MODULE 6.1
Principal Component Analysis (PCA)
A FIRST LOOK

by
S.Lakshmivarahan
School of Computer Science
University of Oklahoma
Norman, OK-73019, USA
varahan@ou.edu

A basic set up

- Let $x \in R^m$ be a random vector in L_2
- Without loss of generality: Assume x is centered
- E(x) = 0 and $Cov(x) = \Sigma \in R^{m \times m}$
- Assume that Σ is SPD

Eigen decomposition of Σ

- Let (λ_i, v_i) be an eigenpair of Σ
- That is,

$$\Sigma v_i = v_i \lambda_i, 1 \leq i \leq m \tag{1}$$

- Define $v = [v_1, v_2, \ldots, v_m] \in R^{m \times m}$, an orthogonal matrix of eigenvectors of Σ.

$$v^T v = v v^T = I_m$$

- Define

$$\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_m) \in R^{m \times m} \tag{2}$$

- Let

$$\lambda_1 > \lambda_2 > \ldots \lambda_m > 0 \tag{3}$$

- Then

$$\Sigma v = v\Lambda, \Sigma = v\Lambda v^T, v\Sigma v^T = \Lambda \tag{4}$$

Eigenvalue and variance

- Define

$$Var(x) = \sum_{i=1}^{m} Var(x_i) = \sum_{i=1}^{m} E(x_i^2) = E(x^T x) \tag{5}$$

- But $E(x^T x) = E[tr(xx^T)]$
$= \text{tr}[\text{E}(\text{xx}^T)]$
$= \text{tr}(\Sigma) = tr(v\Lambda v^T)$
$= \text{tr}(\ \text{v}^T v\Sigma)(\because tr(AB) = tr(BA))$

$$= tr(\Sigma) = \Sigma_{i=1}^{m} \lambda_i \tag{6}$$

- Sum of the variance of the components of x = sum of the eigenvalues of the covariance matrix of x

Statement of the problem

- Find a set of m deterministic, orthonormal vectors $\{\xi_1, \xi_2, \xi_3, \ldots, \xi_m\}$ where each $\xi_i \in R^m$ and a set of m uncorrelated random coefficients $\{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_m\}$ such that

$$(1) \quad x = \alpha_1\xi_1 + \alpha_2\xi_2 + \ldots + \alpha_m\xi_m \tag{7}$$

and

$$(2) \quad var(\alpha_i\xi_i) = \lambda_i, 1 \leq i \leq m \tag{8}$$

- Such a set of orthonormal vectors called the principal patterns and the set of associated random coefficients are called the principal components.

Rayleigh Quotient and its properties

- To this end, let $A \in R^{m \times m}$ be an SPD and $\eta \in R^m$
- Given A, define a functional, $r_A(\eta) : R^m \rightarrow R$ as

$$r_A(\eta) = \frac{\eta^T A \eta}{\eta^T \eta} \tag{9}$$

- Let (μ, ξ) be an eigenpair of A: $A\xi = \mu\xi$
- Then

$$r_A(\xi) = \frac{\xi^T A \xi}{\xi^T \xi} = \mu \tag{10}$$

- Since the eigenvectors are normalized: $\xi^T \xi = 1$,

$$r_A(\xi) = \xi^T A \xi = \mu \tag{11}$$

Rayleigh Quotient for A = Σ

- Recall that v_i is a normalized eigenvector corresponding to the eigenvalue λ_i, $1 \leq i \leq m$ of the covariance matrix Σ of the random vector, x.
- From (11):

$$r_{\Sigma}(v_i) = v_i^T \Sigma v_i = \lambda_i \tag{12}$$

- In view of the ordering of the eigenvalues in (3):

$$r_{\Sigma}(v_1) > r_{\Sigma}(v_2) > \ldots > r_{\Sigma}(v_m) \tag{13}$$

A linear functional of x

- Consider the eigenpair (λ_i, v_i) of Σ:
  $\Sigma v_i = v_i \lambda_i, v_i^T v_i = 1 \quad for \quad 1 \leq i \leq m$
- Define a new random variable

$$\alpha_i = v_i^T x \tag{14}$$

  which is a linear functional of x
- Then,

$$E(\alpha_i) = E(v_i^T x) = v_i^T E(x) = 0 \tag{15}$$

$$\begin{aligned} var(\alpha_i) = E(\alpha_i^2) = E[(v_i^T x)(v_i^T x)] = v_i^T E(xx^T) v_i \\ = v_i^T \Sigma v_i = \lambda_i \end{aligned} \tag{16}$$

Properties of the random vector $\alpha_i v_i$

- $$E(\alpha_i v_i) = E[(v_i^T x)v_i] = E[(v_i^T x)]v_i = 0 \tag{17}$$
- Let the vector $v_i = (h_1, h_2, \ldots, h_m)^T$ with $v_i^T v_i = 1$
- Then,

$$\begin{aligned} var(\alpha_i v_i) = \sum_{j=1}^{m} var(\alpha_i h_j) &= \sum_{j=1}^{m} h_j^2 var(\alpha_i) \\ &= \lambda_i \sum_{j=1}^{m} h_j^2 = \lambda_i \end{aligned} \tag{18}$$

Correlation between α_i and α_j

- $$cov(\alpha_i, \alpha_j) = E(\alpha_i \alpha_j) = E[(v_i^T x) v_j^T x] = v_i^T E(xx^T) v_j \\ = v_i^T \Sigma v_j = 0 \quad (19)$$

  since $v^T \Sigma v = \Lambda$, a diagonal matrix
- That is, α_i and α_j are uncorrelated for $i \neq j$

Solution to the problem stated above

- By setting $\xi_i = v_i$ in (7), we first identify the required set of m orthogonal eigenvectors of Σ that constitute a basis for R^m
- By setting $\alpha_i = (v_i^T x)$ in (7) we identify the required set of uncorrelated random coefficient such that $var(v_i^T x) = \lambda_i$
- Indeed, the required expansion of x is given by

$$x = (v_1^T x)v_1 + (v_2^T x)v_2 + \cdots + (v_m^T x)v_m \qquad (20)$$

Projection matrix along an eigenvector of Σ

- Recall that the orthogonal projection matrix, $P_h \in R^{m \times m}$ along the direction $h \in R^m$ is given by

$$P_h = h(h^T h)^{-1} h^T \tag{21}$$

- Setting $h = v$ a normalized eigenvector of Σ,

$$P_v = v(v^T v)^{-1} v^T = vv^T \in R^{m \times m} \tag{22}$$

Projection of x along the eigenvector of v

- $$P_v x = (vv^T)x = v(v^T x) = (v^T x)v \tag{23}$$
- Consequently, each of the m terms in the summand on the right hand side of (20) is a vector resulting from the orthogonal projections of x along the m orthogonal basis of R^m which are eigenvectors of Σ

Equivalent representations of points in R^m

- The random vector $X \in R^m$ denotes a point R^m which in the standard orthogonal basis is given by

$$X = \sum_{i=1}^{m} X_i e_i \tag{24}$$

  where $e_i \in R^m$ in the i^{th} unit vector

- In the new orthonormal basis of principle patterns defined by the eigenvectors of $cov(x) = \Sigma$, the point x is given by

$$X = \sum_{i=1}^{m} \alpha_i v_i \quad and \quad \alpha_i = X^T v_i \tag{25}$$

- Thus, the same point in R^m admits two labels x and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$, the Principal components

