Canonical Correlation Analysis

CCA is used to reduce the feature dimensions of two correlated data sets. Let \( X \) and \( Z \) be two feature sets of a sample, e.g., \((x_i, z_i) \in (X, Z)\) is the profile of student \( i \), where \( x_i \in \mathbb{R}^{p_1} \) contains academic features (major, GPA, etc) and \( z_i \in \mathbb{R}^{p_2} \) contains personal features (age, gender, etc). CCA finds a projection vector \( w \in \mathbb{R}^{p_1} \) for \( X \) and a projection vector \( u \in \mathbb{R}^{p_2} \) for \( Z \) so that the correlation between projected features \( w^T x \) and \( u^T z \) is maximized. The correlation is

\[
\text{corr}(w^T x, u^T z) := \frac{\text{cov}(w^T x, u^T z)}{\sqrt{\text{var}(w^T x)} \sqrt{\text{var}(u^T z)}} = \frac{w^T \text{cov}(x, z) u}{\sqrt{w^T \text{var}(x) w} \sqrt{u^T \text{var}(z) u}}.
\]

It can be maximized by solving an equivalent optimization problem

\[
\begin{aligned}
\max_{w, u} & \quad w^T \text{cov}(x, z) u \\
\text{s.t.} & \quad w^T \text{var}(x) w = u^T \text{var}(z) u = 1.
\end{aligned}
\]

Applying the Lagrange multiplier, we have the Lagrange function

\[
J(w, u) = w^T \text{cov}(x, z) u - \lambda_1 (w^T \text{var}(x) w - 1) - \lambda_2 (u^T \text{var}(z) u - 1).
\]

Since \( x, z \) are vectors, their covariance and variance are matrices. Let \( \Sigma_{x,z} = \text{cov}(x, z) \in \mathbb{R}^{p_1 \times p_2} \) be their covariance matrix, \( \Sigma_x = \text{var}(x) \in \mathbb{R}^{p_1 \times p_1} \) be the variance matrix of \( x \) and \( \Sigma_z = \text{var}(z) \in \mathbb{R}^{p_2 \times p_2} \) be the variance matrix of \( z \). All matrices can be estimated from sample. Then

\[
J(w, u) = w^T \Sigma_{x,z} u - \lambda_1 (w^T \Sigma_x w - 1) - \lambda_2 (u^T \Sigma_z u - 1).
\]

Applying the critical point method, we have

\[
\frac{\partial J}{\partial w} = \Sigma_{x,z} u - 2\lambda_1 \Sigma_x w = 0 \quad \implies \quad \Sigma_{x,z} u = 2\lambda_1 \Sigma_x w,
\]

and

\[
\frac{\partial J}{\partial u} = \Sigma_{x,z} w - 2\lambda_2 \Sigma_z u = 0 \quad \implies \quad \Sigma_{x,z} w = 2\lambda_2 \Sigma_z u.
\]

The next analysis shows \( \lambda_1 = \lambda_2 \). Left-multiplying \( w^T \) on both sides of (5) gives

\[
w^T \Sigma_{x,z} u = 2\lambda_1 w^T \Sigma_x w = 2\lambda_1.
\]

Similarly, left-multiplying \( u^T \) on both sides of (6) gives

\[
u^T \Sigma_{x,z} w = 2\lambda_2 u^T \Sigma_z u = 2\lambda_2.
\]

Combining (7) and (8), we have

\[
2\lambda_1 = w^T \Sigma_{x,z} u = u^T \Sigma_{x,z} w = 2\lambda_2 \quad \implies \quad \lambda_1 = \lambda_2 =: \lambda.
\]

Now we can write (5) and (6) jointly in a matrix form

\[
\begin{bmatrix} \Sigma_{x,z} & 0 \\ 0 & \Sigma_{x,z} \end{bmatrix} \begin{bmatrix} u \\ w \end{bmatrix} = 2\lambda \begin{bmatrix} 0 & \Sigma_x \\ \Sigma_z & 0 \end{bmatrix} \begin{bmatrix} u \\ w \end{bmatrix}
\]

(10) is a generalized eigenvalue problem, where \([u; w]^T\) is an eigenvector and based on (7) \( \lambda \) is the largest eigenvalue. The rest analysis is similar to PCA.