PCA, Clustering and Classification

By H. Bjørn Nielsen strongly inspired by Agnieszka S. Juncker
Motivation: Multidimensional data

<table>
<thead>
<tr>
<th></th>
<th>Pat1</th>
<th>Pat2</th>
<th>Pat3</th>
<th>Pat4</th>
<th>Pat5</th>
<th>Pat6</th>
<th>Pat7</th>
<th>Pat8</th>
</tr>
</thead>
<tbody>
<tr>
<td>209619_at</td>
<td>7758</td>
<td>4705</td>
<td>5342</td>
<td>7443</td>
<td>8747</td>
<td>4933</td>
<td>7950</td>
<td>5031</td>
</tr>
<tr>
<td>32541_at</td>
<td>280</td>
<td>387</td>
<td>392</td>
<td>238</td>
<td>385</td>
<td>329</td>
<td>337</td>
<td>163</td>
</tr>
<tr>
<td>206398_s_at</td>
<td>1050</td>
<td>835</td>
<td>1268</td>
<td>1723</td>
<td>1377</td>
<td>804</td>
<td>1846</td>
<td>1180</td>
</tr>
<tr>
<td>219281_at</td>
<td>391</td>
<td>593</td>
<td>298</td>
<td>265</td>
<td>491</td>
<td>517</td>
<td>334</td>
<td>387</td>
</tr>
<tr>
<td>207857_at</td>
<td>1425</td>
<td>977</td>
<td>2072</td>
<td>1184</td>
<td>939</td>
<td>814</td>
<td>658</td>
<td>593</td>
</tr>
<tr>
<td>211338_at</td>
<td>37</td>
<td>27</td>
<td>28</td>
<td>38</td>
<td>33</td>
<td>16</td>
<td>36</td>
<td>23</td>
</tr>
<tr>
<td>213539_at</td>
<td>124</td>
<td>197</td>
<td>454</td>
<td>116</td>
<td>162</td>
<td>113</td>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>221497_x_at</td>
<td>120</td>
<td>86</td>
<td>175</td>
<td>99</td>
<td>115</td>
<td>80</td>
<td>83</td>
<td>119</td>
</tr>
<tr>
<td>213958_at</td>
<td>179</td>
<td>225</td>
<td>449</td>
<td>174</td>
<td>185</td>
<td>203</td>
<td>186</td>
<td>185</td>
</tr>
<tr>
<td>210835_s_at</td>
<td>203</td>
<td>144</td>
<td>197</td>
<td>314</td>
<td>250</td>
<td>353</td>
<td>173</td>
<td>285</td>
</tr>
<tr>
<td>209199_s_at</td>
<td>758</td>
<td>1234</td>
<td>833</td>
<td>1449</td>
<td>769</td>
<td>1110</td>
<td>987</td>
<td>638</td>
</tr>
<tr>
<td>217979_at</td>
<td>570</td>
<td>563</td>
<td>972</td>
<td>796</td>
<td>869</td>
<td>494</td>
<td>673</td>
<td>1013</td>
</tr>
<tr>
<td>201015_s_at</td>
<td>533</td>
<td>343</td>
<td>325</td>
<td>270</td>
<td>691</td>
<td>460</td>
<td>563</td>
<td>321</td>
</tr>
<tr>
<td>203332_s_at</td>
<td>649</td>
<td>354</td>
<td>494</td>
<td>554</td>
<td>710</td>
<td>455</td>
<td>748</td>
<td>392</td>
</tr>
<tr>
<td>204670_x_at</td>
<td>5577</td>
<td>3216</td>
<td>5323</td>
<td>4423</td>
<td>5771</td>
<td>3374</td>
<td>4328</td>
<td>3515</td>
</tr>
<tr>
<td>208788_at</td>
<td>648</td>
<td>327</td>
<td>1057</td>
<td>746</td>
<td>541</td>
<td>270</td>
<td>361</td>
<td>774</td>
</tr>
<tr>
<td>210784_x_at</td>
<td>142</td>
<td>151</td>
<td>144</td>
<td>173</td>
<td>148</td>
<td>145</td>
<td>131</td>
<td>146</td>
</tr>
<tr>
<td>204319_s_at</td>
<td>298</td>
<td>172</td>
<td>200</td>
<td>298</td>
<td>196</td>
<td>104</td>
<td>144</td>
<td>110</td>
</tr>
<tr>
<td>205049_s_at</td>
<td>3294</td>
<td>1351</td>
<td>2080</td>
<td>2066</td>
<td>3726</td>
<td>1396</td>
<td>2244</td>
<td>2142</td>
</tr>
<tr>
<td>202114_at</td>
<td>833</td>
<td>674</td>
<td>733</td>
<td>1298</td>
<td>862</td>
<td>371</td>
<td>886</td>
<td>501</td>
</tr>
<tr>
<td>213792_s_at</td>
<td>646</td>
<td>375</td>
<td>370</td>
<td>436</td>
<td>738</td>
<td>497</td>
<td>546</td>
<td>406</td>
</tr>
<tr>
<td>203932_at</td>
<td>1977</td>
<td>1016</td>
<td>2436</td>
<td>1856</td>
<td>1917</td>
<td>822</td>
<td>1189</td>
<td>1092</td>
</tr>
<tr>
<td>203963_at</td>
<td>97</td>
<td>63</td>
<td>77</td>
<td>136</td>
<td>85</td>
<td>74</td>
<td>91</td>
<td>61</td>
</tr>
<tr>
<td>203978_at</td>
<td>315</td>
<td>279</td>
<td>221</td>
<td>260</td>
<td>227</td>
<td>222</td>
<td>232</td>
<td>141</td>
</tr>
<tr>
<td>203753_at</td>
<td>1468</td>
<td>1105</td>
<td>381</td>
<td>1154</td>
<td>980</td>
<td>1419</td>
<td>1253</td>
<td>554</td>
</tr>
<tr>
<td>204891_s_at</td>
<td>78</td>
<td>71</td>
<td>152</td>
<td>74</td>
<td>127</td>
<td>57</td>
<td>66</td>
<td>153</td>
</tr>
<tr>
<td>209365_s_at</td>
<td>472</td>
<td>519</td>
<td>365</td>
<td>349</td>
<td>756</td>
<td>528</td>
<td>637</td>
<td>828</td>
</tr>
<tr>
<td>209604_s_at</td>
<td>772</td>
<td>74</td>
<td>130</td>
<td>216</td>
<td>108</td>
<td>311</td>
<td>80</td>
<td>235</td>
</tr>
<tr>
<td>211005_at</td>
<td>49</td>
<td>58</td>
<td>129</td>
<td>70</td>
<td>56</td>
<td>77</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>219686_at</td>
<td>694</td>
<td>342</td>
<td>345</td>
<td>502</td>
<td>960</td>
<td>403</td>
<td>535</td>
<td>513</td>
</tr>
<tr>
<td>38521_at</td>
<td>775</td>
<td>604</td>
<td>305</td>
<td>563</td>
<td>542</td>
<td>543</td>
<td>725</td>
<td>587</td>
</tr>
<tr>
<td>217853_at</td>
<td>367</td>
<td>168</td>
<td>107</td>
<td>160</td>
<td>287</td>
<td>264</td>
<td>273</td>
<td>113</td>
</tr>
<tr>
<td>217028_at</td>
<td>4926</td>
<td>2667</td>
<td>3542</td>
<td>5163</td>
<td>4683</td>
<td>3281</td>
<td>4822</td>
<td>3978</td>
</tr>
<tr>
<td>201137_s_at</td>
<td>4733</td>
<td>2846</td>
<td>1834</td>
<td>5471</td>
<td>5079</td>
<td>2330</td>
<td>3345</td>
<td>1460</td>
</tr>
<tr>
<td>202284_s_at</td>
<td>600</td>
<td>1823</td>
<td>1657</td>
<td>1177</td>
<td>972</td>
<td>2303</td>
<td>1574</td>
<td>1731</td>
</tr>
<tr>
<td>201999_s_at</td>
<td>897</td>
<td>959</td>
<td>800</td>
<td>808</td>
<td>297</td>
<td>1014</td>
<td>998</td>
<td>663</td>
</tr>
</tbody>
</table>
Outline

