MLA

8/05/15. PRINCIPLE COMPONENT ANALYSIS

- For data with many variables/dimensions it is often difficult to comprehend or visualize inherent associations
- PCA thought of as a method for re-expressing the data so as to reveal its internal structure and explain its variation through the use of a few linear combinations of the original variables
- Used as either a dimension reduction technique or as a method for identifying associations among variables

- Aim of PCA is to describe the variation in a set of correlated variables $X_1 \ldots X_m$ in terms of a new set of uncorrelated variables $Y_1 \ldots Y_p$ hopefully $p \ll m$ and where each Y_i is a linear combination of the $X_1 \ldots X_m$
- New variables 'principle components' derived in decreasing order of importance so that first PC Y_1 accounts for more variation in the original data than any other possible linear combination of $X_1 \ldots X_m$
- Second PC is chosen to account for as much of the remaining variation as possible subject to constraint that it is uncorrelated with Y_1 and so on...
- Hope is that the first few pc will account for a substantial amount of variation in the original data, and as such can be used as a convenient lower dimensional summary of it

Constrained Optimization:
Find a to maximize $a^T Q a$ subject to $a^T a = 1$

Lagrange multipliers
- Given a function $f(x)$, gradient of f, ∇F (the collection of partial derivatives), indicates the direction of the steepest slope. Level curves (lines where $f(x)$ is constant) run perpendicular to the gradient

- Here we want location where $\nabla f = \lambda \nabla g$
- In other words, find x so that $\nabla F - \lambda \nabla g = 0$
- Gives us $m+1$ eqns and $m+1$ unknowns

- Let $p = a^T Q a - \lambda (a^T a - 1)$
- Solution found by computing partial derivs dp/da_k $k = 1, 2 \ldots m$ and $dp/d\lambda$
- Seek to solve $m+1$ eqns when they are set to 0
- Note: $dp/d\lambda = -(a^T a - 1)$, solving ensures constraint is satisfied

See shit IN NOTES.

- First PC of data is linear combination of the variables that have greatest variance
- Corresponds to taking a linear combination of the variables where the weights are given by the eigenvector Q with largest e value. This e value represents the variance of the linear combination
- All PC's are orthogonal to each other

- Requires an eigenvalue decomposition of the covariance matrix in order to find the linear combinations of the data variables with greatest variance
- Each E value of our matrix can be interpreted as the variance of a linear combination of the variables with weights given by e-vectors
- Value of a particular e-value divided by the total sum of e-values is the proportion of the variance explained by the associated PC

- As PCA seeks to maximise variance it can be sensitive to scale differences across variables
- Standardising ensures that the data be expressed in comparable units
- Divide by sample standard deviation ⇒ variance equal to one → correlation matrix
- Replace time by -tm.

- Keep adding until a fixed proportion of variance is included
- Find a kink in the scree plot (variance explained against PC number) means that marginal additional variance explained is reducing as a function of PC

i.e benefit of including an additional PC may not longer be worth the extra cost of model complexity.

MLA

04/05/15 EXAM PAPER 2014 Q1 DAVID WESTBROOKE

1 A. Use/Aim of PCA.
- Dimension reduction - included dimensions are orthogonal and ordered to account for as much variation as possible.
- Visualisation
- Lower Dimensional Summary.
- Helps identify relationships in data
- Potentially assists other clustering or classification algorithms (performance)

B Lagrange Multipliers

$P = \underline{a}^T \Sigma \underline{a} - \lambda(\underline{a}^T\underline{a} - 1)$

Solve $dP/da_k = 0$ for $k = 1 \dots m$

$dP/d\lambda = 0 \Rightarrow a^T a - 1 = 0 \Rightarrow$ constraint satisfied

$\Rightarrow \sum_{i=1}^{m} a_i \tilde{b}_{ik} - \lambda a_k = 0$

$\Rightarrow a^T \Sigma - \lambda a^T \Rightarrow a$ is an eigenvector of Σ

C i. Standardisation of data
- If not standardised, one component will account for most variance and figure more prominently in the solution, swamping out all other values. PCA will focus on that variable

ii. Number of PC's
- Could select 2 or 3 3-account for more variation (96.4%) but 2 dimension is easier to graph (visualise) and interpret.

iii. PC1: high score = high birth rate, high death rate, high ID and small LEM and LEF.
Low score = opposite
Interpreted as health care - independent of GNP

PC2: High score = High loading on GNP ⇒ wealth
High score for poor country, low score for rich country

C iv. (-0.3, -1.1, +0.5, 0.8, 0.9, -0.6)

$0.43(-0.3) + 0.37(-1.1) + 0.46(-0.5) - 0.47(0.8) - 0.48(0.9) + 0.08(-0.6) = -1.622$

$-0.3(-0.3) + 0.14(-1.1) + 0.06(-0.5) - 0.8(0.8) - 0.1(0.9) - 0.99(-0.6) = 0.282$

(-1.622, 0.282)

MLA

15

Which linear combination of $a^T X$ of the variables in the data has maximum variance, subject to the constraint $a^T a = 1$?

