High-dimensional principal component analysis with heterogeneous missingness

Tengyao Wang
University College London

Statistics Seminar, University of Kent
24 Oct 2019
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.
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)
Suppose the (partially observed) matrix $Y \in \mathbb{R}^{n \times d}$ is of the form

$$Y = UV_K^T + Z,$$

where $V_K \in \mathbb{R}^{d \times K}$ has orthonormal columns and $U$ is a random $n \times K$ matrix (with $n > K$) having i.i.d. rows with mean zero.
Suppose the (partially observed) matrix $Y \in \mathbb{R}^{n \times d}$ is of the form

$$Y = UV_K^T + Z,$$

where $V_K \in \mathbb{R}^{d \times K}$ has orthonormal columns and $U$ 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 $Y_{\Omega} := Y \circ \Omega$. We observe the pair $(Y_{\Omega}, \Omega)$ and wish to estimate $\text{Col}(V_K)$. 
Suppose the (partially observed) matrix $Y \in \mathbb{R}^{n \times d}$ is of the form
\[ Y = UV_K^\top + Z, \]
where $V_K \in \mathbb{R}^{d \times K}$ has orthonormal columns and $U$ 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 $Y_\Omega := Y \circ \Omega$. We observe the pair $(Y_\Omega, \Omega)$ and wish to estimate $\text{Col}(V_K)$.

Performance of an estimator $\hat{V}_K$ measured by the loss function
\[ L(\hat{V}_K, V_K) := \| \sin \Theta(\hat{V}_K, V_K) \|_F, \]
where $\Theta(U, V)$ is the matrix of principal angles between $\text{Col}(U)$ and $\text{Col}(V)$. 

Tengyao Wang
Consider the $p$-homogeneous setting, where $\Omega_{ij} \sim \text{Bern}(p)$. Then

$$P := \mathbb{E}(\omega_1 \omega_1^\top) = p^2 \{1_d 1_d^\top - (1 - p^{-1}) I_d\}.$$

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

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

This ensures that $\mathbb{E}(G \mid Y) = n^{-1} Y^\top Y$. 
Consider the \( p \)-homogeneous setting, where \( \Omega_{ij} \overset{iid}{\sim} \text{Bern}(p) \). Then

\[
P := \mathbb{E}(\omega_1 \omega_1^\top) = p^2 \{1_d 1_d^\top - (1 - p^{-1}) I_d\}.
\]

Its elementwise inverse is

\[
W := p^{-2} \{1_d 1_d^\top - (1 - p) I_d\},
\]

and we can define the weighted sample covariance matrix

\[
G := \left( \frac{1}{n} Y_\Omega^\top Y_\Omega \right) \circ W.
\]

This ensures that \( \mathbb{E}(G \mid Y) = n^{-1} Y^\top Y \). Define \( \hat{W} \) by replacing \( p \) in \( W \) with

\[
\hat{p} := (nd)^{-1} \|\Omega\|_1,
\]

and set

\[
\hat{G} := \left( \frac{1}{n} Y_\Omega^\top Y_\Omega \right) \circ \hat{W}.
\]

The IPW estimator of \( V_K \) is given by the top \( K \) eigenvectors of \( \hat{G} \), denoted \( \hat{V}_K \).

(Cai and Zhang, 2018a; Cho, Kim and Rohe, 2017).
Assumptions

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

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

and a version that is invariant to invertible affine transformations:

\[
\|x\|_{\psi^*_r} := \sup_{u \in S^{d-1}} \frac{\|u^\top (x - \mathbb{E}x)\|_{\psi_r}}{\text{Var}^{1/2}(u^\top x)}.
\]

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

\( \text{(A2)} \) \( \|u_1\|_{\psi^*_2} \leq \tau \);

\( \text{(A3)} \) \( Z = (z_{ij})_{i \in [n], j \in [d]} \) has i.i.d. entries with \( \mathbb{E}z_{11} = 0 \), \( \text{Var}z_{11} = 1 \) and \( \|z_{11}\|_{\psi^*_2} \leq \tau \);