• Dimension reduction
  – PCA
  – Clustering

• Classification

• Example: study of childhood leukemia
Childhood Leukemia

- Cancer in the cells of the immune system
- Approx. 35 new cases in Denmark every year
- 50 years ago – all patients died
- Today – approx. 78% are cured
- Riskgroups
  - Standard
  - Intermediate
  - High
  - Very high
  - Extra high
- Treatment
  - Chemotherapy
  - Bone marrow transplantation
  - Radiation
# Prognostic Factors

<table>
<thead>
<tr>
<th></th>
<th>Good prognosis</th>
<th>Poor prognosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immunophenotype</td>
<td>precursor B</td>
<td>T</td>
</tr>
<tr>
<td>Age</td>
<td>1-9</td>
<td>≥10</td>
</tr>
<tr>
<td>Leukocyte count</td>
<td>Low (&lt;50*10^9/L)</td>
<td>High (&gt;100*10^9/L)</td>
</tr>
<tr>
<td>Number of chromosomes</td>
<td>Hyperdiploidy (&gt;50)</td>
<td>Hypodiploidy (&lt;46)</td>
</tr>
<tr>
<td>Translocations</td>
<td>t(12;21)</td>
<td>t(9;22), t(1;19)</td>
</tr>
<tr>
<td>Treatment response</td>
<td>Good response</td>
<td>Poor response</td>
</tr>
</tbody>
</table>
Study of Childhood Leukemia

