Principal Component Analysis

• Uses:
  – Data Visualization
  – Data Reduction
  – Data Classification
  – Trend Analysis
  – Factor Analysis
  – Noise Reduction

• Examples:
  – How many unique “sub-sets” are in the sample?
  – How are they similar / different?
  – What are the underlying factors that influence the samples?
  – Which time / temporal trends are (anti)correlated?
  – Which measurements are needed to differentiate?
  – How to best present what is “interesting”?
  – Which “sub-set” does this new sample rightfully belong?

Data Presentation

• Example: 53 Blood and urine measurements (wet chemistry) from 65 people (33 alcoholics, 32 non-alcoholics).

• Matrix Format

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>8.0000</td>
<td>4.8200</td>
<td>14.1000</td>
<td>41.0000</td>
<td>85.0000</td>
<td>29.0000</td>
<td>34.0000</td>
</tr>
<tr>
<td>A2</td>
<td>7.3000</td>
<td>5.0200</td>
<td>14.7000</td>
<td>43.0000</td>
<td>86.0000</td>
<td>29.0000</td>
<td>34.0000</td>
</tr>
<tr>
<td>A3</td>
<td>4.3000</td>
<td>4.4800</td>
<td>14.1000</td>
<td>41.0000</td>
<td>91.0000</td>
<td>32.0000</td>
<td>35.0000</td>
</tr>
<tr>
<td>A4</td>
<td>7.5000</td>
<td>4.4700</td>
<td>14.9000</td>
<td>45.0000</td>
<td>101.0000</td>
<td>33.0000</td>
<td>33.0000</td>
</tr>
<tr>
<td>A5</td>
<td>7.3000</td>
<td>5.5200</td>
<td>15.4000</td>
<td>46.0000</td>
<td>84.0000</td>
<td>28.0000</td>
<td>33.0000</td>
</tr>
<tr>
<td>A6</td>
<td>6.9000</td>
<td>4.8600</td>
<td>16.0000</td>
<td>47.0000</td>
<td>97.0000</td>
<td>33.0000</td>
<td>34.0000</td>
</tr>
<tr>
<td>A7</td>
<td>7.8000</td>
<td>4.6800</td>
<td>14.7000</td>
<td>43.0000</td>
<td>92.0000</td>
<td>31.0000</td>
<td>34.0000</td>
</tr>
<tr>
<td>A8</td>
<td>8.6000</td>
<td>4.8200</td>
<td>15.8000</td>
<td>42.0000</td>
<td>88.0000</td>
<td>33.0000</td>
<td>37.0000</td>
</tr>
<tr>
<td>A9</td>
<td>5.1000</td>
<td>4.7100</td>
<td>14.0000</td>
<td>43.0000</td>
<td>92.0000</td>
<td>30.0000</td>
<td>32.0000</td>
</tr>
</tbody>
</table>

• Spectral Format

![Spectral Format Graph]
Data Presentation

- Better presentation than ordinant axes?
- Do we need a 53 dimension space to view data?
- How to find the ‘best’ low dimension space that conveys maximum useful information?
- One answer: Find Principal Components
Principal Components

- All principal components (PCs) start at the origin of the ordinant axes.
- First PC is direction of maximum variance from origin.
- Subsequent PCs are orthogonal to 1st PC and describe maximum residual variance.

Principal Component Analysis

- Factor data, \( \mathbf{R} \), into 3 matrices.
  - \( \mathbf{R}_{\text{samples x spectra}} = \mathbf{USV}^T \)
- Columns of \( \mathbf{V} \)
  - describe directions of maximum variance
  - linear combinations of ordinant spectral axes
  - are orthonormal
- Columns of \( \mathbf{U} \)
  - describe relationship among samples
  - projection of each spectra onto column from \( \mathbf{V} \)
  - are orthonormal
- Matrix \( \mathbf{S} \)
  - Diagonal
  - Contains scale of \( \mathbf{R} \)
Principal Component Analysis

\[ R = U S V^T \]

- **Preprocessing**
  - Data Translation
  - Data Scaling
- **Axis Rotation**
  - SVD
  - Varimax Rotation
- **Determining Significant Factors**
  - Statistical Tests
  - Empirical Tests
- **Interpretation**
  - Outlier Detection
  - Variable Selection
Preprocessing

