Data Analytics for Materials Science

27-737

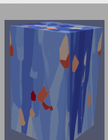
A.D. (Tony) Rollett, Amit Verma, Richard A. LeSar (Iowa State Univ.)

Dept. Materials Sci. Eng., Carnegie Mellon University

Canonical Correlation Analysis (CCA)

Lecture 1, part 1

- “Generation of Statistically Representative Synthetic Three-dimensional Microstructures”, Sudipto Mandal, Jacky Lao, Sean Donegan, and Anthony D. Rollett, *Scripta materialia* **146** 128-132 (2018); doi: 10.1016/j.scriptamat.2017.11.034.
- “Application of canonical correlation analysis to a sensitivity study of constitutive model parameter fitting” S. Mandal, B.T. Gockel, and A.D. Rollett, *Materials and Design* **132**, 30- 43 (2017); doi.org/10.1016/j.matdes.2017.06.050.
- “Data Analytics using Canonical Correlation Analysis and Monte Carlo Simulation” (reference number: NPJCOMPUMATS-00217), J. Rickman, Y. Wang, A.D. Rollett, M.P. Harmer and C. Compson, *(Nature) Computational Materials* 3 26 (2017); doi:10.1038/s41524-017-0028-9. Short link: http://rdcu.be/tWJp.
- “Parsing abnormal grain growth”, A. Lawrence, J.M. Rickman, M.P. Harmer, A.D. Rollett, *Acta Materialia*, **103**, 681-687 (2016); doi: 10.1016/j.actamat.2015.10.034.
- “Grain-boundary character distribution and correlations with electrical and optoelectronic properties of CuInSe2 thin films”, D. Abou-Ras, N. Schäfer, T. Rissom, M.N. Kelly, J. Haarstrich, C. Ronning, G.S. Rohrer, A.D. Rollett, *Acta Materialia* **118** 244–252 (2016); dx.doi.org/10.1016/j.actamat.2016.07.042.
- Gittins, *Canonical Analysis: A Review with Applications in Ecology* (Springer-Verlag, Berlin, 1985).
- Hair et al., *Canonical Correlation Analysis - supplement to: Multivariate Data Analysis*, 6th Ed. (Pearson Prentice Hall, 2006).
- Jobson, *Applied multivariate data analysis: volume II: Categorical and Multivariate Methods,* (Springer Science & Business Media, 2012).



Some references

Suppose we have two sets of data, $\mathbf{X}$ and $\mathbf{Y}$, each with multiple data types, from the same set of *N* observations

- assume p variables in $\mathbf{X}$ and q variables in $\mathbf{Y}$
- $\mathbf{X}$ is thus an $N \times p$ dimensional matrix and $\mathbf{Y}$ is an $N \times q$ dimensional matrix

The $\mathbf{X}$ variables considered as independent data, i.e., they are considered as *input* or *predictor* variables.

The $\mathbf{Y}$ variables are dependent data, i.e., they are *output*.

The basic idea is that the $\mathbf{Y}$ variables are considered to occur in *response* to the $\mathbf{X}$ variables and correlations may exist in both sets.

CCA provides a way to find the *correlations* between the **X** variables and the **Y** variables.

CCA does this by finding two sets of basis vectors, one for **X** and the other for **Y**, such that the **correlations between the *projections* of the variables onto these basis vectors are mutually maximized.**

It has some similarities to PCA.

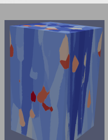
The grain data was based on 2337 grains from an unnamed superalloy with 11 measured variables for each grain:

$$\left\{ b/a,\ c/a,\ a,\ b,\ c,\ x_c,\ y_c,\ z_c,\ D_{eq.},\ N_{Neighbors},\ \Omega_3 \right\}$$

- the position of each grain is $\left\{ x_c,\ y_c,\ z_c \right\}$
- the size of each grain is described by $\left\{ a,\ b,\ c,\ D_{eq.}, N_{Neighbors} \right\}$
- the shape of each grain is captured by $\left\{ b/a,\ c/a,\ \Omega_3 \right\}$

Using PCA was limited by $\left\{ x_c,\ y_c,\ z_c \right\}$ independent variables and the rest being dependent variables: we could not easily extract out possible position dependence. CCA could let us do that. (By the way, we still may not see anything that is useful.)

Data from M. Groeber, AFRL from a DREAM3D data file.



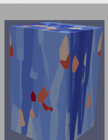
Grain data (from the USAF Materials Lab)

From a dataset supplied by Prof. Francis Wagner (Univ. of Lorraine, Metz) we have a number of experiments (18) in which processing parameters (anneal time, anneal temperature (°C), rolling direction, and test direction) and measured appropriate materials properties (such as yield stress, strain at peak stress, % recrystallized, grain diameter, …) were varied.

Our question is:

- What processing parameters have the most influence on the properties of the processed material.
- Which properties are most closely linked to the processing parameters?

We will use this problem as our first example of the use of CCA.



