Canonical Correlation Analysis (CCA)

Foundations of Data Analysis

March 27, 2020

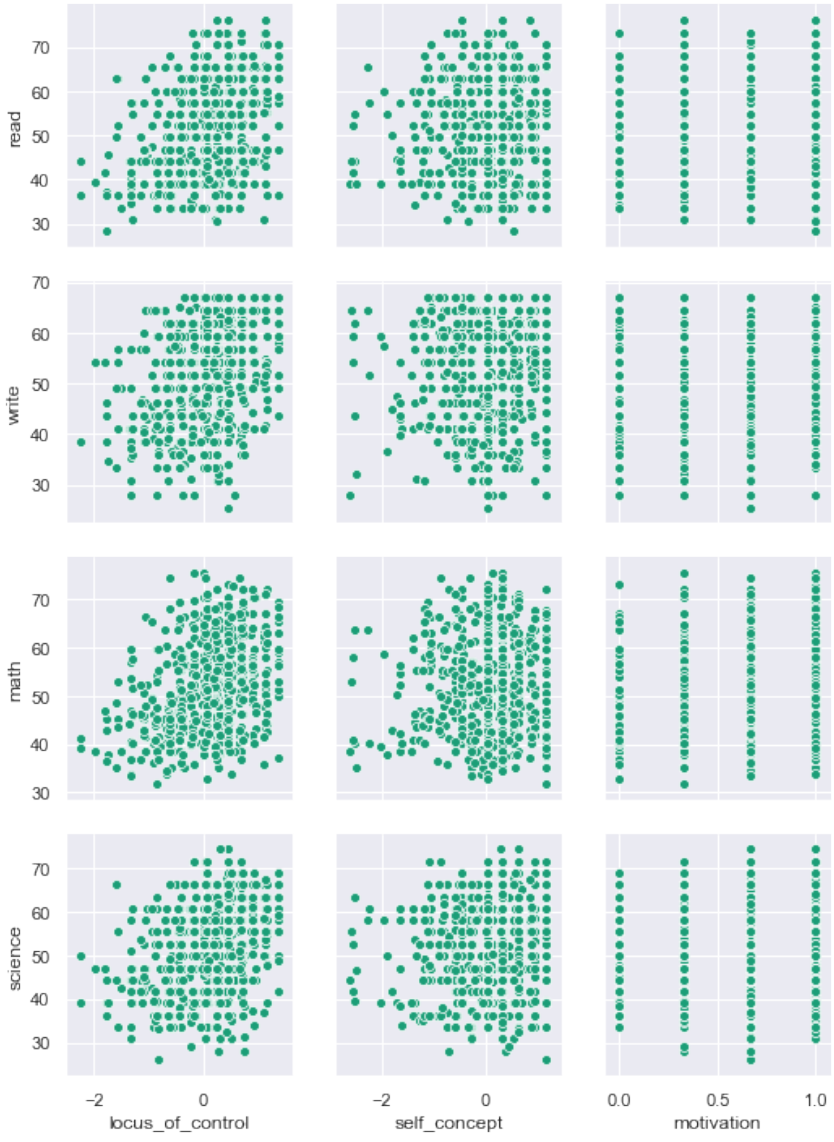
Example: Psych Measures vs. Test Scores

	locus_of_control	self_concept	motivation	read	write	math	science	female
0	-0.84	-0.24	1.00	54.8	64.5	44.5	52.6	1
1	-0.38	-0.47	0.67	62.7	43.7	44.7	52.6	1
2	0.89	0.59	0.67	60.6	56.7	70.5	58.0	0
3	0.71	0.28	0.67	62.7	56.7	54.7	58.0	0
4	-0.64	0.03	1.00	41.6	46.3	38.4	36.3	1

Question: Are these psychological measures related to scores on standardized tests?

(see `CCA.ipynb` for code details)

Pairwise Relationships



Pairwise Correlations:

0.3735	0.0606	0.2106
0.3588	0.0194	0.2542
0.3372	0.0535	0.1950
0.3246	0.0698	0.1156

Canonical Correlation Analysis (CCA)

Group your data table into two sets of variables:

$$X : n \times d_X \qquad Y : n \times d_Y$$

Find a single dimension in X and single dimension in Y that are maximally correlated

Math of CCA

Unit vector in X data: $u \in \mathbb{R}^{d_X}, \|u\| = 1$
Unit vector in Y data: $v \in \mathbb{R}^{d_Y}, \|v\| = 1$