• Mean Centering
  – Translates center of data cloud to origin
  – For $R_{(I,J)}$, subtract mean response of the $I$ samples from each of the $J$ variables.
  – $R_{ij} = R_{ij} - \bar{r}_j$
  – $R_{mc} = R - (\text{ones}(I,1) \times \text{mean}(R))$

• Variance Scaling
  – Normalize each axis to same Euclidean length. Each variable will have same least-squares weight.
  – For $R_{(I,J)}$, subtract mean response of the $I$ samples from each of the $J$ variables.
  – $R_{ij} = R_{ij} / s_j$
  – $R_{vs} = R - (\text{ones}(I,1) \times \text{std}(R))$
Preprocessing

• Autoscaling
  – Translates and stretches axes such that each variables has a mean 0 and standard deviation 1.
  – First mean center than variance scale the data

• Unit Area
  – Area under each spectrum is set to 1. Adjusts for changes in sampling size and lamp intensity
  – Divide each sample by sum of the measurements in each sample
  – \( R_{ij} = R_{ij} / S_{j=1} R_{ij} \)
  – \( Rua = R ./ (sum(R')*ones(1,J)) \)

• Unit Length
  – Length of each spectrum is set to 1. Places all data on unit circle
  – Divide each sample by sum of squared values of the measurements in each sample
  – \( R_{ij} = R_{ij} / S_{j=1} R_{ij}^2 \)
  – \( Rul = R ./ (sum(R.^2)*ones(1,J)) \)
Rotation of Axes

- Consider
  - $R_{\text{new}} = RQ$
  - $R_{\text{new}} = RQ^T$
  - $\Theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$
- $Q$ is a rotation matrix
  - $R_{\text{new}} = RQ$ Rotates Axes
  - $R_{\text{new}} = RQ^T$ Rotates Data
- $R_{\text{new}}$ is location of data on new axes

Axis Rotation

- Goal: Find new axis set, $Q$, such that:
  1. All axes are orthogonal (orthonormal)
  2. The sum of the squared distances from all points to the first axis is minimized: $||R - Q_1Q_1^TR||$
  3. Given $Q_{n-1}$ exist, find $Q_n$ such that 1 and 2 are satisfied and $||R - Q_1,nQ_{1,n}^TR||$ is minimized.

In other words:
- Now, $X = XI^T$ where columns of $I$ are orthonormal axes (directions) and rows in $X$ are distances along each unit axes.
- Want, $X = PQ^T$ where columns of $Q$ are orthonormal axes (directions) satisfying 1 - 3 and rows in $P$ are distances along the new axes.
- Eventually want: $X = USV^T$ where $V=Q$ and $US = P$ with columns of $V$ and $U$ being orthonormal and $S$ being a diagonal matrix.
Three methods for Rotation

• Eigenvalue Problem (EP)
• Nonlinear Iterative Partial Least Squares (NIPLS)
• Singular Value Decomposition (SVD)

Eigenvector Rotation

• Given,
  – \( C = R^T R \) which is a correlation matrix if \( R \) is mean centered and unit length.
  – \( R_{\text{new}} = RQ \) and
  – \( R_{\text{new}}^T = Q^T R^T \)
• Want
  – \( R_{\text{new}}^T R_{\text{new}} = Q^T R^T R Q = Q^T C Q = L \)
  – where \( L \) is diagonal

• This implies that
  – \( Q^T Q = Q^{-1} Q \)
  – so, \( CQ = QL \)
• This sets up an eigenvalue problem
  – \( CQ = QL \)
Eigenvector Rotation

• The EP, $\mathbf{CQ} = \mathbf{QL}$, can be solved 1 at a time:
  – $\mathbf{CQ} = \mathbf{IQ}$
  – $\mathbf{CQ} - \mathbf{IQ} = 0$
  – $(\mathbf{C} - \mathbf{I})\mathbf{R} = 0$

• Solution only exists if $|\mathbf{C} - \mathbf{I}| = 0$
• So, solve for roots of $\mathbf{I}$ then solve for columns of $\mathbf{R}$

Eigenvector Rotation

Example

\[
\mathbf{R} = \begin{bmatrix} 4 & 2 \\ 6 & 4 \\ 8 & 6 \end{bmatrix}
\]

Find $\mathbf{C} = \mathbf{R}_{\text{ss}}^T \mathbf{R}_{\text{ss}}$

\[
\begin{bmatrix}
\mathcal{\gamma}_{\mathbf{C}} & 0 & \mathcal{\gamma}_{\mathbf{C}} \\
0 & \mathcal{\gamma}_{\mathbf{C}} & \mathcal{\gamma}_{\mathbf{C}} \\
\mathcal{\gamma}_{\mathbf{C}} & 0 & \mathcal{\gamma}_{\mathbf{C}}
\end{bmatrix}
= \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}
\]

