Lecture 1. PCA and MDS – A Geometric View

Yuan Yao

HKUST

Outline

Principal Component Analysis

- Horn's Parallel Analysis: Random Permutation Test

Multidimensional Scaling (MDS)

Positive Definite Functions and Kernels

- Kernel PCA

Sufficient Dimensionality Reduction

- PCA as Sufficient Dimensionality Reduction

Supervised PCA

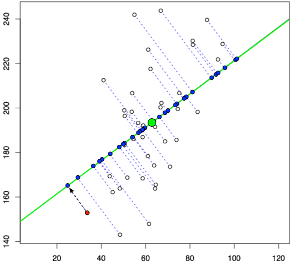
- Linear Discriminant Analysis
- Sliced Inverse Regression
- Localized SIR

Geometric Embedding

- A Fundamental Problem in Data Representation
- Unstructured data ↦ Euclidean Space
  - PCA: high dim ↦ low dim affine space
  - MDS: metric ↦ Euclidean space
- Simple cases for ‘representation’ learning (w.r.t. deep learning)
- image, speech, text, video . . .

Principal Component Analysis (PCA)

- Given n sample points in $\mathbb{R}^p$, i.e. $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$
- Can you find a low dimensional affine representation?



Best k-affine space approximation of data

- Let $X = [x_1, \dots, x_n] \in \mathbb{R}^{p \times n}$.
- Consider
$$\min_{\beta, \mu, U} I := \sum_{i=1}^{n} \|x_i - (\mu + U\beta_i)\|^2 \tag{1}$$
where $U \in \mathbb{R}^{p \times k}$, $U^T U = I_k$, and $\sum_{i=1}^{n} \beta_i = 0$ (nonzero sum of β_i can be represented by μ).

Finding optimal $\hat{\mu}$, $\hat{\beta}$

- Taking the first order optimality condition:

$$\frac{\partial I}{\partial \mu} = -2\sum_{i=1}^{n}(x_i - \mu - U\beta_i) = 0 \Rightarrow \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial I}{\partial \beta_i} = (x_i - \mu - U\beta_i)^T U = 0 \Rightarrow \hat{\beta}_i = U^T(x_i - \hat{\mu}_n)$$

Finding optimal $\hat{U}$

- Plugging in the expression of $\hat{\mu}_n$ and $\hat{\beta}_i$

$$\begin{aligned} I &= \sum_{i=1}^{n} \|x_i - \hat{\mu}_n - UU^T(x_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|x_i - \hat{\mu}_n - P_k(x_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|y_i - P_k(y_i)\|^2, \quad y_i := x_i - \hat{\mu}_n \end{aligned}$$

where $P_k = UU^T$ is a projection operator satisfying the idempotent property $P_k^2 = P_k$.

Finding optimal $\hat{U}$

- Denote $Y = [y_1|y_2|\cdots|y_n] \in \mathbb{R}^{p\times n}$, then the original problem is

$$\begin{aligned}
\min_U \sum_{i=1}^n \|y_i - P_k(y_i)\|^2 &= \min \operatorname{tr}[(Y - P_kY)^T(Y - P_kY)] \\
&= \min \operatorname{tr}[Y^T(I - P_k)(I - P_k)Y] \\
&= \min \operatorname{tr}[YY^T(I - P_k)^2] \\
&= \min \operatorname{tr}[YY^T(I - P_k)] \\
&= \min[\operatorname{tr}(YY^T) - \operatorname{tr}(YY^TUU^T)] \\
&= \min[\operatorname{tr}(YY^T) - \operatorname{tr}(U^TYY^TU)].
\end{aligned}$$

Above we use cyclic property of trace and idempotent property of projection.

Finding optimal $\hat{U}$

- Since Y does not depend on U, the problem above is equivalent to

$$\max_{UU^T=I_k} \frac{1}{n}\operatorname{tr}(U^T Y Y^T U) = \max_{UU^T=I_k} \operatorname{tr}(U^T \hat{\Sigma}_n U) \qquad (2)$$

where $\hat{\Sigma}_n = \frac{1}{n} Y Y^T = \frac{1}{n}(X - \hat{\mu}_n \mathbf{1}^T)(X - \hat{\mu}_n \mathbf{1}^T)^T$ is the sample variance matrix.

- the sample covariance matrix, which is positive semi-definite, has the eigenvalue decomposition $\hat{\Sigma}_n = \hat{U}\hat{\Lambda}\hat{U}^T$, where $\hat{U}^T\hat{U} = I$, $\Lambda = \mathbf{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$, and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_n \geq 0$. Then

$$\max_{UU^T=I_k} \operatorname{tr}(U^T \hat{\Sigma}_n U) = \sum_{i=1}^{k} \hat{\lambda}_i$$

- PCA is given by top-k eigenvectors of sample covariance matrix, i.e. top-k (left) singular vectors of Y

PCA

- **Input**: data matrix $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$
- **Output**: Euclidean k-dimensional coordinates $Z \in \mathbb{R}^{k \times n}$ of data.
- **Procedure**:
  - Centering: $Y = XH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$
  - Singular Value Decomposition $Y = USV^T$, $S = \mathbf{diag}(\sigma_j)$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(n,p)}$
  - PCA is given by top-k SVD (S_k, U_k): $U_k = (u_1, \ldots, u_k) \in \mathbb{R}^{p \times k}$, with embedding coordinates $Z_k = U_k^T Y = S_k V_k^T$, i.e.

$$Z_{ji} = u_j^T(x_i - \hat{\mu}).$$

How much variances in data explained by PCA?