Projected data:

$$\langle x_i - \mu_X, u \rangle, \qquad \langle y_i - \mu_Y, u \rangle$$

or, using centered data matrices, $\tilde{X}, \tilde{Y}$:

$$\tilde{X}u, \qquad \tilde{Y}v$$

Maximize correlation in projected data:

$$(u', v') = \arg\max_{u,v} \operatorname{corr}(\tilde{X}u, \tilde{Y}v)$$

Math of CCA

Let $\Sigma_{XX} = \text{cov}(X, X), \Sigma_{YY} = \text{cov}(Y, Y),$
$\Sigma_{XY} = \text{cov}(X, Y)$

Goal: find u, v that maximize the correlation:

$$\rho = \frac{u^T \Sigma_{XY} v}{\sqrt{u^T \Sigma_{XX} u} \sqrt{v^T \Sigma_{YY} v}}$$

CCA Solution

u is the eigenvector with largest eigenvalue of

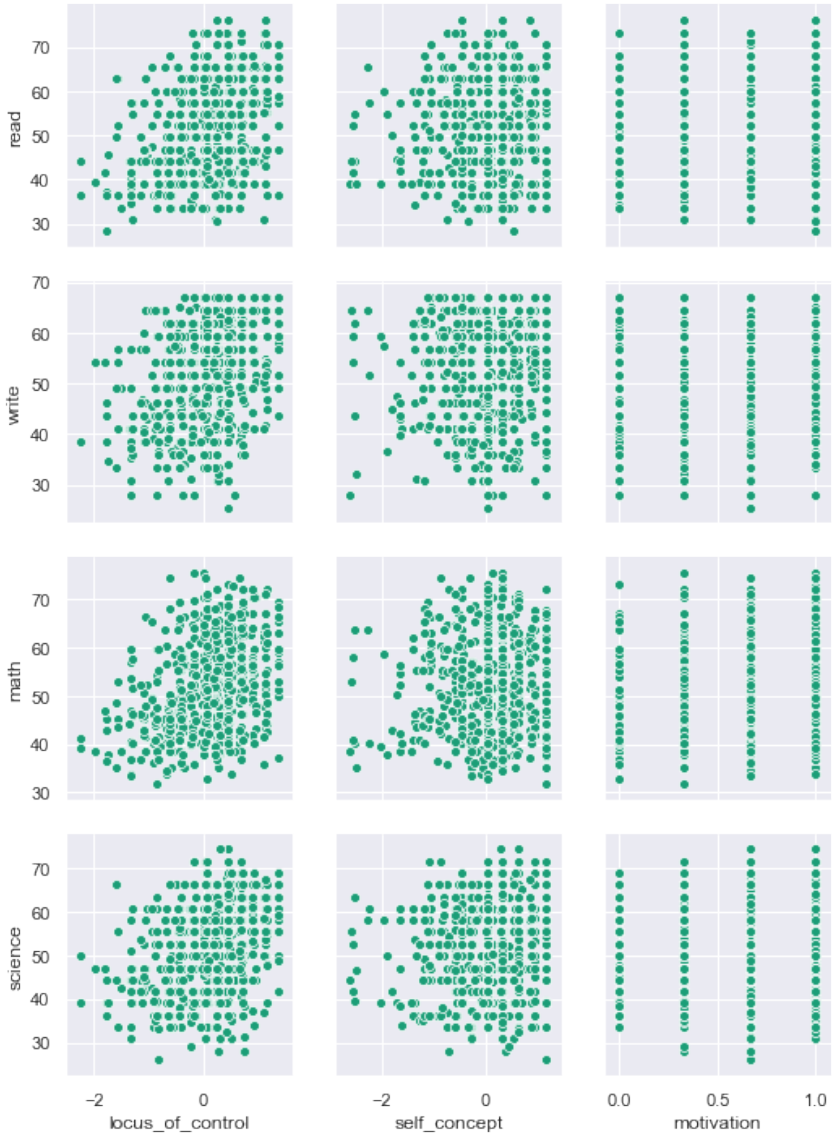
$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$$

v is the eigenvector with largest eigenvalue of

$$\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

Just like PCA, we can then proceed to find the dimensions with the **second most** correlation, etc.

Example: Psych vs. Test



Pairwise Correlations:

0.3735	0.0606	0.2106
0.3588	0.0194	0.2542
0.3372	0.0535	0.1950
0.3246	0.0698	0.1156

Example: Psych vs. Test

```
Psych Canonical Components:
                    0         1
Control      0.876809 -0.429422
Self        -0.174754 -0.496948
Motivation   0.447959  0.755662

Academic Canonical Components:
                    0         1
Read         0.617204  0.012375
Write        0.743148  0.676109
Math         0.253335  0.021393
Science     -0.051115 -0.926878

First Canonical Correlation = 0.4464364824283061
Second Canonical Correlation = 0.15335902492287964
```

Example: Psych vs. Test