Statistical Properties for Multilinear Principal Component Analysis

Su-Yun Huang, Ting-Li Chen, I-Ping Tu
Institute of Statistical Science, Academia Sinica, Taiwan

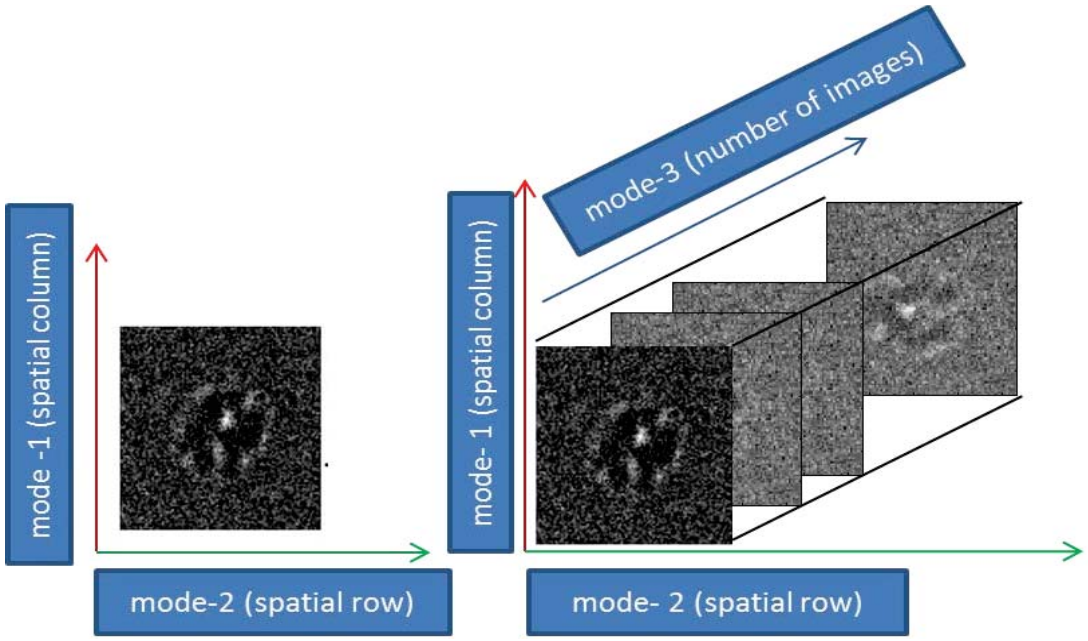
Hung Hung
Epidemiology & Preventive Medicine, National Taiwan University

Pei-Shien Wu

August 25 in LinStat2014

Motivating Data Example

♠ cryo-em images ♠ image clustering



common preprocessing: dimension reduction using PCA

5000 or more images, each is of 130×130 pixels, $\mathbf{dim = 16,900}$

- PCA is probably the simplest and most commonly used dimension reduction tool in many real data applications.

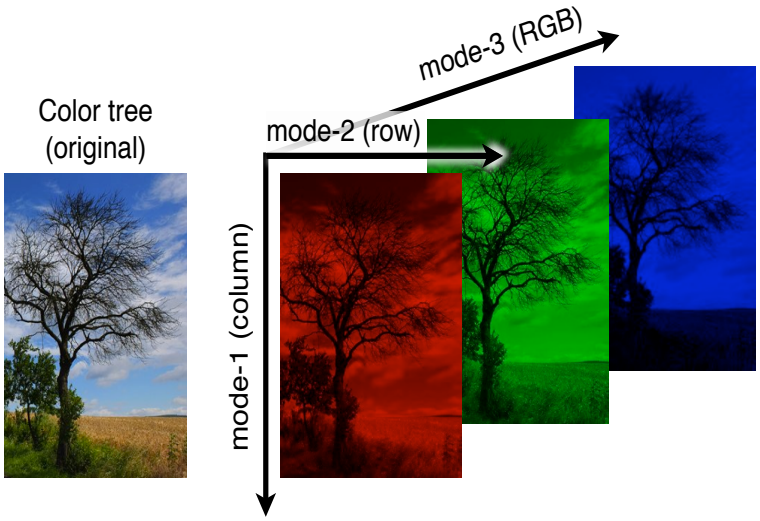
- When data are tensor structured (or arrays), such as **images here as order-2 tensors**, we need more efficient dimension reduction tool.

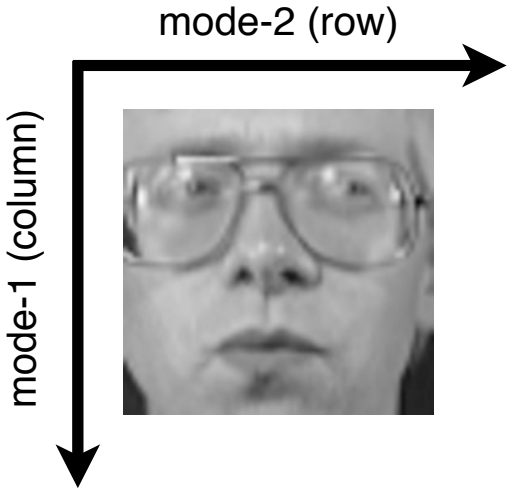
In this talk,

- MPCA (multilinear principal component analysis),

- Statistical aspects of MPCA for tensor data.

Array (tensor) data





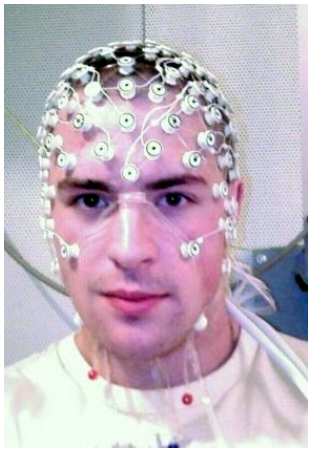
color image as an order-3 tensor BW image as an order-2 tensor

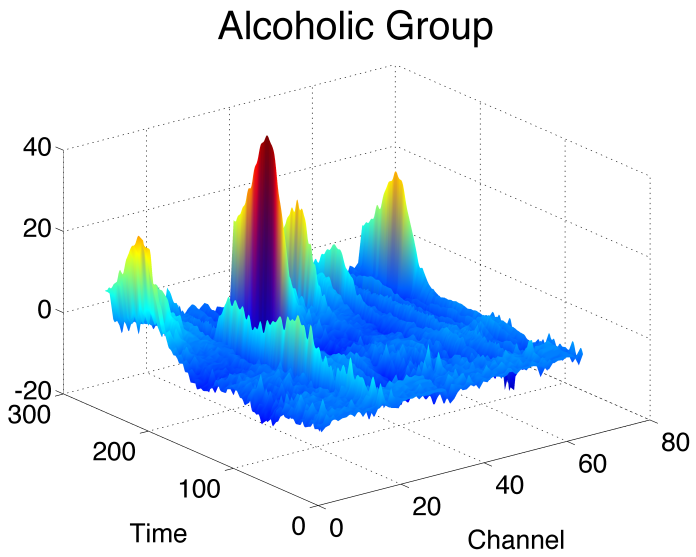
Array (tensor) data

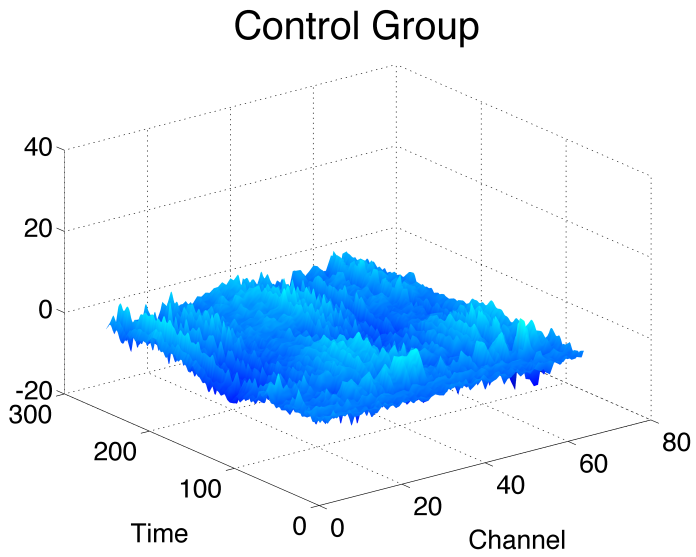
EEG Data (**image by D. Myers**)