The importance or variance of j-th principal component is characterized by the j-th eigenvalue. Given the eigenvalues, the following quantities are often used to measure the variances.

- Total variance:

$$\operatorname{tr}(\hat{\Sigma}_n) = \sum_{i=1}^{p} \hat{\lambda}_i;$$

- Percentage of variance explained by top-k principal components:

$$\sum_{i=1}^{k} \hat{\lambda}_i / \operatorname{tr}(\hat{\Sigma}_n);$$

- Generalized variance as total volume:

$$\det(\hat{\Sigma}_n) = \prod_{i=1}^{p} \hat{\lambda}_i.$$

Example: PCA of Handwritten Digits

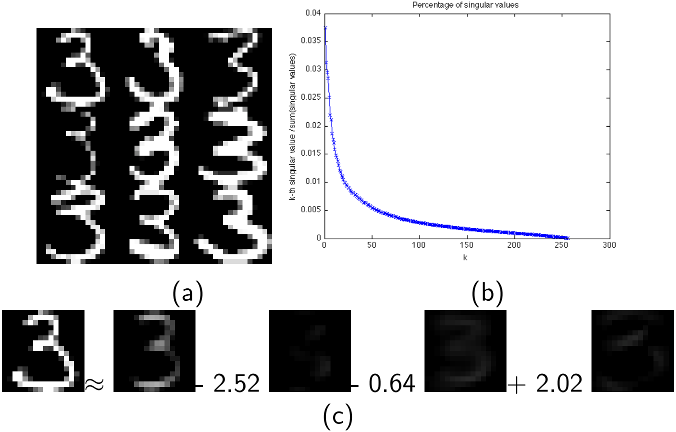


Figure: (a) random 9 images. (b) percentage of singular values over total sum. (c) approximation of the first image by top 3 principle components (singular vectors).

How many principal components?

- No universal rule, depending on applications.
- Rule of thumb: choose k such that

$$\sum_{i=1}^{k} \hat{\lambda}_i / \operatorname{tr}(\hat{\Sigma}_n) > q, \quad \text{e.g.} \quad q = 0.95$$

- *Horn's Parallel Analysis

Horn's Parallel Analysis

Random permutation test:

- Randomly permute sample features/variables for decorrelation
- Compute singular values of random matrices

$$
\begin{aligned}
& X = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p,1} & X_{p,2} & \cdots & X_{p,n} \end{bmatrix} \\
\mapsto \quad & X^1 = \begin{bmatrix} X_{1,\pi_1(1)} & X_{1,\pi_1(2)} & \cdots & X_{1,\pi_1(n)} \\ X_{2,\pi_2(1)} & X_{2,\pi_2(2)} & \cdots & X_{2,\pi_2(n)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p,\pi_p(1)} & X_{p,\pi_p(2)} & \cdots & X_{p,\pi_p(n)} \end{bmatrix} \\
\mapsto \quad & \hat{\lambda}_j^1
\end{aligned}
$$

Horn's Parallel Analysis

- Repeat such procedure for R times, we can get R set singular values. They can be put together as a matrix

$$\begin{bmatrix} \widehat{\lambda}_1^1 & \widehat{\lambda}_2^1 & \cdots & \widehat{\lambda}_p^1 \\ \widehat{\lambda}_1^2 & \widehat{\lambda}_2^2 & \cdots & \widehat{\lambda}_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\lambda}_1^R & \widehat{\lambda}_2^R & \cdots & \widehat{\lambda}_p^R \end{bmatrix}.$$

- Define the p-value for the i-th eigenvalue, and only keep eigenvalues whose p-value is smaller than a threshold, e.g.

$$\mathrm{pval}_i = \frac{1}{R}\#\{\widehat{\lambda}_i^r > \widehat{\lambda}_i\},$$

Keep $\widehat{\lambda}_i$ if $\mathrm{pval}_i < 0.05$.

Example

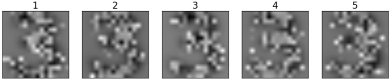
- 

Figure: Examples of randomly permuted data.

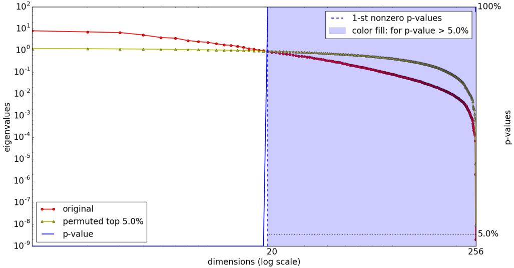
- 

Figure: Results of parallel analysis on PCA. Considering the exponential decay of eigenvalues and to emphasize the top eigenvalues, log scale are adopted for both axes. The top 5% singular values of the parallel data matrices are draw as reference.

Example

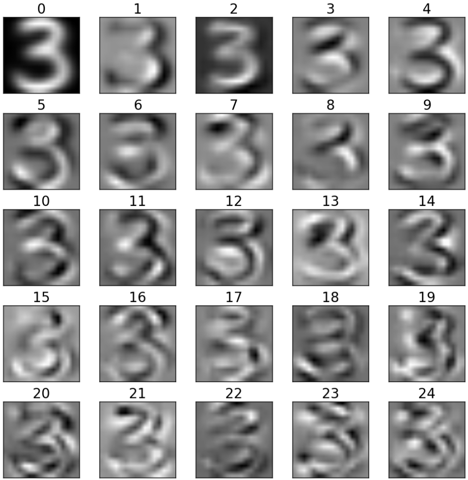


Figure: Images of the sample mean (image No.0) and the top 24 principal components (top 19 are suggested by parallel analysis). It shows that Horn's parallel analysis is conservative when data are concentrated around submanifolds.