Principal component transform

- $$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_m^T \end{bmatrix} x = v^T x \quad (26)$$

  That is, the principal component α is obtained by a linear transform v^T of x as $\alpha = v^T x$

- Conversely,

$$x = v\alpha \quad (27)$$

- While $var(x) = var(\alpha)$, the principal component has the additional property that

$$var(\alpha_1) > var(\alpha_2) > \ldots var(\alpha_m) > 0 \quad (28)$$

Covariance of x and α

- Components of x are correlated but those of α are uncorrelated
- $$\begin{aligned} cov(x,\alpha) &= E(x\alpha^T) - E(x)E(\alpha^T) \\ &= E(x\alpha^T) \quad (\because E(x) = 0) \\ &= E(xx^T v) = \Sigma v = v\Sigma v^T v \\ &= v\Sigma \end{aligned} \tag{29}$$

Correlation between x_i and α_i

- $$\Omega_{ij} = cor(x_i, \alpha_j) = \frac{cov(x_i, \alpha_j)}{[var(x_i) var(\alpha_j)]^{1/2}}$$
- From (29):
$$cov(x_i, \alpha_j) = (v\Sigma)_{ij} = v_{ij}\lambda_j \quad (30)$$
- $var(\alpha_j) = \lambda_j$ from (16)
- Hence,
$$\Omega_{ij} = \frac{\sqrt{\lambda_j} v_{ij}}{[var(x_i)]^{1/2}} \quad (31)$$

An interpretation of Ω_{ij}

- Verify

$$\sum_{j=1}^{m} \Omega_{ij}^2 = \frac{1}{var(x_i)} \sum_{j=1}^{m} \lambda_j v_{ij}^2 = \frac{(v \Lambda v^T)_{ii}}{var(x_i)} = 1 \qquad (32)$$

- Hence, Ω_{ij}^2 denotes the fraction of the variance of x_i explained by the principal component α_j

Example 1

- Let m=2 and $x \sim N(0, \Sigma)$ where
$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \text{ and } \rho > 0$$
- $var(x_1) = 1 = var(x_2), cov(x_1, x_2) = \rho$
- Verify that
$$\lambda_1 = 1 + \rho \text{ and } \lambda_2 = 1 - \rho$$
are the eigenvalues of Σ
- Verify that
$$v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$
are the corresponding eigenvectors of Σ

Example 1(Continued)

- The principal component transform matrix, using(26) is given by

$$v^T = \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

- The principal components are

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = v^T x = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- $\alpha_1 = \frac{1}{\sqrt{2}}(x_1 + x_2), \alpha_2 = \frac{1}{\sqrt{2}}(x_1 - x_2)$

Variance of principal components

- $var(\alpha_1) = var(\frac{1}{\sqrt{2}}(x_1 + x_2)) = \frac{1}{2}E[(x_1 + x_2)^2]$
  $= \frac{1}{2}[var(x_1) + var(x_2) + 2cov(x_1, x_2)]$
  $= 1 + \rho = \lambda_1$
- $var(\alpha_1) = v_1^T \Sigma v_1 = \frac{1}{2}\begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\begin{bmatrix}1 \\ 1\end{bmatrix} = 1 + \rho$
- Verify
  $$var(\alpha_2) = 1 - \rho = \lambda_2$$

A comparison of variances of components of x and α

- x is such that $var(x_1) = 1 = var(x_2)$
- α is such that $var(\alpha_1) = 1 + \rho, var(\alpha_2) = 1 - \rho$
  $var(\alpha_1) > var(\alpha_2)$, since $\rho > 0$

Covariance between x and α

- From (29)

$$cov(x, \alpha) = v\Sigma = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$
$$= \frac{1}{\sqrt{2}} \begin{bmatrix} 1+\rho & 1+\rho \\ 1-\rho & \rho-1 \end{bmatrix}$$

Correlations Ω_{ij} between x_i and α_i

- From (31), since $var(x_1) = var(x_2) = 1$

$$\Omega_{11} = \sqrt{\lambda_1} v_{11} = (\frac{1+\rho}{2})^{1/2}$$
$$\Omega_{12} = \sqrt{\lambda_2} v_{12} (\frac{1-\rho}{2})^{1/2}$$
$$\Omega_{21} = \sqrt{\lambda_1} v_{21} = (\frac{1+\rho}{2})^{1/2}$$
$$\Omega_{22} = \sqrt{\lambda_2} v_{22} (\frac{1-\rho}{2})^{1/2}$$

- $\Omega_{11}^2 + \Omega_{12}^2 = \frac{1}{2}[1 + \rho + 1 - \rho] = 1$
  $\Omega_{21}^2 + \Omega_{22}^2 = \frac{1}{2}[1 + \rho + 1 - \rho] = 1$

Example 2

- Let $x \in R^m$ and $x \sim N(\mu, \Sigma), \Sigma$ - SPD
- Let $\Sigma = v\Lambda v^T$ be the eigen decomposition
- Principal components, $\alpha = v^T(x - \mu)$
- $E(\alpha_i) = 0, var(\alpha_i) = \lambda_i, 1 \leq i \leq m$
- $cov(\alpha_i, \alpha_j) = 0$ for $i \neq j$
- $var(\alpha_i) \geq var(\alpha_2) \geq \cdots \geq var(\alpha_m) > 0$

Example 2(continued)

- $var(x) = tr(\Sigma) = \sum_{i=1}^{m} var(\alpha_i)$
- $\prod_{i=1}^{m} var(\alpha_i) = \prod_{i=1}^{m} \lambda_i = det(\Sigma)$

GOAL OF SDA

- Recall that the ultimate goal of statistical any data analysis is to explain the observed spread as measured by the variance in the data
- The total variance in the data can be modeled as the sum of the natural/inherent variation in the signal component and that of the additive noise that corrupts the signal

A look back

- The above analysis rests on two basic observations
  - (a) The total variance in the given random vector $x \in R^m$ is equal to the sum of the non-negative eigenvalues of the covariance matrix, Σ and x
  - (b) The variance of the projection of x along an eigenvector of Σ is qual to the associated eigenvalue
  - (c) The resulting additive decomposition of x as a linear combination of uncorrelated components in (20) is critical to the use of principal component analysis

A natural partitioning of variance

- Under the assumption that the eigenvalues of Σ are distinct, by ordering them in the decreasing order, it is immediate that the i^{th} component $(x^T v_i)$ of x inherits the fraction $(\lambda_i / \sum_{i=1}^{m} \lambda_i)$ of the total variance in x
- Given any β small(such as 0.01,0.05,etc..,) we can find an integer k,$(1 \leq k \leq m)$ such that

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \geq 1 - \beta \tag{33}$$

Reconstruction

- That is, the first k eigen direction corresponding the k largest eigenvalues together inherit atleast $(1-\beta)$ times the total variance
- Using these k components, we can reconstruct the signal component(see Exercise 1)

$$\hat{x} = \sum_{i=1}^{k} (x^T v_i) v_i \tag{34}$$

- The rest given by

$$x - \hat{x} = \sum_{i=k+1}^{m} (x^T v_i) v_i \tag{35}$$

is often treated as the noise component in x

Special case

- In the event that the components of x are uncorrelated, then Σ would be a diagonal matrix consisting of the eigenvalues which are the variance associated with the components
- In this case

$$x = \sum_{i=1}^{m} \lambda_i e_i \tag{36}$$

where $e_i \in R^m$ is the standard i^{th} unit vector with 1 in the i^{th} location and the zero elsewhere

