

Chapter 2

Dimension reduction techniques - 6

2.1

2.2

2.3

2.4 Canonical correlation

As we mentioned briefly in Chapter 1, the CCA seeks to identify and quantify the associations between two sets of variables. Examples include relating Government policy variables with economic goal variables and associating college performances of a group of students with their pre-college performances etc. Here, we focus on the correlation between a linear combination of one set of variables to a linear combination in another set. We first seek to maximize this correlation. Then, we seek to find a pair of linear combinations uncorrelated with the initial pair and having the largest correlation among those, and so on. This is analogous to what we do in PCA in spirit, but the main focus is on a correlation. They will be called the canonical variables and the successive correlations as canonical correlations. The technique attempts to concentrate on a high-dimensional relationship between two sets of variables into a few pairs of canonical variables.

As before, we have $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ a $(p+q)$ dimensional random vector with two sets of p and q variables. We assume $p \leq q$ throughout. We assume $\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ has mean zero and its variance is partitioned as $\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Clearly, Σ_{12} , the $p \times q$ dimensional matrix measures the association between the two sets. The main task of CCA is to summarize the association between \mathbf{x} and \mathbf{y} using a few linear combinations and their correlations.

Let $U = a^t \mathbf{x}$ and $V = b^t \mathbf{y}$ be a typical pair of linear combinations of \mathbf{x} and \mathbf{y} . Then,

$$\text{Corr}(U, V) = \frac{a^t \Sigma_{12} b}{\sqrt{a^t \Sigma_{11} a} \sqrt{b^t \Sigma_{22} b}}$$

The first pair of canonical variables (U_1, V_1) will be the pair with unit variance and maximum correlation. The second pair will be the pair (U_2, V_2) with unit variance and maximum correlation among all choices uncorrelated with (U_1, V_1) and so on. The k th pair will be uncorrelated with the first $k-1$ pairs.

Result : Define $\Psi_1 = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}$ and $\Psi_2 = \Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1/2}$. Then, they have the same set of positive eigenvalues. We'll call them $\tilde{\rho}_1^2 \geq \dots \geq \tilde{\rho}_p^2 \geq 0$. Let (e_1, \dots, e_p) and (f_1, \dots, f_p) be the corresponding eigenvectors of Ψ_1 and Ψ_2 . Namely, $\Psi_1 = \sum_{i=1}^p \tilde{\rho}_i^2 e_i e_i^t$ and $\Psi_2 = \sum_{i=1}^p \tilde{\rho}_i^2 f_i f_i^t$. Then, $U_i = e_i^t \Sigma_{11}^{-1/2} \mathbf{x}$ and $V_i = f_i^t \Sigma_{22}^{-1/2} \mathbf{y}$ are the canonical variables and $\text{Corr}(U_i, V_i) = \tilde{\rho}_i$. e_i -s and f_i -s are related by the fact that each f_i is proportional to $\Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1/2} e_i$.

If we standardize the variables \mathbf{x} and \mathbf{y} beforehand, they will also give the same canonical correlations and the same canonical variables. In fact, the variables $\Sigma_{11}^{-1/2} \mathbf{x}$ and $\Sigma_{22}^{-1/2} \mathbf{y}$ will also yield the same set. So, it is clear that the correlations within the two groups do not effect the canonical correlation. Any non-singular linear transformations of \mathbf{x} and \mathbf{y} will share the same property. That is, the CCA is invariant under any rotation, reflection or amplification of the original variables within the two sets.

Interpretation : The interpretations of canonical variables are primarily in terms of the standardized variables. These variables can be identified in terms of the original variables. The correlation between the canonical variables and the original variables can be used to determine which variable contributes most to the relationship between the two sets. Moreover, CCA generalizes other correlation methods such as the two-variable product-moment correlation and multiple correlation. The quantity $\tilde{\rho}_k^2$ gives us the proportion of variance of the canonical variable U_k explained by the set \mathbf{y} and called the k th shared variance between the two sets. The largest value $\tilde{\rho}_1^2$ is a measure of the set overlap.

In contrary to PCA, there is no certainty that the first few canonical variables may capture the variability of the individual sets of variables. It may happen as well, that by leaving out the smaller eigenvalues and associated canonical variables, we actually lose a lot of features of the data. Unfortunately, that is an artifact of the CCA.

Sample canonical correlations : As before, we have a sample of n realizations on $p + q$ variables $\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix}$ and construct the sample covariance matrix $\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$. The linear combinations with maximum correlations from those two sets of realizations can be computed using the similar eigenvalue technique. However, there is no certainty that the canonical variables will represent the individual variables in terms of variability and special features and the inference should be restricted to the measure of association between the two sets of variables only.