Summary

- Data matrix: $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$
  - Centering: $Y = XH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$
- Singular Value Decomposition $Y = USV^T$, $S = \mathbf{diag}(\sigma_j)$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(n,p)}$
  - PCA is given by top-k left SVD (S_k, U_k): $U_k = (u_1, \ldots, u_k) \in \mathbb{R}^{p \times k}$, with embedding coordinates $U_k S_k$
  - What about right SVD? — Multidimensional Scaling (MDS), or Kernel PCA

Outline

Principal Component Analysis
- Horn's Parallel Analysis: Random Permutation Test

Multidimensional Scaling (MDS)

Positive Definite Functions and Kernels
- Kernel PCA

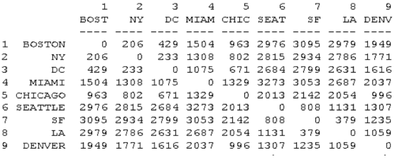
Sufficient Dimensionality Reduction
- PCA as Sufficient Dimensionality Reduction

Supervised PCA
- Linear Discriminant Analysis
- Sliced Inverse Regression
- Localized SIR

Multidimensional Scaling

The problem of classical MDS or isometric Euclidean embedding is:

- *given pairwise distances between data points, can one find a system of Euclidean coordinates for those points whose pairwise distances meet given constraints*?



```
                 1    2    3    4    5    6    7    8    9
              BOST   NY   DC MIAM CHIC SEAT   SF   LA DENV
              ---- ---- ---- ---- ---- ---- ---- ---- ----
1   BOSTON       0  206  429 1504  963 2976 3095 2979 1949
2       NY     206    0  233 1308  802 2815 2934 2786 1771
3       DC     429  233    0 1075  671 2684 2799 2631 1616
4    MIAMI    1504 1308 1075    0 1329 3273 3053 2687 2037
5 CHICAGO      963  802  671 1329    0 2013 2142 2054  996
6 SEATTLE     2976 2815 2684 3273 2013    0  808 1131 1307
7       SF    3095 2934 2799 3053 2142  808    0  379 1235
8       LA    2979 2786 2631 2687 2054 1131  379    0 1059
9   DENVER    1949 1771 1616 2037  996 1307 1235 1059    0
```

Metric MDS

- Consider a forward problem: given a set of points $x_1, x_2, ..., x_n \in \mathbb{R}^p$, let

$$X = [x_1, x_2, ..., x_n]^{p \times n}.$$

The distance between point x_i and x_j satisfies

$$d_{ij}^2 = \|x_i - x_j\|^2 = (x_i - x_j)^T (x_i - x_j) = {x_i}^T x_i + {x_j}^T x_j - 2{x_i}^T x_j.$$

- Now we are considering the inverse problem: *given only d_{ij}, can one find a $\{y_i \in \mathbb{R}^k : i = 1 \ldots, n\}$ for some k satisfying the constraint $d_{ij} = \|y_i - y_j\|$?*

Classical Metric MDS method

- transform squared distance matrix $D = [d_{ij}^2]$ to an inner product form, which is positive semi-definite and often called as kernel matrix;

- compute the eigen-decomposition for this inner product form (kernel matrix).

Classical MDS method

- The key observation is that the two-side centering transform of squared distance matrix D gives the Gram matrix (inner product matrix or kernel matrix) of centered data matrix, i.e.

$$-\frac{1}{2}HDH^T = (XH)^T(XH) =: \widehat{K}. \tag{3}$$

where $H := I - \frac{1}{n}\mathbf{1}\cdot\mathbf{1}^T = H^T$ with $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^n$ is the *Householder centering matrix*.

Classical MDS method

- To see this, let K be the inner product or kernel matrix

$$K = X^T X, \quad X = [x_i] \in \mathbb{R}^{p \times n}$$

with $k = \mathbf{diag}(K_{ii}) \in \mathbb{R}^n$.

- Note that

$$D = (d_{ij}^2) = k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K.$$

- The following lines established the fact that

$$-\frac{1}{2} H \cdot D \cdot H^T = H^T K H = (XH)^T (XH).$$

Classical MDS method

- In fact, note that

$$-\frac{1}{2}H \cdot D \cdot H^T \quad = \quad -\frac{1}{2}H \cdot (k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K) \cdot H^T$$

- Since $k \cdot \mathbf{1}^T \cdot H^T = k \cdot \mathbf{1}(I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T) = k \cdot \mathbf{1} - k(\frac{\mathbf{1}^T \cdot \mathbf{1}}{n}) \cdot \mathbf{1} = 0$, we have $H \cdot k\ \mathbf{1} \cdot H^T = H \cdot \mathbf{1} \cdot k^T \cdot H^T = 0$. This implies that

$$-\frac{1}{2}H \cdot D \cdot H^T = H \cdot K \cdot H^T = HX^TXH^T = (XH)^T(XH),$$

since $H = H^T$, which establishes (3).