- Diagnostic bone marrow samples from leukemia patients
- Platform: Affymetrix Focus Array
  - 8763 human genes
- Immunophenotype
  - 18 patients with precursor B immunophenotype
  - 17 patients with T immunophenotype
- Outcome 5 years from diagnosis
  - 11 patients with relapse
  - 18 patients in complete remission
Principal Component Analysis (PCA)

- used for visualization of complex data
- developed to capture as much of the variation in data as possible
Principal components

• 1. principal component (PC1)
  – the direction along which there is greatest variation

• 2. principal component (PC2)
  – the direction with maximum variation left in data, orthogonal to the 1. PC

• General about principal components
  – linear combinations of the original variables
  – uncorrelated with each other
Principal components
PCA - example
PCA on all Genes
Leukemia data, precursor B and T

Plot of 34 patients, dimension of 8973 genes reduced to 2
Outcome: PCA on all Genes
Principal components - Variance

Variance (%)

PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10

25 | 15 | 10 | 6  | 4  | 3  | 2  | 1  | 1  | 1
Clustering methods

• Hierarchical
  – agglomerative (bottom-up)
    eg. UPGMA
  – divisive (top-down)

• Partitioning
  – eg. K-means clustering
Hierarchical clustering

- Representation of all pairwise distances
- Parameters: none (distance measure)
- Results:
  - in one large cluster
  - hierarchical tree (dendrogram)
- Deterministic
Hierarchical clustering
– UPGMA Algorithm

• Assign each item to its own cluster
• Join the nearest clusters
• Reestimate the distance between clusters
• Repeat for 1 to n
Hierarchical clustering
Hierarchical clustering

Data with clustering order and distances

Dendrogram representation
Leukemia data - clustering of patients
K-means clustering

- Partition data into K clusters
- Parameter: Number of clusters (K) must be chosen
- Randomized initialization:
  - different clusters each time
K-means - Algorithm

- Assign each item a class in 1 to K (randomly)
- For each class 1 to K
  - Calculate the centroid (one of the K-means)
  - Calculate distance from centroid to each item
- Assign each item to the nearest centroid
- Repeat until no items are re-assigned (convergence)
K-means clustering, K=3
K-means clustering, K=3
K-means clustering, K=3
Comparison of clustering methods

• Hierarchical clustering
  – Distances between all variables
  – Timeconsuming with a large number of genes
  – Advantage to cluster on selected genes

• K-mean clustering
  – Faster algorithm
  – Does not show relations between all variables
Distance measures

- Euclidian distance

\[ d(x_i, y_i) = \left( \sum_{i=1}^{N} (x_i - y_i)^2 \right)^{1/2} \]

- Vector angle distance

\[ d(x_i, y_i) = (1 - \cos \alpha) = 1 - \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}} \]

- Pearson's distance

\[ d(x_i, y_i) = (1 - CC) = 1 - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \]
Comparison of distance measures

Euclidean  Vector angle  Pearson
Classification

- Feature selection
- Classification methods
- Cross-validation
- Training and testing
Reduction of input features

• Dimension reduction
  – PCA

• Feature selection (gene selection)
  – Significant genes: t-test
  – Selection of a limited number of genes
# Microarray Data

<table>
<thead>
<tr>
<th>Class</th>
<th>precursorB</th>
<th>T</th>
<th>Patient3</th>
<th>Patient4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Patient1</td>
<td>Patient2</td>
<td>Patient3</td>
<td>Patient4</td>
<td>...</td>
</tr>
<tr>
<td>Gene1</td>
<td>1789.5</td>
<td>963.9</td>
<td>2079.5</td>
<td>3243.9</td>
<td>...</td>
</tr>
<tr>
<td>Gene2</td>
<td>46.4</td>
<td>52.0</td>
<td>22.3</td>
<td>27.1</td>
<td>...</td>
</tr>
<tr>
<td>Gene3</td>
<td>215.6</td>
<td>276.4</td>
<td>245.1</td>
<td>199.1</td>
<td>...</td>
</tr>
<tr>
<td>Gene4</td>
<td>176.9</td>
<td>504.6</td>
<td>420.5</td>
<td>380.4</td>
<td></td>
</tr>
<tr>
<td>Gene5</td>
<td>4023.0</td>
<td>3768.6</td>
<td>4257.8</td>
<td>4451.8</td>
<td></td>
</tr>
<tr>
<td>Gene6</td>
<td>12.6</td>
<td>12.1</td>
<td>37.7</td>
<td>38.7</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Gene8793</td>
<td>312.5</td>
<td>415.9</td>
<td>1045.4</td>
<td>1308.0</td>
<td></td>
</tr>
</tbody>
</table>
Outcome: PCA on Selected Genes
Outcome Prediction:
CC against the Number of Genes

