CAP 5768: Introduction to Data Science

Giri NARASIMHAN

www.cis.fiu.edu/~giri/teach/5768.html
PCA Recap

From Johnson & Wichern, Applied multivariate statistical analysis, 6th Ed
PCA

- Tool for Dimensionality Reduction
  - Reduces impact of curse of dimensionality
- Tool for finding Subspace in which data lies
- Summarization of data to find important variables
PCA Animation

https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues
Matrices & Transformations

- Linear Transformations
  
  \[ Ax = y \]
Data as Matrices

\[
\mathbf{X} = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{bmatrix} = \begin{bmatrix}
x_1' \\
x_2' \\
\vdots \\
x_n'
\end{bmatrix} \leftarrow 1\text{st (multivariate) observation}
\]

\[
\begin{bmatrix}
x_1' \\
x_2' \\
\vdots \\
x_n'
\end{bmatrix} \leftarrow n\text{th (multivariate) observation}
\]
Eigenvalues and Eigenvectors

- $Ax = \lambda x$, for square matrices $A$
- Characteristic Eq: $|A - \lambda I| = 0$
The scalar $x'Ax$ is called quadratic form

$$Q(x) = x'Ax,$$

$$Q(x) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij}x_i x_j.$$
Spectral Decomposition

For symmetric square matrices $A$, the spectral decomposition is:

$$A = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_k e_k e_k'$$
Spectral Decomposition ... 2

\[ A = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i' = \mathbf{P} \Lambda \mathbf{P}' \]

\[ \mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k] \]

\[ \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \]
Dimension Reduction Revisited

- If we take $r$ eigenvectors, then
  - $P_r = [e_1, e_2, \ldots, e_r]$, and

- $A$ can be approximated by taking $r$ eigenvectors

$$P \Lambda P' = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_r \end{pmatrix}$$

$$(k \times r) \quad (r \times r) \quad (r \times k)$$
Singular Value Decomposition

- **Spectral Decomp. for sq. symm. matrices**
- **Non-sq. asymmetric matrices?**
  - Use sq. root of eigenvalues of AA’
  - Singular values of A

\[
A = U \Lambda V' \\
(\text{m} \times k) \quad (\text{m} \times m) (\text{m} \times k) (k \times k)
\]
Dimensionality Reduction

Given $m \times k$ matrix $A$, we can approximate it by $m \times s$ matrix $B$ with $s < k = \text{rank}(A)$. Then

\[
B = \sum_{i=1}^{s} \lambda_i u_i v_i^t
\]

Here we are picking $s$ singular values from SVD.
Central Limit Theorem
How to make data “normal”?

- Let $X_1, X_2, \ldots, X_n$ be independent observations from any distribution with mean $\mu$ and variance $\Sigma$. Then

$$\sqrt{n} \left( \bar{X} - \mu \right)$$

has an approximate $N_p(0, \Sigma)$ distribution.

- Sample size, $n$, must be large relative to $p$. 
Q-Q plot

<table>
<thead>
<tr>
<th>Ordered observations $x_{(j)}$</th>
<th>Probability levels $(j - \frac{1}{2})/n$</th>
<th>Standard normal quantiles $q_{(j)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-1.00$</td>
<td>.05</td>
<td>$-1.645$</td>
</tr>
<tr>
<td>$-1.10$</td>
<td>.15</td>
<td>$-1.036$</td>
</tr>
<tr>
<td>$.16$</td>
<td>.25</td>
<td>$-1.674$</td>
</tr>
<tr>
<td>$.41$</td>
<td>.35</td>
<td>$-1.385$</td>
</tr>
<tr>
<td>$.62$</td>
<td>.45</td>
<td>$.125</td>
</tr>
<tr>
<td>$.80$</td>
<td>.55</td>
<td>$.125</td>
</tr>
<tr>
<td>$1.26$</td>
<td>.65</td>
<td>$.385</td>
</tr>
<tr>
<td>$1.54$</td>
<td>.75</td>
<td>$.674</td>
</tr>
<tr>
<td>$1.71$</td>
<td>.85</td>
<td>$1.036$</td>
</tr>
<tr>
<td>$2.30$</td>
<td>.95</td>
<td>$1.645$</td>
</tr>
</tbody>
</table>

The image shows a Q-Q plot with ordered observations and corresponding probability levels and standard normal quantiles. The plot compares the empirical distribution with the expected normal distribution.
Chi-Square Plots

\[ d_j^2 = (x_j - \bar{x})' S^{-1} (x_j - \bar{x}), \quad j = 1, 2, \ldots, n \]

<table>
<thead>
<tr>
<th>j</th>
<th>(d_j^2)</th>
<th>(q_{0.2}(\frac{j - \frac{1}{2}}{10}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.59</td>
<td>.10</td>
</tr>
<tr>
<td>2</td>
<td>.81</td>
<td>.33</td>
</tr>
<tr>
<td>3</td>
<td>.83</td>
<td>.58</td>
</tr>
<tr>
<td>4</td>
<td>.97</td>
<td>.86</td>
</tr>
<tr>
<td>5</td>
<td>1.01</td>
<td>1.20</td>
</tr>
<tr>
<td>6</td>
<td>1.02</td>
<td>1.60</td>
</tr>
<tr>
<td>7</td>
<td>1.20</td>
<td>2.10</td>
</tr>
<tr>
<td>8</td>
<td>1.88</td>
<td>2.77</td>
</tr>
<tr>
<td>9</td>
<td>4.34</td>
<td>3.79</td>
</tr>
<tr>
<td>10</td>
<td>5.33</td>
<td>5.99</td>
</tr>
</tbody>
</table>
Chi-Square Distribution

- Squared Generalized Distances
  \[ d_j^2 = (x_j - \bar{x})' S^{-1} (x_j - \bar{x}), \quad j = 1, 2, \ldots, n \]
  - If X is multivariate normal and n and n-p are large, then the squared distances behave like a chi-squared plot or gamma plot.
Detecting Outliers

- Visual detection

- Harder in multivariate case. Why?
  - May be univariate or multivariate outlier
Bivariate Outliers
Multivariate Outliers

- Some outliers are hard to detect
- Look for large values of

\[(x_j - \bar{x})' S^{-1} (x_j - \bar{x}).\]
Outlier detection

- Dot plots for each variable
- Scatter plot for each pair of variables
- Calculate z-values and examine for outliers
  \[ z_{jk} = \frac{x_{jk} - \bar{x}_k}{\sqrt{s_{kk}}} \]
- Calculate gen sq distances & look for outliers
  \[ (x_j - \bar{x})' S^{-1} (x_j - \bar{x}) . \]
### Other Transforms for Normality

**HELPFUL TRANSFORMATIONS TO NEAR NORMALITY**

<table>
<thead>
<tr>
<th>Original Scale</th>
<th>Transformed Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counts, $y$</td>
<td>$\sqrt{y}$</td>
</tr>
<tr>
<td>Proportions, $\hat{p}$</td>
<td>$\text{logit}(\hat{p}) = \frac{1}{2} \log \left( \frac{\hat{p}}{1 - \hat{p}} \right)$ (4-33)</td>
</tr>
<tr>
<td>Correlations, $r$</td>
<td>Fisher’s $z(r) = \frac{1}{2} \log \left( \frac{1 + r}{1 - r} \right)$</td>
</tr>
</tbody>
</table>