29/09/14 Multivariate Analysis Slides 2 Extra Notes

Vector notation

n rows.
m columns,
n = total number of observational units
m = number of variables

$$X = \begin{pmatrix} x_{11} & x_{12} & & x_{1m} \\ \vdots & \cdot & & \\ x_{n1} & x_{n2} & & x_{nm} \end{pmatrix}$$

age height etc

A^T is transpose of A

$$[A^T]_{ij} = A_{ji}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

Variance, Covariance, and Correlation.

$$\begin{aligned} Var(x) &= E[(X-\mu)^2] \qquad Var(x) = Cov(X,X) \\ &= E[x^2 - 2XE[X] + (E[X])^2] \\ &= E[x^2] - 2E[x]E[x] + (E[x])^2 \\ &= E[x^2] - (E[x])^2 \end{aligned}$$

$$\begin{aligned} Cov[x,y] &= E[(x-E[x])(y-E[y])] \\ &= E[xy - xE[y] - E[x]y + E[x]E[y]] \\ &= E[xy] - E[x]E[y] - E[x]E[y] + E[x]E[y] \\ &= E[xy] - E[x]E[y] \end{aligned}$$

$$Cor[X,y] = \frac{Cov[X,y]}{\sqrt{Var[x]\,Var[y]}}$$

Covariance Matrix

Record variance and covariance of a set of random variables $X = (X_1, X_2 \dots X_m)$ using a matrix.

$$\underset{\text{(cov matrix)}}{\Sigma} = \begin{bmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & E[(X_1-\mu_1)(X_2-\mu_2)] & \dots & E[(X_1-\mu_1)(X_n-\mu_n)] \\ \vdots & \ddots & & \vdots \\ E[(X_n-\mu_n)(X_1-\mu_1)] & \dots & & E[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}$$

$Cov(X,X) = Var(X)$ so matrix also represents variance

Independence

Independent iff $P(X_1=x_1, X_2=x_2) = P(X_1=x_1)\,P(X_2=x_2)$

For two independent random variables X_1 and X_2 $E[X_1X_2] = E[X_1]E[X_2]$

$$E[X_1X_2] = \Sigma x_1 x_2 \, P(X_1=x_1, X_2=x_2)$$
$$\sum_{x_1}\sum_{x_2} x_1 x_2 \, P(X_1=x_1, X_2=x_2)$$
$$\sum_{x_1} x_1 P(X_1=x_1) \sum_{x_2} x_2 P(X_2=x_2)$$
$$= E[X_1]\,E[X_2]$$

Linear Combinations

Suppose a and b are constants and random variable X has expected value μ and variance σ^2

$$E[aX+b] = \Sigma (aX+b)\,P(X=x)$$
$$\sum_x aX\,P(X=x) + \Sigma b\,P(X=x)$$
$$a\,\Sigma x P(X=x) + b\,\Sigma P(X=x)$$
$$a\,E[x] + b$$

29/09/14 Multivariate Analysis Slides 2 Extra ...

$Var[aX+b] = E[(aX+b)^2] - (E[aX+b])^2$

$= E[a^2x^2 + 2abX + b^2] - (aE[x]+b)^2$

$= a^2E[x^2] + 2abE(x) + b^2 - a^2E^2[x] - 2abE(x) - b^2$

$= a^2E[x^2] - a^2E^2[x]$

$= a^2 Var[x] = a^2\sigma^2$

Let X_1 and X_2 be two independent random variable with mean μ_1, μ_2 and variance σ_1^2, σ_2^2. If a_1 and a_2 are constants: $E[a_1X_1 + a_2X_2] = a_1E[X_1] + a_2E[X_2]$