Lagrange Multipliers

- given a function $f(x)$, gradient of f, ∇f (the collection of partial derivatives) indicates the direction of steepest slope
- Level curves (line where $f(x)$ is constant) run perpendicular to the gradient

- A constraint $g(x) = c$ represents a plane that cuts through the variable space
- When direction of gradient of constraint equals direction of gradient of f function, then that location constituted a local maximum

- Here we want location where $\nabla f_{\text{curve}} = \lambda \nabla g_{\text{plane}}$
- $\Rightarrow$ Find x so $\nabla f - \lambda \nabla g = 0$
- Gives d $m+1$ equations and $m+1$ unknowns

- Use Lagrange Multiplier to find $\max a^T Q a$ $a^T a - 1 = 0$
- Let $P = a^T Q a - \lambda (a^T a - 1)$
- Compute partial derivatives $\frac{dP}{da_k}$ for $k = 1, 2 \ldots m$ and $\frac{dP}{d\lambda}$
- Solve $m+1$ equations when setting them $= 0$
- Notice $dP/d\lambda = -(a^T a - 1)$, guaranteed that constraint is satisfied

- Write $P = \sum_{i=1}^{m} a_i^2 q_{ii} + \sum_{i=1}^{m} \sum_{j \neq i} a_i a_j q_{ij} - \lambda \left(\sum_{i=1}^{m} a_i^2 - 1 \right)$

$$\frac{dP}{da_k} = 2 a_k q_{kk} + \sum_{j \neq k} a_j q_{kj} + \sum_{i \neq k} a_i q_{ik} - 2\lambda a_k$$

$$= 2 a_k q_{kk} + 2 \sum_{i \neq k} a_i q_{ik} - 2\lambda a_k$$

So $\frac{dP}{da_k} = 0$ $\sum_{i=1}^{m} a_i q_{ik} - \lambda a_k = 0$

For $k=1 \dots m$ $\frac{\partial \mathcal{L}}{\partial a_k} = 0$ implies

$$(a_1 a_2 \dots a_m) \begin{pmatrix} q_{1k} \\ \vdots \\ q_{mk} \end{pmatrix} = \lambda a_k$$

we have $a^T Q = \lambda a^T$

Taking transpose of both sides $Q^T a = Qa = \lambda a$.

- That is, a is an eigenvector of Q. Constraint tells us it is a unit vector.

- Due to maximizer part of problem we know it is the eigenvector with largest eigenvalue

$$Var(a^T x) = a^T Q a = a^T \lambda a = \lambda a^T a = \lambda$$

MLA

INTRODUCTION AND PRINCIPLE COMPONENT ANALYSIS

- Multivariate data arise when two or more attributes are recorded for each set of objects (e.g. height/weight/age). Variables can be continuous, binary, categegorical etc.

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

n rows by m columns
n - number of observations/units
m - number of variables.

- $\mu = E[x] = \sum_x x P(X=x)$ if X is discrete random variable with probability mass function $P(x)$

- let $X_1 \cdots X_m$ be random variables.
  - $Var[X_i] = E[X_i^2] - [E[X_i]]^2$
  - $Covar[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])]$
  - $Corr[X_i, X_j] = \frac{Cov[X_i, X_j]}{\sqrt{Var[X_i]\, Var[X_j]}}$ correlation is "normalised" covariance

- $\Sigma = \begin{pmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & & \vdots \\ s_{m1} & \cdots & s_{mm} \end{pmatrix}$ covariance matrix with $s_{ij} = Cov(X_i, X_j)$, $s_{ii} = Var[X_i]$

- Two random variables X_1, X_2 are independent iff $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) P(X_2 = x_2)$ and so $E[X_1 X_2] = E[X_1] E[X_2]$

Linear Combination

- $E[aX + b] = aE[X] + b = a\mu + b$
- $Var[aX + b] = a^2 Var[X] = a^2 \sigma^2$
- If X_1, X_2 independent then
  $E[a_1 X_1 + a_2 X_2] = a_1 E[X_1] + a_2 E[X_2] = a_1 \mu_1 + a_2 \mu_2$
  $Var[a_1 X_1 + a_2 X_2] = a_1^2 Var[X_1] + a_2^2 Var[X_2] = a_1 \sigma_1^2 + a_2 \sigma_2^2$
- If X_1, X_2 not independent then
  $Var[a_1 X_1 + a_2 X_2] = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + 2 a_1 a_2 Cov[X_1, X_2]$

- $a_1X_1 + \dots a_mX_m = (a_1 \dots a_m)\begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix} = a^TX$

$E[a^TX] = a^T\mu$ with $\mu = (\mu_1 \dots \mu_m)$

- Similarly $Var[a_1X_1 + \dots + a_mX_m) = Var[a^TX] = a^T\Sigma a$

- Let A be a $m \times m$ matrix. Then λ is an eigenvalue of A if $\exists\ v \neq 0$: $Av = \lambda v$. The vector v is said to be an eigenvector of a corresponding eigen value λ
- Can find eigenvalue by solving $\det(A - \lambda I) = 0$