The Classical MDS Algorithm

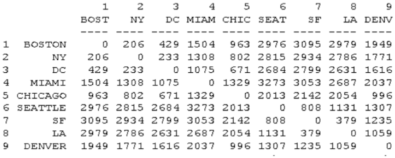
- **Input**: A squared distance matrix $D^{n\times n}$ with $D_{ij} = d_{ij}^2$.
- **Output**: Euclidean k-dimensional coordinates $Z_k \in \mathbb{R}^{k\times n}$ of data.
- **Procedure**:
  - Compute $\widehat{K} = -\frac{1}{2} H \cdot D \cdot H^T$, with the Househölder matrix H.
  - Compute Eigenvalue decomposition $\widehat{K} = \widehat{V}\widehat{\Lambda}\widehat{V}^T$ with $\widehat{\Lambda} = \mathbf{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n)$ where $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_n \geq 0$;
  - Choose top k nonzero eigenvalues and corresponding eigenvectors, set the embedding coordinates $Z_k = \widehat{\Lambda}_k^{\frac{1}{2}} \widehat{V}_k^T$ where

    $$\widehat{V}_k = [\widehat{v}_1, \ldots, \widehat{v}_k], \quad \widehat{v}_k \in \mathbb{R}^n,$$

    $$\widehat{\Lambda}_k = \mathbf{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k),$$

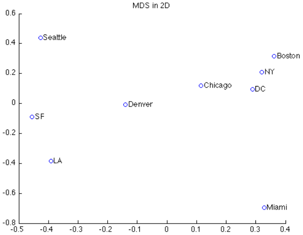
    with $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots \geq \widehat{\lambda}_k \geq 0$.

Example

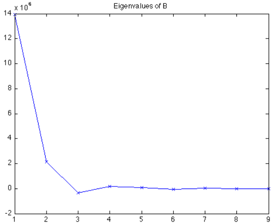


```
                 1    2    3    4    5    6    7    8    9
              BOST   NY   DC MIAM CHIC SEAT   SF   LA DENV
              ---- ---- ---- ---- ---- ---- ---- ---- ----
1  BOSTON        0  206  429 1504  963 2976 3095 2979 1949
2      NY      206    0  233 1308  802 2815 2934 2786 1771
3      DC      429  233    0 1075  671 2684 2799 2631 1616
4   MIAMI     1504 1308 1075    0 1329 3273 3053 2687 2037
5 CHICAGO      963  802  671 1329    0 2013 2142 2054  996
6 SEATTLE     2976 2815 2684 3273 2013    0  808 1131 1307
7      SF     3095 2934 2799 3053 2142  808    0  379 1235
8      LA     2979 2786 2631 2687 2054 1131  379    0 1059
9  DENVER     1949 1771 1616 2037  996 1307 1235 1059    0
```

(a)



(b)



(c)

Remark: Nonmetric MDS

- Given a set of points $x_i \in \mathbb{R}^p$ $(i = 1, 2, \cdots, n)$; form a data Matrix $X^{p\times n} = [X_1, X_2 \cdots X_n]^T$, when p is large, especially in some cases larger than n, we want to find k-dimensional projection with which pairwise distances of the data point are preserved as well as possible.

- That is to say, if we know the original pairwise distance $d_{ij} = \|X_i - X_j\|$ or data distances with some disturbance $\tilde{d}_{ij} = \|X_i - X_j\| + \epsilon$, we want to find $Y_i \in \mathbb{R}^k$ s.t.:

$$\min_{Y_i \in \mathbb{R}^k} \sum_{i,j} (\|Y_i - Y_j\|^2 - \tilde{d}_{ij}^2)^2. \tag{4}$$

  Without loss of generality, we set $\sum_i Y_i = 0$, *i.e.* putting the origin as data center. This is called *nonmetric* MDS since such general $\tilde{d}_{ij}$ is not necessarily a distance.

Outline

- Principal Component Analysis
  - Horn's Parallel Analysis: Random Permutation Test
- Multidimensional Scaling (MDS)
- Positive Definite Functions and Kernels
  - Kernel PCA
- Sufficient Dimensionality Reduction
  - PCA as Sufficient Dimensionality Reduction
- Supervised PCA
  - Linear Discriminant Analysis
  - Sliced Inverse Regression
  - Localized SIR

Positive Definite Matrix

- Definition (Positive Semi-definite Matrix)

  Suppose $A^{n\times n}$ is a real symmetric matrix, then A is called positive semi-definite (p.s.d.), denoted by $A \succeq 0$, if $\forall v \in \mathbb{R}^n, v^T A v \geq 0$.

  - Positive semi-definiteness completely characterizes the inner product matrices: $A \succeq 0 \iff A = Y^T Y$ for some Y.

- Property

  Suppose $A^{n\times n}$, $B^{n\times n}$ are real symmetric matrix, $A \succeq 0$, $B \succeq 0$. Then we have:

  (a) $A + B \succeq 0$;
  (b) $A \circ B \succeq 0$;

  where $A \circ B$ is called Hadamard product and $(A \circ B)_{i,j} := A_{i,j} \cdot B_{i,j}$.

- Definition (Conditionally Negative Definite Matrix)

  Let $A^{n \times n}$ be a real symmetric matrix. A is conditionally negative definite (c.n.d.), if for $\forall v \in \mathbb{R}^n$ such that $\mathbf{1}^T v = \sum_{i=1}^{n} v_i = 0$, there holds $v^T A v \leq 0$.