- Consequently, unless the condition number $\frac{\lambda_1}{\lambda_m}$ is large or the ratio $\frac{\lambda_1 - \lambda_m}{\lambda_1}$ is close to 1, we may not get a k-mode approximation to x for small values of k

Quality of signal approximation

- The quality of the approximation in (34) can be measured by quantifying the variance of the difference $(x - \hat{x})$ in (35)
- Recall from (14)-(16) that $E(x^T v_i) = 0$ and $var(x^T v) = \lambda_i$
- Hence,

$$var(x - \hat{x}) = E[(x - \hat{x})^T (x - \hat{x})] \tag{37}$$

Quality of signal(continued)

- Substituting (35) in (37) and simplifying:

$$var(x - \hat{x}) = E[(\sum_{i=k+1}^{m}(x^T v_i)v_i)(\sum_{j=k+1}^{m}(x^T v_j)v_j)^T]$$
$$= E[\sum_{i=k+1}^{m}(x^T v_i)v_i)^2] \quad [\because v_i \perp v_j]$$
$$= \sum_{i=k+1}^{m} E(x^T v_i)^2 = \sum_{i=k+1}^{m} \lambda_i \qquad (38)$$

Inherent Optimality

- From the ordering of λ_i's in (3), it is immediate that the sum on the right hand side of (38) is the sum of the least (m-k) values of λ and hence is a minimum for every k

References

- Principal component analysis PCA was first introduced by K.Pearson(1902) "On lines and planes of closest fit to systems of points in space", Philosophical magazine, 2:599-572, and by
  H. Hotelling(1935) "The most predictable criterion", Journal of Educational psychology, 26: 139-142
- The role of PCA or Empirical Orthogonal Function(EoF) within the context of meteorological weather prediction was first analyzed by E.Lorenz(1956)"Empirical orthogonal functions ans statistical weather prediction" statistical Forecast Project Report 1, Department of Meteorology, MIT
- H.Von Storch and F.W.Zwiers (1999) Statistical Analysis in Climate Research, Cambridge university press

Exercises

- Consider the expression in (34):

$$\hat{x} = (x^T v_1)v_1 + (x^T v_2)v_2 + \cdots + (x^T v_l)v_k \qquad (39)$$

using $(x^T v_i)v_i = v_i(x^T v_i) = v_i(v_i^T x) = (v_i v_i^T)x$ rewrite $\hat{x}$ in (39) as

$$\hat{x} = [(v_1 v_1^T) + (v_2 v_2^T) + \cdots + (v_k v_k^T)]x \qquad (40)$$

$$= \begin{bmatrix} v_1 & v_2 & \dots & v_k \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_k^T \end{bmatrix} x$$

$$= V[I - k]V^T x$$

where $V \in R^{m \times m}$ and $k = Diag(k_{11}, k_{22}, \ldots, k_{mm})$ with

$$k_{ii} = 0 \text{ for } 1 \leq i \leq k$$
$$= 1 \text{ for } k+1 \leq 1 \leq m$$

MODULE 6.2
Principal Component Analysis (PCA)
: As an optimal process

by
S.Lakshmivarahan
School of Computer Science
University of Oklahoma
Norman, OK-73019, USA
varahan@ou.edu

PCA an optimal process

- Given a random vector $x \in R^m$, find the principal patterns an components as an optimization process
- There are two key steps in this development
- First, given a k-dimensional subspace s_k of R^m for $1 \leq k \leq m$, find the best approximation $\hat{x}(k)$ of x where $\hat{x}(k) \in s_k$
- Second, find the specific subspace s_k that inherits the maximum variance of x

Specification of s_k

- Let s_k be spanned by the orthonormal columns of $H \in R^{m \times k}$ given by
- From

$$H = [h_1, h_2, \ldots, h_k] \tag{1}$$

$$\begin{aligned} h_i^T h_j &= 0 \text{ for } i \neq j \\ &= 1 \text{ for } i = j \end{aligned}$$

- Verify

$$H^T H = I_k \tag{2}$$

-

$$\begin{aligned} HH^T &\in R^m - symmetric \\ (HH^T)^2 &= HH^T - idempotent \end{aligned} \tag{3}$$

Best approximation $\hat{x}(k)$

- Let

$$\hat{x}(k) = H\alpha, \alpha \in R^K \tag{4}$$

- By orthogonal projection theorem: using(2)

$$\alpha = (H^T H)^{-1} H^T x = H^T x \tag{5}$$

- That is

$$\alpha_i = h_i^T x \quad for \quad 1 \leq i \leq k \tag{6}$$

Error in the projection

- $$e = x - \hat{x}(h) = x - H\alpha = x - HH^T x = (I_m - HH^T)x \quad (7)$$

Variance of the error

- $$\begin{aligned} var(e) &= E[e^T e] \\ &= E[x^T(I - HH^T)^T(I - HH^T)x] \\ &= E[x^T(I - HH^T)^2 x] \\ &= E[x^T(I - HH^T)x] \\ \\ &= E(x^T x) - E[(x^T H)(H^T x)] \end{aligned} \tag{8}$$

Variance of the error

- $$E(x^T x) = var(x) = tr(\Sigma) \tag{9}$$
- $$E[(x^T H)(H^T x)] = E[\alpha^T \alpha] \quad (using (5))$$
$$= \sum_{i=1}^{k} E(\alpha_i^2) \tag{10}$$
- $$E(\alpha_i^2) = E[h_i^T x h_i^T x] = E[h_i^T x x^T h_i]$$
$$= h_i^T E(xx^T) h_i$$
$$= h_i^T \Sigma h_i \tag{11}$$

Variance of the error

- Substituting (9),(10) and (11) in (8):

$$var(e) = tr(\Sigma) - \sum_{i=1}^{k} h_i^T \Sigma h_i \tag{12}$$

- Since $tr(\Sigma)$ is fixed, $var(e)$ is a minimum when the second term on the right hand side of (12) that represents the total variance of the k principal component $\alpha_i, 1 \leq i \leq k$ is a maximum

Optimization problem

- Let

$$Q = \sum_{i=1}^{k} h_i^T \Sigma h_i \tag{13}$$

- Goal is to maximize (13) subject to two conditions on $h_i, 1 \leq i \leq k$:

$$h_i^T h_i = 1 \quad \textit{and} \quad h_i^T h_j = 0 \quad \textit{for} \quad i \neq j \tag{14}$$

Lagrangian, L

- Build the Lagrangian

$$L(H, \mu, \eta) = \sum_{i=1}^{k} h_i^T \Sigma h_i + \sum_{i=1}^{k} \mu_i (1 - h_i^T h_i) + \sum_{i \neq j} \eta_{ij} h_i^T h_j \tag{15}$$

