Principal Component Analysis (PCA)

Data Science 101 Team
PCA I: putting a line through the data

- Scatterplots with lines through the data are one of the most common plots we see.
- When you studied regression, you did study one way of putting lines through the data. We are going to look at a different approach, which treats the different variables symmetrically.
- We will start with two dimensional datasets, but we are really interested in an tool that can be applied to summarize the relation between many variables.
An historical dataset

- Galton in 1885 compiles a table with the height of 928 adult children, classified by height of “midparents”
- All female heights were re-scaled by multiplying by 1.08, and midparent heights were computed averaging the heights of father and rescaled mother.
- It is working with this dataset that Galton develops the concepts of regression and correlation.
### Galton data

<table>
<thead>
<tr>
<th>Height of the midparent in inches</th>
<th>&lt;61.7</th>
<th>62.2</th>
<th>63.2</th>
<th>64.2</th>
<th>65.2</th>
<th>66.2</th>
<th>67.2</th>
<th>68.2</th>
<th>69.2</th>
<th>70.2</th>
<th>71.2</th>
<th>72.2</th>
<th>73.2</th>
<th>&gt;73.7</th>
<th>Total no. of adult children</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;73.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>72.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>71.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>43</td>
</tr>
<tr>
<td>70.5</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td>18</td>
<td>14</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>68</td>
</tr>
<tr>
<td>69.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>183</td>
</tr>
<tr>
<td>68.5</td>
<td>1</td>
<td></td>
<td>7</td>
<td>11</td>
<td>16</td>
<td>25</td>
<td>31</td>
<td>34</td>
<td>48</td>
<td>21</td>
<td>18</td>
<td>4</td>
<td>3</td>
<td></td>
<td>219</td>
</tr>
<tr>
<td>67.5</td>
<td></td>
<td></td>
<td>3</td>
<td>5</td>
<td>14</td>
<td>15</td>
<td>36</td>
<td>38</td>
<td>28</td>
<td>38</td>
<td>19</td>
<td>11</td>
<td>4</td>
<td></td>
<td>211</td>
</tr>
<tr>
<td>66.5</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>17</td>
<td>17</td>
<td>14</td>
<td>13</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>78</td>
</tr>
<tr>
<td>65.5</td>
<td>1</td>
<td></td>
<td>9</td>
<td>5</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>66</td>
</tr>
<tr>
<td>64.5</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>&lt;64.0</td>
<td>1</td>
<td></td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Totals</td>
<td>5</td>
<td>7</td>
<td>32</td>
<td>59</td>
<td>48</td>
<td>117</td>
<td>138</td>
<td>120</td>
<td>167</td>
<td>99</td>
<td>64</td>
<td>41</td>
<td>17</td>
<td>14</td>
<td>928</td>
</tr>
</tbody>
</table>
Galton data - available in R in the “psych” package

```r
summary(galton)
```

<table>
<thead>
<tr>
<th></th>
<th>child</th>
<th>parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>61.70</td>
<td>64.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>66.20</td>
<td>67.50</td>
</tr>
<tr>
<td>Median</td>
<td>68.20</td>
<td>68.50</td>
</tr>
<tr>
<td>Mean</td>
<td>68.09</td>
<td>68.31</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>70.20</td>
<td>69.50</td>
</tr>
<tr>
<td>Max.</td>
<td>73.70</td>
<td>73.00</td>
</tr>
</tbody>
</table>
Galton data

Using ‘sunflowerplot’

Child Height
Mid−parent Height
Means, conditional and non

To predict a child’s height given a parent’s height, look at the average height of children that have parents of the same height.
Subtracting the means and using regression lines

Centered Galton's data

Child Height

Mid-parent Height
Since this is still a new type of dataset, one of the goals of the paper should be to document the issues that come up when analyzing the resequencing data. So, I think that some observations on how different approaches perform (with a somewhat redundant analysis) would be useful.

We selected regions with proven association to at least one phenotype. So, while it is important to control for multiple comparisons (because we look at multiple phenotypes etc), there is room for putting the question of proving significant association aside and simply studying what is the best possible model for the phenotypes in the relevant regions.

The current analysis is based on two parallel tracks that do not merge: we analyze separately common (MAF $> 0.01$) and rare (MAF $< 0.01$) variants; we used different tests and control for multiple comparison within these two sets of analysis. We never consider models where rare and common variants are jointly considered (at least this is my understanding of how the burden tests are being carried out, but I may be wrong).

