MCI5004: Molecular Biomarkers in Clinical Research

Principal Component Analysis in Biomarker Discovery

Wong Limsoon
Plan

• PCA
• PCA in biomarker selection

• Batch effects
• PCA for isolating batch effects

• PCA at the level of protein complexes / biological pathway subnetworks
PRINCIPAL COMPONENT ANALYSIS (PCA)
PCA, intuitively

Credit: Alessandro Giuliani

https://georgemdalas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction
PCA, a la Pearson (1901)

For example:—Let $P_1, P_2, \ldots, P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots, x_n, y_n$, and perpendicular distances $p_1, p_2, \ldots, p_n$ from a line $AB$. Then we shall make

$$U = S(y^2) = \text{a minimum.}$$

If $y$ were the dependent variable, we should have made

$$S(y' - y)^2 = \text{a minimum.}$$

Credit: Alessandro Giuliani
PCA, in modern English 😊

Introduction

- Technique quite old: Pearson (1901) and Hotelling (1933), but still one of the most used multivariate techniques today.
- Main idea:
  - Start with variables $X_1, \ldots, X_p$.
  - Find a rotation of these variables, say $Y_1, \ldots, Y_p$ (called principal components), so that:
    - $Y_1, \ldots, Y_p$ are uncorrelated. Idea: they measure different dimensions of the data.
    - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \geq \text{Var}(Y_p)$. Idea: $Y_1$ is most important, then $Y_2$, etc.

Definition of PCA

- Given $X = (X_1, \ldots, X_p)'$.
- We call $a'X$ a standard linear combination (SLC) if $\sum a_i^2 = 1$.
- Find the SLC $a_{(1)}' = (a_{11}, \ldots, a_{p1})$ so that $Y_1 = a_{(1)}'X$ has maximal variance.
- Find the SLC $a_{(2)}' = (a_{12}, \ldots, a_{p2})$ so that $Y_2 = a_{(2)}'X$ has maximal variance, subject to the constraint that $Y_2$ is uncorrelated to $Y_1$.
- Find the SLC $a_{(3)}' = (a_{13}, \ldots, a_{p3})$ so that $Y_3 = a_{(3)}'X$ has maximal variance, subject to the constraint that $Y_3$ is uncorrelated to $Y_1$ and $Y_2$.
- Etc...
Nice free Excel add-on


SIZE AND SHAPE VARIATION IN THE PAINTED TURTLE.¹
A PRINCIPAL COMPONENT ANALYSIS

Pierre Jolicoeur and James E. Mosimann²

Walker Museum, University of Chicago
and
Institut de Biologie, Université de Montréal

(Received for publication July 11, 1960)

Credit: Alessandro Giuliani
<table>
<thead>
<tr>
<th>24 Males</th>
<th>24 Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>width</td>
</tr>
<tr>
<td>93</td>
<td>74</td>
</tr>
<tr>
<td>94</td>
<td>78</td>
</tr>
<tr>
<td>96</td>
<td>80</td>
</tr>
<tr>
<td>101</td>
<td>84</td>
</tr>
<tr>
<td>102</td>
<td>85</td>
</tr>
<tr>
<td>103</td>
<td>81</td>
</tr>
<tr>
<td>104</td>
<td>83</td>
</tr>
<tr>
<td>106</td>
<td>83</td>
</tr>
<tr>
<td>107</td>
<td>82</td>
</tr>
<tr>
<td>112</td>
<td>89</td>
</tr>
<tr>
<td>113</td>
<td>88</td>
</tr>
<tr>
<td>114</td>
<td>86</td>
</tr>
<tr>
<td>116</td>
<td>90</td>
</tr>
<tr>
<td>117</td>
<td>90</td>
</tr>
<tr>
<td>117</td>
<td>91</td>
</tr>
<tr>
<td>119</td>
<td>93</td>
</tr>
<tr>
<td>120</td>
<td>89</td>
</tr>
<tr>
<td>120</td>
<td>93</td>
</tr>
<tr>
<td>121</td>
<td>95</td>
</tr>
<tr>
<td>125</td>
<td>93</td>
</tr>
<tr>
<td>127</td>
<td>96</td>
</tr>
<tr>
<td>128</td>
<td>95</td>
</tr>
<tr>
<td>131</td>
<td>95</td>
</tr>
<tr>
<td>135</td>
<td>106</td>
</tr>
</tbody>
</table>