- 122 people (77 alcoholic, 45 control)
- 256 time points, 64 channels

  $X_i : 256 \times 64$, order-2 tensor







Data structure -arrays (tensors)

- Each observation is an order-m tensor,

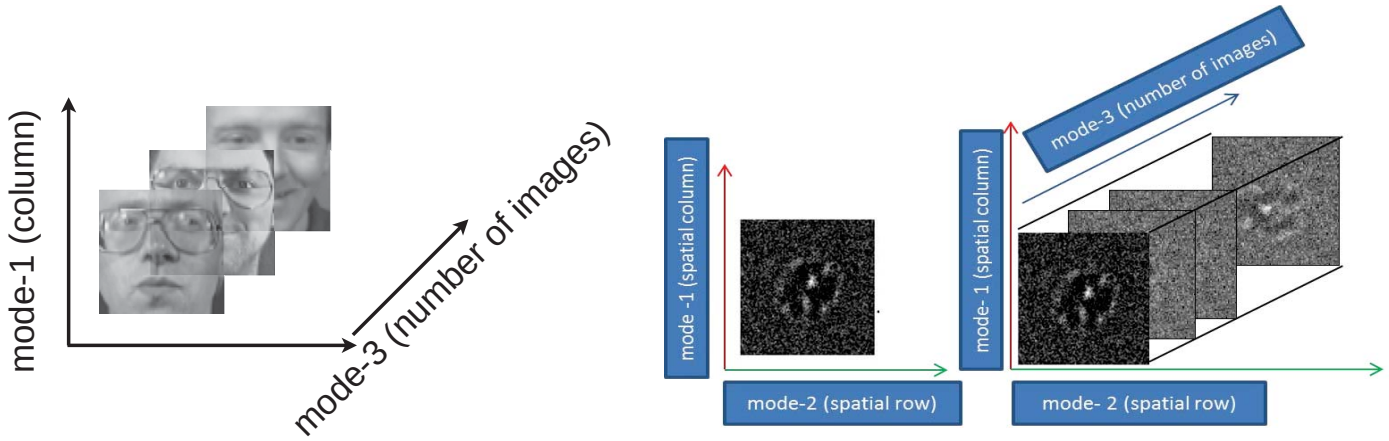
$$\boldsymbol{\mathcal{X}}_i = \boldsymbol{\mathcal{Z}}_i + \boldsymbol{\mathcal{E}}_i \quad \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_m}, \quad i = 1, \ldots, n,$$

  where $\boldsymbol{\mathcal{Z}}_i$ is the signal component and $\boldsymbol{\mathcal{E}}_i$ is the noise part. Collectively, $\{\boldsymbol{\mathcal{X}}_i\}_{i=1}^n$ form an order-$(m+1)$ data tensor.

- When $m = 2$, we have **matrix-variate observations**.

Data tensor (array)

For illustration simplicity, we present the case where observations are matrices, i.e., $m = 2$.



MPCA & HOSVD model for matrix data

- Each observation is an **order-2 tensor**, i.e., a matrix,

$$\begin{aligned} X_i &= Z_i + \epsilon_i, \quad \in \mathbb{R}^{p\times q}, \quad i = 1, \ldots, n, \\ &= M + AU_iB^\top + \epsilon_i\,. \end{aligned}$$

- The signal part is assumed a tensor structure,

$$Z = M + \boldsymbol{A}\,U\,\boldsymbol{B}^\top, \quad \text{vec}(Z) = \text{vec}(M) + (\boldsymbol{B}\otimes\boldsymbol{A})\,\text{vec}(U),$$

$M \in \mathbb{R}^{p\times q}$: mean tensor, $\mathcal{O}$: orthogonal matrices,

$A \in \mathcal{O}^{p\times p_0}$, $B \in \mathcal{O}^{q\times q_0}$ (often $p_0 \ll p$, $q_0 \ll q$),

$U \in \mathbb{R}^{p_0\times q_0}$: a coefficient tensor whose entries are random.

- The noise component $\text{vec}(\epsilon_i) \overset{\text{iid}}{\sim} N(0, I_m)$, $i = 1, \ldots, n$, $m = pq$.

- General order-k tensor model: $(\boldsymbol{B}\otimes\boldsymbol{A}) \overset{\textbf{replaced by}}{\longleftarrow\!\!-\!\!-\!\!-} (\boldsymbol{A_k}\otimes\cdots\otimes\boldsymbol{A_1})$

Notation

- Matrix-variate observations: $X_i \in \mathbb{R}^{p \times q}$.
- $\| \ \|_F$: Frobenius norm.
- $\mathcal{O}_{p \times \tilde{p}}$ consists of matrices $M \in \mathbb{R}^{p \times \tilde{p}}$ such that $M^\top M = I_{\tilde{p}}$.
- A_0, B_0 (true), $\quad \hat{A}$, $\hat{B}$ (estimate).
- $\tilde{p}$: reduced rank for column subspace dimensionality.
- $\tilde{q}$: reduced row subspace dimensionality.

High order SVD (De Lathauwer et al., 2000)

mode-1 unfolding

$\clubsuit\ \mathbf{X}_{(1)} = \left[X_1 - \bar{X}, \ldots, X_n - \bar{X}\right]_{p \times nq}$  $\ldots$

$\widehat{\boldsymbol{A}}^{(\mathbf{hosvd})} \in \mathbb{R}^{p \times \tilde{p}}$ consists of $\tilde{p}$ **leading eigenvectors** of $\mathbf{X}_{(1)}\mathbf{X}_{(1)}^{\top}$.

mode-2 unfolding

$$\clubsuit\ \mathbf{X}_{(2)} = \begin{bmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}^\top_{q \times np}$$

mode-2 unfolding

$\cdots$

$\widehat{B}^{(\mathbf{hosvd})} \in \mathbb{R}^{q \times \tilde{q}}$ consists of $\tilde{q}$ **leading eigenvectors** of $\mathbf{X}_{(2)}\mathbf{X}_{(2)}^\top$.

Multilinear PCA (best rank-$(\tilde{p}, \tilde{q})$ approximation)

(De Lathauwer et al., 2000; Lu et al., 2008)

MPCA eigen-system satisfies the following stationary conditions.

