Dimension Reduction

2018 Statistics Summer School

I-Ping Tu

Institute of Statistical Science, Academia Sinica

July 31, 2018

The 'Learning Evolution' of Statistics

- Mean (medium,...)$\rightarrow$ Mean Vector
- Variance $\rightarrow$ Covariance Matrix
- What can you do about the covariance matrix?
- Ans: Eigenvalue Decomposition!

Review of Covariance Matrix

- Let $x_1, \cdots, x_n$ be length-p observation vectors
- WLOG, let their mean be length-p 0-vector.
- Let the data matrix $X = (x_1, \cdots, x_n)$ be a p by n matrix.
- The sample covariance matrix

  $\Sigma = XX^T/(n-1) = \sum_{i=1}^{n} x_i x_i^T/(n-1)$.

A Proof of $XX^T = \sum_{i=1}^{n} x_i x_i^T$

- By $X = (x_1, \cdots, x_n)$, we have $X_{ij} = (x_j)_i$.

$$
\begin{aligned}
(XX^T)_{jk} &= \sum_{i=1}^{n} X_{ji} X_{ki} \\
&= \sum_{i=1}^{n} (x_i)_j (x_i)_k \\
&= \sum_{i=1}^{n} (x_i x_i^T)_{jk} \\
&= (\sum_{i=1}^{n} x_i x_i^T)_{jk}
\end{aligned}
$$

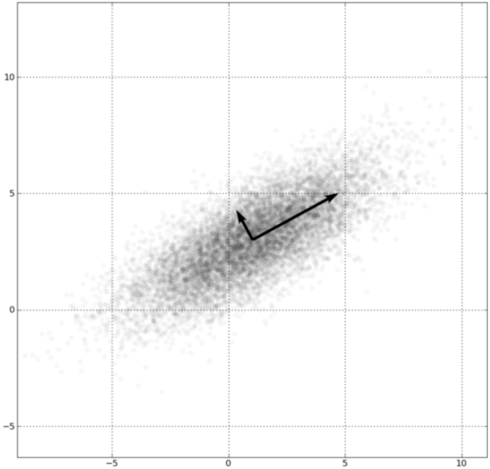
An Intuition Definition for PCA

- PCA: Principal Component Analysis
- The underlying statistical philosophy is "Larger variance captures more information."

An Intuition Definition for PCA

- Find a direction vector ($\in R^p$) p_1, such that the variance of the data $\{x_i\}_{i\leq n}$ projected to this direction $\{x_i^T p_1\}_{i\leq n}$ has maximum variance.
- Then find p_2 orthonormal to p_1, such that the variance of $\{x_i^T p_2\}_{i\leq n}$has maximum variance.
- $\cdots$ find p_k orthonormal to $p_1, \cdots, p_{k-1}$, such that the variance of $\{x_i^T p_k\}_{i\leq n}$ has maximum variance.

An Example with $p = 2$



Eigenvalue Decomposition

- Given a symmetric p.d. matrix $\Sigma_{p\times p}$, Σ can be decomposed as

$$\Sigma = \sum_{i=1}^{p} \lambda_i u_i u_i^t,$$

where $u_1, \cdots, u_p$ are orthonormal and $\lambda_1 \geq \cdots \geq \lambda_p$. Furthermore, u_i can be uniquely decided up to $+/-$ sign if eigenvalue λ_i is distinct.

- $$\Sigma u_i = \lambda_i u_i$$

- $$\Sigma U = U \Lambda$$

The Intuition Definition vs Eigenvalue Decomposition

- $p_1 = \mathrm{argmax}_q \sum_{i=1}^{n} (x_i^T q)^2$
- Will $p_1 = u_1$?

