_{\boldsymbol{X} \in \mathbb{R}^{n \times d}} \left\{ \frac{1}{2} \|\boldsymbol{Y}_{\boldsymbol{\Omega}} - \boldsymbol{X}_{\boldsymbol{\Omega}}\|_{\text{F}}^2 + \lambda \|\boldsymbol{X}\|_* \right\}$$

(Mazumder, Hastie and Tibshirani, 2010).

Noiseless case



Noiseless case



Simulations: Noisy case

Now generate $\boldsymbol{z}_i \sim N_d(\mathbf{0}, \boldsymbol{I}_d)$, independent of all other data, and set $\boldsymbol{\Sigma}_{\boldsymbol{u}} = \nu^2 \boldsymbol{I}_2$ where $\nu \in \{20, 40, 60\}$, corresponding to

$$\mathrm{SNR} := \mathrm{tr}\,\mathrm{Cov}(\boldsymbol{y}_1)/\mathrm{tr}\,\mathrm{Cov}(\boldsymbol{z}_1) \in \{1.6, 6.4, 14.4\}.$$

Also compare with `hardImpute` (Mazumder, Hastie and Tibshirani, 2010), which retains only a fixed number of top singular values in each iteration of matrix imputation; i.e. `softImpute` with $\lambda = 0$.

For `softImpute`, use oracle choice of λ for each repetition.

Simulations: Noisy case

		$\nu = 20$	$\nu = 40$	$\nu = 60$
(H1)	`hardImpute`	$0.444_{(0.001)}$	$0.251_{(0.001)}$	$0.186_{(0.0005)}$
	`softImpute(oracle)`	$0.186_{(0.0004)}$	$0.095_{(0.0002)}$	$0.064_{(0.0002)}$
	`primePCA_init`	$0.306_{(0.001)}$	$0.266_{(0.001)}$	$0.259_{(0.001)}$
	`primePCA`	$0.171_{(0.0004)}$	$0.084_{(0.0002)}$	$0.056_{(0.0001)}$
(H2)	`hardImpute`	$0.473_{(0.001)}$	$0.291_{(0.001)}$	$0.236_{(0.001)}$
	`softImpute(oracle)`	$0.308_{(0.001)}$	$0.185_{(0.001)}$	$0.141_{(0.001)}$
	`primePCA_init`	$0.399_{(0.002)}$	$0.357_{(0.001)}$	$0.349_{(0.001)}$
	`primePCA`	$0.232_{(0.001)}$	$0.115_{(0.001)}$	$0.077_{(0.0005)}$
(H3)	`hardImpute`	$0.479_{(0.001)}$	$0.385_{(0.001)}$	$0.427_{(0.001)}$
	`softImpute(oracle)`	$0.374_{(0.001)}$	$0.222_{(0.001)}$	$0.170_{(0.001)}$
	`primePCA_init`	$0.486_{(0.001)}$	$0.449_{(0.001)}$	$0.442_{(0.001)}$
	`primePCA`	$0.290_{(0.001)}$	$0.145_{(0.001)}$	$0.097_{(0.0004)}$
(H4)	`hardImpute`	$0.174_{(0.0005)}$	$0.089_{(0.0003)}$	$0.062_{(0.0003)}$
	`softImpute(oracle)`	$0.121_{(0.0002)}$	$0.062_{(0.0001)}$	$0.042_{(0.0001)}$
	`primePCA_init`	$0.203_{(0.001)}$	$0.175_{(0.0005)}$	$0.169_{(0.0004)}$
	`primePCA`	$0.116_{(0.0003)}$	$0.058_{(0.0002)}$	$0.038_{(0.0001)}$

Million Song Dataset

- Original data has $110{,}000$ users (rows) and $163{,}206$ songs (columns); entries represent number of times a song was played by a particular user.
- Proportion of non-missing entries in the matrix is 0.008%.