- $\widehat{A}^{(\text{mpca})} \in \mathbb{R}^{p \times \tilde{p}}$ consists of $\tilde{p}$ leading eigenvectors of

$$\mathbf{X}_{(1)} \left(I_n \otimes \widehat{\boldsymbol{B}}^{(\mathbf{mpca})} \widehat{\boldsymbol{B}}^{(\mathbf{mpca})\top} \right) \mathbf{X}_{(1)}^{\top}.$$

- $\widehat{B}^{(\text{mpca})} \in \mathbb{R}^{q \times \tilde{q}}$ consists of $\tilde{q}$ leading eigenvectors of

$$\mathbf{X}_{(2)} \left(I_n \otimes \widehat{\boldsymbol{A}}^{(\mathbf{mpca})} \widehat{\boldsymbol{A}}^{(\mathbf{mpca})\top} \right) \mathbf{X}_{(2)}^{\top}.$$

HOSVD & MPCA

Extracting leading eigenvectors from each mode using

♣ $\mathbf{X}_{(1)}$ & $\mathbf{X}_{(2)}$ (HOSVD)

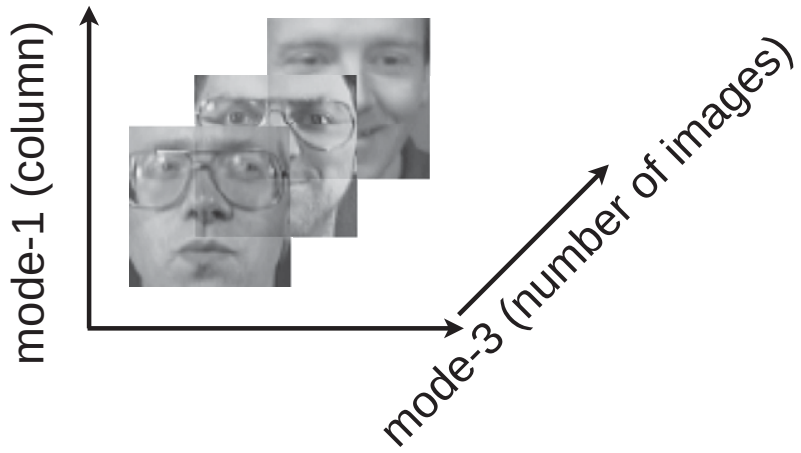
♠ $\mathbf{X}_{(1)}(\boldsymbol{I_n} \otimes \hat{\boldsymbol{B}})$ & $\mathbf{X}_{(2)}(\boldsymbol{I_n} \otimes \hat{\boldsymbol{A}})$ (MPCA)

$\widehat{A}^{(\text{mpca})}$ and $\widehat{B}^{(\text{mpca})}$ are more efficient than $\widehat{A}^{(\text{hosvd})}$ and $\widehat{B}^{(\text{hosvd})}$ under some technical conditions (Hung et al., 2012)

We will focus on MPCA and use notation $\widehat{A}$ and $\widehat{B}$.
(superscript dropped)

Experimental Study

Experimental study

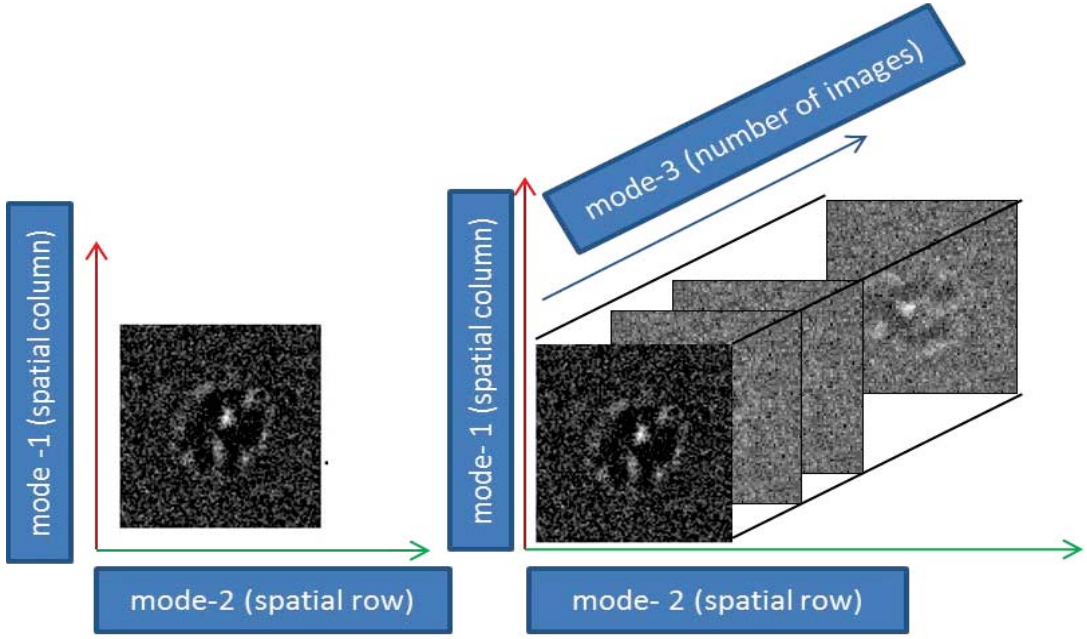


Compare MPCA and PCA on **Olivetti Faces data**.

Another data example:

cryo-em image clustering

(Chen et al., 2014, "γ-SUP", which is a self-updating clustering algorithm based on minimum γ-divergence with application to cryo-EM images)



Experimental study on Olivetti Faces data set

- 400 face images of 64×64: partition them to 100-300 training-test sets.
- Both MPCA and PCA are applied on the training images to produce basis to reconstruct the test images.
  - MPCA: 24 row and 24 column eigenvectors are used to generate 576 basis (24 is selected by hypothesis test for 95% explained-variation)
  - PCA: 576 ($= 24 \times 24$) eigenvectors

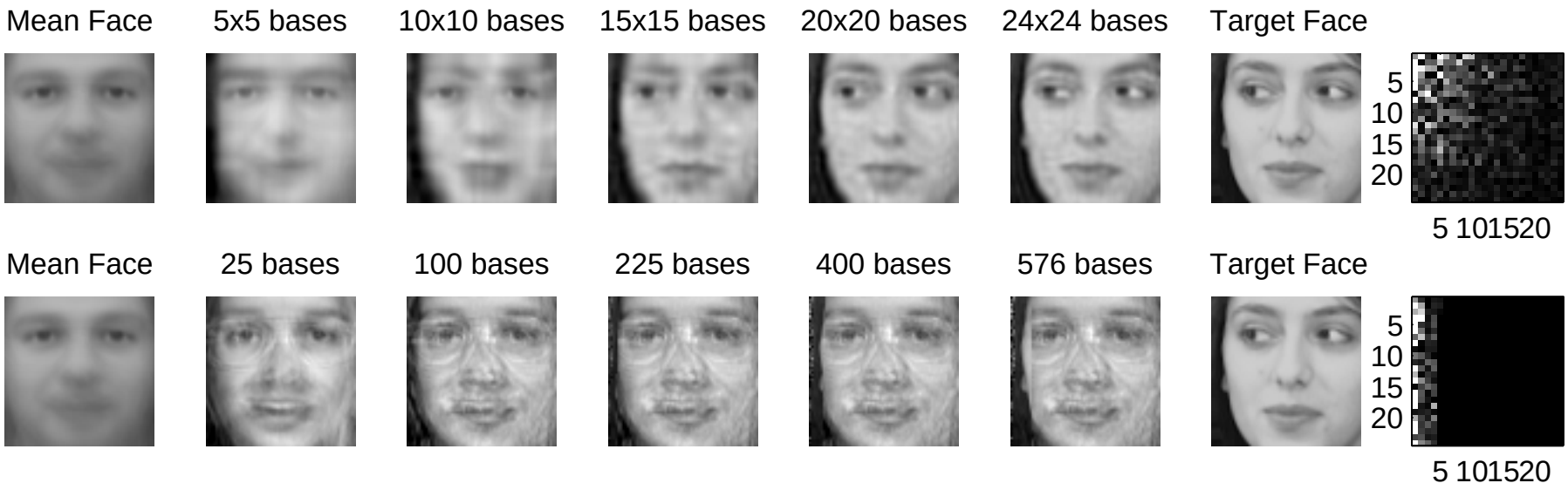
500 replicates, for random partition into training-test subsets, are performed to compare the mean test error.