The Intuition Definition vs Eigenvalue Decomposition

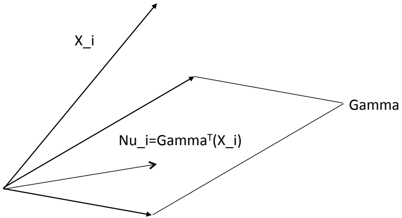
$$
\begin{aligned}
p_1 &= \underset{|q|=1}{\operatorname{argmax}} \sum_{i=1}^{n} (x_i^T q)^2 = \underset{|q|=1}{\operatorname{argmax}} \sum_{i=1}^{n} (q^T x_i)(x_i^T q) \\
&= \underset{|q|=1}{\operatorname{argmax}} \sum_{i=1}^{n} q^T (x_i x_i^T) q = \underset{|q|=1}{\operatorname{argmax}}\, q^T (\sum_{i=1}^{n} x_i x_i^T) q \\
&= \underset{|q|=1}{\operatorname{argmax}}\, q^T (X X^T) q = \underset{|q|=1}{\operatorname{argmax}}\, q^T (\sum_{i=1}^{p} \lambda_i u_i u_i^t) q \\
&= \underset{|q|=1}{\operatorname{argmax}} \sum_{i=1}^{p} \lambda_i (u_i^t q)^2 \\
&= u_1
\end{aligned}
$$

PCA Model

- $X \in \mathbb{R}^p$ is a random vector.
- $X = \Gamma\nu + \varepsilon$.
- $\min_{\Gamma \in \mathcal{O}_{p\times p'}, \nu \in \mathbb{R}^{p'}} E[||X - \Gamma\nu||_F^2]$.
- Sample version
    - Assume the mean $\bar{X}$ has been taken off.
    - $\min_{\Gamma \in \mathcal{O}_{p\times p'}, \nu_i \in \mathbb{R}^{p'}, i \leq n} \frac{1}{n} \sum_{i=1}^n ||X_i - \Gamma\nu_i||^2$.

PCA vs Linear Regression: A Geometric Point of View, after Γ is chosen.

Geometric interpretation

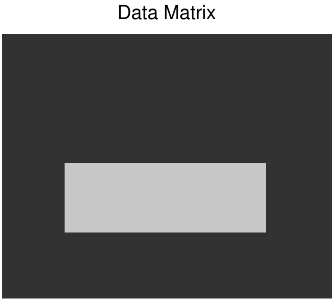


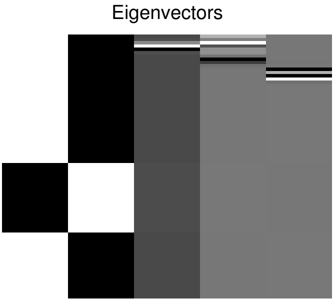
Ordinary Least-Squares

PCA Demo: Data Matrix ($p \times n$)

PCA Demo:

Data Matrix vs the first 5 Eigenvectors





PCA Demo: Data Matrix ($p \times n$)

PCA Demo:

Data Matrix vs the first 5 Eigenvectors

Data Matrix

Eigenvectors

What's SVD

- Given a $p \times n$ real value matrix X, $X = UDV^T$,
- U: a $p \times p$ orthonormal matrix,
- V: a $p \times p$ orthonormal matrix,
- $D : p \times q$ and D's nonzero elements only appear at diagonal with size $\min(p, q)$.

SVD: An Optimization Formula

- Given rank k, SVD can be formulated as an optimization problem.

$$(U_{1:k,\cdot}, D_{1:k,1:k}, V_{1:k,\cdot}) = \underset{A \in \mathcal{O}_{p\times k}, B \in \mathcal{O}_{q \times k}, D \in R^{kxk}}{\operatorname{argmin}} \|X - ADB^T\|_F^2$$

A Neat Formula of SVD

- Let the column vectors of U and V be $u_1, \cdots, u_k$ and $v_1, \cdots, v_k$.
- Let the diagonal elements of D be $d_1, \cdots, d_k$.
- Then, $\mathbf{X} = \mathbf{UDV}^T = \sum_{i=1}^{k} d_i u_i v_i^T$.
- Each $u_i v_i^T$ represents a rank one $p \times q$ matrix.