Credit: Alessandro Giuliani
Pearson Correlation Coefficients,

<table>
<thead>
<tr>
<th></th>
<th>length</th>
<th>width</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>1.00000</td>
<td>0.97831</td>
<td>0.96469</td>
</tr>
<tr>
<td>width</td>
<td>0.97831</td>
<td>1.00000</td>
<td>0.96057</td>
</tr>
<tr>
<td>height</td>
<td>0.96469</td>
<td>0.96057</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

Width = 19.94 + 0.605*Length

Credit: Alessandro Giuliani
Principal components

<table>
<thead>
<tr>
<th></th>
<th>PC1 (98%)</th>
<th>PC2 (1.4%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>0.992</td>
<td>-0.067</td>
</tr>
<tr>
<td>Width</td>
<td>0.990</td>
<td>-0.100</td>
</tr>
<tr>
<td>Height</td>
<td>0.986</td>
<td>0.168</td>
</tr>
</tbody>
</table>

PC1 = 33.78*Length + 33.73*Width + 33.57*Height
PC2 = -1.57*Length – 2.33*Width + 3.93*Height

- Presence of an overwhelming size component explaining system variance comes from the presence of a ‘typical’ common shape
- Displacement along pc1 = size variation (all positive terms)
- Displacement along pc2 = shape deformation (both positive and negative terms)
<table>
<thead>
<tr>
<th>unit</th>
<th>sex</th>
<th>Length</th>
<th>Width</th>
<th>Height</th>
<th>PC1(size)</th>
<th>PC2(shape)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T25</td>
<td>F</td>
<td>98</td>
<td>81</td>
<td>38</td>
<td>-1.15774</td>
<td>0.80754832</td>
</tr>
<tr>
<td>T26</td>
<td>F</td>
<td>103</td>
<td>84</td>
<td>38</td>
<td>-0.96544</td>
<td>-0.1285916</td>
</tr>
<tr>
<td>T27</td>
<td>F</td>
<td>103</td>
<td>86</td>
<td>42</td>
<td>-0.7822</td>
<td>1.37433475</td>
</tr>
<tr>
<td>T28</td>
<td>F</td>
<td>105</td>
<td>86</td>
<td>40</td>
<td>-0.82922</td>
<td>0.28526912</td>
</tr>
<tr>
<td>T29</td>
<td>F</td>
<td>109</td>
<td>88</td>
<td>44</td>
<td>-0.55001</td>
<td>1.4815252</td>
</tr>
<tr>
<td>T30</td>
<td>F</td>
<td>123</td>
<td>92</td>
<td>50</td>
<td>0.027368</td>
<td>2.47830153</td>
</tr>
<tr>
<td>T31</td>
<td>F</td>
<td>123</td>
<td>95</td>
<td>46</td>
<td>-0.05281</td>
<td>0.05403839</td>
</tr>
<tr>
<td>T32</td>
<td>F</td>
<td>133</td>
<td>99</td>
<td>51</td>
<td>0.418589</td>
<td>0.88961967</td>
</tr>
<tr>
<td>T33</td>