	MPCA	PCA
Mean	452	2870
SD	4	43

The error is defined as the Frobenius norm for two images.

20 test faces randomly drawn (rows 1-2), reconstructions by MPCA (rows 3-4) and PCA (rows 5-6).

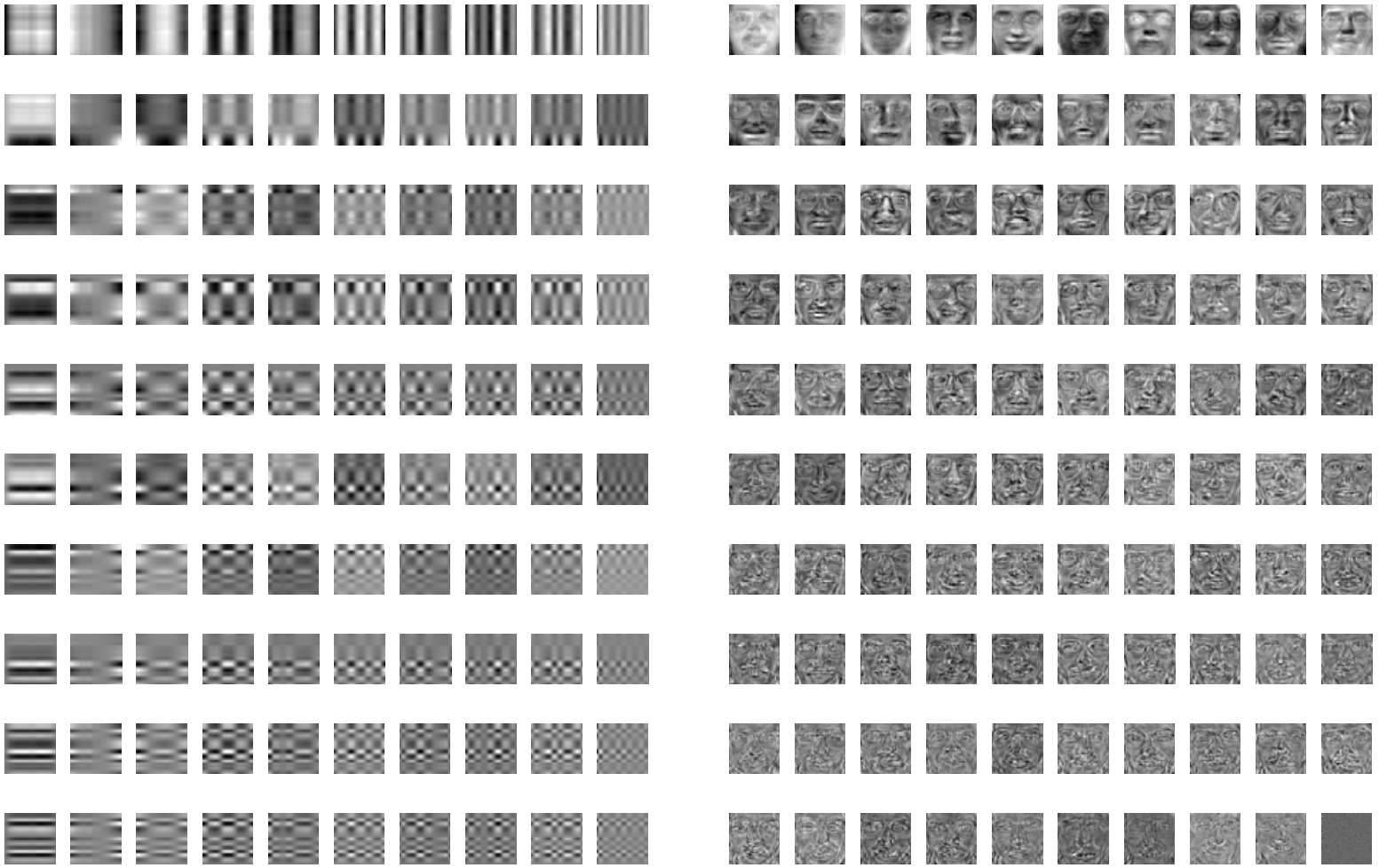
Different performances of MPCA and PCA



Test image reconstruction, MPCA (top) and PCA (bottom).

The image turns its view to left with 10×10 basis elements;
the pupil turns to the left with 15×15 basis elements;
nostrils and folds of eyelids show up with 20×20 basis elements;
the facial curves become clear when 24×24 basis elements.

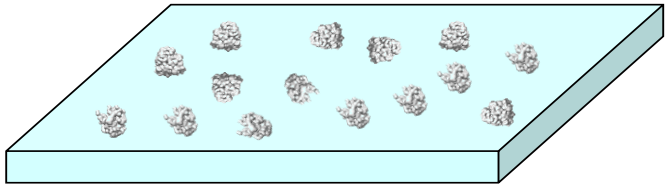
Leading 100 bases, MPCA ($b \otimes a \in \mathbb{R}^{4096}$) vs PCA ($\gamma \in \mathbb{R}^{4096}$)



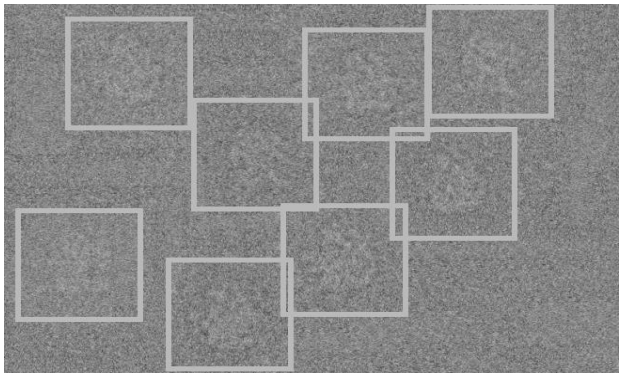
MPCA: more module oriented. PCA: too much information in a basis.