Necessary condition(NC) for a maximum

- $$\bigtriangledown_{h_i} L = 0$$

$$= 2\Sigma h_i - 2\mu_i h_i + \Sigma_{j \neq i} \eta_{ij} h_j \qquad (16)$$

NC for maximum

- Multiplying both sides on the left by h_i^T and exploiting the orthonormality of h_i's:

$$0 = 2h_i^T \Sigma h_i - 2\mu_i h_i^T h_i + \sum_{j \neq i} \eta_{ij} h_i^T h_j$$

$$= 2[h_i^T \Sigma h_i - \mu_i] \tag{17}$$

- Hence

$$h_i^T \Sigma h_i = \mu_i \quad or \quad \Sigma h_i = \mu_i h_i \tag{18}$$

Relation to eigen structure of Σ

- From (18): (μ_i, h_i) are the eigen pair of Σ
- Since we are interested in the maximum of the sum

$$\sum_{i=1}^{k} h_i^T \Sigma h_i = \sum_{i=1}^{k} \mu_i \tag{19}$$

it follows that (λ_i, h_i) are the eigen pairs of Σ corresponding to the k largest eigenvalues of Σ where we assume that

$$\mu_1 > \mu_2 > \cdots > \mu_k > \ldots \mu_m > 0 \tag{20}$$

Langrangian multipliers η_{ij}

- Multiplying both sides of (16) on the left by $h_p, p \neq i$

$$0 = 2h_p^T \Sigma h_i - 2\mu_i h_p^T h_i + \sum_{j \neq i} \eta_{ij} h_p^T h_j \tag{21}$$

- Since $H^T \Sigma H = Diag(\mu_1, \mu_2, \ldots, \mu_k), h_p^T \Sigma h_i = 0$ for $p \neq i$
- Since $p \neq i$, the only term that survives for $j = p \neq i$ which is η_{ip}
- Hence

$$\eta_{ip} = 0 \tag{22}$$

Langrangian multipliers η_{ij}

- By running p over the set $1, 2, \ldots, k$ and $p \neq i$, for (k-1) values of $p \neq i$, we get

$$\eta_{ip} = 0 \tag{23}$$

- By repeating this argument for each i, it follows that all η_{ij} for $i \neq j$ are all zeros

Summary

- By choosing the k-orthonormal columns of H to be the k-orthonormal eigenvectors corresponding to the k largest eigenvalues of Σ, we maximize the sum

$$\sum_{i=1}^{k} h_i^T \Sigma h_i = \sum_{i=1}^{k} \mu_i \tag{24}$$

PC expansion for x

- Recall $v \in R^{m \times m}$ and $\Lambda = Diag(\lambda_1, \ldots, \lambda_m)$:

$$v^T \Sigma v = \Lambda \quad or \quad \Sigma = v \Lambda v^T \tag{25}$$

- Setting $k = m, H = v$,

$$\mu_i = \lambda_i \tag{26}$$

- From (4):

$$x = \hat{x}(m) = v\alpha \quad and \quad \alpha = v^T x \tag{27}$$

which is the same as in Module 6.1

Example

- An example may help illustrate the derivation of the necessary condition for maximum in (16)
- Set $k = 3$

Three equations for $i = 1, 2, 3$

- From (16) we get

$$0 = 2\Sigma h_1 - 2\mu_1 h_1 + \eta_{12} h_2 + \eta_{13} h_3 \tag{28}$$

$$0 = 2\Sigma h_2 - 2\mu_2 h_2 + \eta_{21} h_1 + \eta_{23} h_3 \tag{29}$$

$$0 = 2\Sigma h_3 - 2\mu_3 h_3 + \eta_{31} h_1 + \eta_{32} h_2 \tag{30}$$

Conditions from (28)

- Multiplying both sides of (28) in turn on the left by h_1^T, h_2^T and h_3^T, using orthonormality we get
- $0 = 2h_1^T \Sigma h_1 - 2\mu_1 h_1^T h_1 \implies \Sigma h_1 = \mu_1 h_1$
- $0 = 2h_2^T \Sigma h_1 - 2\mu_1 h_2^T h_1 + \eta_{12} h_2^T h_2 + \eta_{13} h_2^T h_3$ we get $\eta_{12} = 0$
- $0 = 2h_3^T \Sigma h_3 - 2\mu_1 h_3^T h_1 + \eta_{12} h_3^T h_1 + \eta_{13} h_3^T h_3$ we get $\eta_{13} = 0$

Conditions from (29)

- Multiplying both sides of (29) in turn on the left by h_1^T, h_2^T and h_3^T, we get
$$\Sigma h_2 = \mu_2 h_2, \eta_{21} = 0, \eta_{23} = 0$$
- Similar action on (30) gives
$$\Sigma h_3 = \mu_3 h_3, \eta_{31} = 0, \eta_{32} = 0$$

MODULE 6.3
Data Matrix : Generation

by
S.Lakshmivarahan
School of Computer Science
University of Oklahoma
Norman, OK-73019, USA
varahan@ou.edu

Why data matrix ?

- Analysis thus far assumed the knowledge of the properties of a random vector $x \in R^m$
- In real world applications, we do not know these second-order properties
- Have access only to an ensemble of realization of x obtained through direct measurements
- First step: organize this ensemble data in the form of a data matrix where each column is a realization of x

Data matrix : $x \in R^{m \times n}$

- $$x = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & \dots & j & \dots & n \end{array} \\ \begin{bmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{bmatrix} \end{array}$$

- Each column of x refer to an object
- Each row of x refer to an attribute of the object

Data matrix

- The j^{th} column, x_{*j} refer to the profile of the j^{th} object
- The i^{th} row, x_{i*} refer to the values of the i^{th} attribute of all the objects
- x_{ij} is the i^{th} attribute of the j^{th} object

Example 1 - Classification of Students

- Objects refer to n students in a class
- The m attributes refer to the grades in a set of m-courses taken by each of the n student
- x_{ij} is the grade of the j^{th} student in the i^{th} course

Example 2 - Classification of Humans

- Objects refer to a set of n humans
- Attributes may refer to height, color of skin, weight, length of the torso, education, head size, color of the eye, blood group

Example 3 - Classification of Models

- The set of n objects may denote a set of initial condition for a class of models
- The attribute may denote the solution of the model on a 2-D grid with $m = m_x m_y$ points

Example 4 - Meteorology

- In weather analysis/prediction, it is of interest to understand the variation of the (geopotential) height of the atmosphere at different pressure levels, say 900, 700, 500, 300, 100 mbar

100 _____ 5
300 _____ 4
500 _____ 3
700 _____ 2
900 _____ 1

m = 5 levels

- Geopotential, ϕ is defined as the work required to raise unit mass from the surface of the earth to height h:
$\phi(h) = \int_0^h g dh$

