1 PCA & CCA

Principal components analysis (PCA) considers interrelationships within a set of variables.

- The original purpose of PCA is to reduce a large number (p) of variables to a much smaller number (m) of PCs while retaining as much as possible of the variation in the p original variables.
- This technique is especially useful if $m \ll p$ and if the m PCs can be readily interpreted.
- If m is very much smaller than p, then the reduction of dimensionality alone may justify the use of PCA, even if the PCs have no clear meaning.
- The results of a PCA are much more satisfying if intuitively reasonable interpretations can be given to some or all of the m retained PCs.
  - Although in many examples the PCs can be readily interpreted, this is by no means universally true.
  - And some interpretations owe a lot to the analyst's imagination.
  - Careful thought should go into any interpretation.
  - In some cases, transformation of variables before analysis may improve the chances of a simple interpretation.

Canonical correlation analysis (CCA) considers relationships between two sets of variables.

- For example, in psychology, an investigator may measure a number of aptitude variables and a number of achievement variables on a sample of students and wish to say something about the relationship between "aptitude" and "achievement".

2 A example of PCs as a small number of interpretable variables

- One type of application where PCA has been found useful is identification of the most important sources of variation in anatomical measurements for various species.
- Typically, a large number of measurements are made on individuals of a species, and a PCA is done.
- The first PC almost always has positive coefficients for all variables and simply reflects overall "size" of the individuals.
- Later PCs usually contrast some of the measurements with others, and can often be interpreted as defining certain aspects of "shape" that are important for the species.
- A small data set with seven measurements.
  - A class of students, 15 women and 13 men.
  - The results of the PCA is done separately for women and men.
  - The seven measurements: lengths of hand, wrist, height, forearm, head, chest, waist.

- The PCA was done on the correlation matrix.
- When variables are on very different scales or have very different variances, a principal components analysis of the data should not be performed on the covariance matrix, but on the correlation matrix.
- The covariance matrix gives greater weight to larger, and hence more variable, measurements, such as height and chest girth, and less weight to smaller measurements such as wrist girth and hand length.

Table 4.1. First three PCs: student anatomical measurements.

Component number		1	2	3
			Women	
Hand	Coefficients	0.33	0.56	0.03
Wrist		0.26	0.62	0.11
Height		0.40	−0.44	−0.00
Forearm		0.41	−0.05	−0.55
Head		0.27	−0.19	0.80
Chest		0.45	−0.26	−0.12
Waist		0.47	0.03	−0.03
Eigenvalue		3.72	1.37	0.97
Cumulative percentage of total variation		53.2	72.7	86.5
			Men	
Hand	Coefficients	0.23	0.62	0.64
Wrist		0.29	0.53	−0.42
Height		0.43	−0.20	0.04
Forearm		0.33	−0.53	0.38
Head		0.41	−0.09	−0.51
Chest		0.44	0.08	−0.01
Waist		0.46	−0.07	0.09
Eigenvalue		4.17	1.26	0.66
Cumulative percentage of total variation		59.6	77.6	87.0

Table 4.2. Simplified version of the coefficients in Table 4.1.

Component number	1	2	3
		Women	
Hand	+	+	
Wrist	+	+	
Height	+	−	
Forearm	+		−
Head	+	(−)	+
Chest	+	(−)	
Waist	+		
		Men	
Hand	+	+	+
Wrist	+	+	−
Height	+	(−)	
Forearm	+	−	+
Head	+		−
Chest	+		
Waist	+		

- $Var(Y_1) = \lambda_1, Var(Y_2) = \lambda_2, \ldots, Var(Y_p) = \lambda_p$.
- The jth PC accounts for a proportion P_j of the total variation of the original data, where $P_j = \frac{\lambda_j}{tr(\Sigma)}$.
- The first m principal components, where $m < p$ account for a proportion $P^{(m)}$ of the total variation of the original data, where $P^{(m)} = \sum_{j=1}^{m} P_j = \frac{\sum_{j=1}^{m} \lambda_j}{tr(\Sigma)}$.
- The first PC accounts for 53% (women) or 60% (men) of the total variation.
- The second PC accounts for slightly less than 20% of the total variation for both sexes.
- The third PC contributes $9\% - 14\%$ of total variation.
- Overall, the first three PCs account for nearly 87% of total variation.