cryo-EM image analysis

Cryo Imaging

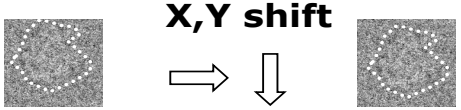


Particle Boxing



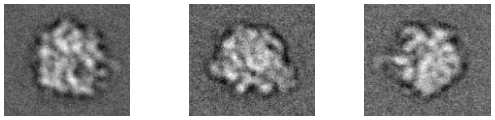
Alignment

X,Y shift

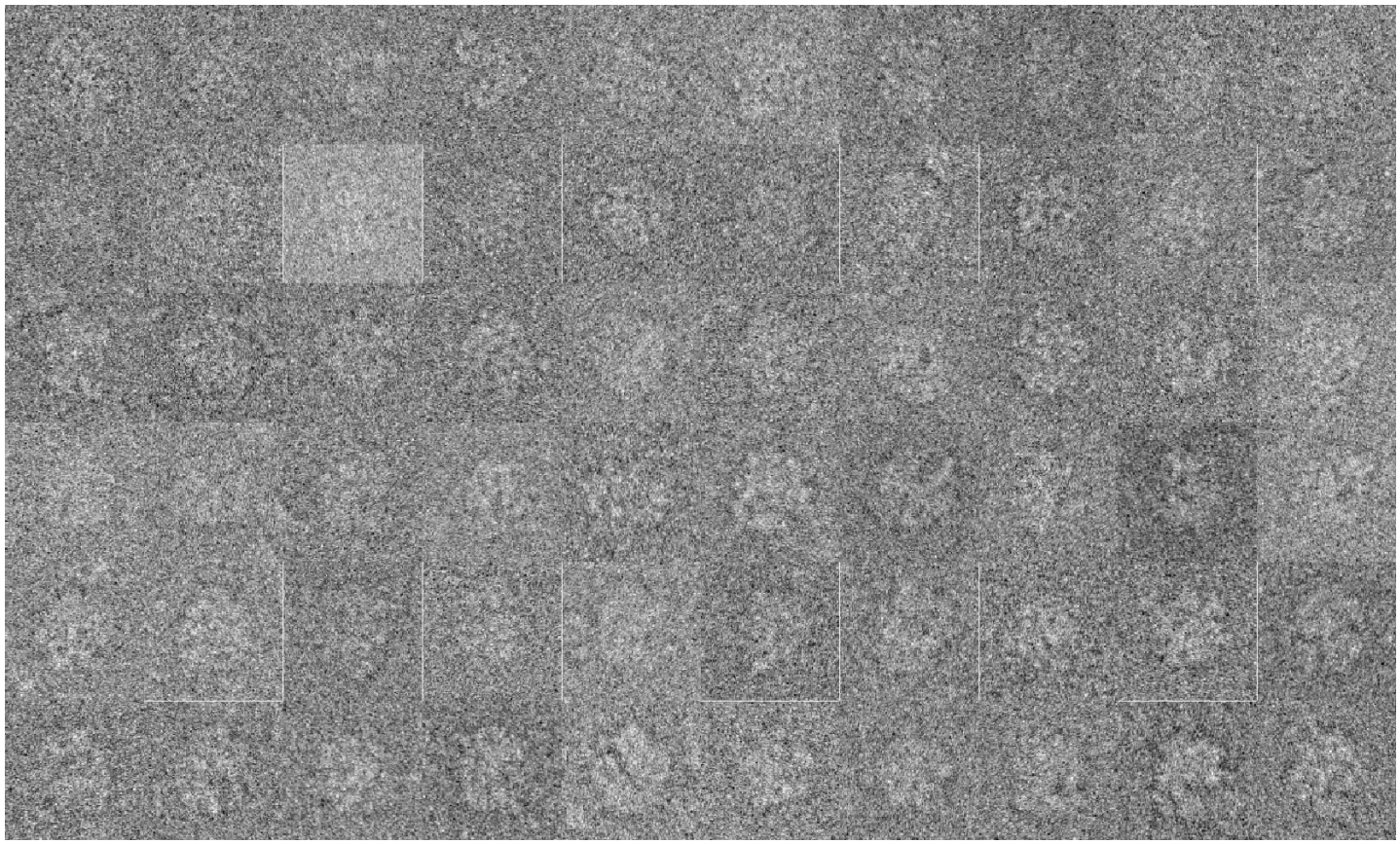


In-plane rotation

Clustering and Average

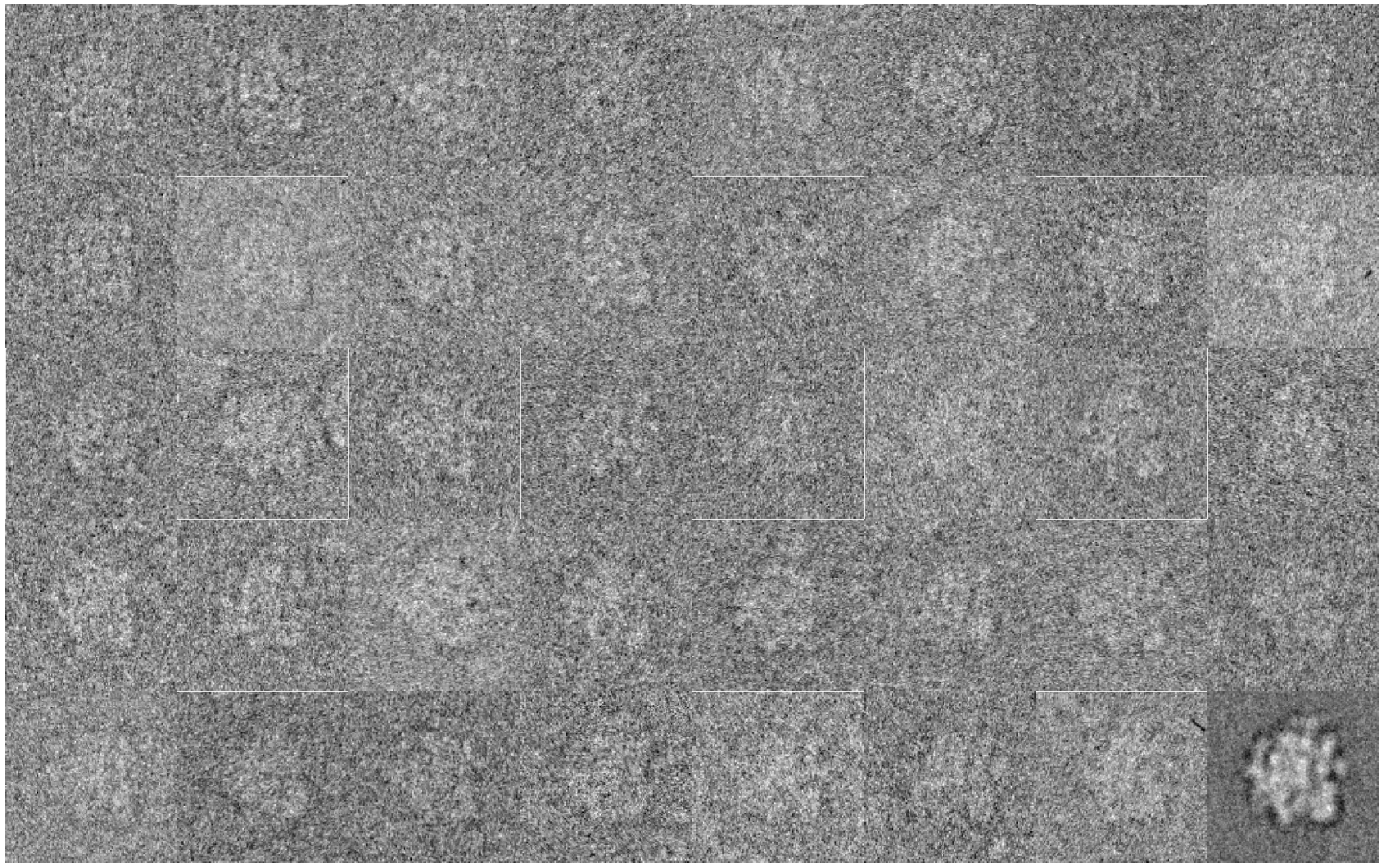


5000 Ribosome cryo-EM images



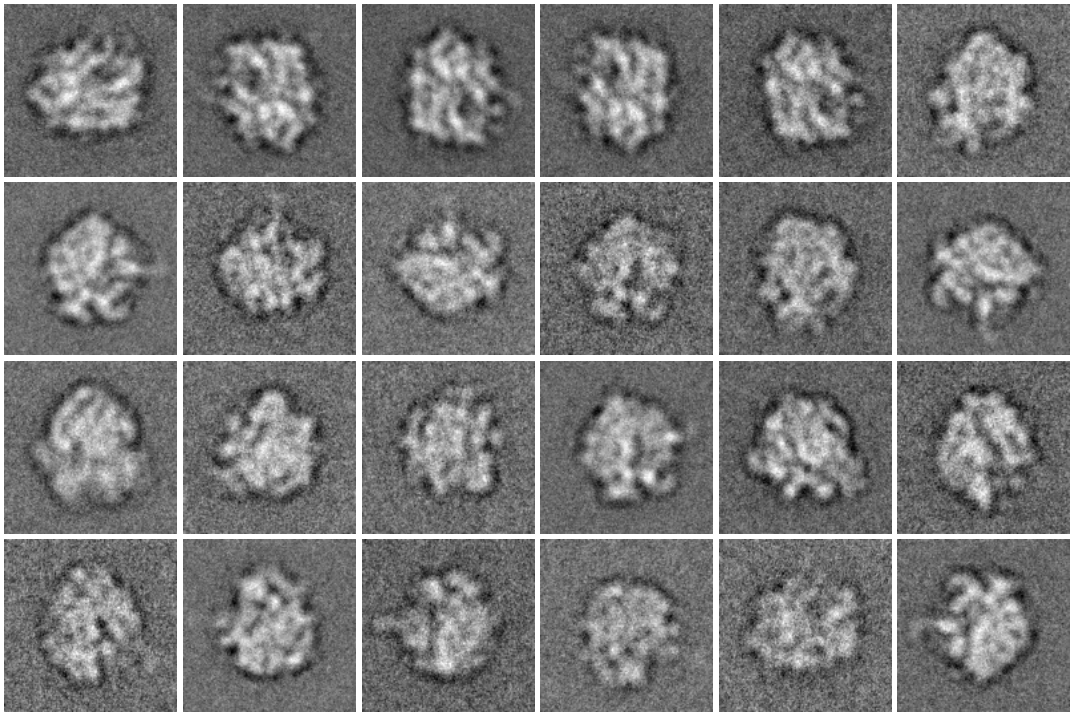
MPCA for dimension reduction, then followed by clustering analysis on reduced core matrices.

One cluster of images

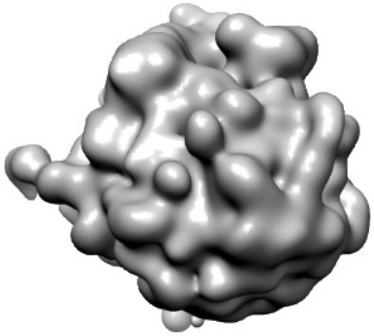


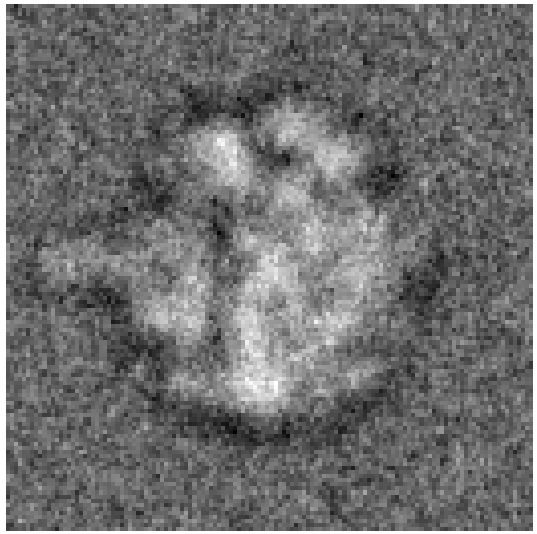
denoising by taking cluster average

24 Cluster Averages for Ribosome Data



One Interesting Example





Statistical Theory

Why MPCA is successful?

Usual probabilistic PCA model on $\mathrm{vec}(X)$

- Assume $\mathrm{vec}(X-\mu) = \widetilde{\Gamma}\nu + \mathrm{vec}(\varepsilon)$, $\mu = E(X)$: mean
  - $\widetilde{\Gamma}_{m\times r}$: Basis (PCs) of interest, $m = pq$, $r \ll m$;
  - ν: PCA scores (random r-vector);
  - ν and ε are stochastically independent;
  - ε: random error with $E(\mathrm{vec}(\varepsilon)) = 0$, $\mathrm{Cov}(\mathrm{vec}(\varepsilon)) = \sigma^2 I$.
- $\left\{X_i \in \mathbb{R}^{p\times q}\right\}_{i=1}^n$: data set, $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$
- Minimize $\frac{1}{n}\sum_{i=1}^n \|\mathrm{vec}(X_i - \bar{X}) - \widetilde{\Gamma}\nu_i\|^2$ over $\widetilde{\Gamma} \in \mathcal{O}_{m\times r}$, $\nu \in \mathbb{R}^r$
  - $\widetilde{\Gamma}$: **leading eigenvectors** of $\frac{1}{n}\sum \mathrm{vec}(X_i - \bar{X})\mathrm{vec}(X_i - \bar{X})^T$

MPCA model & rationale