Example 4 - Continued

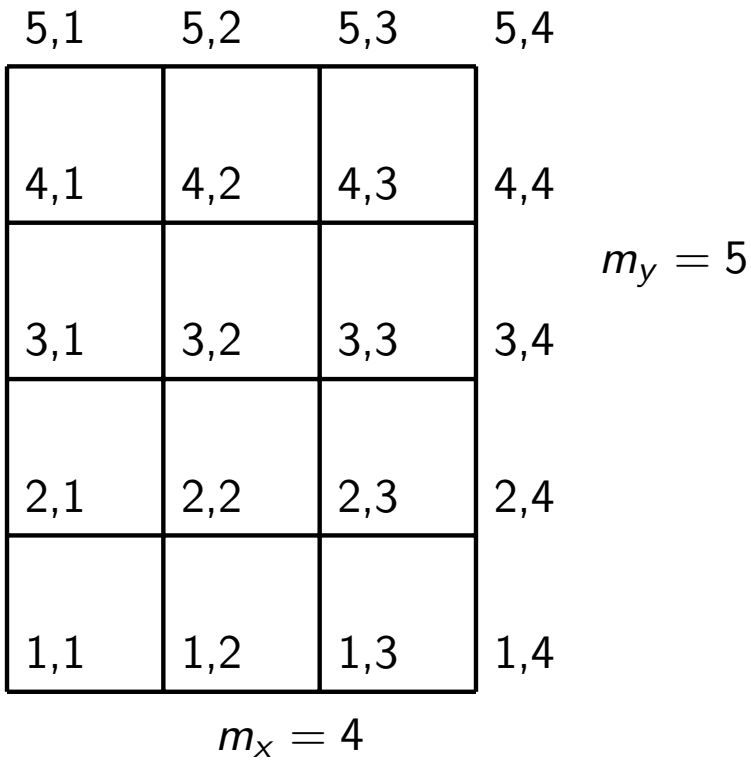
- Baloons with instruments for measuring pressure, height, temperature, humidity etc are hoisted from a given location
- Once a day, for 120 days with 15 days before the start and 15 days after the and of a given season - say winter in northern hemisphere for 10 successive years
- Here $m = 5, n = 120 \times 10 = 1,200$ days

Example 5 - Climate Analysis

- Spatio-temporal distribution of sea surface temperature(SST) across the globe, distribution of the concentration of green house gases etc, is of great interest in climate studies
- For simplicity, consider a 2-D version of this problem
- Pick a domain of interest and embed an uniform 2-D grid with m_x number of points along the east-west and m_y number of points, along the north-south direction for a total of $m = m_x m_y$ points
- Here x is a random vector of size m

2-D grid numbering

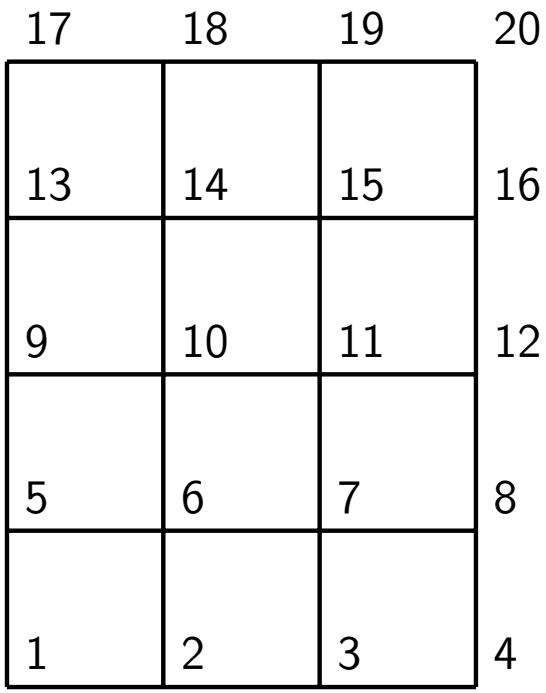
- Consider a grid with $m_x = 4$ and $m_y = 5$ for a total 20 points
- Points are labeled with two indices (p,q) where p refers to the level and q refers to the node at that level
- 4,3 is the third node at the fourth level



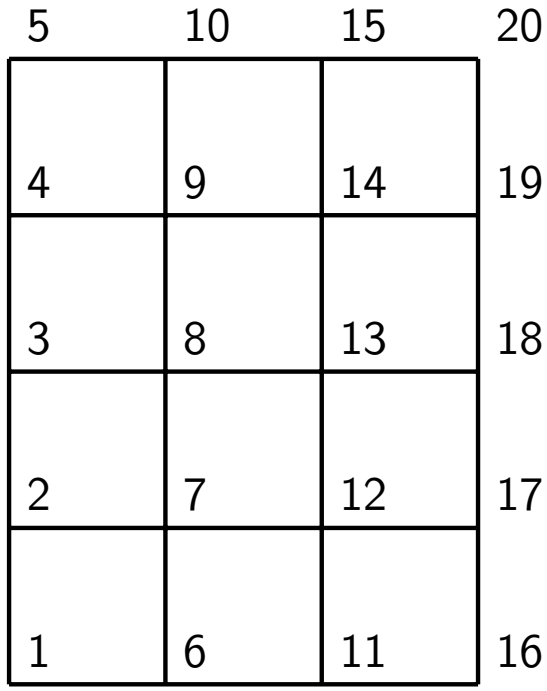
Row / Column major order

- From computing perspective, it is useful to number the nodes using a simple index so that the data across the grid can be stored in an 1-D array
- Two possibilities: map (p,q) to a single integer
  Row major order : $k = (p-1)m_x + q$
  Column major order : $s = (q-1)m_y + p$

Re-numbered 2-D grid



- Row-major order
- (p, q)= (4, 3) $\Leftrightarrow k = 15$



- Column-major order
- (p, q)= (4, 3) $\Leftrightarrow s = 14$

Generation of data matrix

- Let $m_x = 31$ and $m_y = 16$ with $m = 496$
- Let $a_1 \in N(0, \sigma_1^2)$, $a_2 \in N(0, \sigma_2^2)$, $\varepsilon(x, y) \in N(0, \sigma_3^2)$
- Define, for $1 \leq t \leq 100$

$$g_1(x, y, t) = a_1(t) cos(\frac{\pi x}{30}) cos(\frac{\pi y}{15}) \quad (1)$$

$$g_2(x, y, t) = a_2(t) cos(\frac{\pi x}{15}) cos(\frac{\pi y}{7}) \quad (2)$$

- Let

$$g(x, y, t) = g_1(x, y, t) + g_2(x, y, t) + \varepsilon(x, y, t) \quad (3)$$

First column of the Data matrix (without noise $\varepsilon(x, y, t)$)

- Set t = 1, generate $a_1(1)$ and $a_2(1)$ by setting $\sigma_1^2 = 0.6$ and $\sigma_2^2 = 0.3$
- For $0 \leq x \leq 30$ and $0 \leq y \leq 15$, compute $g_1(x, y, 1), g_2(x, y, 1)$ and $g(x, y, 1) = g_1(x, y, 1) + g_2(x, y, 1)$ to obtain a column vector using column major order, for example $Z_{*1} \in R^{496}$ which is the first column of the data matrix, $Z \in R^{496 \times n}$

Entire matrix $Z \in R^{496 \times 100}$

- For $t = n = 2, 3, \ldots, 100$, repeat the above process by generating a new pair, $(a_1(t), a_2(t))$ of random numbers and compute the t^{th} column of Z for $2 \leq n \leq 100$
- Clearly, each row corresponds to a grid point and each column to an instant in time
- The j^{th} column gives the profile of the variable of interest across the grid at time j
- The i^{th} row gives the distribution of the variable at a grid point across time.

