Canonical Correlations

General goal: explore correlation structure between two sets of variables $\mathbf{X}_1$ and $\mathbf{X}_2$.

Begin with population definitions.

Assume

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$$

where $\mathbf{X}_1$ has p_1 components and $\mathbf{X}_2$ has p_2 components.

Partition variance covariance matrix of $\mathbf{X}$ as usual into $\mathbf{\Sigma}_{ij}$.

Which linear combination of $\mathbf{X}_1$ entries is most correlated with which linear combination of $\mathbf{X}_2$ entries?

Consider vectors $\mathbf{a}$, $\mathbf{b}$.

$$\text{Corr}(\mathbf{a}^T\mathbf{X}_1, \mathbf{b}^T\mathbf{X}_2) = \frac{\mathbf{a}^T\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}^T\boldsymbol{\Sigma}_{11}\mathbf{a}\mathbf{b}^T\boldsymbol{\Sigma}_{22}\mathbf{b}}}$$

Scale invariant as function of either $\mathbf{a}$ or $\mathbf{b}$.

So: maximize $\mathbf{a}^T\boldsymbol{\Sigma}_{12}\mathbf{b}$ subject to two conditions:

$$\mathbf{a}^T\boldsymbol{\Sigma}_{11i}\mathbf{a} = 1 = \mathbf{b}^T\boldsymbol{\Sigma}_{22}\mathbf{b}.$$

Two Lagrange multipliers gives equations:

$$\boldsymbol{\Sigma}_{12}\mathbf{b} = \lambda_1\boldsymbol{\Sigma}_{11}\mathbf{a}$$
$$\boldsymbol{\Sigma}_{21}\mathbf{a} = \lambda_2\boldsymbol{\Sigma}_{22}\mathbf{b}$$

Manipulate to get

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{a} = \lambda_1\lambda_2\boldsymbol{\Sigma}_{11}\mathbf{a}$$

Homework problem: maximize $\mathbf{a}^T\mathbf{x}$ subject to $\mathbf{x}^T\mathbf{Q}\mathbf{x} = 1$ over $\mathbf{x}$.

Equivalent to maximizing

$$\frac{\mathbf{a}^T\mathbf{x}}{\sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x}}}$$

Solution is

$$\mathbf{x} = \mathbf{Q}^{-1}\mathbf{a}$$

Maximum value is

$$\sqrt{\mathbf{a}^T\mathbf{Q}^{-1}\mathbf{a}}$$

Apply to our problem to maximize over $\mathbf{b}$ with $\mathbf{a}$ fixed. Get

$$\sqrt{\frac{\mathbf{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\mathbf{a}}{\mathbf{a}^T\boldsymbol{\Sigma}_{11}\mathbf{a}}}$$

Generalized eigenvalue problems: suppose $\mathbf{A}$, $\mathbf{B}$ symmetric and suppose $\mathbf{B}$ not singular. Find solutions of

$$(\mathbf{A} - \lambda\mathbf{B})\mathbf{v} = 0$$

for non-zero $\mathbf{v}$. (Notice solution set is a vector space.)

Write $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}$ for symmetric $\mathbf{B}^{1/2}$.

Define $\mathbf{w} = \mathbf{B}^{1/2}\mathbf{v}$, multiply basic equation by $\mathbf{B}^{-1/2}$ to get

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{w} = \lambda\mathbf{w}$$

Thus solutions are of form $(\lambda, \mathbf{B}^{-1/2}\mathbf{w})$ where $(\lambda, \mathbf{w})$ is an eigenvalue-eigenvector pair for

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{w}.$$

Equivalently: find eigenvalues of $\mathbf{B}^{-1}\mathbf{A}$ or of $\mathbf{A}\mathbf{B}^{-1}$.

Corresponding maximization problem: maximize

$$\frac{\mathbf{v}^t\mathbf{A}\mathbf{v}}{\mathbf{v}^t\mathbf{B}\mathbf{v}}$$

Maximum value is largest λ.

Application to our problem:

Having maximized over $\mathbf{b}$ now must maximize correlation squared by maximizing

$$\frac{\mathbf{a}^T \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{a}}{\mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a}}$$

so $\mathbf{B} = \mathbf{\Sigma}_{11}$ and $\mathbf{A} = \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$.

Maximal squared correlation is largest eigenvalue of

$$\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}.$$

First canonical correlates: find largest eigenvalue of

$$\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$$

and $\mathbf{a}_1$ the corresponding eigenvector. Then

$$\mathbf{b}_1 = \mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{a}_1$$

Maximal value of squared correlation is largest eigenvalue

Corresponding $\mathbf{a}_1$, $\mathbf{b}_1$ are eigenvectors of

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

and

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$$

Usually normalized.

Second canonical correlates: repeat maximization but require:

Independence of $\mathbf{a}_2^T\mathbf{X}_1$ and $\mathbf{a}_1^T\mathbf{X}_1$ and of $\mathbf{b}_2^T\mathbf{X}_2$ and $\mathbf{b}_1^T\mathbf{X}_2$.