2.5 Correspondence Analysis

The Correspondence Analysis is a graphical procedure for representing associations in a table of frequencies or counts (contingency table). From such a table of I rows and J columns, the plot of CA shows $I + J$ points correspond-

ing to the rows and columns, in possibly different symbols in a 2-dimension plot. The measure of such a representation is measured by a quantity called inertia, which gives an idea of the information in each dimension.

In the plot, if two row points are close together, that indicates that they have a similar profile (conditional distribution) across the columns. The same holds if some column points are seen close. If two row points or two column points are far, that means their profiles are significantly different. Finally, if a row point and a column point are close together, that can be interpreted as their combination occurring more frequently than expected from an independence model. In case the rows and the columns are nearly independent, all the row points will be close together (since their profile will be similar) and the column points will be close as well, both close to the origin.

Example 1 : In a study of archaeological study in Southwest USA, $J = 4$ different types of pottery are found across $I = 7$ sites. The contingency table represents frequencies of different types of potteries found in the different sites. The site profiles are then constructed across the pottery types, and the pottery type profiles are constructed across the sites. The profiles indicate that some sites have similar pottery profiles, and some of them are different. The correspondence analysis is carried out to find out such notions, in particular some pottery types are more frequent in specific sites and so on.

Example 2 : We have a botanical problem known as gradient analysis. This concerns the quantification of the notion that certain species of flora prefer certain types of habitat, and that their presence in a particular location can be taken as an indicator of the local conditions. Thus one species of grass may prefer wet conditions, while another might prefer dry conditions. This will involve counting the number of each species in those particular locations and then use the contingency table for further analysis.

2.5.1 Formulation

Let $\mathbf{X} = ((x_{ij}))$ denote the $I \times J$ frequency table. We also assume that \mathbf{X} has full column rank J and $I \geq J$. The rows and columns of \mathbf{X} correspond to two different categorical variables, and the entries x_{ij} denote the relative frequencies of the (i, j) th cell. Let $n = \sum_{i,j} x_{ij}$ be the total count. Denote $p_{ij} = \frac{x_{ij}}{n}$, the relative frequencies, and P the corresponding $I \times J$ matrix. Define the vectors of row totals and column totals as r and c , such that $r_i = \sum_j p_{ij}$ and $c_j = \sum_i p_{ij}$, and the corresponding diagonal matrices

$$D_r = \text{diag}(r), \quad D_c = \text{diag}(c)$$

Correspondence Analysis is formulated as a weighted least square problem, to find $\tilde{P} = ((\tilde{p}_{ij}))$ such that $\text{rank}(\tilde{P}) = q$ (a small specified number), which minimizes

$$\sum_i \sum_j \frac{(p_{ij} - \tilde{p}_{ij})^2}{r_i c_j} = \text{tr}(B - \tilde{B})(B - \tilde{B})^t$$

such that $B = D_r^{-1/2} P D_c^{-1/2}$ and $\tilde{B} = D_r^{-1/2} \tilde{P} D_c^{-1/2}$. To solve this problem, we need a singular value decomposition of B .

Singular value decomposition : Any matrix $B_{I \times J}$ with rank s can be written as $B = \sum_{k=1}^s \lambda_k u_k v_k^t$ such that $\{u_i\}$ and $\{v_i\}$'s are orthogonal vectors of dimension I and J respectively. It follows that $B v_i = \lambda_i u_i$ and $u_i^t B = \lambda_i v_i^t$.

Writing $B = \sum_{k=1}^s \lambda_k u_k v_k^t$, the quantity $\text{tr}(B - \tilde{B})(B - \tilde{B})^t$ is minimized by $\tilde{B} = \sum_{k=1}^q \lambda_k u_k v_k^t$, and consequently, $\tilde{P} = D_r^{1/2} \tilde{B} D_c^{1/2}$. For $q = 1$, the approximation is given by $\tilde{P} = r c^t$ (Check !!). For a 2-dimension representation, we need $q = 3$, since the first singular value only gives the centering.

Define $F^i = \lambda_i D_r^{-1/2} u_i$, $G^i = \lambda_i D_c^{-1/2} v_i$ for $i = 1, 2$. The 2-dimensional plot of correspondence is then obtained from the co-ordinates of (F^2, F^3) and (G^2, G^3) respectively. For example, the first row is represented by the point (F_1^2, F_1^3) , the second row by (F_2^2, F_2^3) and so on. the first column is shown as the point (G_1^2, G_1^3) . This 2 dimensional plot gives an idea about the correspondence between rows, between columns and between row-column etc.

Total inertia is a measure of the variation in the count data and measures the deviation of the table from the independence model. It is given by

$$\sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = \sum_{k=2}^s \lambda_k^2$$

the inertia associated with the best reduced rank 3 approximation is then given by $\lambda_2^2 + \lambda_3^2$ and can be expressed as a percentage of the total inertia.