Noisy data matrix $\bar{Z} \in R^{496 \times 100}$

- For each time index $t, 1 \leq t = n \leq 100$, generate the spatial noise vector $\eta(t) \in R^m, m = 496$ where the components are uncorrelated gaussian noise with mean zero and variance $\sigma_3^2 = 0.2$
- Repeating this process N = 100 times, create a matrix $\eta \in R^{m \times n}$
- Create a noisy data matrix :

$$\bar{Z} = Z + \eta \tag{4}$$

where t was obtained earlier

Noiseless and noisy data matrices

- The matrix $Z \in R^{496\times 100}$ is a data matrix that represent a 100 member ensemble of realization of the 2-D field variable g(x,y,t) defined in (3) by setting $\varepsilon(x, y, t) = 0$
- The matrix $\bar{Z} \in R^{496\times 100}$ is a data matrix that represents a 100 member ensemble of the noisy realization of the field variable g(x,y,t) in (3) with the noise matrix η added to Z, that is, $\bar{Z} = Z + \eta$
- These two matrices will be used to test the methodology

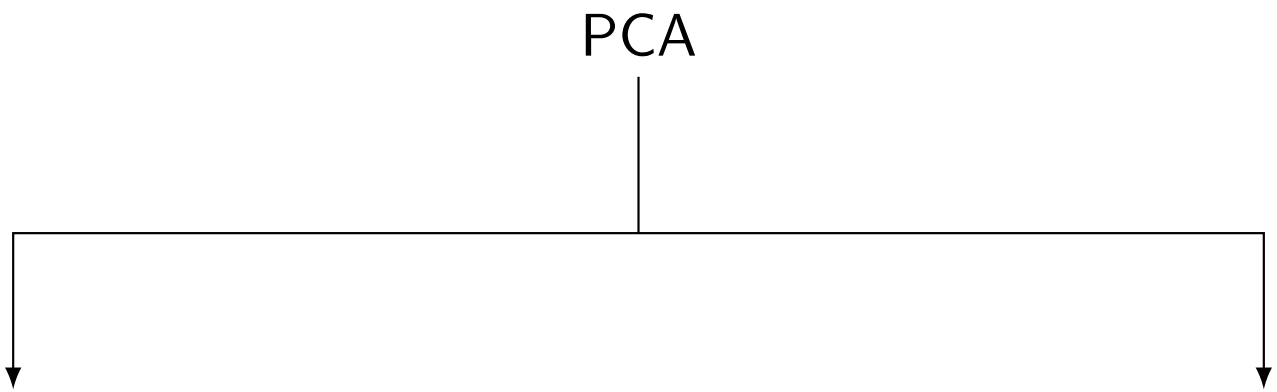
References

- G. Eshel(2012) Spatiotemporal Data Analysis, Princeton University press, contains a good discussion of several examples of spatiotemporal data analysis using EOF

MODULE 6.4
PCA using data matrix: Empirical Orthogonal Function (EoF)

by
S.Lakshmivarahan
School of Computer Science
University of Oklahoma
Norman, OK-73019, USA
varahan@ou.edu

Population vs. sample based PCA



- Population PCA
- $x \in \mathbb{R}^m$, random vector
- μ, Σ - known
- $\Sigma = v \Lambda v^T$
- $x = \sum_{i=1}^{m} (x^T v_i) v_i$
- Reconstructed x $= \hat{x} = \sum_{i=1}^{k} (x^T v_i) v_i$

- Sample based PCA
- μ, Σ - not known
- Work with data matrix, $x \in R^{m \times n}$
- Estimate $\hat{\mu}, \hat{\Sigma}$
- PCA based on $\hat{\mu}, \hat{\Sigma}$ called EoF based analysis

Data matrix

- Assume that the raw data matrix, $X \in R^{m \times n}$ is given
- First step towards EoF analysis is to extract the underlying covariance/correlation structure of data
- This calls for transforming the data:
  - centering
  - normalizing, if the units across the rows of z are widely different

Compute row mean

- Let $J_n = (1, 1, \ldots, 1)^T \in R^n$ be a column vector of all 1's. $J_4 = (1, 1, 1, 1)^T$
- Let $M = (M_1, M_2, \ldots, M_m)^T$ be the row mean vector where

$$M_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} = \frac{1}{n} (X_{i*}) J_n \tag{1}$$

- Then

$$M = \frac{1}{n} X J_n \tag{2}$$

Centering the data

- Let

$$\tilde{X} = [\tilde{X}_{ij}] \in R^{m \times n} \tag{3}$$

be the centered data matrix where

$$\tilde{X}_{ij} = X_{ij} - M_i, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n \tag{4}$$

- Then

$$\tilde{X} = [X - M{J_n}^T] \tag{5}$$

where $M{J_n}^T$ is the $m \times n$ outer product matrix

Normalized data

- Let ${s_i}^2$ be the sample variance of the i^{th} row of X. Then

$$ {s_i}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - M_i)^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\tilde{X}_{ij})^2 \qquad (6) $$

- The normalized data matrix, $\hat{X}$ is given by

$$ \hat{X} = [\hat{X}_{ij}] \; and \; \hat{X}_{ij} = \frac{\tilde{X}_{ij}}{s_i} \qquad (7) $$

Normalized data matrix

- Define a diagonal matrix

$$D = diag(s_1^2, s_2^2, \ldots, s_m^2) \tag{8}$$

consisting of m sample variances across the diagonal of D

- Then

$$\hat{X} = D^{-1/2}\tilde{X} \tag{9}$$

where

$$D^{1/2} = Diag(s_1, s_2, \ldots, s_m) \tag{10}$$

Covariance matrix, $C \in R^{m \times m}$

- Let $C = cov(x)$ be the sample covariance of the data in the matrix X
- That is,

$$C_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} \tilde{X_{ik}} \tilde{X_{jk}} \text{ for } i \neq j$$
$$= \frac{1}{n-1} \sum_{k=1}^{n} (\tilde{X_{ik}}^2) = s_i^2 \text{ for } i = j \tag{11}$$

- Then

$$C = \frac{1}{n-1} \tilde{X} (\tilde{X})^T \tag{12}$$

Correlation matrix, $R \in R^{m \times m}$

- Let $R = cov(\hat{x})$ be the sample correlation matrix
- That is,

$$\begin{aligned} R_{ij} &= \frac{1}{n-1}\sum_{k=1}^{n}(\hat{X_{ik}}\hat{X_{jk}}) \quad for \quad i \neq j \\ &= \frac{1}{n}\sum_{k=1}^{n}(\hat{X_{ik}})^2 = 1 \quad for \quad i = j \end{aligned} \tag{13}$$

- Then

$$\begin{aligned} cor(z) = cov(\hat{z}) = \frac{1}{n-1}\hat{X}(\hat{X})^T &= D^{-1/2}\frac{\tilde{X}(\tilde{X})^T}{(n-1)}D^{-1/2} \\ &= D^{-1/2}CD^{-1/2} \end{aligned} \tag{14}$$

- Clearly:

$$|R_{ij} \leq 1| \tag{15}$$

A prelude to SVD analysis

- Module 1.2 contains the theoretical basis for SVD analysis of a general matrix, $H \in R^{m \times n}$
- Recall that if we multiply H by a constant $\alpha > 0$, the non-zero eigenvalues of $H^T H$ and HH^T get multiplied by α^2 and the singular value of H by α
- A quick review of Module 1.2 reveal that there are a number of ways in which the above theory can be applied for the SVD analysis of the data matrix, X