- How many components are needed to provide an adequate summary of a given data set? Retain just enough components to explain some specified large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested.
- These three PCs are large enough.
- When we interpret PCs, as with other types of tabular data, it is usually only the general pattern of the coefficients that is really of interest, not values.
- + or − indicates a coefficient whose absolute value is greater than half the maximum coefficient (absolute value) for the relevant PC.
- (+) or (−) indicates a coefficient whose absolute value is between a quarter and a half of the largest absolute value of the PC.
- This kind of simple representation is often helpful in interpreting PCs, particularly if a PCA is done on a large number of variables.

Component number	1	2	3
		Women	
Hand	+	+	
Wrist	+	+	
Height	+	−	
Forearm	+		−
Head	+	(−)	+
Chest	+	(−)	
Waist	+		
		Men	
Hand	+	+	+
Wrist	+	+	−
Height	+	(−)	
Forearm	+	−	+
Head	+		−
Chest	+		
Waist	+		

- All the coefficients of the first PC and the seven variables are positive. It clearly measures overall "size" for both women and men.
- The second PC contrasts hand and wrist measurements with height, implying that, after overall size has been accounted for, the main source of variation is between individuals with large hand and wrist measurements relative to their heights, and individuals with the converse relationship.
- For women, head and chest measurements also have some contribution to this PC.
- For men, the forearm measurement, which is closely related to height, partially replaces height in this PC.
- The third PC differ more between the sexes but retain some similarity.
- For women, it is almost a contrast between head and forearm measurements.
- For men, in addition, hand and wrist measurements appear with the same signs as forearm and head.
- There is another experiments containing 3000 people which have very similar results with this small experiments.
- The PCs describe the major directions of variation within a sample, regardless of the sample size.

3 Example of CCA

- An example with 15 variables measured on 272 sand and mud samples taken from various locations in one Bay of England. (Jeffers)
- Environment: Eight variables are chemical or physical properties of the sand or mud samples.
- Species: Seven variables measure the abundance of seven groups of invertebrate species.

Table 9.3. Coefficients for the first two canonical variates in a canonical correlation analysis of species and environmental variables.

		First canonical variates	Second canonical variates
Environment variables	x_1	0.03	0.17
	x_2	0.51	0.52
	x_3	0.56	0.49
	x_4	0.37	0.67
	x_5	0.01	−0.08
	x_6	0.03	0.07
	x_7	−0.00	0.04
	x_8	0.53	−0.02
Species variables	x_9	0.97	−0.19
	x_{10}	−0.06	−0.25
	x_{11}	0.01	−0.28
	x_{12}	0.14	0.58
	x_{13}	0.19	0.00
	x_{14}	0.06	0.46
	x_{15}	0.01	0.53
Canonical correlation		0.559	0.334

- The first canonical variate for species is dominated by a single species.
- The first canonical variate for environmental variables involves non-trivial coefficients for four of the variables.
- The second pair of canonical variates has fairly large coefficients for environmental variables and three species.
- The canonical correlation is 0.559 and 0.334.

Combine CCA and PCA.

- Jeffers also looks at PCs for the environmental and species variables separately, and concludes that four and five PCs, respectively. (Which have account for most of the variation in each group.)

- He continue to look at the between-group correlations for each set of retained PCs. That is, do a CCA only based on the retained PCs.
- The canonical correlation is 0.420 and 0.258, compared with 0.559 and 0.334 when all the variables are used.
- The first two canonical variates for the environmental variables and the first canonical variate for the species variables are each dominated by a single PC.
- The second canonical variate for the species variables has two non-trivial coefficients.
- Thus, the canonical variates for PCs look easier to interpret than those based on the original variables.
- Note that, even if only one PC occurs in a canonical variate, the PC itself is not necessarily an easily interpreted entity.
- Thus, the between-group relationships found by CCA of the retained PCs are different in some examples from found from CCA on the original variables.