$= a_1\mu_1 + a_2\mu_2$

$= \sum_{x,y} (aX + by)\, p(X=x, y=y)$

$= a\sum_{x,y} xp(x,y) + b\sum_{x,y} yp(x,y)$ aside $\sum_x p(x,y) = p(y)$

$= a\sum_x xp(x) + b\sum_y yp(y)$

$= aE[x] + bE[y]$

$Var[a_1X_1 + a_2X_2]$

If independent $\Rightarrow$ $= E[(aX+by - E(aX+by))^2]$

$= E[(aX+bY - [aE(x) + bE[y]])^2]$

$= E[[a(X - E[x]) + b(Y - E(y))]^2]$

$= E[a^2(X-E[x])^2 + b^2(Y-E[y])^2 + 2ab(X-E(x))(Y-E(y))]$

$= a^2E[(x-E[x])^2] + b^2E[(y-E[y])^2] + 2abE[(x-E[x])(y-E[y])]$

$= a^2 var[x] + b^2 var[y] + 2ab\, Cov(x,y)$ if not independent

If independent $\Rightarrow$ $E[xy - xE[y] - E[x]y + E[x]E[y]]$

$E[x,y] - E[x]E[y] - E[x]E[y] + E[x]E[y]$

if independent $E[x]E[y]$

$= 0$

$\Rightarrow Var[a_1X_1 + a_2X_2] = a^2 var[x] + b^2 var[y]$

$$Cov[aX+b, cY+d] = E[(aX+b)(cY+d)] - E[aX+b]E[cY+d]$$
$$= E[acXY + adX + bcY + bd] - (aE[X]+b)(cE[Y]+d)$$
$$= acE[XY] - acE[X]E[Y]$$
$$= ac\,Cov[X,Y]$$

$Cov[aX+bY, cW+dZ] =$
$\cancel{E[(aX+bY) - E(aX+bY)\,((cW+dZ) - E[cW+dZ])]}$
$E[(aX+bY)(cW+dZ)] - E[aX+bY]\,E[dW+dZ]$
$E[acXW + adXZ + bcYW + bdYW] - (aE[X]+bE[Y])(cE[Y]+dE[Z])$
$acE[XW] + adE[XZ] + bcE[YW] + bdE[YW] - acE[X]E[W] - bdE[Y]E[Z]$
$-bcE[Y]E[W] - bdE[Y][W]$

$= ac\,Cov[X,W] + ad\,Cov[X,Z] + bc\,Cov[Y,W] + bd\,Cov[Y,Z]$

$E[a_1X_1 + a_2X_2 \dots + a_mX_m] = a_1\mu_1 + a_2\mu_2 + \dots + a_m\mu_m$

Let $a = (a_1, a_2 \dots a_m)^T$ vector of constants
$X = (X_1, X_2 \dots X_m)^T$ vector of r.v's

We can write $a_1X_1 + a_2X_2 + \dots a_mX_m = (a_1 a_2 \dots a_m)\begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} = a^TX$.

Hence we can write $E[a^TX] = a^T\mu$ where $\mu = (\mu_1, \mu_2 \dots \mu_m)$

Matrix representation : Variance
Similarly $Var[a_1X_1 + a_2X_2 + \dots + a_mX_m] = Var[a^TX]$ and

$Var[a^TX] = a_1^2Var[X_1] + a_2^2Var[X_2] + \dots + a_m^2Var[X_m] + a_1a_2Cov[X_1,X_2]$
$+ \dots + a_1a_m Cov[X_1, X_m] + \dots + a_{m-1}a_m Cov[X_{m-1}, X_m]$

$$= \sum_{i=1}^{m} a_i^2 Var[X_i] + \sum_{i=1}^{m}\sum_{j\neq i}^{m} a_i a_j Cov[X_i, X_j]$$

from covariance matrix

$$= \sum_{i=1}^{m} a_i^2 S_{ii} + \sum_{i=1}^{m}\sum_{j\neq i}^{m} a_i a_j S_{ij}$$

covariance matrix

$$= a^T\Sigma a$$

9/04/14

5.