SVD using second moment matrix

- Set $H = \frac{1}{\sqrt{n}} X \in R^{m \times n}$ - the raw data matrix
- Grammians: $H^T H = \frac{1}{n} X^T X \in R^{n \times n}$
  $HH^T = \frac{1}{n} X X^T \in R^{m \times m}$

SVD using covariance matrix

- Set $H = \frac{1}{\sqrt{n}}\tilde{X} \in R^{m \times n}$ - centered data or anomaly matrix
- Grammians: $H^T H = \frac{1}{n}(\tilde{X})^T \tilde{X} \in R^{n \times n}$ - covariance
  $HH^T = \frac{1}{n}\tilde{X}(\tilde{X})^T \in R^{m \times m}$

SVD using normalized matrix

- Set $H = \frac{1}{\sqrt{n}}\hat{X} \in R^{m \times n}$ - normalized data matrix
- Grammians: $H^T H = \frac{1}{n}(\hat{X})^T \hat{X} - correlation$
  $HH^T = \frac{1}{n}\hat{X}(\hat{X})^T$

SVD Analysis of $H \in R^{m \times n}$: case 1 Let $m > n$

- Consider the smaller of the two grammians: $H^T H \in R^{n \times n}$
- Let (λ_i, v_i) be an eigen pair of $H^T H$: $(H^T H)v_i = \lambda_i v_i$ where

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0 \tag{16}$$

- Define $v = [v_1, v_2, \ldots, v_n] \in R^{n \times n}, vv^T = v^T v = I_n$
- $\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_n) \in R^{n \times n}$
- Then

$$(H^T H)v = v\Lambda \tag{17}$$

Case 1 - continued

- Define

$$u_i = \frac{1}{\sqrt{\lambda_i}} Hv \in R^m, 1 \leq i \leq n \tag{18}$$

- Verify: (λ_i, u_i) be an eigenpair of $HH^T \in R^{m \times m}$
- Let $u = [u_1, u_2, \ldots, u_n] \in R^{m \times n}, u^T u = I_n$
- Then

$$(HH^T)u = u\Lambda \tag{19}$$

Case 1 - continued

- From (18)

$$Hv_i = u_i \lambda_i^{1/2} for 1 \leq i \leq n \tag{20}$$

- Hence

$$Hv = u\Lambda^{1/2} \tag{21}$$

- The SVD of H : $H = u\Lambda^{1/2} v^T$

$$= \sum_{i=1}^{n} \lambda_i^{1/2} u_i v_i^T \tag{22}$$

SVD of $H \in R^{m \times n}$: Case 2 Let $n > m$

- Consider the smaller of the two grammians: $HH^T \in R^{m \times m}$
- Let $(\lambda_i, u_i), 1 \leq i \leq m$ be an eigen pair of HH^T, that is $(HH^T)u_i = u_i\lambda_i$ where

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0 \tag{23}$$

- If $u = [u_1, u_2, \ldots, u_m] \in R^{m \times m}, uu^T = u^T u = I_m$ then

$$(HH^T)u = u\Lambda \tag{24}$$

where

$$\Lambda = Diag(\lambda_1, \lambda_2, \ldots, \lambda_m) \in R^{m \times m} \tag{25}$$

Case 2 - continued

- Define

$$v_i = \frac{1}{\sqrt{\lambda_i}} H^T u_i \in R^n, 1 \leq i \leq m \tag{26}$$

- Then $(H^T H)v_i = \frac{1}{\sqrt{\lambda_i}}(H^T H)H^T u_i = \frac{1}{\sqrt{\lambda_i}} H^T (HH^T)u_i$

$$= \frac{1}{\sqrt{\lambda_i}} H^T u_i \lambda_i = v_i \lambda_i \tag{27}$$

  that is (λ_i, v_i) be an eigen pair of $H^T H$

- Setting $v = [v_1, v_2, \ldots, v_m] \in R^{n \times m}$, we get

$$(H^T H)v = v\Lambda \tag{28}$$

Case 2 - continued

- From (26)

$$H^T u_i = v_i \lambda_i^{1/2} for 1 \leq i \leq m \tag{29}$$

- Using $u \in R^{m \times m}, u^T u = uu^T = I, v \in R^{n \times m}$ (29) becomes

$$H^T = v \Lambda^{1/2} u^T$$

- SVD of H :

$$H = u \Lambda^{1/2} v^T$$

$$= \sum_{i=1}^{m} \lambda_i u_i v_i^T \tag{30}$$

Use of SVD to approximate H

- The SVD of H in (22) and (30) can be used in two distinct ways to approximate H
- First: We can use SVD to decompose H into a signal and a noise components
- Second: We can use SVD to reduce the dimension m of the data matrix H to $d < m$ to obtain $\bar{H}_1 \in R^{d \times n}$ that is an approximation to H when m is large

FIRST: Reconstruction of signal:Case 2: $n > m$

- For definiteness, consider case 2 when $n > m$
- Let $0 \leq \beta \leq 1$ be a given(small) real number
  eg: $\beta = 0.1, 0.05, 0.01$ etc
- Let k be the smallest integer such that

$$\sum_{i=1}^{k} \lambda_i \geq (1-\beta) \sum_{i=1}^{m} \lambda_i \tag{31}$$

  where λ_i's ordered as in (15) and (23)
- The signal component H_1 is given by

$$H_1 = \sum_{i=1}^{k} \lambda_i u_i v_i^T \tag{32}$$

- The noise component H_2 is given by

$$H_2 = H - H_1 = \sum_{i=k+1}^{m} \lambda_i u_i v_i^T \tag{33}$$

Signal - noise decomposition of H

- Recall:

$$H = u\Lambda^{1/2}v^T \tag{34}$$

where $u \in R^{m\times n}, \Lambda^{1/2} \in R^{n\times n}, v \in R^{n\times n}$ and $H \in R^{m\times n}$

- For the k in (31), define partitions of u, v and $\Lambda^{1/2}$

$$u = [u_1, u_2], u_1 \in R^{m\times k}, u_2 \in R^{m\times(n-k)}$$
$$v = [v_1, v_2], v_1 \in R^{n\times k}, v_2 \in R^{n\times(n-k)}$$

$$\Lambda^{1/2} = \begin{bmatrix} \Lambda_1^{1/2} & 0 \\ 0 & \Lambda_1^{1/2} \end{bmatrix} \Lambda_1^{1/2} \in R^{k\times k}, \Lambda_2^{1/2} \in R^{(n-k)\times(n-k)}$$

Decomposition of H

- Then

$$H = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \Lambda_1^{1/2} & 0 \\ 0 & \Lambda_1^{1/2} \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix}$$

$$= u_1 \Lambda_1^{1/2} v_1^T + u_2 \Lambda_2^{1/2} v_2^T \quad (35)$$

$$= H_1 + H_2 \quad (36)$$

- H_1 is the signal component and H_2 is the noise component

A measure of the quality of approximation

- In approximating H by H_1 in (35), we need to develop a measure to quantify the goodness of the approximation
- To this end, assume that the data matrix H is a full-rank matrix, that is,

$$Rank(H) = min\{m, n\} = m \tag{37}$$

- It turns out that the signal part H_1 defined in (32) and (35) enjoys the property of being the "best" rank-k approximations to H in the sense that the noise component H_2 has an inherent minimality under a suitably defined matrix norm

