Principal Component Analysis & Data Reduction

Hugh R. Wilson

- Purpose of PCA
- Simple example
- Derivation of PCA
- Dimensionality reduction
- More complex example
- How many dimensions?
- Experiment on learning face PCA
- Nonlinear PCA
Reading

• Today: Principal components: Kutz, pp. 387 (bottom) - 393
• Wed: Linear discriminant analysis: Kutz, pp. 442 - 445
Purpose of PCA

• Simplify data by reducing dimensionality
• Perfect correlation between 2 measurements:
  • Redundancy
  • Example: nose length = eye separation
  • One dimension is irrelevant
• Remove least significant dimensions
• All PCs mutually orthogonal: independent information
• Remove noise (small variances in highest PCs)
• Equivalent to rotation of axes
Gram-Schmidt Orthogonalization

• Two linearly independent column vectors A & B
• Assume vectors normalized to unit length
• Covariance: $A'B = \rho_{AB}$
• $A'$ is transpose of A
• Make dimensions independent (orthogonal)
• $\cos(AB\_angle) = \rho_{AB}$
• $B_{new} = B - \rho_{ABA}$
• $A' B_{new} = A'B - A'A\rho = \rho - \rho = 0$
• Can repeat with vectors C, D, etc. as long as all are linearly independent
• $C_{new} = C - \rho_{ACA} - \rho_{BCB_{new}}$
Data for PCA

- Make N measurements on each of M specimens, $M \geq N$
- Eg: nose length, eye separation, mouth width, nose width
- Eg: multiple disease symptom measurements on people
- Create data matrix:
  - M rows, one for each specimen
  - N columns, one for each measurement
- Data represented as M points in N dimensional space
- Can we simplify?
PCA in 2 Dimensions

- Two measurements that are correlated
- PC1 explains 92.5% of the variance
- PC2 explains 7.5% of the variance
PCA Approach

• Use covariance (or correlation) of measurements to reduce dimensionality
• Require PCA dimensions to be orthogonal: independent info
• First PC: accounts for largest % of variance
• Second PC: accounts for next largest % of variance
PCA Derivation

• Subtract means from data, $D$
• Find vector $v$ to maximize variance of projected data:
  • $\text{var}(Dv) = (1/(N-1))(Dv)'(Dv) = (1/(N-1))(v'D')(Dv) = v'Sv$
  • $S$ is the data covariance matrix: $(1/(N-1))D'D$
• Normalization constraint: $v'v = 1$
• Use Lagrange multiplier for constraint & maximize:
  • $v'Sv + \lambda(1 - v'v)$
  • $(d/dv')(v'Sv + \lambda(1 - v'v)) = 0$
  • $Sv = \lambda v$
• PCs are eigenvectors of covariance matrix
• Variance explained is $\lambda$
• Vector with largest $\lambda$ is PC1, second largest $\lambda$ is PC2, etc.
• Equivalent to minimizing mean squared error of representation in lower dimensional space
PCA in Matlab™

- Data: M rows of subjects x N measurements on each (M > N)
- Subtract mean
- Calculate eigenvalues of data covariance
- Project data onto P < N PCs (must flip PC matrix left-right)

\[ Dta = Dta - \text{ones}(M,1) \times \text{mean}(Dta) \]

\[ [PC,Var] = \text{eig}\left(\text{cov}(Data)\right) \]

\[ PClr = \text{fliplr}(PC) \]

\[ PCproj = Dta \times PClr(:,1:P) \]
How Many PCs?

- Decide how much variance is important (e.g. 75%, 90%)
- Cut off when variance explained drops below average/PC
- Plot variance per component & look for break (scree plot)
Face Identification
Synthetic Faces

Face geometry represented by 39 measurements
# Synthetic Face PCA

- Each face: 39 measurements

<table>
<thead>
<tr>
<th>PC#</th>
<th>Variance</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>0.2345</td>
<td>0.2345</td>
</tr>
<tr>
<td>2.0000</td>
<td>0.2192</td>
<td>0.4537</td>
</tr>
<tr>
<td>3.0000</td>
<td>0.1013</td>
<td>0.5549</td>
</tr>
<tr>
<td>4.0000</td>
<td>0.0818</td>
<td>0.6368</td>
</tr>
<tr>
<td>5.0000</td>
<td>0.0654</td>
<td>0.7022</td>
</tr>
<tr>
<td>6.0000</td>
<td>0.0637</td>
<td>0.7658</td>
</tr>
<tr>
<td>7.0000</td>
<td>0.0465</td>
<td>0.8123</td>
</tr>
<tr>
<td>8.0000</td>
<td>0.0384</td>
<td>0.8507</td>
</tr>
<tr>
<td>9.0000</td>
<td>0.0292</td>
<td>0.8799</td>
</tr>
<tr>
<td>10.0000</td>
<td>0.0211</td>
<td>0.9010</td>
</tr>
<tr>
<td>11.0000</td>
<td>0.0198</td>
<td>0.9208</td>
</tr>
<tr>
<td>12.0000</td>
<td>0.0151</td>
<td>0.9359</td>
</tr>
<tr>
<td>13.0000</td>
<td>0.0125</td>
<td>0.9484</td>
</tr>
</tbody>
</table>

10 components: 90% of variance

Remaining 29: noise or irrelevant
• PCA reveals data structure determined by covariance
• Calculation for N measurements, M samples (M > N):
  • Subtract means from measurements
  • Data covariance matrix CV
  • \([\text{Vector}, \text{EV}] = \text{eig}(\text{CV})\)
• Dimensionality reduction
• May facilitate separating categories (next lecture)
• Can be implemented by neural learning networks
• Nonlinear manifolds in data require embellishments