Million Song Dataset

- Original data has $110{,}000$ users (rows) and $163{,}206$ songs (columns); entries represent number of times a song was played by a particular user.
- Proportion of non-missing entries in the matrix is 0.008%.
- Restrict attention to songs that have at least 100 listeners ($1{,}777$ songs in total). This improves the proportion of non-missing entries to 0.23%.

Million Song Dataset

- Original data has $110{,}000$ users (rows) and $163{,}206$ songs (columns); entries represent number of times a song was played by a particular user.
- Proportion of non-missing entries in the matrix is 0.008%.
- Restrict attention to songs that have at least 100 listeners ($1{,}777$ songs in total). This improves the proportion of non-missing entries to 0.23%.
- Quantiles of the number of listeners for each song:

0%	50%	60%	70%	80%	90%	100%
100	154	178	214	272.8	455.6	5043

Quantiles of the total play counts of each user:

0%	50%	60%	70%	80%	90%	100%
0	6	9	14	21	38	1114

Million Song Dataset

- Original data has $110{,}000$ users (rows) and $163{,}206$ songs (columns); entries represent number of times a song was played by a particular user.
- Proportion of non-missing entries in the matrix is 0.008%.
- Restrict attention to songs that have at least 100 listeners ($1{,}777$ songs in total). This improves the proportion of non-missing entries to 0.23%.
- Quantiles of the number of listeners for each song:

0%	50%	60%	70%	80%	90%	100%
100	154	178	214	272.8	455.6	5043

  Quantiles of the total play counts of each user:

0%	50%	60%	70%	80%	90%	100%
0	6	9	14	21	38	1114

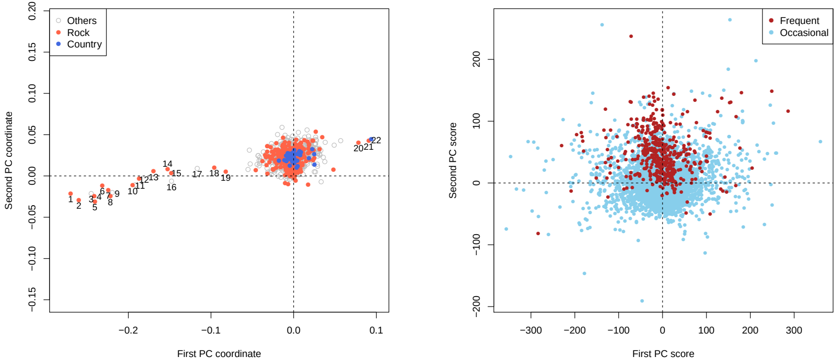
- Quantiles of non-missing matrix entry values:

0%	50%	60%	70%	80%	90%	100%
1	1	2	3	5	8	500

  To guard against excessive influence from outliers, discretise play counts:

Play count	1	2 – 3	4 – 6	7 – 10	≥ 11
Level of interest	1	2	3	4	5

Million Song Dataset



Plots of the first two principal components $\widehat{V}_2^{\text{prime}}$ (left) and the associated scores $\{\widehat{\boldsymbol{u}}_i\}_{i=1}^n$ (right).

Outlier songs

ID	Title	Artist	Genre
1	Your Hand In Mine	Explosions In The Sky	Rock
2	All These Things That I've Done	The Killers	Rock
3	Lady Marmalade	Christina Aguilera / Lil' Kim/ Mya / Pink	Pop
4	Here It Goes Again	Ok Go	Rock
5	I Hate Pretending (Album Version)	Secret Machines	Rock
6	No Rain	Blind Melon	Rock
7	Comatose (Comes Alive Version)	Skillet	Rock
8	Life In Technicolor	Coldplay	Rock
9	New Soul	Yael Naïm	Pop
10	Blurry	Puddle Of Mudd	Rock
11	Give It Back	Polly Paulusma	Pop
12	Walking On The Moon	The Police	Rock
13	Face Down (Album Version)	The Red Jumpsuit Apparatus	Rock
14	Savior	Rise Against	Rock
15	Swing Swing	The All-American Rejects	Rock
16	Without Me	Eminem	Rap
17	Almaz	Randy Crawford	Pop
18	Hotel California	Eagles	Rock
19	Hey There Delilah	Plain White T's	Rock
20	Revelry	Kings Of Leon	Rock
21	Undo	Björk	Rock
22	You're The One	Dwight Yoakam	Country

Summary

- Heterogeneous missingness is ubiquitous.