![Graph showing the correlation coefficient (CC) against the number of top ranking genes.](image-url)
Linear discriminant analysis

- Assumptions:
  - Data eg. Gaussian distributed
  - Variances and covariances the same for classes
Nearest Centroid

- Calculation of a centroid for each class

\[ \bar{x}_{ik} = \frac{\sum_{j \in C_k} x_{ij}}{n_k} \]

- Calculation of the distance between a test sample and each class centroid

- Class prediction by the nearest centroid method
K-Nearest Neighbor (KNN)

• Based on distance measure
  – For example Euclidian distance

• Parameter $k =$ number of nearest neighbors
  – $k=1$
  – $k=3$
  – $k=...$

• Prediction by majority vote for odd numbers
Support Vector Machines

- Machine learning
- Relatively new and highly theoretic
- Works on non-linearly separable data

- Finding a hyperplane between the two classes by minimizing the distance between the hyperplane and closest points
## Comparison of Methods

<table>
<thead>
<tr>
<th>Linear discriminant analysis</th>
<th>Neural networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest centroid</td>
<td>Support vector machines</td>
</tr>
<tr>
<td>Simple method</td>
<td>Advanced methods</td>
</tr>
<tr>
<td>Based on distance calculation</td>
<td>Involve machine learning</td>
</tr>
<tr>
<td>Good for simple problems</td>
<td>Several adjustable parameters</td>
</tr>
<tr>
<td>Good for few training samples</td>
<td>Many training samples required</td>
</tr>
<tr>
<td>Distribution of data assumed</td>
<td>Flexible methods</td>
</tr>
</tbody>
</table>

**KNN**
Cross-validation

Data: 10 samples

**Cross-5-validation:**
- Training: 4/5 of data (8 samples)
- Testing: 1/5 of data (2 samples)
- -> 5 different models

**Leave-one-out cross-validation (LOOCV)**
- Training: 9/10 of data (9 samples)
- Testing: 1/10 of data (1 sample)
- -> 10 different models
Validation

• Definition of
  – true and false positives
  – true and false negatives

<table>
<thead>
<tr>
<th>Actual class</th>
<th>B</th>
<th>B</th>
<th>T</th>
<th>T</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted class</td>
<td>B</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>B</td>
</tr>
</tbody>
</table>

| TP | FN | TN | FP |
Accuracy

• Definition: \( \frac{TP + TN}{TP + TN + FP + FN} \)

• Range: 0 – 100%
Matthews correlation coefficient

- Definition: \( \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TN+FN)(TN+FP)(TP+FN)(TP+FP)}} \)

- Range: (-1) – 1
Sensitivity and Specificity

Sensitivity: the ability to detect "true positives"
\[
\frac{TP}{TP + FN}
\]

Specificity: the ability to avoid "false positives"
\[
\frac{TP}{TP + FP}
\]
Overview of Classification

Expression data

Subdivision of data for cross-validation into training sets and test sets

Feature selection (t-test)
Dimension reduction (PCA)

Training of classifier:
- using cross-validation
  - choice of method
  - choice of optimal parameters

Testing of classifier

Independant test set
Important Points

• Avoid overfitting

• Validate performance
  – Test on an independant test set
  – by using cross-validation

• Include feature selection in cross-validation

Why?
  – To avoid overestimation of performance!
  – To make a general classifier
Study of Childhood Leukemia: Results

• **Classification of immunophenotype (precursorB og T)**
  - 100% accuracy
    • During the training
    • When testing on an independant test set
  - Simple classification methods applied
    • K-nearest neighbor
    • Nearest centroid

• **Classification of outcome (relapse or remission)**
  - 78% accuracy (CC = 0.59)
  - Simple and advanced classification methods applied
Risk classification in the future?

**Patient:**
- Clinical data
- Immunophenotyping
- Morphology
- Genetic measurements
- Microarray technology

**Prognostic factors:**
- Immunophenotype
- Age
- Leukocyte count
- Number of chromosomes
- Translocations
- Treatment response

**Risk group:**
- Standard
- Intermediate
- High
- Very high
- Extra high

**Custom designed treatment**
Summary

• Dimension reduction important to visualize data
  – Principal Component Analysis
  – Clustering
    • Hierarchical
    • Partitioning (K-means)
      (distance measure important)

• Classification
  – Reduction of dimension often necessary (t-test, PCA)
  – Several classification methods available
  – Validation