Recall PCA model: $\text{vec}(X - \mu) = \widetilde{\Gamma}\nu + \text{vec}(\varepsilon)$, no structure on span($\widetilde{\Gamma}$) except for the orthonormality.

$$\text{MPCA}: \qquad X - \mu = A_0 U B_0^T + \varepsilon.$$

- **Column** basis $A_0 \in \mathcal{O}_{p \times p_0}$ and **row** basis $B_0 \in \mathcal{O}_{q \times q_0}$.
- Random coordinate (or MPCA score) U,
  $E(UU^T)$, $E(U^TU)$: nonsingular, distinct characteristic roots.
- U and ε: independent; $E(\varepsilon) = 0$ and $\text{Cov}\{\text{vec}(\varepsilon)\} = \sigma^2 I$
- $\text{vec}(A_0 U B_0^T) = (\boldsymbol{B_0 \otimes A_0})\text{vec}(U)$

  MPCA: $\Gamma := B_0 \otimes A_0$, **structured Γ (cf. non-structured $\widetilde{\Gamma}$)**

- PCA: $\text{vec}(X-\mu) = \widetilde{\Gamma}\nu + \text{vec}(\varepsilon)$, no structure on span($\widetilde{\Gamma}$) except for in $\mathcal{O}_{m\times r}$, $m = pq$. **Order $O(pq)$ vs order $O(p+q)$.**

- **Kronecker envelope** span$(B_0 \otimes A_0)$: unique minimal subspace such that $\widetilde{\Gamma} \subseteq$ span$(B_0 \otimes A_0)$ (Li et al., 2010)

  - You can always have $\widetilde{\Gamma} \subseteq$ span$(I_q \otimes I_p)$, or span$(Q \otimes P)$.

  - $\widetilde{\Gamma} = (B_0 \otimes A_0)G \Rightarrow \widetilde{\Gamma}\nu = (B_0 \otimes A_0)G\nu$

  - $G\nu$ folded into a $p_0 \times q_0$ matrix U, then

  - $\widetilde{\Gamma}\nu = (B_0 \otimes A_0)G\nu = \text{vec}(A_0 U B_0^T) \leftarrow$ this is MPCA model

- MPCA: $\Gamma = B_0 \otimes A_0$, structured Γ; it uses a **structured larger** subspace to enclose $\widetilde{\Gamma}$. span$(B_0 \otimes A_0) \supseteq$ span$(\widetilde{\Gamma})$;

  fewer parameters for span$(B_0 \otimes A_0)$, $pp_0 + qq_0 - \frac{p_0(p_0+1)+q_0(q_0+1)}{2}$