Summary

- Heterogeneous missingness is ubiquitous.
- The way in which the heterogeneity interacts with the underlying structure of interest is crucial.

Summary

- Heterogeneous missingness is ubiquitous.
- The way in which the heterogeneity interacts with the underlying structure of interest is crucial.
- `primePCA` iteratively projects observed entries of data matrix onto column space of current estimate to impute missing entries, then updates estimate by computing leading right singular space of imputed matrix.

Summary

- Heterogeneous missingness is ubiquitous.
- The way in which the heterogeneity interacts with the underlying structure of interest is crucial.
- `primePCA` iteratively projects observed entries of data matrix onto column space of current estimate to impute missing entries, then updates estimate by computing leading right singular space of imputed matrix.
- With an incoherence condition, the error of `primePCA` converges to zero at geometric rate in the noiseless setting.

Summary

- Heterogeneous missingness is ubiquitous.
- The way in which the heterogeneity interacts with the underlying structure of interest is crucial.
- `primePCA` iteratively projects observed entries of data matrix onto column space of current estimate to impute missing entries, then updates estimate by computing leading right singular space of imputed matrix.
- With an incoherence condition, the error of `primePCA` converges to zero at geometric rate in the noiseless setting.
- Theoretical guarantees depend on average, as opposed to worst-case, properties of the missingness mechanism.

Summary

- Heterogeneous missingness is ubiquitous.
- The way in which the heterogeneity interacts with the underlying structure of interest is crucial.
- `primePCA` iteratively projects observed entries of data matrix onto column space of current estimate to impute missing entries, then updates estimate by computing leading right singular space of imputed matrix.
- With an incoherence condition, the error of `primePCA` converges to zero at geometric rate in the noiseless setting.
- Theoretical guarantees depend on average, as opposed to worst-case, properties of the missingness mechanism.

Main reference:

- Zhu, Z., Wang, T. and Samworth, R. J. (2019) High-dimensional principal component analysis with heterogeneous missingness. `https://arxiv.org/abs/1906.10125`.

Other references

- Anderson, T. W. (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.*, **52**, 200–203.
- Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2017) Linear and conic programming estimators in high dimensional errors-in-variables models. *J. Roy. Statist. Soc., Ser. B*, **79**, 939–956.
- Cai, T. T. and Zhang, A. (2018a) Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.*, **46**, 60–89.
- Cai, T. T. and Zhang, L. (2018b) High-dimensional linear discriminant analysis: optimality, adaptive algorithm, and missing data. `arXiv:1804.03018`.
- Cho, J., Kim, D. and Rohe, K. (2017) Asymptotic theory for estimating the singular vectors and values of a partially-observed low rank matrix with noise. *Statist. Sinica*, **27**, 1921–1948.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B*, **39**, 1–38.
- Elsener, A and van de Geer, S. (2018) Sparse spectral estimation with missing and corrupted measurements. `arXiv:1811.10443`.

Other references

- Ford, B. L. (1983) An overview of hot-deck procedures. In W. G. Madow, I. Olkin and D. B. Rubin (Eds.) *Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies*, 185–207. Academic Press, New York.
- Loh, P.-L. and Tan, X. L. (2018) High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electron. J. Statist.*, **12**, 1429–1467.
- Loh, P.-L. and Wainwright, M. J. (2012) High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.*, **40**, 1637–1664.
- Lounici, K. (2014) High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, **20**, 1029–1058.
- Mazumder, R., Hastie, T. and Tibshirani R. (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.
- Rubin, D. B. (2004) *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons, Hoboken.
- Shen, D., Shen, H., Zhu, H. and Marron, J. (2016) The statistics and mathematics of high dimension low sample size asymptotics. *Statist. Sinica*, **26**, 1747–1770.