\( \text{(A4)} \) \( \|y_{21}^j\|_{\psi_1} \leq M \) for all \( j \in [d] \).
For $r \in \mathbb{N}$ and a $d$-dimensional random vector $\mathbf{x}$, define its Orlicz norm

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

and a version that is invariant to invertible affine transformations:

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

(A1) $\mathbf{U}$, $\mathbf{Z}$ and $\Omega$ are independent;
(A2) $\|\mathbf{u}_1\|_{\psi^*_2} \leq \tau$;
(A3) $\mathbf{Z} = (z_{ij})_{i \in [n], j \in [d]}$ has i.i.d. entries with $\mathbb{E} z_{11} = 0$, $\text{Var} z_{11} = 1$ and $\|\mathbf{z}_{11}\|_{\psi^*_2} \leq \tau$;
(A4) $\|y_{1j}^2\|_{\psi_1} \leq M$ for all $j \in [d]$. 

Tengyao Wang 7/29
**Upper bound**

**Theorem.** Assume (A1)–(A4) and that $n, d \geq 2, dp \geq 1$. Let $\lambda_j$ denote the $j$th largest eigenvalue of $\Sigma_u$. If $n \geq d \log^2 d \log^2 n / (\lambda_1 p + \log d)$, then

$$\mathbb{E} L(\hat{V}_K, V_K) \lesssim_{M, \tau} \frac{1}{\lambda_K} \left( \frac{K d (\lambda_1 p + \log d) \log^2 d}{n p^2} \right)^{1/2}.$$
**Upper bound**

**Theorem.** Assume (A1)–(A4) and that $n, d \geq 2$, $dp \geq 1$. Let $\lambda_j$ denote the $j$th largest eigenvalue of $\Sigma_u$. If $n \geq d \log^2 d \log^2 n/(\lambda_1 p + \log d)$, then

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

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 \( \lambda_j \) denote the \( j \)th largest eigenvalue of \( \Sigma_u \). If \( n \geq d \log^2 d \log^2 n/(\lambda_1 p + \log d) \), then

\[
\mathbb{E} L(\hat{V}_K, V_K) \lesssim_M \tau \frac{1}{\lambda_K} \left( \frac{K d (\lambda_1 p + \log d) \log^2 d}{np^2} \right)^{1/2}.
\]

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,

\[
\mathbb{E} L(\hat{V}_K, V_K) \lesssim_M \tau \begin{cases} 
\frac{1}{\lambda_K} \left( \frac{K d \log^3 d}{np^2} \right)^{1/2} & \text{if } \lambda_1 p \lesssim \log d, \\
\frac{\lambda_1^{1/2}}{\lambda_K} \left( \frac{K d \log^2 d}{np} \right)^{1/2} & \text{if } \lambda_1 p \gtrsim \log d.
\end{cases}
\]
Minimax lower bound

Let \( \mathcal{P}_{n,d}(\lambda_1, p) \) denote the class of distributions of pairs \( (Y_{\Omega}, \Omega) \) satisfying (A1), (A2), (A3) with \( K = 1 \). Since we are now working with vectors instead of matrices, we write \( v \) in place of \( V_1 \).

**Theorem.** There exists a universal constant \( c > 0 \) such that

\[
\inf_{\hat{v}} \sup_{P \in \mathcal{P}_{n,d}(\lambda_1, p)} \mathbb{E}_P L(\hat{v}, v) \geq c \min \left\{ \frac{1}{\lambda_1} \left( \frac{d(\lambda_1 p + 1)}{np^2} \right)^{1/2}, 1 \right\},
\]

where the infimum is taken over all estimators \( \hat{v} = \hat{v}(Y_{\Omega}, \Omega) \) of \( v \).