<td>F</td>
<td>133</td>
<td>102</td>
<td>51</td>
<td>0.498425</td>
<td>0.33681756</td>
</tr>
<tr>
<td>T34</td>
<td>F</td>
<td>133</td>
<td>102</td>
<td>51</td>
<td>0.498425</td>
<td>0.33681756</td>
</tr>
<tr>
<td>T35</td>
<td>F</td>
<td>134</td>
<td>100</td>
<td>48</td>
<td>0.341684</td>
<td>-0.774911</td>
</tr>
<tr>
<td>T36</td>
<td>F</td>
<td>136</td>
<td>102</td>
<td>49</td>
<td>0.476898</td>
<td>-0.8289156</td>
</tr>
<tr>
<td>T37</td>
<td>F</td>
<td>137</td>
<td>98</td>
<td>51</td>
<td>0.457949</td>
<td>0.7671682</td>
</tr>
<tr>
<td>T38</td>
<td>F</td>
<td>138</td>
<td>99</td>
<td>51</td>
<td>0.501055</td>
<td>0.50628189</td>
</tr>
<tr>
<td>T39</td>
<td>F</td>
<td>140</td>
<td>105</td>
<td>53</td>
<td>0.7822</td>
<td>1.37433475</td>
</tr>
<tr>
<td>T40</td>
<td>F</td>
<td>147</td>
<td>108</td>
<td>57</td>
<td>1.129025</td>
<td>0.99544</td>
</tr>
<tr>
<td>T41</td>
<td>F</td>
<td>149</td>
<td>107</td>
<td>55</td>
<td>1.055392</td>
<td>0.1285916</td>
</tr>
<tr>
<td>T42</td>
<td>F</td>
<td>153</td>
<td>107</td>
<td>56</td>
<td>1.161368</td>
<td>0.2215593</td>
</tr>
<tr>
<td>T43</td>
<td>F</td>
<td>155</td>
<td>115</td>
<td>63</td>
<td>1.687277</td>
<td>1.86930869</td>
</tr>
<tr>
<td>T44</td>
<td>F</td>
<td>158</td>
<td>115</td>
<td>62</td>
<td>1.696753</td>
<td>1.17117077</td>
</tr>
<tr>
<td>T45</td>
<td>F</td>
<td>159</td>
<td>118</td>
<td>63</td>
<td>1.833086</td>
<td>1.00956637</td>
</tr>
<tr>
<td>T46</td>
<td>F</td>
<td>162</td>
<td>124</td>
<td>61</td>
<td>1.962232</td>
<td>-1.261771</td>
</tr>
<tr>
<td>T47</td>
<td>F</td>
<td>177</td>
<td>132</td>
<td>67</td>
<td>2.662548</td>
<td>-1.0787317</td>
</tr>
<tr>
<td>T48</td>
<td>F</td>
<td>155</td>
<td>117</td>
<td>60</td>
<td>1.620491</td>
<td>0.0960818</td>
</tr>
<tr>
<td>T49</td>
<td>M</td>
<td>93</td>
<td>74</td>
<td>37</td>
<td>-1.46649</td>
<td>2.0125924</td>
</tr>
<tr>
<td>T50</td>
<td>M</td>
<td>94</td>
<td>78</td>
<td>35</td>
<td>-1.42356</td>
<td>0.26342486</td>
</tr>
<tr>
<td>T51</td>
<td>M</td>
<td>96</td>
<td>80</td>
<td>35</td>
<td>-1.33735</td>
<td>-0.258445</td>
</tr>
<tr>
<td>T52</td>
<td>M</td>
<td>101</td>
<td>84</td>
<td>39</td>
<td>-0.98842</td>
<td>0.4920881</td>
</tr>
<tr>
<td>T53</td>
