SIGNAL & DATA ANALYSIS IN NEUROSCIENCE 2017

PRINCIPAL COMPONENT ANALYSIS

Ayala Matzner

biu.sigproc@gmail.com

2

PCA and Dimensionality Reduction

- Principal component analysis (PCA) is a technique that is useful for the compression and classification of data.
- The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample's information.
- By information we mean the variation present in the sample, given by the correlations between the original variables. The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

Math review: eigen-values and eigen-vectors

- Definition: Let T[MxN] be a linear transformation. X[Mx1] is an eigenvectror of T & λ is an eigen-value of T if:
  $T \cdot X = \lambda \cdot X$
- How to find eigen-values and vectors:
  To extract eigen-values solve: $|T - \lambda I| = 0$
  To extract eigen-vector X_i solve: $T \cdot X_i = \lambda \cdot X_i$
  $(T - \lambda I) \cdot X_i = 0$
  Note that there are ∞ solutions for X_i, as a result we will see dependent vectors.

PCA: algorithm

1. Input: $Y_{[nxm]}$, n- # of trials, m - # of samples in each trial.
2. Subtract the mean from each dimension of X = Y -E(Y) – meaning for each sample point/variable dimension we subtract the mean (calculated over trials)
3. Derive eigen-values & vectors of X'·X (Covariance matrix. Actually a real covariance matrix is obtained by dividing X'X by N, but that is not important for this algorithm).
4. Sort diagonal eigen-values (highest → lowest) and adequate eigen-vectors Vi[mx1] - the principal components.
5. To find the projection of trial k X[k, :] on principal component Vi[mx1] :
   ρk,i = X[k, :]·Vi[mx1]
   Reconstructed signal: X[k, :] = ∑i ρk,i · V'i[mx1]

Example

- Example: Each observation (e.g. trial) contains 2 variables (dimensions) – e.g. value sampled at times t1, t2:

```
              t1   t2                  t1    t2                                t1    t2
A = trial 1: [ 2    6 ]   E(A) = [ 2      5]   B=A-E(A) =  trial 1:  [  0     1 ]
    trial 2: [1.5   5 ]                                    trial 2:  [-0.5    0 ]
    trial 3: [2.5   4 ]                                    trial 3:  [ 0.5   -1 ]

C = B'*B =                  t1    t2
t1: [0  -0.5  0.5 ]       [  0     1 ]            t1     t2
t2: [ 1    0   -1 ]   *   [-0.5    0 ]     =   t1 [ 0.5  -0.5 ]
                          [ 0.5   -1 ]         t2 [-0.5    2  ]

λ1,2:   det(C – λI) = 0
[ 0.5  -0.5 ]  -  [ λ   0  ] = 0     [ 0.5- λ    -0.5 ]  = 0
[-0.5    2  ]     [ 0   λ  ]        , [ -0.5      2- λ ]       ,   det(C) = ad-bc = λ² – 2.5 λ + 0.75 = 0,
```

$\lambda_1 = 0.3486$

$\lambda_2 = 2.1514$

Example contd.

V_1: C- λI = 0, λ_1 =0.3486

```
[ 0.5- 0.3486     -0.5 ]  = 0   [ 0.1514    -0.5   ] [ V1₁ ]    = 0
[ -0.5         2- 0.3486]      , [ -0.5      1.6514 ] [ V1₂ ]
```

$0.1514\ V1_1 - 0.5\ V1_2 = 0$, $V1_2 = 0.3028 * V1_1$ => V1 = [1] [0.3028]

V1 Normalized to 1: $\sqrt{V1_{1norm}^2 + V1_{2norm}^2} = 1$, $V1_{norm} = a * V1$, $a = \frac{1}{\sqrt{V1_1^2 + V1_2^2}} = 0.9571$, $V1_{norm}$ = [0.9571] [0.2898]

V_2: C- λI = 0, λ_2 =2.1514

```
[ 0.5- 2.1514     -0.5 ]  = 0   [ -1.6514    -0.5   ] [ V2₁ ]    = 0
[ -0.5         2- 2.1514]      , [ -0.5      -0.1514 ] [ V2₂ ]
```

$-0.5\ V2_1 - 0.1514\ V2_2 = 0$, $V2_1 = -0.3028 * V2_2$ => V2 = [- 0.3028] [1]

V2 Normalized to 1: $\sqrt{V2_{1norm}^2 + V2_{2norm}^2} = 1$, $V2_{norm} = a * V2$, $a = \frac{1}{\sqrt{V2_1^2 + V2_2^2}} = 0.9571$,

$V1_{norm}$ = [-0.2898] [0.9571]

*Matlab: [vectors values] = eig(C)

PCA: Assumptions and limitations

- Assumptions:
  - Mean and variance (moments 1&2) are sufficient statistics. Large variances are important.
- Limitations:
  - Linear transformation.
- Properties:
  - PCs are orthogonal.

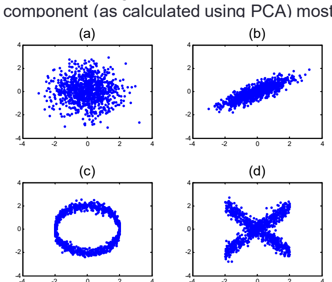
Example : Exam 2007

- The shape of spikes recorded in the *Animalis Simplistics* may be described by: $A(n) = \alpha(n)V_1 + \beta(n)V_2$
  (A, V1 & V2 are vectors of length τ).
  The values of α are distributed normally (μ=0,σ=1) and $\beta(n) = 2\alpha(n)$.
  After performing PCA on the spikes, the number (k) of non-zero eigenvalues is:

a. k = 0
b. k = 1
c. k = 2
d. k > 2
e. k = n

Example exam 2006

- For which of the following two variable distributions is the first principal component (as calculated using PCA) most useful?



ICA: Independent component analysis

- Definition: ICA separates the data into statistically independent components
- Requirements:
  - Sources are independent
  - Sources are non-Gaussian (at most one can be Gaussian)
- Algorithms: no closed solution.
  - Maximize non-Gaussianity of recovered sources, using
    - Kurtosis (=0 for a Gaussian)
    - Entropy (maximal for a Gaussian)
  - Minimize mutual information between recovered signal components – less information==more independence
  - Maximum likelihood – includes information on priors for the sources. (so not completely blind..)

PCA vs. ICA

- PCA finds direction of maximal variance
- ICA finds direction of maximal independence in non-Gaussian data (higher-order statistics).
- ICA can be used for blind source separation.

- ICA demo: http://www.cis.hut.fi/projects/ica/cocktail/cocktail_en.cgi