Dimensionality required

MPCA requires less parameters

Number of required parameters at $(p, q) = (10, 10)$ *and* $p_0 = 5$

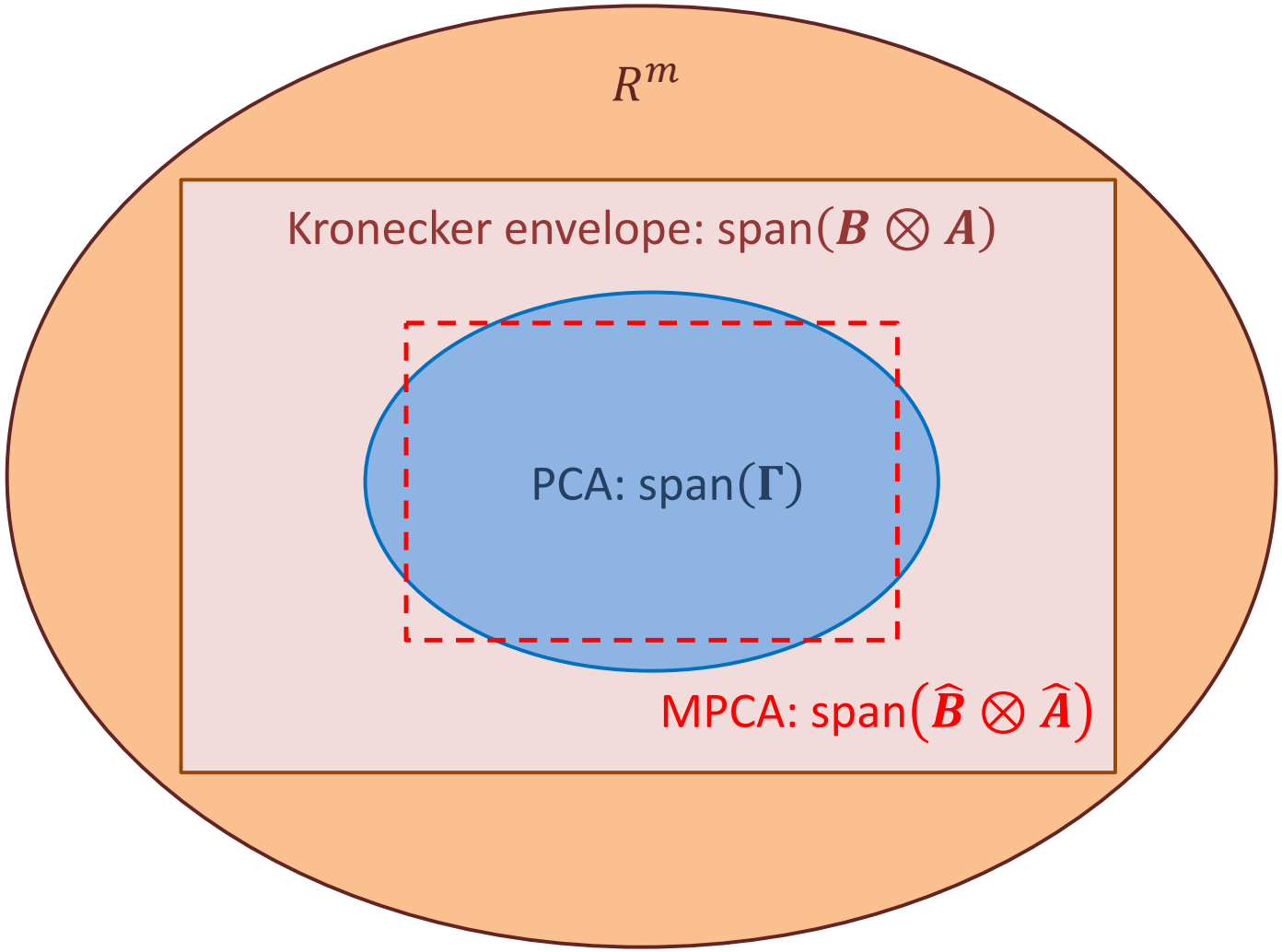
q_0	1	2	3	4	5
MPCA	44	52	59	65	70
PCA	485	945	1380	1790	2175

$\rightarrow$ More efficient for small sample size

MPCA is a **bigger subspace**, which is a Kronecker envelope, to contain a smaller PCA subspace. Because of its **Kronecker product structure**, it requires **fewer parameters** to specify this subspace.

This is an approach of trading bias for variance.

Pictorial illustration for Kronecker envelope, $m = p \times q$.



Chen, T.L. et al. (2014). γ-SUP: a clustering algorithm for cryo-electron microscopy images of asymmetric particles. *Ann. Applied Statist.* → MPCA application to cryo-em image clustering

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21, 1253-1278. → high-order SVD

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000b). On the best rank-1 and rank-$(R_1, R_2, \ldots, R_N)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21, 1324-1342. → best specified-rank tensor decomposition

Hung, H., Wu, P.S., Tu, I.P. and Huang, S.Y. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, 99, 569-583. → MPCA in statistical framework

Hung, H. and Wang, C.C. (2012). Matrix variate logistic regression model with application to EEG data. to appear in *Biostatistics*. → a special-case tensor regression model

Li, B., Kim, M.K. and Altman, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *Annals of Statistics*, 38, 1094-1121.→ Kronecker envelope

Lu, H., Plataniotis, K.N. and Venetsanopoulos, A.N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks*, 19, 18-39. → MPCA

Tyler, D.E. (1981). Asymptotic inference for eigenvectors. *Annals of Statistics*, 9, 725-736. → Technique for asymptotics

Ye, J. (2005). Generalized low rank approximations of matrices. *Machine Learning*, 61, 167-191, 2005. → implementation algorithm, GLRAM (iterative alternating LS)

Zhang, D. and Zhou, Z.H. (2005). $(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, 69, 224-231. → $(2D)^2$PCA, before the appearing of MPCA

Thank You for Your Attention

Inclusion properties between (A, B) and (A_0, B_0)

Proposition 1. *(a) If* $\tilde{p} \geq p_0$ *and* $\tilde{q} \geq q_0$, *then* $\text{span}(A) \supseteq \text{span}(A_0)$ *and* $\text{span}(B) \supseteq \text{span}(B_0)$.

(b) If $\tilde{p} < p_0$ *and* $\tilde{q} \geq q_0$, *then* $\text{span}(A) \subsetneq \text{span}(A_0)$ *and* $\text{span}(B) \supseteq \text{span}(B_0)$.

(c) If $\tilde{p} \geq p_0$ *and* $\tilde{q} < q_0$, *then* $\text{span}(A) \supseteq \text{span}(A_0)$ *and* $\text{span}(B) \subsetneq \text{span}(B_0)$.

(d) If $\tilde{p} < p_0$ *and* $\tilde{q} < q_0$, *then* $\text{span}(A) \subsetneq \text{span}(A_0)$ *and* $\text{span}(B) \subsetneq \text{span}(B_0)$.

- MPCA targets $\text{span}(A_0)$ & $\text{span}(B_0)$ when $(\tilde{p}, \tilde{q}) = (p_0, q_0)$.
- With $\tilde{p} > p_0$ or $\tilde{q} > q_0$ being over-specified, MPCA subspace still contains $\text{span}(A_0)$ or $\text{span}(B_0)$ as proper subspace.
- When $(\tilde{p}, \tilde{q}) < (p_0, q_0)$, MPCA finds subspaces of the true.