A Transformation Point of View: $\mathbf{X} = UDV^T$

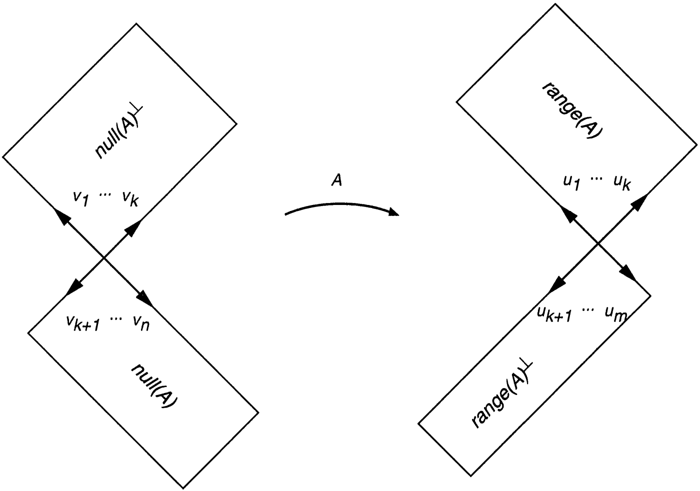


Figure 1. Strang's diagram.

What does SVD say?

Any real value transformation matrix from R^q to R^p can be decomposed to 3 steps:

- A rotation matrix V in R^q.
- A scale matrix D.
- A rotation matrix U in R^p.

A Neat Formula of SVD

- Let the column vectors of U and V be $u_1, \cdots, u_k$ and $v_1, \cdots, v_k$.
- Let the diagonal elements of D be $d_1, \cdots, d_k$.
- Then, $\mathbf{X} = \mathbf{UDV}^T = \sum_{i=1}^k d_i u_i v_i^T$.
- SVD finds the k vector-pairs with one to one mapping from R^q to R^p.
- Each $u_i v_i^T$ represents a rank one matrix with dimension $p \times q$.

SVD vs PCA

- If by SVD, $X = UDV^T$, where U is an orthonormal matrix such that $UU^T = I_p$
- By PCA, U is the eigenmatrix of $XX^T = UD^2U^T$.
- Furthermore, V is the eigenmatrix of $X^TX = VD^2V^T$.

Application of SVD: Dimension Reduction

- $\mathbf{X}_{p\times q} = \mathbf{U}_{p\times p}\mathbf{D}_{p\times q}\mathbf{V}_{q\times q} \approx \tilde{\mathbf{U}}_{p\times k}\tilde{\mathbf{D}}_{k\times k}\tilde{\mathbf{V}}^T_{k\times q}$
- k : selected rank.

196 X 257

Rank =100

Another Point of View

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{k} d_i u_i v_i^T$$

- A linear combination of bases: $\{u_i v_i^T,\ 1 \leq i \leq k\}$.
- Each $u_i v_i^T$ is a rank one matrix.