- Lemma (Young/Householder-Schoenberg'1938)

  For any signed probability measure α $(\alpha \in \mathbb{R}^n, \sum_{i=1}^{n} \alpha_i = 1)$,

$$B_\alpha = -\frac{1}{2} H_\alpha C H_\alpha^T \succeq 0 \iff C \text{ is c.n.d.}$$

  where H_α is Househölder centering matrix: $H_\alpha = I - \mathbf{1} \cdot \alpha^T$.

Theorem (Classical MDS)

Let $D^{n\times n}$ be a real symmetric matrix and

$$C = D - \frac{1}{2}d \cdot \mathbf{1}^T - \frac{1}{2}\mathbf{1} \cdot d^T, \text{ with } d = \mathbf{diag}(D).$$

Then the following holds.

1. $B_\alpha = -\frac{1}{2}H_\alpha D H_\alpha^T = -\frac{1}{2}H_\alpha C H_\alpha^T$ for $\forall\alpha$ as a signed probability measure;
2. $C_{i,j} = B_{i,i}(\alpha) + B_{j,j}(\alpha) - 2B_{i,j}(\alpha)$;
3. D c.n.d. $\iff$ C c.n.d.;
4. C c.n.d. $\Rightarrow C$ is a squared distance matrix (i.e. $\exists Y^{n\times k}$ s.t. $C_{i,j} = \sum_{m=1}^{k}(y_{i,m} - y_{j,m})^2$).

Schoenberg Transform

Theorem (Schoenberg Transform)

Given D a squared distance matrix, $C_{i,j} = \Phi(D_{i,j})$. Then

$$C \text{ is a squared distance matrix } \iff \Phi \text{ is a Schoenberg Transform.}$$

- Definition (Schoenberg Transform)

  The Schoenberg Transform $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is defined by

$$\Phi(t) := \int_0^\infty \frac{1 - \exp(-\lambda t)}{\lambda} g(\lambda) d\lambda, \tag{5}$$

  where $g(\lambda)$ is some nonnegative measure on $[0, \infty)$ s.t

$$\int_0^\infty \frac{g(\lambda)}{\lambda} d\lambda < \infty.$$

Schoenberg Transform

- Examples of Schoenberg Transforms include
  - $\Phi_0(t) = t$ with $g_0(\lambda) = \delta(\lambda)$;
  - $\Phi_1(t) = \dfrac{1 - \exp(-at)}{a}$ with $g_1(\lambda) = \delta(\lambda - a)$ $(a > 0)$;
  - $\Phi_2(t) = \ln(1 + t/a)$ with $g_2(\lambda) = \exp(-a\lambda)$;
  - $\Phi_3(t) = \dfrac{t}{a(a+t)}$ with $g_3(\lambda) = \lambda \exp(-a\lambda)$;
  - $\Phi_4(t) = t^p$ $(p \in (0,1))$ with $g_4(\lambda) = \dfrac{p}{\Gamma(1-p)}\lambda^{-p}$.

Isometric Hilbert Embedding

- Definition (Positive Semi-definite Functions)

  A symmetric function $k(x,y) = k(y,x)$ is called *positive definite* if for all finite x_i, x_j,

  $$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0, \quad \forall c_i, c_j$$

  with equality $=$ holds iff $c_i = c_j = 0$. In other words the function k restricted on $\{(x_i, x_j) : i, j = 1, \ldots, n\}$ is a positive definite matrix.

- Theorem (Schoenberg 38)

  A separable space M with a metric function $d(x,y)$ can be isometrically imbedded in a Hilbert space H, if and only if the family of functions $e^{-\lambda d^2}$ are *positive definite* for all $\lambda > 0$ (in fact we just need it for a sequence of λ_i whose accumulate point is 0).

Complete Monotonicity and Positive Definiteness

- Note that Schoenberg transform satisfies $\Phi(0) = 0$,

  $$\Phi'(t) = \int_0^\infty \exp(-\lambda t) g(\lambda) d\lambda \geq 0,$$

  $$\Phi''(t) = -\int_0^\infty \exp(-\lambda t) \lambda g(\lambda) d\lambda \leq 0,$$

  and so on. In other words, Φ is a *completely monotonic function* defined by $(-1)^n \Phi^{(n)}(x) \geq 0$, with additional constraint $\Phi(0) = 0$.

- e^{-t} is completely monotone. Schoenberg connects positive definite and completely monotone functions.

Theorem (Schoenberg, 1938)

A function ϕ is *completely monotone* on $[0, \infty)$ if and only if $\phi(d^2)$ is positive definite and radial on $\mathbb{R}^k$ for all k.

Mercer Kernel and RKHS

- Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a compact Euclidean domain.

- A *Mercer kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, is a continuous symmetric real-valued function which is *positive definite*, often called a reproducing kernel.