- If v is an eigen vector corresponding to eigenvalue λ, then then $u = \alpha v$ will also be an eigenvector corresponding to λ.
- V is a unit eigenvector if $\sum_{i=1}^{m} v_i^2 = v^Tv = 1$. We can turn any eigenvector v into a unit eigenvector by multiplying it by $\alpha = 1/\sqrt{v^Tv}$
- 2 vectors u, v are orthogonal if $u^Tv = v^Tu = 0$
- 2 vectors are orthonormal if they are orthogonal and $u^Tu = v^Tv = 1$

- If λ is an eigenvalue of covariance matrix Σ then $\Sigma v = \lambda v$ where v is the eigenvector corresponding to λ. Hence

$$v^T\Sigma v = v^T\lambda v = \lambda v^Tv \text{ so}$$

$$\lambda = \frac{v^T\Sigma v}{v^Tv}$$

- An $m \times m$ covariance matrix Σ has m orthonormal eigenvectors

4/15.

PRINCIPLE COMPONENT ANALYSIS

- For data with many variables/dimensions it is often difficult to comprehend/visualize inherent associations - two or more variables could be highly correlated.
- PCA can be thought of as a method for re-expressing the data so as to reveal its internal structure and explain its variation through the use of a few linear combinations of the original values

- The aim of pca is to describe the variation of a set of correlated variables $x_1, \dots x_m$ in terms of a new set of uncorrelated variables $y_1 \dots y_p$, hopefully with $p << m$ and each y_i a linear combination of $x_1, \dots x_m$
- The new variables or principle components are derived in decreasing order of importance, so that the 1st PC accounts for more variation in the original data than any other possible linear combination of $x_1, \dots x_m$
- The second PC y_2 is chosen to account for as much of the remaining variation as possible subject to the constraint that it be uncorrelated with y_1 and so on
- Hope is that first few PC's will account for a substantial amount of the variation in the original data and as such can be used as a convenient lower dimension summary of it.

- If we have a dataset of n observations, each consisting of m measurements then the sample mean of each variable is $\bar{x}_i = \sum_{r=1}^{n} x_{ri}/n$ and the sample covariance matrix is a matrix whose terms are defined to be $q_{ij} = \frac{1}{n-1} \sum_{r=1}^{n} (x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j)$
  for $i = 1 \dots n \quad j = 1 \dots m$

- Then calculate the eigenvalues and corresponding eigenvectors of the sample covariance matrix

- We use Lagrange multipliers to find the maximum value of $a^T Q a$ subject to $a^T a = 1$. We eventually get $a^T Q = \lambda a^T$, which taking the transpose gives $Qa = \lambda a \rightarrow$ a is an eigenvector of Q and by the constraint it a unit eigenvector. Also due to maximization part of problem we know it is eigenvector with larger eigenvalue

$$Var[a^T x] = a^T Q a = a^T \lambda a = \lambda a^T a = \lambda T = \lambda$$

- The 1st PC of the dataset is the linear combination of the variables that has the greatest variance. This corresponds to taking a linear combination of the variables, where the weights are given by the eigenvector of Q with larger eigenvalue. This eigenvalue also represents the variance of the linear combination.
- The 2nd PC is the linear combination of the variables where the weights are given by the eigenvector of Q corresponding to the 2nd largest eigenvalue and this eigenvalue represents the variance of this linear combination and so on.

- PCA required an eigen value analysis of the covariance matrix in order to find the linear combination of the data variables with greatest variance
- These linear combinations are called PC's - constructed so they are uncorrelated to each other. PC's have decreasing variance.
- The proportion of the variation in the data variable that is explained by a PC is equal to that computed associated eigenvalue divided by the sum of all eigen value. We say that the value of a particular evalue divided by the total sum of eigenvalues is the proportion of the variance explained by the associated PC

How Do We Interpret PCA output?

- The 1st row in the table shows the standard deviation of each PC (the standard deviation is square root of e-value associated with component)
- 2nd row shows proportion of the variance in the data explained by each PC : $sd^2 / \Sigma sd^2$

PRINCIPLE COMPONENT ANALYSIS (cont'd)

How to interpret PCA output cont.
- 3rd row shows the cumulative proportion of the variation obtained for each PC

- An indicator of the number of PC's required to adequately summarize the data can be infered by examining the proportion of the variance explained by the PC's

- Consider the new variable(PC) which capture variation of data. PC1 (column 1. 11 elements are the coefficients/loadings of each original variable on the PC. It matters if the loadings have opposite sign but not which is positive and neglect Magnitude of loadings /important.

- Results differ if scale differs (mm vs. cm). Standardizing ensures that the data are expressed in comparable units.
- One way to do this is to make each variable have variance equal to 1. To do this, divide value of each variable by its sample Standard deviation. We could then do PCA on covariance matrix of the transformed data. The covariance matrix of a set of variable with variance equal to 1 is a correlation matrix

How to chose appropriate number of PC's?
- No correct answer ⇒ rule of thumb
- Keep adding until a fixed proportion of variance is included
- Find a kink in Scree plot. (variance v. PC numb)
This means that the marginal additional variance explained is reduced as a function of PC. i.e. that the added benefit of including an additional PC may no longer be worth the extra cost of model complexity (remember we are seeking dimension reduction)