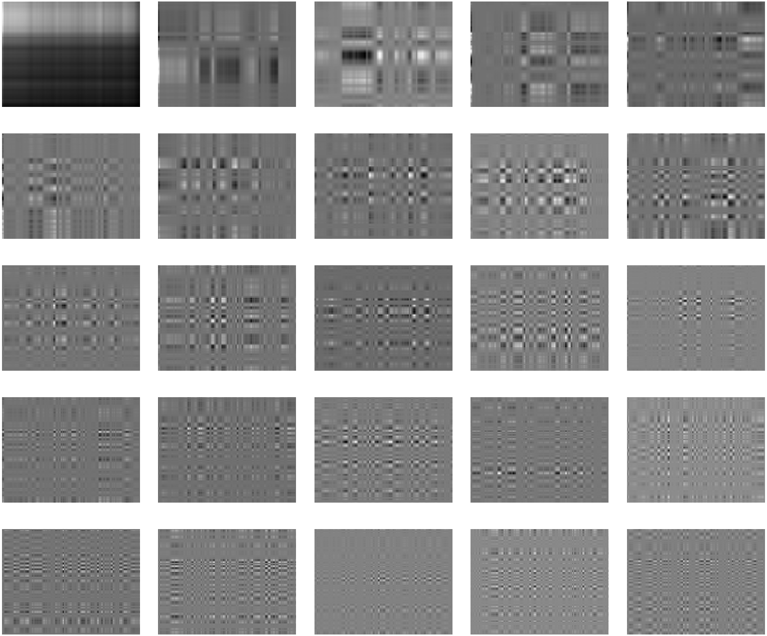
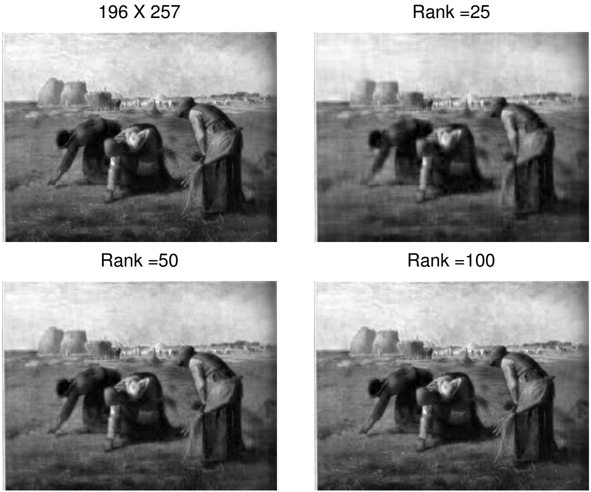


Figure: The first 25 rank-one bases.

SVD Rank Demo: The more, the better.



SVD Rank Demo: The more, the better?

196 X 257

Rank =25

Rank =50

Rank =100

Gleaners and Angelus

Angelus

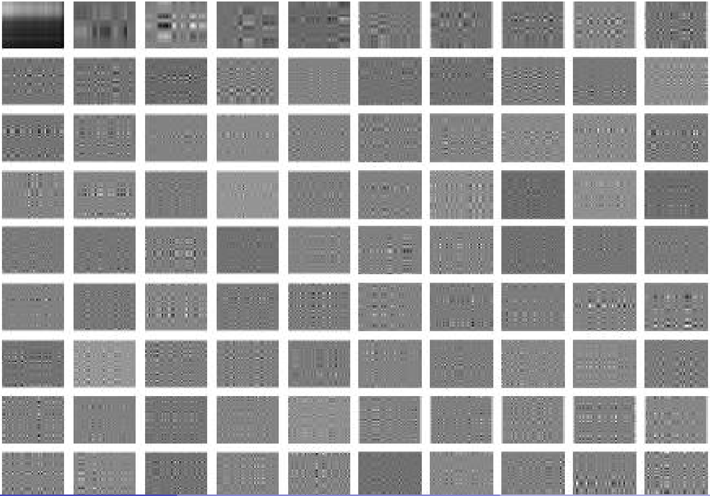
Gleaner

Angelus Reconstructed by Gleaners

Gleaner

Angelus Reconstructed by Gleaner

Gleaners Basis 10 x 10



MPCA, HOSVD: multi-matrix versions of SVD

- We have matrix $A_1, \cdots, A_n$, all from R^q to R^p.
- We want to stick on one set of U and V by sacrificing some bias.
- We relax the condition on the diagonal property of D.

Multilinear Principal Component Analysis (MPCA)

- We define **column** basis $A \in \mathcal{O}_{p\times\tilde{p}}$, **row** basis $B \in \mathcal{O}_{q\times\tilde{q}}$.
- SVD: Find A, B(with $\tilde{p} = \tilde{q}$) and a diagonal matrix D that

$$\min_{A,B,D} \|X - ADB^T\|_F^2$$

.