- Reproducing kernel Hilbert space $\mathcal{H}_K$ is constructed as follows.
  - A Mercer kernel K induces a function $K_x : \mathcal{X} \to \mathbb{R}$ $(x \in \mathcal{X})$ defined by $K_x(t) = K(x, t)$ for $t \in \mathcal{X}$
  - An inner product between two functions K_x and $K_{x'}$ can be defined as the bilinear form $\langle K_x, K_{x'} \rangle_{\mathcal{H}_K} = K(x, x')$ $(x, x' \in \mathcal{X})$ due to the positive definite K.
  - Take the completion of the $\operatorname{span}\{K_x : x \in \mathcal{X}\}$ with respect to the inner product as the unique linear extension of the bilinear form $\langle K_x, K_{x'} \rangle_{\mathcal{H}_K} = K(x, x')$ $(\forall x, x' \in \mathcal{X})$
  - The most important property of RKHS is the *reproducing property*: for all $f \in \mathcal{H}_K$ and $x \in \mathcal{X}$, $f(x) = \langle f, K_x \rangle_{\mathcal{H}_K}$

Covariance operator

- Let L^2_ρ be the Hilbert space of square integrable functions on $\mathcal{X}$ with respect to the probability measure $\rho_{\mathcal{X}}$.
- Define a linear operator $L_K : L^2_\rho \to L^2_\rho$ by

$$L_K(f)(x) = \int_X K(x,t) f(t) d\rho_X.$$

- The operator $L_K : L^2_\rho \to L^2_\rho$ is compact with a discrete spectrum, i.e. an orthonormal eigensystem $(\lambda_k, \phi_k)_{k\in\mathbb{N}}$, such that $L_K \phi_k = \lambda_k \phi_k$.
- The restriction of L_K on $\mathcal{H}_K$ induces an operator $L_K|_{\mathcal{H}_K} : \mathcal{H}_K \to \mathcal{H}_K$, which is called as the *covariance operator* of $\rho_{\mathcal{X}}$ in $\mathcal{H}_K$.

Spectral Representation of Mercer's Kernel

- Theorem (Mercer's Theorem)

  Let $\mathcal{X}$ be a compact domain or a manifold, $\rho_{\mathcal{X}}$ a Borel measure on $\mathcal{X}$, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel. Let λ_k be the k-th eigenvalue of L_K and $\{\phi_k\}_{k \in \mathbb{N}}$ the corresponding eigenvectors. For all $x, t \in \mathcal{X}$,

  $$K(x,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi(t) \tag{6}$$

  where the convergence is absolute (for each $x, t \in \mathcal{X} \times \mathcal{X}$) and uniform (on $\mathcal{X} \times \mathcal{X}$).

Kernel PCA

Definition (Kernel PCA/MDS)

Given a data sample of $\{x_i : i = 1, \ldots, n\}$ drawn independently and identically distributed from $\rho_{\mathcal{X}}$, the kernel matrix $K = (k(x_i, x_j) : i, j = 1, \ldots, n)$ is a positive definite matrix. Then the following procedure gives a k-dimensional Euclidean embedding of data.

(a) Find the top-k eigen-decomposition of the following centred matrix

$$\widehat{K} = HKH^T, \quad \text{where } K = (k(x_i, x_j) : i, j = 1, \ldots, n).$$

(b) Embed the data in the same way as classical MDS Algorithm.

Summary: PCA and MDS

- Data matrix: $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$
  - Centering: $Y = XH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$
- Singular Value Decomposition $Y = USV^T$, $S = \mathbf{diag}(\sigma_j)$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(n,p)}$
  - PCA is given by top-k (left) SVD (S_k, U_k): $U_k = (u_1, \ldots, u_k) \in \mathbb{R}^{p \times k}$, with embedding coordinates $U_k S_k$
  - MDS is given by top-k (right) SVD (S_k, V_k): $V_k = (v_1, \ldots, v_k) \in \mathbb{R}^{n \times k}$, with embedding coordinates $V_k S_k$
  - Kernel PCA (MDS): for $K \succeq 0$, $K_c = HKH^T$, $K_c = U\Lambda U^T$ gives MDS embedding $U_k \Lambda_k^{1/2} \in \mathbb{R}^{n \times k}$

PCA

- PCA is unsupervised learning of data
  - It only analyzes X, without Y
  - Invented by Pearson (1901) and Hotelling (1933)
- Supervised PCA?
  - Dennis Cook (2001): sufficient dimensionality reduction
  - Fisher's Linear Discriminant Analysis (1920s) and Ker-Chao Li's Sliced Inverse Regression (1991)

Outline

Principal Component Analysis

Horn's Parallel Analysis: Random Permutation Test

Multidimensional Scaling (MDS)

Positive Definite Functions and Kernels

Kernel PCA

Sufficient Dimensionality Reduction

PCA as Sufficient Dimensionality Reduction

Supervised PCA

Linear Discriminant Analysis

Sliced Inverse Regression

Localized SIR

Sufficient Dimensionality Reduction

Definition (Cook 2005)

A **sufficient dimension reduction** Γ ($\Gamma \in \mathbb{R}^{p\times d}$, $\Gamma^T\Gamma = I_d$) refers to the setting that the conditional distribution of $Y|X$ is the same as the distribution of $Y|\Gamma^T X$ for all X, i.e.

$$\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X).$$

- Example: in regression $Y = f(X, \varepsilon)$, for some unknown function f, sufficient dimensionality reduction implies that $Y = f(\Gamma^T X, \varepsilon)$.

Sufficient Dimensionality Reduction

Definition (Cook 2005)

A **sufficient dimension reduction** Γ ($\Gamma \in \mathbb{R}^{p \times d}$, $\Gamma^T \Gamma = I_d$) refers to the setting that the conditional distribution of $Y|X$ is the same as the distribution of $Y|\Gamma^T X$ for all X, i.e.

$$\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X).$$

- Example: in regression $Y = f(X, \varepsilon)$, for some unknown function f, sufficient dimensionality reduction implies that $Y = f(\Gamma^T X, \varepsilon)$.