Euclidean norm and energy of a vector

- Let $a \in R^n$
- The euclidean norm of the vector denoted by $||a||$ is given by

$$||a|| = (a_1^2 + a_2^2 + \cdots + a_n^2)^{1/2} \tag{38}$$

- The square of this norm,$||a||^2$ is a measure of the generalized energy associated with a
- Clearly $||a|| = 0$ exactly when $a = 0$

Frobenius norm and energy of a matrix

- Let $A \in R^{m \times n}$
- The Frobenius norm of the matrix A, denoted by $||A||_F$ is given by

$$||A_F|| = [\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2]^{1/2} \tag{39}$$

- $||A||_F^2$ is a measure of the energy associated with A
- $||A||_F = 0$ exactly when $A = 0$

Outer product matrix

- Let $u \in R^m$ and $v \in R^n$
- Then

$$B = uv^T = [u_i v_i] \tag{40}$$

  is a rank one matrix
- Let $m = 3$ and $n = 2$. Then

$$B = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} u_1 v_1 & u_1 v_2 \\ u_2 v_1 & u_2 v_2 \\ u_3 v_1 & u_3 v_2 \end{bmatrix}$$

Energy of an outer product matrix

- $$||B||_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} u_i^2 v_j^2$$

$$= \sum_{i=1}^{m} u_i^2 \sum_{j=1}^{n} v_j^2 = ||u||^2 ||v||^2 \quad (41)$$

- Let u and v are unit vectors, then

$$||B||_F^2 = ||uv^T||_F^2 \quad (42)$$

A property of Frobenius norm

- Let $A \in R^{m \times n}$. Then

$$||A||_F^2 = tr(AA^T) \tag{43}$$

- Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad AA^T = \begin{bmatrix} a^2 + b^2 & ac + bd \\ ac + bd & c^2 + d^2 \end{bmatrix}$$

- Verify (43)

Energy in noise component H_2

- From (35) and (36)

$$\begin{aligned} ||H_2||_F^2 &= t_r(H_2 H_2^T) \\ &= tr(u_2 \Lambda_2^{1/2} v_2^T v_2 \Lambda_2^{1/2} u_2^T) \\ &= tr(u_2 \Lambda_2 u_2^T) \quad [\because v_2^T v_2 = I_{n-k}] \\ &= tr(\textstyle\sum_{i=k+1}^{m} \lambda_i u_i v_i^T) \quad [\because tr(u_i u_i^T) = 1] \end{aligned}$$

$$= \sum_{i=k+1}^{m} \lambda_i tr(u_i u_i^T) = \sum_{i=k+1}^{m} \lambda_i \qquad (44)$$

Optimality of the signal component H_1

- In the light of the ordering of the λ_i's in (23)
  $||H_2||_F^2$ = sum of the least (m-k) eigenvalues of the smaller Grammian HH^T
- Hence, $||H_2||_F^2$ is a minimum for any k that satisfies (31)
- A similar arguments applies for the case 1 : $n < m$
- We encourage the reader to fillout the details

Second: Dimension reduction(DR): case 1 : $m > n$

- From (15):

$$H = u\Lambda^{1/2}v^T \tag{45}$$

- Recall:

$$\begin{gathered} u \in R^{m\times n}, \Lambda^{1/2} \in R^{n\times n}, V \in R^{n\times n} \\ u^T u = I_n, vv^T = v^T v = I_n \end{gathered} \tag{46}$$

- (45) then becomes:

$$\bar{H} = u^T H = u^T u\Lambda^{1/2}v^T = \Lambda^{1/2}v^T \tag{47}$$

An useful partitioning of u, v, Λ

- $u = [u_1, u_2], u_1 \in R^{m \times d}, u_2 \in R^{m \times n-d}$
- $v = [v_1, v_2], v_1 \in R^{n \times d}, v_2 \in R^{n \times n-d}$
- $$\Lambda^{1/2} = \begin{bmatrix} \Lambda_1^{1/2} & 0 \\ 0 & \Lambda_2^{1/2} \end{bmatrix}, \quad \Lambda_1^{1/2} \in R^{d \times d}, \Lambda_2^{1/2} \in R^{(n-d) \times (n-d)}$$
- Then $u_1^T u_1 = I_d, u_2^T u_2 = I_{n-d}$

A partitioning of left hand side in (47)

- Substituting these partitions in the left hand side of (47) and simplifying:
- $$\bar{H} = \begin{bmatrix} u_1^T \\ u_2^T \end{bmatrix} \quad H = \begin{bmatrix} u_1^T H \\ u_2^T H \end{bmatrix} = \begin{bmatrix} \bar{H}_1 \\ \bar{H}_2 \end{bmatrix} \tag{48}$$
- $\bar{H}_1 \in R^{d \times n}$ and $\bar{H}_2 \in R^{m-d \times n}$

A partitioning of the right hand side in (47)

- $$\Lambda^{1/2} v^T = \begin{bmatrix} \Lambda_1^{1/2} & 0 \\ 0 & \Lambda_2^{1/2} \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix} = \begin{bmatrix} \Lambda_1^{1/2} v_1^T \\ \Lambda_2^{1/2} v_2^T \end{bmatrix} \quad (49)$$

A root partition of H_1

- Combining (47) - (49) :
-

$$\bar{H} = \begin{bmatrix} u_1^T \\ u_2^T \end{bmatrix} \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \Lambda_1^{1/2} v_1^T \\ \Lambda_2^{1/2} v_2^T \end{bmatrix} = \begin{bmatrix} I_d & 0 \\ 0 & I_{n-d} \end{bmatrix} \begin{bmatrix} \Lambda_1^{1/2} v_1^T \\ \Lambda_2^{1/2} v_2^T \end{bmatrix} = \begin{bmatrix} \Lambda_1^{1/2} v_1^T \\ \Lambda_2^{1/2} v_2^T \end{bmatrix} \quad (50)$$

Approximation of H by H_1 : DR

- $\bar{H_1} \in R^{d \times n}$ is called the d-dimensional approximation to $H \in R^{m \times n}$ where $d < m$
- It turns out that this representation of H by $\bar{H_1}$ has a natural optimality property as proved below

Energy in $\bar{H}_2 \in R^{(m-d)\times n}$

- From the definition of the Frobenius norm,it follows that

$$||\bar{H}||_F^2 = ||\bar{H}_1||_F^2 + ||\bar{H}_2||_F^2 \tag{51}$$

- From (50):

$$\begin{aligned} ||H_2||_F^2 &= tr(\bar{H}_2 \bar{H}_2{}^T) \\ &= tr(\Lambda_2^{1/2} v_2^T v_2 \Lambda_2^{1/2}) \\ &= tr(\Lambda_2) = \sum_{i=d+1}^{n} \lambda_i \end{aligned} \tag{52}$$

Optimality of $\bar{H}_1$

- In the light of the ordering in (15), the right hand side of (52) is the sum of the least (n-d) eigenvalues of the (smaller) Grammian $H^T H \in R^{n \times n}$
- Hence, $\bar{H}_1$ enjoys the inherent optimality property of "optimal reduced dimensional representation of H"