- MPCA: Find simultaneous A, B, and U_i by

$$\min_{A,B,U_i,1\le i\le n} \frac{1}{n}\sum_{i=1}^{n} \|(X_i - \bar{X}) - AU_iB^T\|_F^2$$

- D diagonal and U_i usually not.

Implementation: iterative alternating eigenvalue decompositions(Ye, 2005)

- $\hat{\boldsymbol{A}}$: leading $\tilde{p}$ eigenvectors of $\hat{\Sigma}_{\hat{B}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})P_{\hat{B}}(X_i - \bar{X})^T$.

  solve a small eigenvalue problem: a $p \times p$ matrix, where $P_{\hat{B}} = \hat{B}\hat{B}^T$
- $\hat{\boldsymbol{B}}$: leading $\tilde{q}$ eigenvectors of $\hat{\Sigma}_{\hat{A}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^T P_{\hat{A}}(X_i - \bar{X})$.

  solve a small eigenvalue problem: a $q \times q$ matrix, where $P_{\hat{A}} = \hat{A}\hat{A}^T$.
- **Iterative alternating until convergence.**

Experimental Setting

- 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.

Basis from the training set

- 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.

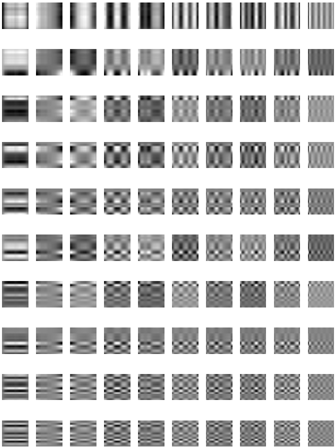
	MPCA	PCA
Mean	452	2870
SD	4	43

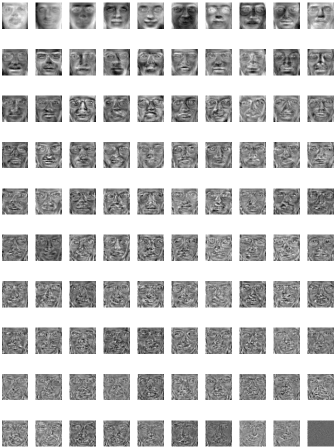
The error is defined as the Frobenius norm for two images.

Visual Comparisons

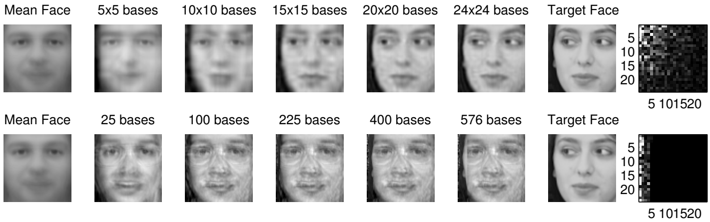
20 test faces randomly drawn (rows 1-2), reconstructions by MPCA (rows 3-4) and PCA (rows 5-6).