Multivariate Analysis Slides 2 Extra Notes

Suppose $U = a^T X$ and $V = b^T X$

$\mathrm{Cov}[U,V] = \sum_{i=1}^{n} a_i b_i s_{ii} + \sum_{i=1}^{n} \sum_{j \neq i}^{n} a_i b_j s_{ij}$

In matrix notation $\mathrm{Cov}[U,V] = a^T \Sigma b = b^T \Sigma a$
because of Symmetry

Eigenvalues and Eigenvectors

$m \times m$ matrix A.
λ is an eigenvalue of A if there exists a non zero vector v such that $Av = \lambda v$

- Vector v is said to be an eigenvector of A corresponding to eigenvalue λ.
- Solve $\det(A - \lambda I) = 0$ to find them

$Av - \lambda I v = 0$

hence $(A - \lambda I)v = 0$

Two vectors u and v are orthogonal if $u \cdot v = 0 = v \cdot u$

Two vectors u and v are orthonormal if they are orthogonal and $u^T u = 1$ and $v^T v = 1$

Eigenvalues of a cov matrix Σ are non neg
If λ is e value of Σ then $\Sigma v = \lambda v$ where v is e vector for corresponding to λ.

hence $v^T \Sigma v = v^T \lambda v = \lambda v^T v$

$\lambda = \frac{v^T \Sigma v}{v^T v}$ non negative

/14 MLA Proofs

$Cov[aX + bY, cW + dZ]$

$= E[(aX + bY - E[aX + bY])(cW - dZ - E[cW + dZ])]$

$= E[a(X - E[X])c(W - E[W]) + a(X - E[X])d(Z - E[Z]) + b(Y - E[Y])c(W - E[W])$

$+ b(Y - E[Y])d(Z - E[Z])]$

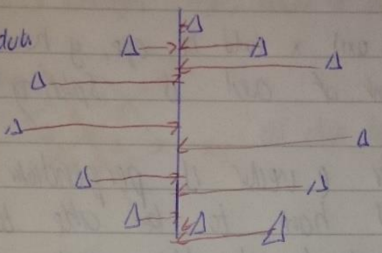
$= ac\,Cov[X, W] + ad\,Cov[X, Z] + bc\,Cov[Y, W] + bd\,Cov[Y, Z]$

1/10/14 Multivariate Analysis Slides 3

It is often useful to measure data in terms of its principle components rather than on a normal x-y axis

What is principle components? They're the underlying structure in the data. They are the directions where there is the most variance, the directions where the data is most spread out.

consider triangles as data



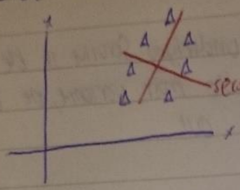
To find the direction where there is most variance, find the straight line where the data is most spread out when projected onto it. A vertical straight line with the points projected ↑ onto it

The data isn't very spread out here therefore it doesn't have a large variance. It's probably not the principle component.

Draw the line which creates the biggest spread out of data - largest variance → this is the principle component

Eigenvectors and Eigenvalues
- We can deconstruct data set into eigenvalues and eigenvectors
- Eigenvector is a direction, eigen value a number telling us how much variance there is in the data in that direction.
- In example above eigenvector ~~with highest e value~~ is e value is number telling us how spread out data is on the line
- E vector with highest e value is therefore the principle component

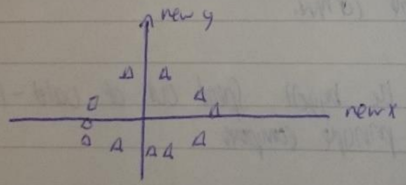
- Amount of e values/vectors corresponds to dimension of data.
- The e vectors put the data into a new set of dimensions, these dimensions have to be equal to the original amount of dimensions



Oval is on x-y axis, x could be age, y hours on internet.
Principle component of oval is splitting it long way

- Turns out other e vector is perpendicular to PC.
- The e vectors have to be able to span the whole xy area, in order to do this most effectively the two directions need to be orthogonal 90° to each other

- The e vectors have given us a much more useful axis to frame the data in.
- We can now ~~reform~~ reframe the data in the new dimensions.



- Note that nothing has been done to data itself. Were just looking at it from a new angle
- E vectors get you from one set of axes to another
- More intuitive to the shape of the data now.
- These directions are where there is most variation and that is where there is more information

3

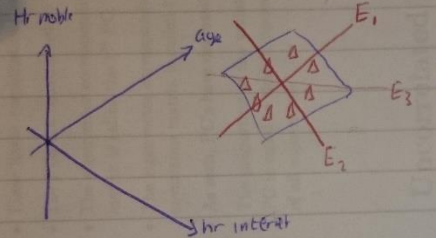
Multivariate Analysis Slides 3 Notes

PCA and e vectors help in analysis of data by dimension reduction

Dimension Reduction

- PCA can be used to reduce data down into its basic components, stripping away any unnecessary parts