- Can you find Γ without knowing f?

Sufficient Dimensionality Reduction

Definition (Cook 2005)

A **sufficient dimension reduction** Γ ($\Gamma \in \mathbb{R}^{p\times d}$, $\Gamma^T\Gamma = I_d$) refers to the setting that the conditional distribution of $Y|X$ is the same as the distribution of $Y|\Gamma^T X$ for all X, i.e.

$$\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X).$$

- Example: in regression $Y = f(X, \varepsilon)$, for some unknown function f, sufficient dimensionality reduction implies that $Y = f(\Gamma^T X, \varepsilon)$.

- Can you find Γ without knowing f?

- Yes! Consider the inverse problem, with conditional distribution $\mathbb{P}(X|Y)$.

An Inverse Model

Example (Inverse model)

For each value in response variable y,

$$X_y = \mu + \Gamma \nu_y + \varepsilon \tag{7}$$

where

- $X_y \in \mathbb{R}^p$,
- $\nu_y \in \mathbb{R}^d$, $d < p$,
- $\Gamma \in \mathbb{R}^{p \times d}$ such that $\Gamma^T \Gamma = I_d$,
- $\varepsilon \sim N_p(0, \sigma^2 I_p)$,
- assume $\sum_y \nu_y = 0$ for removing the degree of freedom in translation.

Sufficient Dimensionality Reduction

Lemma (Cook 2005)

Under the inverse model, $\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X)$*, i.e.* Γ *is a sufficient dimensionality reduction.*

Proof

- First, $X|(Y = y) \sim N_p(\mu + \Gamma\nu_y, \sigma^2 I_p)$.
- By Bayesian formula, we have for any f

$$\begin{aligned} f_{Y|X}(y|x) &\propto f_{X|Y}(x|y) f_Y(y) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|x - \mu - \Gamma\nu_y\|^2\right) f_Y(y) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\nu_y^T\nu_y - 2\nu_y^T\Gamma^T(x - \mu)\right) f_Y(y) \end{aligned}$$

where the last line is given by the orthogonality $\Gamma^T\Gamma = I$.

Proof (continued)

- Similarly, since $\Gamma^T X|(Y=y) \sim N_d(\Gamma^T\mu + \nu_y, \sigma^2 I_d)$, we have

$$\begin{aligned} f_{Y|\Gamma^T X}(y|\Gamma^T x) &\propto f_{\Gamma^T X|Y}(\Gamma^T x|y) f_Y(y) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|\Gamma^T x - \Gamma^T\mu - \nu_y\|^2\right) f_Y(y) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(\nu_y^T\nu_y - 2\nu_y^T\Gamma^T(x-\mu)\right) f_Y(y) \end{aligned}$$

  by the orthogonality $\Gamma^T\Gamma = I$.

- Therefore, $\mathbb{P}(Y|X) = \mathbb{P}(Y|\Gamma^T X)$ of the same density kernels. □

Estimate of Γ

- Can we estimate Γ from finite sample without knowing f?

Estimate of Γ

- Can we estimate Γ from finite sample without knowing f?
- PCA gives the Maximum Likelihood Estimate of Γ

Maximum Likelihood Estimate

- Under the inverse model, the conditional likelihood function

$$f(X_y|\mu, \Gamma, \nu_y) = \frac{1}{\sigma^p \sqrt{(2\pi)^p}} \exp\left[-\frac{1}{2\sigma^2} (X_y - \mu - \Gamma\nu_y)^T (X_y - \mu - \Gamma\nu_y) \right]$$

- MLE

$$\max_{\mu,\Gamma,\nu_y} \prod_y f(X_y|\mu, \Gamma, \nu_y)$$

$$\Leftrightarrow \max_{\mu,\Gamma,\nu_y} -\frac{1}{2\sigma^2} \sum_y \|X_y - \mu - \Gamma\nu_y\|^2 - \sum_y p \log\sigma + C.$$

Maximum Likelihood Estimate (continued)

- MLE solution

$$\widehat{\Gamma} = \arg\min_{\Gamma^T\Gamma=I} \sum_y \|X_y - \hat{\mu} - P_\Gamma(X_y - \hat{\mu})\|^2, \quad P_\Gamma = \Gamma\Gamma^T. \tag{8}$$

  where $\widehat{\mu} = \frac{1}{n}\sum_y X_y$, $\nu_y = \widehat{\Gamma}^T(X_y - \hat{\mu})$.

- If y is of distinct values (e.g. the unknown f is injective), PCA (top d eigen-decomposition of $\widehat{\Sigma}$) gives $\widehat{\Gamma}$.

- If y is of discrete values (e.g. classification), discriminant analysis (eigen-decomposition of $\widehat{\Sigma}_B = \frac{1}{K}\sum_{y=1}^{K}(\hat{\mu}_y - \hat{\mu})(\hat{\mu}_y - \hat{\mu})^T$) gives $\widehat{\Gamma}$.

Maximum Likelihood Estimate (continued)

- In general
$$X_y = \mu + \Gamma \nu_y + \epsilon \tag{9}$$
where $\varepsilon \sim N_p(0, \Sigma)$, $\widehat{\mu}_y = \widehat{E}[X_y|y]$.

- Rescale $Z_y = \Sigma^{-1/2} X_y$.