<td>M</td>
<td>102</td>
<td>85</td>
<td>38</td>
<td>-0.98532</td>
<td>-0.2361914</td>
</tr>
<tr>
<td>T54</td>
<td>M</td>
<td>103</td>
<td>81</td>
<td>37</td>
<td>-1.11528</td>
<td>-0.438547</td>
</tr>
<tr>
<td>T55</td>
<td>M</td>
<td>104</td>
<td>83</td>
<td>39</td>
<td>-0.95555</td>
<td>0.44687352</td>
</tr>
<tr>
<td>T56</td>
<td>M</td>
<td>106</td>
<td>83</td>
<td>39</td>
<td>-0.93287</td>
<td>0.2393841</td>
</tr>
<tr>
<td>T57</td>
<td>M</td>
<td>107</td>
<td>82</td>
<td>38</td>
<td>-0.98269</td>
<td>-0.096172</td>
</tr>
<tr>
<td>T58</td>
<td>M</td>
<td>112</td>
<td>89</td>
<td>40</td>
<td>-0.63393</td>
<td>-0.840209</td>
</tr>
<tr>
<td>T59</td>
<td>M</td>
<td>113</td>
<td>88</td>
<td>40</td>
<td>-0.64405</td>
<td>-0.956061</td>
</tr>
<tr>
<td>T60</td>
<td>M</td>
<td>114</td>
<td>86</td>
<td>40</td>
<td>-0.68078</td>
<td>-0.4047359</td>
</tr>
<tr>
<td>T61</td>
<td>M</td>
<td>116</td>
<td>90</td>
<td>43</td>
<td>-0.42133</td>
<td>0.10845233</td>
</tr>
<tr>
<td>T62</td>
<td>M</td>
<td>117</td>
<td>90</td>
<td>41</td>
<td>-0.48485</td>
<td>-0.9039457</td>
</tr>
<tr>
<td>T63</td>
<td>M</td>
<td>117</td>
<td>91</td>
<td>41</td>
<td>-0.45824</td>
<td>-1.0882131</td>
</tr>
<tr>
<td>T64</td>
<td>M</td>
<td>119</td>
<td>93</td>
<td>41</td>
<td>-0.37202</td>
<td>-1.610083</td>
</tr>
<tr>
<td>T65</td>
<td>M</td>
<td>120</td>
<td>89</td>
<td>40</td>
<td>-0.50198</td>
<td>-1.4175463</td>
</tr>
<tr>
<td>T66</td>
<td>M</td>
<td>120</td>
<td>93</td>
<td>44</td>
<td>-0.23552</td>
<td>-0.2831547</td>
</tr>
<tr>
<td>T67</td>
<td>M</td>
<td>121</td>
<td>95</td>
<td>42</td>
<td>-0.24581</td>
<td>-1.6640875</td>
</tr>
<tr>
<td>T68</td>
<td>M</td>
<td>125</td>
<td>93</td>
<td>45</td>
<td>-0.11305</td>
<td>-0.1986272</td>
</tr>
<tr>
<td>T69</td>
<td>M</td>
<td>127</td>
<td>96</td>
<td>45</td>
<td>-0.00223</td>
<td>-0.9407645</td>
</tr>
<tr>
<td>T70</td>
<td>M</td>
<td>128</td>
<td>95</td>
<td>45</td>
<td>-0.01035</td>
<td>-0.7971646</td>
</tr>
<tr>
<td>T71</td>
<td>M</td>
<td>131</td>
<td>95</td>
<td>46</td>
<td>0.079136</td>
<td>-0.559302</td>
</tr>
<tr>
<td>T72</td>
<td>M</td>
<td>135</td>
<td>106</td>
<td>47</td>
<td>0.477846</td>
<td>-2.4250481</td>
</tr>
</tbody>
</table>

