CCA focuses on discovering the association between 2 sets of variables. Let $X_{p\times1}$ and $Y_{q\times1}$ with $p \leq q$. There are several ways to examine the association between $X$ and $Y$.

1. The correlation matrix
2. The pairwise scatterplots
3. A regression model
4. CCA

CCA finds linear functions of $X$ and $Y$ with max correlation.

$$U_1 = a'_1X \quad V_1 = b'_1Y,$$

$$\vdots$$

$$U_p = a'_pX \quad V_p = b'_pY.$$ (\(U_i, V_i\)) are called canonical variables. Their correlations are called canonical correlations. Let $\rho_i = corr(U_i, V_i)$. We find the canonical variables as follows:

1. Maximize $\rho_1$ subject to $V(U_1) = V(V_1) = 1$
2. Maximize $\rho_2$ subject to $V(U_2) = V(V_2) = 1$ and $cov(U_i, V_j) = 0 \forall i \neq j$.
3. \ldots

The resulting canonical variables do not aim at explaining the variation within their own set of variables.
Let

$$M_1 = R_{11}^{-1/2}R_{12}R_{22}^{-1}R_{21}R_{11}^{-1/2}$$

$$M_2 = R_{22}^{-1/2}R_{21}R_{11}^{-1}R_{12}R_{22}^{-1/2}$$

$M_1$ has eigenvalue and eigenvector pairs ($\lambda_i, e_i$) and $M_2$ has eigenvalue and eigenvector pairs ($\lambda_i, f_i$). Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. The k-th canonical variable pair is

$$U_k = e'_kR_{11}^{-1/2}X \quad V_k = f'_kR_{22}^{-1/2}Y$$
The $k$-th canonical correlation is
\[ corr(U_k, V_k) = \sqrt{\lambda_k} \]

Proof is omitted. Note canonical correlation is not affected by standardization.

Sequentially test if the canonical correlations equal to 0 or not.

$H_0 : \rho_1 = \ldots = \rho_p = 0 \quad H_a : \text{at least one canonical correlation is not 0}$

$H_0 : \rho_1 \neq 0, \rho_2 = \ldots = \rho_p = 0 \quad H_a : \text{at least one of } \rho_2, \ldots, \rho_p \text{ is not 0}$

$\vdots$

$H_0 : \rho_1, \ldots, \rho_{p-1} \neq 0, \rho_p = 0 \quad H_a : \rho_1 \neq 0, \ldots, \rho_p \neq 0$

Several examples: cancorr1.sas, cancorr2.sas, cancorr3.sas, sales.sas, cca-reg.sas, cca-par.sas, a paper as a reading material