Asymptotic properties of MPCA

- MPCA are functions of sample covariance matrix S_n,
  as it finds $\min_{A \in \mathcal{O}_{p\times\tilde{p}}, B\in\mathcal{O}_{q\times\tilde{q}}} \text{trace}\left((B\otimes A)^T S_n (B\otimes A)\right)$.
- X_i iid copies of X with finite 4^{th} moments, $\text{Cov}(\text{vec}(X)) = \Sigma$

  CLT: $\sqrt{n}(S_n - \Sigma) \xrightarrow{d} N$, $\text{vec}(N) \sim \mathcal{N}(0, \Sigma_N)$

  $\Sigma_N = (I_{m^2} + K_{m,m})(\Sigma \otimes \Sigma)$, if $X \sim \mathcal{N}$, K: commutation matrix

Basic technique: based on the CLT result above, plus

- Delta method $\sqrt{n}\,(f(S_n) - f(\Sigma)) \xrightarrow{d} \mathcal{N}(0, \Sigma_f)$,

  where $\Sigma_f = \nabla f(\Sigma) \cdot \Sigma_N \cdot \nabla f(\Sigma)^T$
- **calculation of gradients on structured subsets**

  $\nabla f(\Sigma) := \frac{\partial f}{\partial \text{vec}(\Sigma)^T}$

Weak convergence of MPCA components $(\hat{A}, \hat{B})$

For $(\tilde{p}, \tilde{q}) \leq (p_0, q_0)$,* we have the limiting distribution

$$\sqrt{n}\left(\begin{bmatrix} \mathrm{vec}(\hat{A}) \\ \mathrm{vec}(\hat{B}) \end{bmatrix} - \begin{bmatrix} \mathrm{vec}(A) \\ \mathrm{vec}(B) \end{bmatrix}\right) \xrightarrow{d} D_{H_{\tilde{p},\tilde{q}}}\,\mathrm{vec}(N),$$

where $\left([A; B] := [{}^{A}_{B}]\right)$

$$D_{H_{\tilde{p},\tilde{q}}} := \left[\frac{\partial a_1}{\partial \mathrm{vec}(\Sigma)^T}\ ;\ \cdots\ ;\ \frac{\partial a_{\tilde{p}}}{\partial \mathrm{vec}(\Sigma)^T}\ ;\ \frac{\partial b_1}{\partial \mathrm{vec}(\Sigma)^T}\ ;\ \cdots\ ;\ \frac{\partial b_{\tilde{q}}}{\partial \mathrm{vec}(\Sigma)^T}\right].$$

When $(\tilde{p}, \tilde{q}) = (p_0, q_0)$, for $i = 1, \cdots, p_0$ and $j = 1, \cdots, q_0$,

$$\frac{\partial a_i}{\partial \mathrm{vec}(\Sigma)^T} = \left\{a_i \otimes \mathrm{vec}(P_{B_0}) \otimes (\lambda_i I_p - E[X P_{B_0} X^T])^+\right\}^T (K_{p,q} \otimes I_{pq})$$

$$\frac{\partial b_j}{\partial \mathrm{vec}(\Sigma)^T} = \left\{b_j \otimes \mathrm{vec}(P_{A_0}) \otimes (\xi_j I_q - E[X^T P_{A_0} X])^+\right\}^T (I_{pq} \otimes K_{p,q}).$$

*For $\tilde{p} > p_0$ or $\tilde{q} > q_0$ the eigenvalues are multiple roots and the tensor principal components are not uniquely determined.

Asymptotic efficiency of MPCA

- MPCA and (2D)^{2}PCA target the same subspace
- In favor of MPCA since it is less noise-contaminated.
- MPCA: $E[(X-\mu)P_{B_0}(X-\mu)^T]$, $E[(X-\mu)^T P_{A_0}(X-\mu)]$
  (2D)^{2}PCA: $E[(X-\mu)(X-\mu)^T]$, $E[(X-\mu)^T(X-\mu)]$

Theorem 1. *Let* $(\tilde{p},\tilde{q}) = (p_0,q_0)$ *and let* $(\tilde{A},\tilde{B})$ *be the* (2D)^2*PCA components under* (p_0,q_0). *Assume also the normality of* vec(X).* *Then, ("aCov" stands for asymptotic Cov)*

$$\text{aCov}\Big(\text{vec}(P_{\tilde{B}\otimes\tilde{A}})\Big) - \text{aCov}\Big(\text{vec}(P_{\hat{B}\otimes\hat{A}})\Big) \geq 0.$$

*It is not difficult to get that $\hat{A}$ (or $\hat{B}$) is more efficient than $\tilde{A}$ (or $\tilde{B}$, resp). Difficulty arises when considering $\hat{B}\otimes\hat{A}$ due to the correlation between $\hat{A}$ and $\hat{B}$.

Expression of differentials

Sketch of the proof (Lemma 1):

1. (A, B) satisfies the set of stationary equations

$$\left(\sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \Sigma (b_j \otimes I_p) \right) a_i = \lambda_i a_i, \ i = 1, \cdots, \tilde{p},$$
$$\left(\sum_{i=1}^{\tilde{p}} (I_q \otimes a_i)^T \Sigma (I_q \otimes a_i) \right) b_j = \xi_j b_j, \ j = 1, \cdots, \tilde{q}.$$

2. Let Σ be perturbed to $\Sigma + \epsilon\dot{\Sigma}$ with the corresponding perturbed $a_i + \epsilon\dot{a}_i$, $\lambda_i + \epsilon\dot{\lambda}_i$, and $b_j + \epsilon\dot{b}_j$.

Expression of differentials

3. Equating the terms with order ϵ we deduce that, for $i = 1, \cdots, \tilde{p}$,

$$\begin{aligned} \dot{\lambda}_i &= a_i^T \dot{\Sigma}_B a_i, \\ \dot{a}_i &= \{\lambda_i I_p - \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \Sigma (b_j \otimes I_p)\}^+ \dot{\Sigma}_B a_i, \end{aligned}$$

where

$$\dot{\Sigma}_B = E[X(\dot{B}B^T + B\dot{B}^T)X^T] + \sum_{j=1}^{\tilde{q}} (b_j \otimes I_p)^T \dot{\Sigma} (b_j \otimes I_p)$$

with $\dot{B} = [\ \dot{b}_1, \cdots, \dot{b}_{\tilde{q}}\ ]$ satisfying $\dot{B}^T B + B^T \dot{B} = 0$.

The red part makes the derivations rather complicated. If it vanishes, then the proof is completed by the usual method. When?

Expression of $D_{\Phi(\tilde{p},\tilde{q})}$

Note that $\sum_{i=1}^{\tilde{p}} a_i^T \dot{\Sigma}_B a_i$ can be expressed as

$$\sum_{i=1}^{\tilde{p}} a_i^T E[X(\dot{B}B^T + B\dot{B}^T)X^T]a_i$$

$$= \sum_{j=1}^{\tilde{q}} \left(\dot{b}_j^T E[XP_A X^T] b_j + b_j^T E[XP_A X^T] \dot{b}_j \right) = 0$$

by noting that b_j is an eigenvector of $E[XP_AX^T]$ and $b_j^T \dot{b}_j = 0$. Thus,

$$\sqrt{n} \left(\hat{\Phi}(\tilde{p},\tilde{q}) - \Phi(\tilde{p},\tilde{q}) \right) \xrightarrow{d} D_{\Phi(\tilde{p},\tilde{q})} \, \mathrm{vec}(N),$$

Expression of $D_{H_{\tilde{p},\tilde{q}}}$

When $(\tilde{p},\tilde{q}) = (p_0, q_0)$, there must exist a nonsingular matrix η such that $B_0 = B\,\eta$. From $X = A_0 U B_0^T + \varepsilon$,

$$\begin{aligned} & E[X(\dot{B}B^T + B\dot{B}^T)X^T] \\ = \; & E[A_0 U \eta^T (B^T\dot{B} + \dot{B}^T B)\eta U^T A_0^T] + \sigma^2 \mathrm{trace}(B^T\dot{B} + \dot{B}^T B) I_p \\ = \; & 0 \end{aligned}$$

by noting that $B^T\dot{B} + \dot{B}^T B = 0$. Thus,

$$\sqrt{n}\left(\begin{bmatrix} \mathrm{vec}(\widehat{A}) \\ \mathrm{vec}(\widehat{B}) \end{bmatrix} - \begin{bmatrix} \mathrm{vec}(A) \\ \mathrm{vec}(B) \end{bmatrix}\right) \xrightarrow{d} D_{H_{\tilde{p},\tilde{q}}} \mathrm{vec}(N),$$

This is NOT true when $(\tilde{p},\tilde{q}) < (p_0, q_0)$. We still can't solve this problem!