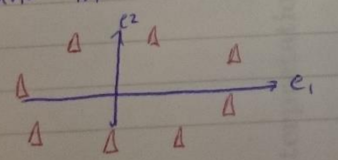
- Example measuring 3 things, age, hours on net and hours on mobile



all data lies on a plane in example

- We find 3 e vectors/values of data set, 2 have large values, one has e value of zero   zero because no height in data
- E_1 is primary component

- Now rearrange data on our new axis E3 is zero so we represent in 2D



- This is dimension reduction
- We have reduced from 3D to 2D

- We can reduce dimensions even if there isn't an e value of zero
- What if we had e-values 10, 8 and 0.1
- E vector corresponding to 0.1 will not have much info so we can discard it in order to make the data set ~~more~~ simpler

$B = \{(0,4),(2,1)\}$, $C = \{(2,1),(3,3)\}$ (choose any given measure and any linkage).

$A\left\{\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}0\\3\end{pmatrix}\right\}$ $B\left\{\begin{pmatrix}0\\4\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix}\right\}$ $C\left\{\begin{pmatrix}2\\1\end{pmatrix}\begin{pmatrix}3\\3\end{pmatrix}\right\}$

	A	B	C
A	0	1	1
B	1	0	0
C	1	0	0

Single linkage and maximum dissimilarity

$d(A,A) = \min\left\{ d\left(\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}1\\2\end{pmatrix}\right)\ d\left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}0\\3\end{pmatrix}\right)\ d\left(\begin{pmatrix}0\\3\end{pmatrix}\begin{pmatrix}0\\3\end{pmatrix}\right)\right\}$

$= \min(0,1,0) = d(B,B) = d(C,C) = 0$

$d(A,C) = \min\ d\left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix}\right)\ d\left(\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}3\\3\end{pmatrix}\right)\ d\left(\begin{pmatrix}0\\3\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix}\right)\ d\left(\begin{pmatrix}0\\3\end{pmatrix}\begin{pmatrix}3\\3\end{pmatrix}\right)$

$= \min(1,2,2,3) = 1$

$d(A,B) = 1$

$d(B,C) = 0$

$\pi_k f(x|\mu_k,\Sigma_k) > \pi_L f(x|\mu_L,\Sigma_L)$

$\Rightarrow \frac{\pi_k}{\cancel{(2\pi)^{m/2}}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right) > \frac{\pi_L}{\cancel{(2\pi)^{m/2}}|\Sigma_L|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_L)\Sigma_L^{-1}(x-\mu_L)\right)$

$\log(\pi_k) - \frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) > \log(\pi_L) - \frac{1}{2}\log(|\Sigma_L|) - \frac{1}{2}(x-\mu_L)^T\Sigma_L^{-1}(x-\mu_L)$

if $\Sigma_k = \Sigma_L = \Sigma \Rightarrow \log(\pi_k) - \frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu_k + \frac{1}{2}\mu_k^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$

$> \log(\pi_L) - \frac{1}{2}x^T\Sigma^{-1}x + \frac{1}{2}x^T\Sigma^{-1}\mu_L + \frac{1}{2}\mu_L^T\Sigma^{-1}x - \frac{1}{2}\mu_L^T\Sigma^{-1}\mu_L$

$\log(\pi_k) + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k > \log(\pi_L) + x^T\Sigma^{-1}\mu_L - \frac{1}{2}\mu_L^T\Sigma^{-1}\mu_L$

$\pi_k = \pi_L = 1/K$ $\quad x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - x^T\Sigma^{-1}\mu_L + \frac{1}{2}\mu_L^T\Sigma^{-1}\mu_L > 0$

$\Rightarrow (x - \frac{1}{2}\mu_k)^T\Sigma^{-1}\mu_k - (x - \frac{1}{2}\mu_L)^T\Sigma^{-1}\mu_L > 0$

$\Rightarrow (x - \frac{1}{2}(\mu_k+\mu_L))^T\Sigma^{-1}(\mu_k - \mu_L) > 0$

DEATH BY MLA