Basis Comparisons: MPCA vs PCA



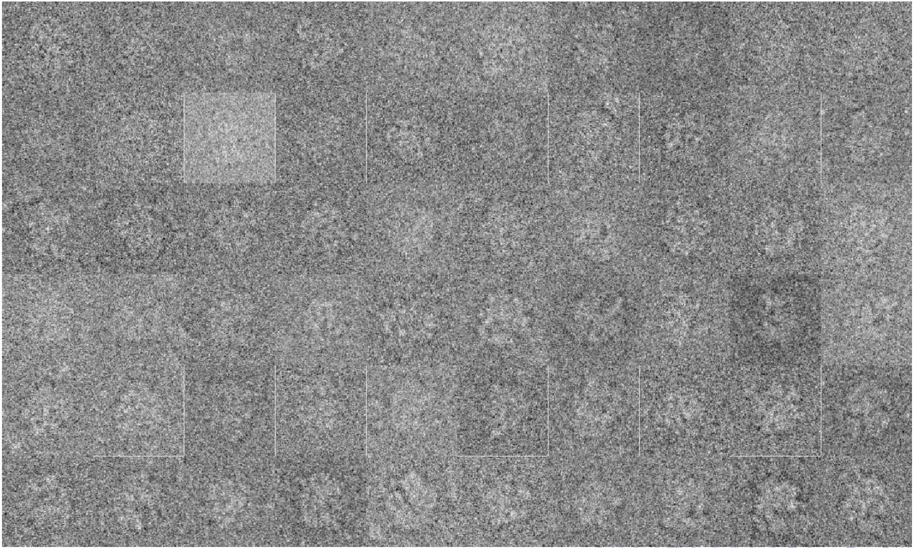


Stepwise Comparison

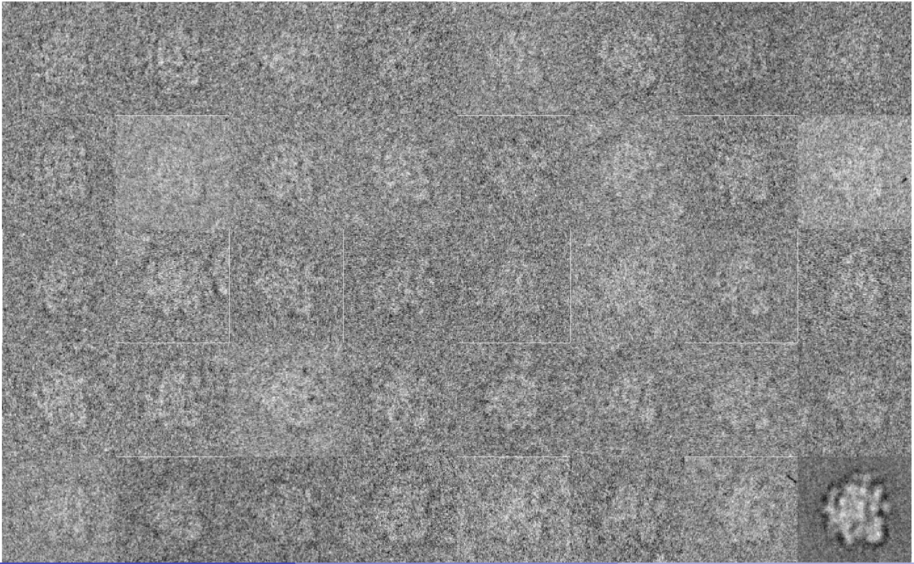


Test image reconstruction, MPCA (top) and PCA (bottom).

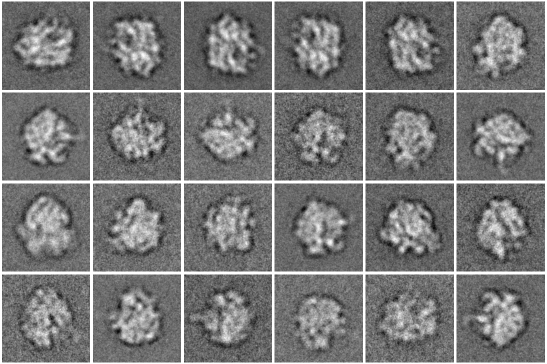
5000 Ribosome cryoEM images



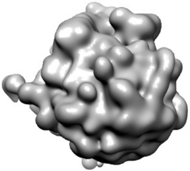
A Cluster Example

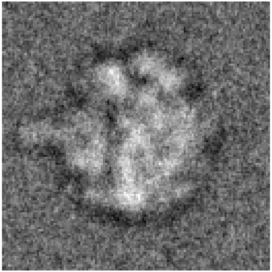


24 Cluster Averages for Ribosome Data



One Interesting Example





Summary: Properties and Applications of

- PCA
- SVD
- MPCA