And so on to get min$\{p_1, p_2\}$ triples:

$$\lambda_i, \mathbf{a}_i, \mathbf{b}_i$$

with λ_i in decreasing order.

With data: just replace $\mathbf{\Sigma}$ by $\mathbf{S}$.

Canonical Correlations example

Data: Table 9.12 in Johnson and Wichern.

Sales data: 3 measurements on the sales performance of 50 salespeople for a large firm and 4 test scores.

Use SAS to do canonical correlation analysis between the first 3 variables and the remaining 4.

Here is the SAS code.

```
data sales;
 infile "T9-12.DAT";
 input growth profit new
         create mech abst math;
proc cancorr;
 var growth profit new;
 with create mech abst math;
run;
```

And here is the output.

```
        Canonical Correlation Analysis
            Adjusted   Approx   Squared
  Canonical Canonical Standard Canonical
   Correln  Correln   Error    Correln
1 0.994483  0.994021  0.001572  0.988996
2 0.878107  0.872097  0.032704  0.771071
3 0.383606  0.366795  0.121835  0.147153
Eigenvalues of INV(E)*H =CanRsq/(1-CanRsq)

 Eigenvalue Difference Proportion  Cumul
1  89.8745   86.5063     0.9621    0.9621
2   3.3682    3.1956     0.0361    0.9982
3   0.1725         .     0.0018    1.0000
```

Notice the first canonical correlates account for most of the correlation between the two sets of variables.

```
   Canonical Correlation Analysis
Test of H0: The canonical correlations in
  current row and all that follow are zero
  Likelihood
      Ratio   Approx F Num DF Den DF  Pr > F
1 0.00214847  87.3915   12   114.06  0.0001
2 0.19524127  18.5263    6    88     0.0001
3 0.85284669   3.8822    2    45     0.0278
Multivariate Statistics and F Approximations
S=3    M=0    N=20.5
Statistic    Value      F  NumDF DenDF Pr > F
Wilks'       0.0022   87.39  12  114.1 0.0001
Pillai's     1.9072   19.63  12  135   0.0001
Hotel'g-Ly 93.4152  324.35  12  125   0.0001
Roy's      89.8745 1011.08   4   45   0.0001
NOTE: F Statistic for Roy's Greatest Root is
an upper bound.
```

Clearly all 3 of the canonical correlations are non-zero. The multivariate tests are of $H_o : \Sigma_{12} = 0$.

Canonical Correlation Analysis

Raw Canonical Coefficients for the 'VAR' Variables

	V1	V2	V3
GROWTH	0.06238	-0.17407	-0.37715
PROFIT	0.02093	0.24216	0.10352
NEW	0.07826	-0.23829	0.38342

Raw Canonical Coefficients for the 'WITH' Variables

	W1	W2	W3
CREATE	0.06975	-0.19239	0.24656
MECH	0.03074	0.20157	-0.14190
ABST	0.08956	-0.49576	-0.28022
MATH	0.06283	0.06832	0.01133

Canonical Correlation Analysis

Standardized Canonical Coefficients for the 'VAR' Variables

	V1	V2	V3
GROWTH	0.4577	-1.2772	-2.7673
PROFIT	0.2119	2.4517	1.0480
NEW	0.3688	-1.1229	1.8067

Standardized Canonical Coefficients for the 'WITH' Variables

	W1	W2	W3
CREATE	0.2755	-0.7600	0.9739
MECH	0.1040	0.6823	-0.4803
ABST	0.1916	-1.0607	-0.5996
MATH	0.6621	0.7199	0.1194

The standardized coefficients are easiest to interpret.

For instance a quantity which is not too different from the average of the 3 sales indices is strongly correlated with a weighted average of the psychological test scores which puts most of the weight on Math.

The second set of correlates focus on the relation between profitability minus the average of the other two sales indices correlated with (Math + Mech) - (Abstract + Creativity).

The last one is not particularly meaningful to me but it is a very small part of the correlation structure between the two sets of variables.

Canonical Structure

Correlations Between the 'VAR' Variables
and Their Canonical Variables

	V1	V2	V3
GROWTH	0.9799	0.0006	-0.1996
PROFIT	0.9464	0.3229	0.0075
NEW	0.9519	-0.1863	0.2434

Correlations Between the 'WITH' Variables
and Their Canonical Variables

	W1	W2	W3
CREATE	0.6383	-0.2157	0.6514
MECH	0.7212	0.2376	-0.0677
ABST	0.6472	-0.5013	-0.5742
MATH	0.9441	0.1975	-0.0942

Canonical Structure

Correlations Between the 'VAR' Variables and the
Canonical Variables of the 'WITH' Variables

	W1	W2	W3
GROWTH	0.9745	0.0006	-0.0766
PROFIT	0.9412	0.2835	0.0029
NEW	0.9466	-0.1636	0.0934

Correlations Between the 'WITH' Variables and
the Canonical Variables of the 'VAR' Variables

	V1	V2	V3
CREATE	0.6348	-0.1894	0.2499
MECH	0.7172	0.2086	-0.0260
ABST	0.6437	-0.4402	-0.2203
MATH	0.9389	0.1735	-0.0361