- Eigen-decomposition of $\Sigma^{-1/2}\widehat{\Sigma}_B\Sigma^{-1/2}$ (with $\widehat{\Sigma}$ for the estimate of Σ) meets Fisher's Linear Discriminant Analysis for $\widehat{\Gamma}$.

- Therefore *PCA/LDA can be also derived as a sufficient dimensionality reduction in supervised learning, even the function f is unknown here*.

Outline

Principal Component Analysis

Horn's Parallel Analysis: Random Permutation Test

Multidimensional Scaling (MDS)

Positive Definite Functions and Kernels

Kernel PCA

Sufficient Dimensionality Reduction

PCA as Sufficient Dimensionality Reduction

Supervised PCA

Linear Discriminant Analysis

Sliced Inverse Regression

Localized SIR

Linear Discriminant Analysis

- Data: $\{X_i, y_i\}_{i=1}^N$ where y_i is discrete in $\{1, 2, \ldots, K\}$ but not ordered

- Compute sample mean and within class means

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad \hat{\mu}_k = \frac{1}{N_k}\sum_{y_i=k} X_i;$$

- Compute Between class covariance matrix

$$\widehat{\Sigma}_B^{p\times p} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^T;$$

- Compute Within class covariance matrix

$$\hat{\Sigma}_W^{p\times p} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{y_i=k}(X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^T;$$

Fisher's Linear Discriminant Analysis

We choose the k-th class such that the following *linear* score function is the largest:

$$\hat{\delta}_k(x) = \hat{\mu}_k^T \hat{\Sigma}^{-1} x - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k, \qquad (10)$$

where given data $(x_i, y_i), i = 1, ..., n$,

- $\hat{\pi}_k = n_k / n$ is the sample proportion of class k where n_k is the number of subjects in class k
- $\hat{\mu}_k$ is the sample mean of class k

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i;$$

- $\hat{\Sigma}$ is the pooled (overall) sample covariance

$$\hat{\Sigma} = \widehat{\Sigma}_B + \widehat{\Sigma}_W = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T,$$

Fisher's LDA

- Fisher's LDA (1920s) aims to capture dominant variations between different classes of data:
  - Compute **generalized Eigen-decomposition** $\widehat{\Sigma}_B = \widehat{\Sigma} U \Lambda U^T$ with $\Lambda = \mathbf{diag}(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$;
  - Choose top-d generalized eigenvectors corresponding to top $d \leq K$ nonzero eigenvalues,

$$U_d = [u_1, \ldots, u_d], \quad u_j \in \mathbb{R}^p.$$

Sliced Inverse Rgression

- Data: $\{X_i, y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ is continuous (or ordered discrete)

- Divide the range of y_i into S non-overlapping slices $H_s (s = 1, ..., S)$. N_s is the number of observations within each slice.

- Compute the sample mean and total covariance matrix

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} X_i, \qquad \hat{\Sigma}^{p \times p} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \hat{\mu})(X_i - \hat{\mu})^T;$$

- Compute the mean of X_i over all slices and Between slices covariance matrix

$$\hat{\mu}_k = \frac{1}{N_s} \sum_{y_i \in H_s} X_i, \qquad \hat{\Sigma}_B^{p \times p} = \frac{1}{K} \sum_{h}^{K} (\hat{\mu_k} - \hat{\mu})(\hat{\mu_k} - \hat{\mu})^T;$$

Li's SIR

- K.-C. Li's Slice Inverse Regression (1991) aims to capture dominant variations between different slices of data:
  - Compute **Generalized Eigen-decomposition** $\hat{\Sigma}_B = \hat{\Sigma} U \Lambda U^T$ with $\Lambda = \mathbf{diag}(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$;
  - Choose top-d generalized eigenvectors corresponding to top $d \leq K$ nonzero eigenvalues,

$$\Gamma_d = [u_1, \ldots, u_d], \quad u_k \in \mathbb{R}^p.$$

Localized Sliced Inverse Rgression

- Data: $\{X_i, y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ is continuous (or ordered discrete)
- Divide the range of y_i into S non-overlapping slices $H_s(s = 1, ..., S)$. N_s is the number of observations within each slice.
- Compute the sample mean $\hat{(}\mu)$ and total covariance $\hat{\Sigma}$ as in SIR
- Compute the **localized** mean of X_i over all slices and **localized** Between-slice covariance matrix

$$\hat{\mu}_{i,loc} = \frac{1}{|s_i|} \sum_{j \in s_i} X_j, \qquad \hat{\Sigma}_{locB} = \frac{1}{N} \sum_i (\hat{\mu}_{i,loc} - \hat{\mu})(\hat{\mu}_{i,loc} - \hat{\mu})^T \ ;$$

where $s_i = \{j : x_j$ belongs to the k nearest neighbours of x_i in $H_s\}$ and s indexes the slice H_s to which i belongs.

LSIR

- Wu-Liang-Mukherjee Localized Slice Inverse Regression (2009) aims to capture nonlinear variations between different slices of data:
  - Compute **Generalized Eigen-decomposition** $\hat{\Sigma}_{locB} = \hat{\Sigma} U \Lambda U^T$ with $\Lambda = \mathbf{diag}(\lambda_1, \lambda_2, ...\lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$;
  - Choose top-d generalized eigenvectors corresponding to top $d \leq K$ nonzero eigenvalues,

$$\Gamma_d = [u_1, \ldots, u_d], \quad u_k \in \mathbb{R}^p.$$