Thus \( \hat{V}_1 \) achieves the minimax optimal rate of estimation up to a poly-logarithmic factor when \( M \) and \( \tau \) are regarded as constants and \( K = 1 \).
Suppose that
\[ P\{ \omega_1 = (1, 0, 1, \ldots, 1)^\top \} = P\{ \omega_1 = (0, 1, 1, \ldots, 1)^\top \} = 1/2. \]

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

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

\[ \mathbb{P}\{\omega_1 = (1, 0, 1, \ldots, 1)\top\} = \mathbb{P}\{\omega_1 = (0, 1, 1, \ldots, 1)\top\} = 1/2. \]

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

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

But the respective leading eigenvectors of \( \Sigma \) and \( \Sigma' \) are \( \alpha \) and \( \alpha' \), 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 \( \tilde{y}_i := y_i \circ \omega_i \).

**Algorithm 1** \( \text{refine}(K, \hat{V}^{(\text{in})}_K, \Omega, Y_\Omega) \), a single step of refinement of current iterate \( \hat{V}^{(\text{in})}_K \)

Input: \( K \in [d], \hat{V}^{(\text{in})}_K \in \mathbb{O}^{d \times K}, \Omega \in \{0,1\}^{n \times d} \) with \( \min_i \| \omega_i \|_1 \geq 1 \), \( Y_\Omega \in \mathbb{R}^{n \times d} \)

Output: \( \hat{V}^{(\text{out})}_K \in \mathbb{O}^{d \times K} \)

1: for \( i \) in \([n]\) do
2: \( J_i \leftarrow \{ j \in [d] : \omega_{ij} = 1 \} \)
3: \( \hat{u}_i \leftarrow (\hat{V}^{(\text{in})}_K)_{J_i}^\top \hat{y}_i, J_i \)
4: \( \hat{y}_{i,J_i^c} \leftarrow \hat{V}^{(\text{in})}_K \hat{u}_i, J_i^c \)
5: \( \hat{y}_{i,J_i} \leftarrow y_i, J_i \)
6: end for
7: \( \hat{Y} \leftarrow (\hat{y}_1, \ldots, \hat{y}_n)^\top \)
8: \( \hat{V}^{(\text{out})}_K \leftarrow \text{top} \; K \; \text{right singular vectors of} \; \hat{Y} \)
Two-to-infinity subspace distance

For $U, V \in \mathbb{O}^{d \times K}$, let $W_1 D_{U,V} W_2^\top$ be an SVD of $V^\top U$ and let $W_{U,V} := W_1 W_2^\top$. Then $W_{U,V}$ solves the Procrustes problem in the sense that

$$W_{U,V} \in \arg \min_{W \in \mathbb{O}^{K \times K}} \| U - V W \|_F.$$ 

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

$$\mathcal{T}(U, V) := \| U - V W_{U,V} \|_{2 \to \infty},$$

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

Tengyao Wang
Proposition. Let $\hat{V}_K^{(out)} := \text{refine}(K, \hat{V}_K^{(in)}, \Omega, Y_\Omega)$. We assume that 
\[ \min_{i \in [n]} \|\omega_i\|_1 > K, \quad \text{that} \quad \min_{i \in [n]} \frac{d^{1/2} \sigma_K((\hat{V}_K^{(in)})_{J_i})}{|J_i|^{1/2}} \geq 1/\sigma_* > 0, \]
and write the SVD of $Y$ as $L \Gamma R^\top$.

Suppose that $Z = 0$, and that both $\|L\|_{2\to\infty} \leq \mu_1 (K/n)^{1/2}$ and 
$\|R\|_{2\to\infty} \leq \mu_2 (K/d)^{1/2}$ hold for some $\mu_1, \mu_2 \geq 1$. Then there exist $c_1, C > 0$, 
depending only on $\mu_1, \mu_2$ and $\sigma_*$, such that whenever 
(i) $\mathcal{T}(\hat{V}_K^{(in)}, V_K) \leq \frac{c_1 \sigma_K(\Gamma)}{K^2 \sigma_1(\Gamma) \sqrt{d}}$, 
(ii) $\rho := \frac{CK^2 \sigma_1(\Gamma) \|\Omega^c\|_{1\to1}}{\sigma_K(\Gamma)n} < 1,$ 
we have that 
\[ \mathcal{T}(\hat{V}_K^{(out)}, V_K) \leq \rho \mathcal{T}(\hat{V}_K^{(in)}, V_K). \]
Algorithm 2 primePCA, an iterative algorithm for estimating $V_K$ given initialiser $\hat{V}_K^{(0)}$