Female turtles are larger and have more exaggerated height 😊

Credit: Alessandro Giuliani
Exercise

• Madrid and Warsaw are at almost the same distance to Latium cities

Are Madrid and Warsaw near each other?

Intuitive points

• PCA gives the axes that orthogonally account for variance in the data
• PCs correspond to explanations / factors giving rise to the variance
• Coefficient of a variable in a PC suggests how relevant that variable is for that PC
PCA IN BIOMARKER SELECTION
PCA in biomarker selection

When PCA is applied e.g. on gene expression data,

- PCs w/ large variance \( \approx \) diff expressed pathways
- Variables with large coefficients in a PC \( \approx \) key genes in the pathway associated with that PC

PCA can be a useful biomarker-selection approach

- E.g., biomarkers \( \approx \) genes w/ high loading
  - Loading of gene \( x = \sum_j | \alpha_{xj} \sigma_j^2 | \), where \( \alpha_{xj} \) is coefficient of \( x \) in PC\(_j\), and \( \sigma_j^2 \) is variance of PC\(_j\)
Example

- Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid > 50
- Diff subtypes respond differently to same Tx
- Over-intensive Tx
  - Development of secondary cancers
  - Reduction of IQ
- Under-intensive Tx
  - Relapse
- The subtypes look similar
- Can we diagnosis the subtypes based on gene expression profiling?
PCA in ALL subtype diagnosis

- **Steps:**
  - Identify genes with high variance
  - Perform PCA on them
  - Plot using PC1 to 3
Induction of hypothesis

• The PCs capture different biological pathways. The values of PCs capture different states of these pathways

• Hypothesis: If patient X has ALL subtype T, X’s biological pathways are in state $S_T$

… and abduction during diagnosis

• Observation: John’s biological pathways are in state $S_T$

• Abduction: John has ALL subtype T
BATCH EFFECTS
What are batch effects?

- Batch effects are unwanted sources of variation caused by different processing date, handling personnel, reagent lots, equipment/machines, etc.

- Batch effects is a big challenge faced in biological research, especially towards translational research and precision medicine.
Sometimes, a gene expression study may involve batches of data collected over a long period of time...

**PCA scatter plot**

- **Samples from diff batches are grouped together, regardless of subtypes and treatment response**

**Image credit: Difeng Dong’s PhD dissertation, 2011**
Paired boxplots of PCs

- It is easier to see which PC is enriched in batch effects by showing, side by side, the distribution of values of each PC stratified by class and suspected batch variables.
Normalization

- **Aim of normalization:**
  - Reduce variance w/o increasing bias

- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix’s

- **Transform data so that distribution of probe intensities is same on all arrays**
  - E.g., \( (x - \mu) / \sigma \)

- **Quantile normalization**

- **Gene fuzzy score, GFS**
Quantile Normalization

- Given \( n \) arrays of length \( p \), form \( X \) of size \( p \times n \) where each array is a column
- Sort each column of \( X \) to give \( X_{sort} \)
- Take means across rows of \( X_{sort} \) and assign this mean to each elem in the row to get \( X'_{sort} \)
- Get \( X_{normalized} \) by arranging each column of \( X'_{sort} \) to have same ordering as \( X \)

- Implemented in some microarray s/w, e.g., EXPANDER
In such a case, batch effect may be severe... to the extent that you can predict the batch that each sample comes!

After quantile normalization

⇒ Need normalization to correct for batch effect

Image credit: Difeng Dong’s PhD dissertation, 2011
Caution: It is difficult to eliminate batch effects effectively

Green and orange are normal samples differing in processing date

a: Before normalization

b: Post normalization

c: Checks on individual genes susceptible to batch effects

d: Clustering after normalization (samples still cluster by processing date)

Leek et al, Nature Reviews Genetics, 11:733-739, 2010
Caution: “Over normalized” signals in cancer samples

A gene normalized by quantile normalization (RMA) was detected as down-regulated DE gene, but the original probe intensities in cancer samples were not diff from those in normal samples.

A gene was detected as an up-regulated DE gene in the non-normalized data, but was not identified as a DE gene in the quantile-normalized data.

Simulated data

• Real one-class data from a multiplex experiment (no batches); n = 8
• Randomly assigned into two phenotype classes D and D*, 100x
• 20% biological features are assigned as differential, and a randomly selected effect size (20%, 50%, 80%, 100% and 200%) added to D*
• Half of D and D* are assigned to batch 1, and the other half assigned to batch 2. A randomly selected batch effect (20%, 50%, 80%, 100% and 200%) is added to all features in batch 1
Batch-effect correction can introduce false positives

- Precision is strongly affected by batch correction via COMBAT
- This means that false positives are added post-batch correction. Data integrity is affected
- Moreover, post-batch correction does not restore performance to where no batch is present
Exercise