The analysis of single variant association conditional on GWAS SNP is important. But it turns out that sometimes it is interesting to condition on other variants and it is a bit ad hoc to decide when and which. Also, criteria for significance have not been established for this specific analysis.
Putting a line through the data

- We can recognize the regression line of mid-parents’ heights on children’s
- And the regression line of children’s heights on parents’
- But there are two additional lines that appear, and they are marked as major and minor axis of the ellipse: what are these?
- To understand more about this, we start working with another dataset, which is “contemporary” to Galton’s and on the same topic, but simpler in that it deals with men only, and it is not discretized
Pearson’s data

Pearson's height data

Father

Son
Pearson’s data

We subtract means from fathers’ and sons’ heights and plot the linear regression lines
Minimal sum of squared errors

Reg. Line of Son on Father

Reg. Line of Father on Son

Reg. Line of Son on Father

Reg. Line of Father on Son

Vertical Errors

Horizontal Errors
Distances of a point to a line

Euclidean distances: symmetric with respect to $x$ and $y$. 

![Graph showing distances of points to a line between fathers' and sons' heights.](image-url)
Finding the line that minimizes SS of Euclidean distances

- One example where we create two variables

\[ x \sim \mathcal{N}(0, 1) \]

\[ y = 2x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \]

- We are going to evaluate the sum of squared euclidean distances between these data points and 8 lines, all with intercept 0 and slopes

\[ -0.5, -1, -2, -4, 4, 2, 1, 0.5 \]
Finding the line that minimizes SS of Euclidean distances

- **slope = -0.5**
  - Sum of SE = 26.51
  - Variable 1
  - Variable 2

- **slope = -1**
  - Sum of SE = 56.2
  - Variable 1
  - Variable 2

- **slope = -2**
  - Sum of SE = 62.65
  - Variable 1
  - Variable 2

- **slope = -4**
  - Sum of SE = 49.13
  - Variable 1
  - Variable 2

- **slope = 4**
  - Sum of SE = 10.49
  - Variable 1
  - Variable 2

- **slope = 2**
  - Sum of SE = 6.99
  - Variable 1
  - Variable 2

- **slope = 1**
  - Sum of SE = 13.22
  - Variable 1
  - Variable 2

- **slope = 0.5**
  - Sum of SE = 13.41
  - Variable 1
  - Variable 2
Finding the line that minimizes SS of Euclidean distances

- So, in our example, the line with minimal \( \text{SSEuclidean Errors} \) was the one that reflected the “true” relation between the two variables.
- Note that so far, we have not actually figured out how to calculate this line (see later and next lecture).
- In the following examples, however, I am going to use the rule that we will discover to

1. Compare to Galton’s results (and looking back at Pearson’s data)
2. Understand some of the properties of this line
Going back to Galton’s data

Galton’s data (centered)
Going back to Galton’s data

The line we found is called Principal Component
Going back to Pearson’s data

Pearson's height data (centered)

Father
Son
Pearson’s work

On lines and planes of closest fit to systems of points in space (1901)
An instructive example

\[ z \sim \mathcal{N}(0, 1) \]

\[ x = z + \epsilon \]

\[ y = z + \eta \]

- There is one underlying variable \( z \) that contributes both to \( x \) and to \( y \).
- There is a noise component \( \epsilon \) and \( \eta \) both in \( x \) and \( y \).
- Let \( \sigma \) be the standard-deviation of \( \epsilon \) and \( \eta \). We consider 8 values for \( \sigma = (1, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1) \).
- We are going to generate 8 datasets following this rule, plot the data, and the two regression lines (red and blue) as well as the line that minimizes the sum of square euclidean distance from the points (green).
An instructive example

Sigma = 1, Slope OR = 1.08
Sigma = 0.7, Slope OR = 1.03
Sigma = 0.6, Slope OR = 0.95
Sigma = 0.5, Slope OR = 1.09

Sigma = 0.4, Slope OR = 1.03
Sigma = 0.3, Slope OR = 1.02
Sigma = 0.2, Slope OR = 1.01
Sigma = 0.1, Slope OR = 1.02
An instructive example

- The slopes of the regression lines change with the change in standard deviation, but the slope of the “Orthogonal Regression” line not so much.
- As the variance of $\epsilon$ and $\eta$ decreases, the difference between the two regression lines diminishes, and they both become closer to the green line.
- Let’s try again and see what happens if we do the same exercise as before, but this time

\[
z \sim \mathcal{N}(0, 1)
\]
\[
x = z + \epsilon
\]
\[
y = z/2 + \eta
\]
An instructive example, II
How do we calculate the line that minimizes SS of Euclidean distances?

» let’s look back at our trial and error attempt and focus on the blue projections, rather than red distances.
An application of Pythagoras’ theorem
An application of Pythagoras’ theorem

- The square of the distance of a point to a line going through the origin + the square of the distance of the projection of the point on this line to the origin = square of the distance of a point to the origin
- Finding the line with minimal sum of squared euclidean distances from the points, is equivalent to finding the line on which the points projections have maximal variance
- This last goal is very interesting and we will explore it further next time