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

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

1: for $i$ in $[n]$ do
2: $\mathcal{J}_i \leftarrow \{j \in [d] : \omega_{ij} = 1\}$
3: end for
4: for $t$ in $[n_{\text{iter}}]$ do
5: $\mathcal{I}^{(t-1)} \leftarrow \{i : \|\omega_i\|_1 > K, \sigma_K((\hat{V}_K^{(t-1)})_{\mathcal{J}_i}) \geq \frac{\|\mathcal{J}_i\|^{1/2}}{d^{1/2}\sigma_*}\}$
6: $\hat{V}_K^{(t)} \leftarrow \text{refine}(K, \hat{V}_K^{(t-1)}, \Omega_{\mathcal{I}^{(t-1)}}, (Y_\Omega)_{\mathcal{I}^{(t-1)}})$ \# refine is defined in Algorithm 1.
7: if $L(\hat{V}_K^{(t)}, \hat{V}_K^{(t-1)}) < \kappa^*$ then break
8: end if
9: end for
10: return $\hat{V}_K = \hat{V}_K^{(t)}$
Theorem. For \( t \in [n_{\text{iter}}] \), let \( \hat{V}_K^{(t)} \) be the \( t \)th iterate of Algorithm 2 with input \( K, \hat{V}_K^{(0)}, \Omega \in \{0, 1\}^{n \times d}, Y_\Omega \in \mathbb{R}^{n \times d}, n_{\text{iter}} \in \mathbb{N}, \sigma_* \in (0, \infty) \) and \( \kappa^* = 0 \). Let
\[
I := \left\{ i : \|\omega_i\|_1 > K, \sigma_K((V_K)_{J_i}) \geq |J_i|^{1/2}/(d^{1/2} \sigma_*) \right\},
\]
where \( J_i := \left\{ j : \omega_{ij} = 1 \right\} \). Let \( Y_I = L \Gamma R^\top \) be an SVD of \( Y_I \). Suppose that both \( \|L\|_2 \rightarrow \infty \leq \mu_1(K/|I|)^{1/2} \) and \( \|R\|_2 \rightarrow \infty \leq \mu_2(K/d)^{1/2} \). Let
\[
Z := \left\{ \sigma_K((V_K)_{J_i})d^{1/2}/|J_i|^{1/2} : i \in [n], \|\omega_i\|_1 > K \right\},
\]
and assume that \( \epsilon := \min_{z \in Z} |z - \sigma_*^{-1}| > 0 \). Then there exist \( c_1, C > 0 \), depending only on \( \mu_1, \mu_2, \sigma_* \) and \( \epsilon \), such that whenever
\[
\mathcal{T}(\hat{V}_K^{(0)}, V_K) \leq \frac{c_1 \sigma_K(Y_I)}{K^2 \sigma_1(Y_I) \sqrt{d}} \quad \text{and} \quad \rho := \frac{C K^2 \sigma_1(Y_I) \|\Omega^c_I\|_1}{\sigma_K(Y_I) |I|} < 1,
\]
we have \( \mathcal{T}(\hat{V}_K^{(t)}, V_K) \leq \rho^t \mathcal{T}(\hat{V}_K^{(0)}, V_K) \) for every \( t \in [n_{\text{iter}}] \).
Consider the following modified weighted sample covariance matrix

\[
\tilde{G} := \frac{1}{n} \sum_{i=1}^{n} \tilde{y}_i \tilde{y}_i^\top \circ \tilde{W},
\]

where for any \( j, k \in [d] \),

\[
\tilde{W}_{jk} := \begin{cases} 
\sum_{i=1}^{n} \omega_{ij} \omega_{ik} & \text{if } \sum_{i=1}^{n} \omega_{ij} \omega_{ik} > 0, \\
0, & \text{otherwise}.
\end{cases}
\]

We take as our initial estimator of \( \tilde{V}_K \) the matrix of top \( K \) eigenvectors of \( \tilde{G} \), denoted \( \tilde{V}_K \).