Autocale $\mathbf{R}$:

Mean Center: $\bar{x}_2 = 3; \bar{x}_2 = 2; \mathbf{R} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$

Variance Scale: $|\bar{x}_2| = \sqrt{2}; |\bar{x}_2| = \sqrt{2}; \mathbf{R} = \begin{bmatrix} \mathcal{\gamma}_{\mathbf{C}} & \mathcal{\gamma}_{\mathbf{C}} \\
\mathcal{\gamma}_{\mathbf{C}} & \mathcal{\gamma}_{\mathbf{C}} \\
\mathcal{\gamma}_{\mathbf{C}} & \mathcal{\gamma}_{\mathbf{C}} \end{bmatrix}$

Diagonalize $\mathbf{C}$

$|\mathbf{C} - \lambda \mathbf{I}| = \begin{bmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix} = 0$
Eigenvector Rotation Example

• Solve for roots
  - \((1 - l)^2 - 1 = 0\)
  - \(l(l-2) = 0\)
  - \(l_1 = 2, l_2 = 0\)
  - Note \(\sum \lambda_i = \text{Trace}[C]\)

• Solve for first eigenvector \((l = 2)\):
  \[
  \begin{bmatrix}
  1 - 2 & 1 \\
  1 & 1 - 2
  \end{bmatrix}
  \begin{bmatrix}
  v_{11} \\
  v_{21}
  \end{bmatrix}
  =
  \begin{bmatrix}
  0 \\
  0
  \end{bmatrix}
  \]
  - \(-v_{11} + v_{21} = 0\)
  - \(v_{11} - v_{21} = 0\)
  - \(v_{11} = v_{21}\)

• Normalize 1st eigenvector to unit length
  \[v_1 = \begin{bmatrix} \sqrt{3} \\ \sqrt{3} \end{bmatrix}\]

• Solve for 2nd eigenvector \((l = 0)\):
  \[
  \begin{bmatrix}
  1 - 0 & 1 \\
  1 & 1 - 0
  \end{bmatrix}
  \begin{bmatrix}
  v_{12} \\
  v_{22}
  \end{bmatrix}
  =
  \begin{bmatrix}
  0 \\
  0
  \end{bmatrix}
  \]
  - \(v_{12} + v_{22} = 0\)
  - \(v_{12} + v_{22} = 0\)
  - \(v_{12} = -v_{22}\)

• Normalize 2nd eigenvector to unit length:
  \[v_2 = \begin{bmatrix} \sqrt{3} \\ -\sqrt{3} \end{bmatrix} \text{ or } \begin{bmatrix} -\sqrt{3} \\ \sqrt{3} \end{bmatrix}\]

• Therefore
  \[V = \begin{bmatrix}
  \sqrt{3} & \sqrt{3} \\
  \sqrt{3} & -\sqrt{3}
  \end{bmatrix}\]

• Find points on new axis
  \[R_{new} = \begin{bmatrix}
  \sqrt{3} & \sqrt{3} & 0 & 0 \\
  0 & 0 & \sqrt{3} & \sqrt{3} \\
  \sqrt{3} & \sqrt{3} & \sqrt{3} & -\sqrt{3}
  \end{bmatrix}
  \begin{bmatrix}
  -1 \\
  0 \\
  0 \\
  1
  \end{bmatrix}\]

• Graphically
  \[v_1\]
  \[v_2\]

Note: \(v_1\) contains 100% of the information
PCA EP Summary

- Translation
  - center and scale \( \mathbf{R} \)
- Covariance matrix
  - \( \mathbf{C} = \mathbf{R}^T \mathbf{R} \)
- Diagonalize \( \mathbf{C} \)
  \[
  |\mathbf{C} - \lambda \mathbf{I}| = \begin{vmatrix} 1 - \lambda & 1 \\ 1 & 1 - \lambda \end{vmatrix} = 0
  \]
- Find rotation matrix
  - \( (\mathbf{C} - \mathbf{I})\mathbf{v} = 0 \)
  - \( \mathbf{V} = [\mathbf{v}_1 | ... | \mathbf{v}_n] \)
- Score matrix:
  - \( \mathbf{R}_{\text{new}} = \mathbf{RV} \)