• Why normalization methods like mean scaling, z-score, and quantile normalization sometimes do not work well?
PCA FOR ISOLATING BATCH EFFECTS
PCA for isolating batch effects

• When a batch effect is observed, it is common practice to apply a batch effect-removal or -correction method

• However, this does not necessarily work well in practice. Moreover, if the data does not fit the correction method’s assumptions, it may lead to false positives

• Instead, we may opt for a more direct strategy by simply removing PCs (usually PC1) enriched in batch effects, and deploying the remaining PCs as features for analysis
Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, *BMC Genomics* 18(Suppl2):142, 2017

- **Batch effects dominate in PC1**

Determine PCs associated with batch using paired boxplots of PCs.
Removal of batch effect-laden PCs removes most batch effects
Samples separately by class post PC1 removal, no batch subgrouping

A and B are different datasets with different batch effects inserted

Batch effects dominate

Class-effect discrimination recovered

(Notation: A/B_D/D*_1/2 refers to the dataset, class and batches respectively)
In short, PC manipulation is helpful for dealing w/ batch effects.
Exercise

• Suggest a modification to the formula below to avoid selecting genes laden with batch effects

PCA can be a useful biomarker-selection approach

• E.g., biomarkers \( \approx \) genes w/ high loading

  – Loading of gene \( x = \sum_j | \alpha_{xj} * \sigma_j^2 | \), where \( \alpha_{xj} \) is coefficient of \( x \) in \( \text{PC}_j \), and \( \sigma_j^2 \) is variance of \( \text{PC}_j \)
Batch Effect-Resistant Feature Selection
What if class and batch effects are strongly confounded?

- Neither batch-effect correction nor PCA work well
- We also do not want to inadvertently lose information on disease subpopulations (which look like batch effects but are meaningful)

⇒ Consider using protein complexes / subnetworks of biological pathways as biomarkers / context for biomarker selection
FSNET

• FSNET --- a protein complex-based feature-selection methods. Use expression rank-based weighting method (viz. GFS) on individual proteins, followed by intra-class-proportion weighting

And for comparison …

• SP is the protein-based two-sample t-test

Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, *BMC Genomics*, 18(Suppl 2):142, 2017
FSNET

• $\beta(g, C)$
  - Proportion of tissues in class C that have protein $g$ among their most-abundant proteins

• Score($S, p, C$)
  - Score of protein complex $S$ and tissue $p$ weighted based on class $C$

• $f_{SNET}(S, X, Y, C)$
  - Complex $S$ is differentially high in sample set $X$ and low in sample set $Y$, weighted based on class $C$, when $f_{SNET}(S, X, Y, C)$ is at largest extreme of t-distribution

\[
\beta(g_i, C_j) = \sum_{p_k \in C_j} \frac{f_S(g_i, p_k)}{|C_j|}
\]

\[
\text{score}(S, p_k, C_j) = \sum_{g_i \in S} f_S(g_i, p_k) \times \beta(g_i, C_j)
\]

\[
f_{SNET}(S, X, Y, C_j) = \frac{\text{mean}(S, X, C_j) - \text{mean}(S, Y, C_j)}{\sqrt{\text{var}(S, X, C_j)/|X| + \text{var}(S, Y, C_j)/|Y|}}
\]
Network-based methods are enriched for class-related variation (Real data)

- **PCA on SP-selected genes:** Class & batch effects are confounded; cf. PC2

- **PCA on FSNET-selected complexes:** Class & batch effects are less confounded in top PCs
Top complex-based features are strongly associated with class, not batch.

• FSNET captures class effects while being robust against batch effects. In contrast, both class and batch variability are present in the top variables selected by SP.
CONCLUDING REMARKS
What have we learned?

- PCA is a useful paradigm for biomarker selection
- PCA is not just a visualization tool; it can also be used for dealing with batch effects
- When class & batch effects are deeply confounded at the level of proteins / genes, it is might be better to analyze at the level of protein complexes / pathway subnetworks
References

• [Batch effects] Leek et al., *Nature Reviews Genetics*, 11:733-739, 2010