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

$$
\lambda_K > C \left\{ \left( \frac{M\tau^2 R \|\tilde{W}\|_1}{n} \right)^{1/2} \xi \log d \right\} + \frac{M \|\tilde{W}\|_{op} \xi \log^2 d}{n},
$$

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

$$
L(\tilde{V}_K, V_K) \leq \frac{2^{9/2} e\tau \mu}{\lambda_K} \left( \frac{KMR}{d} \right)^{1/2} \left( \frac{\xi^{1/2} \|\tilde{W}\|_1^{1/2} \log^{1/2} d}{n^{1/2}} + \frac{\xi \|\tilde{W}\|_F \log d}{n} \right).
$$
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

$$\lambda_K > C \left\{ \left( \frac{M \tau^2 R \| \tilde{W} \|_{1 \to 1} \xi \log d}{n} \right)^{1/2} + \frac{M \| \tilde{W} \|_{\text{op}} \xi \log^2 d}{n} \right\},$$

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

$$L(\tilde{V}_K, V_K) \leq \frac{2^{9/2} e \tau \mu}{\lambda_K} \left( \frac{K MR}{d} \right)^{1/2} \left( \frac{\xi^{1/2} \| \tilde{W} \|_{1/2} \log^{1/2} d}{n^{1/2}} + \frac{\xi \| \tilde{W} \|_F \log d}{n} \right).$$

N.B. The bound depends on $\tilde{W}$ only through the *entrywise* $\ell_1$ and $\ell_2$ norms of the whole matrix.
Simulations: Noiseless case

Fix $n = 2000$, $d = 500$, $K = 2$ and $u_i \sim N_d(0, \Sigma_u)$ where $\Sigma_u = 100 I_2$. Set

$$V_K = \sqrt{\frac{1}{500}} \left( \begin{array}{cc} 1_{250} & 1_{250} \\ 1_{250} & -1_{250} \end{array} \right) \in \mathbb{R}^{500 \times 2}.$$
Simulations: Noiseless case

Fix \( n = 2000, d = 500, K = 2 \) and \( u_i \sim N_d(0, \Sigma_u) \) where \( \Sigma_u = 100I_2 \). Set

\[
V_K = \sqrt{\frac{1}{500}} \begin{pmatrix} 1_{1250} & 1_{1250} \\ 1_{1250} & -1_{1250} \end{pmatrix} \in \mathbb{R}^{500 \times 2}.
\]

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

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

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

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

Fix $n = 2000$, $d = 500$, $K = 2$ and $u_i \sim N_d(0, \Sigma_u)$ where $\Sigma_u = 100I_2$. Set

$$V_K = \sqrt{\frac{1}{500}} \begin{pmatrix} 1 & \frac{1}{250} \\ \frac{1}{250} & -1 & \frac{1}{250} \end{pmatrix} \in \mathbb{R}^{500 \times 2}.$$

(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 $P_1, \ldots, P_n \sim iid U[0, 0.2]$ and $Q_1, \ldots, Q_d \sim iid U[0.05, 0.95]$ 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

$$\hat{Y}^{soft} := \arg \min_{X \in \mathbb{R}^{n \times d}} \left\{ \frac{1}{2} \| Y_{\Omega} - X_{\Omega} \|_F^2 + \lambda \| X \|_* \right\}$$

(Mazumder, Hastie and Tibshirani, 2010).
**Noiseless case**

**primePCA**

![primePCA graph](image1)

**softImpute**

![softImpute graph](image2)

(H1)

(H2)
Noiseless case

primePCA

softImpute

(H3)