PCA by NIPLS

- NIPLS is an iterative methods where 1 factor is calculated at a time.
- Uses model
  - \( \mathbf{R} = \mathbf{PQ}^T \)
- 1. Start with normalized guess of scores for \( n^{\text{th}} \) factor \( \mathbf{p} \) where \( \|\mathbf{p}^T\mathbf{p}\| = 1 \)
- 2. Calculate R-block loadings, \( \mathbf{q} \)
  \[
  \mathbf{q} = \frac{\mathbf{R}^T\mathbf{p}}{\|\mathbf{p}^T \mathbf{RR}^T \mathbf{p}\|}
  \]
- 3. Calculate new X-block score vector, \( \mathbf{p} \)
  - \( \mathbf{p} = \mathbf{Rq} \)
- 4. Check for convergence,
  - \( \text{abs}(\|\mathbf{p}^T\mathbf{p}\|_{\text{new}} - \|\mathbf{p}^T\mathbf{p}\|_{\text{old}}) < 10^{-6} \),
  - If not converged loop back to 2.
- 5. If converged, variance (information) described by this PC is subtracted from \( \mathbf{R} \)
  - \( \mathbf{R}_{\text{new}} = \mathbf{R} - \mathbf{pq}^T \)
- 6. If more PCs are needed, return to 1 with \( \mathbf{R}_{\text{new}} \) (from 5).
PCA by Singular Value Decomposition

- \( X_{(I \times J)} \) has a singular value decomposition
- \( X_{(I \times J)} = U_{(I \times I)} S_{(I \times J)} V_{(J \times J)}^T \)
  - where
    - \( U^T U = I_{I,I} \)
    - \( V^T V = I_{J,J} \)
    - \( U \) are eigenvectors of \( X^T X \)
    - \( V \) are eigenvectors of \( X'X \)
    - \( S \) are square roots of eigenvalues from \( XX^T \) or \( X'X \)
- Proof:
  - \( X = USV^T \) and \( X^T = VSU^T \)
  - \( X'X = VSU'USV^T \)
  - \( X'X = VS^2 V^T \)
  - \( X'XV = VS^2 \)
- Pseudoinverse:
  - \( X^+ = (USV^T)^+ = VS^{-1} U^T \)
- Projection:
  - \( X^+ \cdot X = VS^{-1} U^T USV^T = VV^T \)
  - \( XX^+ = USV^T VS^{-1} U^T = UU^T \)

Notes Regarding Factor Analysis

- Any matrix can be factored into
  - \( R = PQ^T \)
  - \( R = USV^T \)
- Infinite ways to factor \( R \)
- PCA (orthogonal factors) is just one possibility
- Not all factors are meaningful
  - just mathematical constructs
- But, some factors (models) are useful!
- Determining ‘significant’ factors is difficult.
- Noise filtering is possible by eliminating ‘non-significant’ factors
- Can be used for data and variable reduction
Example

- Variables: 53 blood and urine measurements
- Objects: 65 people (33 “alcoholics” in treatment; 32 “non-alcoholics”)
- PCA on un-normalized data
  - Why doesn’t it work?

Example

- Look at raw data
  - some variables have large mean
  - others have large variance
  - this variance does not translate into ‘predictive’ variance
  - some variables have very low variance
  - these may be more useful
  - but are swamped by other variances
Example

- PCA on mean centered data
- Still no classification
- Why?
  - Look at loadings
  - A couple of high variance measurements dominate the data set

Example

- PCA on autoscaled data
- Decent classification
  - only PC1 is useful
- Loadings not dominated by any one variable
- Questions
  - How much variance does each PC describe?
  - Are all variables needed?
  - Any outliers?
Example

• Factored data $\mathbf{R} = \mathbf{USV}^T$
• Recall that scale is in $\mathbf{S}$
• Total variance $SS_i^2$
• Percent variance captured by $i$th PC is
  $- S_i^2 / SS_i^2$

Example

• Squaring a loading gives an estimate of the leverage each variable has on determining the loading.
• Variables with small leverage can be eliminated with little impact on the overall model
Example

PCA on autoscaled data (24 variables w/ $\lambda^2 > .015$):

PCA on autoscaled data (10 variables w/ $\lambda^2 > .040$):

Determining Number of PC

- No single best method
- Subjective
- Knowledge of data and errors are helpful

- Some tactics:
  - Variance described
  - Inspection of scores and loadings
  - Experimental error
  - Empirical
  - Statistical
  - Cross validation