(H4)
Simulations: Noisy case

Now generate $z_i \sim N_d(0, I_d)$, independent of all other data, and set

$$\Sigma_u = \nu^2 I_2$$

where $\nu \in \{20, 40, 60\}$, corresponding to

$$\text{SNR} := \frac{\text{tr Cov}(y_1)}{\text{tr Cov}(z_1)} \in \{1.6, 6.4, 14.4\}.$$

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 $\lambda$ for each repetition.
## Simulations: Noisy case

<table>
<thead>
<tr>
<th></th>
<th>( \nu = 20 )</th>
<th>( \nu = 40 )</th>
<th>( \nu = 60 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(H1)</td>
<td>hardImpute</td>
<td>0.444(_{0.001})</td>
<td>0.251(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>softImpute(oracle)</td>
<td>0.186(_{0.0004})</td>
<td>0.095(_{0.0002})</td>
</tr>
<tr>
<td></td>
<td>primePCA_init</td>
<td>0.306(_{0.001})</td>
<td>0.266(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>primePCA</td>
<td>0.171(_{0.0004})</td>
<td>0.084(_{0.0002})</td>
</tr>
<tr>
<td>(H2)</td>
<td>hardImpute</td>
<td>0.473(_{0.001})</td>
<td>0.291(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>softImpute(oracle)</td>
<td>0.308(_{0.001})</td>
<td>0.185(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>primePCA_init</td>
<td>0.399(_{0.002})</td>
<td>0.357(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>primePCA</td>
<td>0.232(_{0.001})</td>
<td>0.115(_{0.001})</td>
</tr>
<tr>
<td>(H3)</td>
<td>hardImpute</td>
<td>0.479(_{0.001})</td>
<td>0.385(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>softImpute(oracle)</td>
<td>0.374(_{0.001})</td>
<td>0.222(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>primePCA_init</td>
<td>0.486(_{0.001})</td>
<td>0.449(_{0.001})</td>
</tr>
<tr>
<td></td>
<td>primePCA</td>
<td>0.290(_{0.001})</td>
<td>0.145(_{0.001})</td>
</tr>
<tr>
<td>(H4)</td>
<td>hardImpute</td>
<td>0.174(_{0.0005})</td>
<td>0.089(_{0.0003})</td>
</tr>
<tr>
<td></td>
<td>softImpute(oracle)</td>
<td>0.121(_{0.0002})</td>
<td>0.062(_{0.0001})</td>
</tr>
<tr>
<td></td>
<td>primePCA_init</td>
<td>0.203(_{0.001})</td>
<td>0.175(_{0.0005})</td>
</tr>
<tr>
<td></td>
<td>primePCA</td>
<td>0.116(_{0.0003})</td>
<td>0.058(_{0.0002})</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Quantile</th>
<th>0%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>154</td>
<td>178</td>
<td>214</td>
<td>272.8</td>
<td>455.6</td>
<td>5043</td>
</tr>
</tbody>
</table>

Quantiles of the total play counts of each user:

<table>
<thead>
<tr>
<th>Quantile</th>
<th>0%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>6</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>38</td>
<td>1114</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Quantile</th>
<th>0%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>100</td>
<td>154</td>
<td>178</td>
<td>214</td>
<td>272.8</td>
<td>455.6</td>
<td>5043</td>
</tr>
</tbody>
</table>

- Quantiles of the total play counts of each user:

<table>
<thead>
<tr>
<th>Quantile</th>
<th>0%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>0</td>
<td>6</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>38</td>
<td>1114</td>
</tr>
</tbody>
</table>

- Quantiles of non-missing matrix entry values:

<table>
<thead>
<tr>
<th>Quantile</th>
<th>0%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>500</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Play count</th>
<th>1</th>
<th>2 – 3</th>
<th>4 – 6</th>
<th>7 – 10</th>
<th>≥ 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of interest</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Plots of the first two principal components $\mathbf{\hat{V}}^\text{prime}_2$ (left) and the associated scores $\{\mathbf{\hat{u}}_i\}_{i=1}^n$ (right).
## Outlier songs

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

Main reference:
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.
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.
- \texttt{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 \texttt{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.
- \textit{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 \textit{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:

Other references


Other references