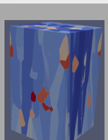
Example 1: processing and properties for cp-Ti

Ech.	serie	Rolling	TT	ReX	D_my	1/√D	Tr //	YS	L_P1	σe_max	ε*	σ_max - σe	L_P2	εr	
					(in µm)			MPa	%	MPa	%	MPa	%	%	
R1	1	75% //RD	740°C-2h	100%	11.7	0.2924	*RD*	315	-	466	19.5	151	22.5	35.0	
R2	1	75% //RD	500°C-1h	91%	1.8	0.7454	*RD*	420	1.9	510	19.6	90	21.4	32.4	
R3	1	75% //RD	470°C-2h	62%	1.2	0.9129	*RD*	490	-	600	14.6	110	20.4	27.6	
R4	1	75% //RD	500°C-40min	80.00%	1.7	0.767	*RD*	439	1.0	538	17.3	99		33.7	
R5	1	75% //RD	550°C-40min	98.00%	2.8	0.5976	*RD*	375	1.8	485	21	110		40.2	
RR1	2	75% //RD	740°C-2h	98.30%	9.7	0.3211	*RD*	**332**	~ 0	**468**	**14**	**136**	14.6	**24.3**	
RR2	2	75% //RD	500°C-1h	69.20%	1.4	0.8452	*RD*	404		480	7.2		9.5	**13.5**	<< too thin
RR3	2	75% //RD	470°C-2h	**46.00%**	1.0	1	*RD*	**544**	~ 0	645	10.3	101	17.5	**22.1**	
RR4	2	75% //RD	500°C-40min				*RD*	**526**		674	12.3	148		**22.6**	<< essai méca. loupé
RT1	2	75% //RD	740°C-2h	98.30%	9.7	0.3211	*TD*	**379**	1.5	**456**	**9**	**77**	14.7	**28.5**	
RT2	2	75% //RD	500°C-1h	69.20%	1.4	0.8452	*TD*	**511**		**523.5**	**0.3**	12.5	10.5	**21.2**	
RT3	2	75% //RD	470°C-2h	**46.00%**	1.0	1	*TD*	621		672	1.1	51	4.9	14.3	
RT4	2	75% //RD	500°C-40min				*TD*	522	-	579	2.2	57		14.4	
T1	1	75% //TD	740°C-2h	100%	10.9	0.3029	*RD'*	300	-	425	19.0	125	24.0	35.0	
T2	1	75% //TD	500°C-1h	98%	1.8	0.7454	*RD'*	400	3.7	450	19.5	50	29.4	43.0	
T3	1	75% //TD	470°C-2h	80%	1.2	0.9129	*RD'*	485	2.1	512	17.0	27	26.2	36.5	
T4	1	75% //TD	500°C-40min	94.00%	1.7	0.767	*RD'*	390	3.0	463	19.3	73		44.5	
T5	1	75% //TD	550°C-40min	98.00%	2.6	0.6202	*RD'*	375	3.9	436	21	61		44.8	

Tr // : direction of extension (tensile test with 'small' samples)
L_P1: length (in %) of the plateau after the YS
σe_max: max Eng. stress
ε*: elongation (in %) for the max stress (Considere's criterium)
L_P2: 'pseudo plateau' around σe_max (-5%)
εr : elongation at break

Spreadsheet from Prof. Francis Wagner

Anneal_time	Anneal_Temperature_	Rolling_Direction	Test_Direction	Yield_Stress	L_P1	Eng_Stress_max	Strain_at_peak_stress	stress_max-s_yield	Strain_to_failure	per_cent_recrystallized	Grain_Diameter	1/sqrtD	L_P2
120	740	1	1	315	0	466	19.5	151	35	1	11.7	0.29235267	22.5
60	500	1	1	420	1.9	510	19.6	90	32.4	0.91	1.8	0.74535599	21.4
120	470	1	1	490	0	600	14.6	110	27.6	0.62	1.2	0.91287093	20.4
40	500	1	1	439	1	538	17.3	99	33.7	0.8	1.7	0.76696499	0
40	550	1	1	375	1.8	485	21	110	40.2	0.98	2.8	0.59761431	0
120	740	1	1	332	0	468	14	136	24.3	0.98	9.7	0.32108065	14.6
60	500	1	1	404	0	480	7.2	90	13.5	0.69	1.4	0.84515426	9.5
120	470	1	1	544	0	645	10.3	101	22.1	0.46	1	1	17.5
40	500	1	1	526	0	674	12.3	148	22.6	0.8	1.7	0.76696499	0
120	740	1	2	379	1.5	456	9	77	28.5	0.98	9.7	0.32108065	14.7
60	500	1	2	511	0	523.5	0.3	12.5	21.2	0.69	1.4	0.84515426	10.5
120	470	1	2	621	0	672	1.1	51	14.3	0.46	1	1	4.9
40	500	1	2	522	0	579	2.2	57	14.4	0.8	1.7	0.76696499	0
120	740	2	1	300	0	425	19	125	35	1	10.9	0.30289127	24
60	500	2	1	400	3.7	450	19.5	50	43	0.98	1.8	0.74535599	29.4
120	470	2	1	485	2.1	512	17	27	36.5	0.8	1.2	0.91287093	26.2
40	500	2	1	390	3	463	19.3	73	44.5	0.94	1.7	0.76696499	0
40	550	2	1	375	3.9	436	21	61	44.8	0.98	2.6	0.62017367	0

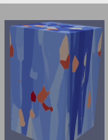


Simplified dataset

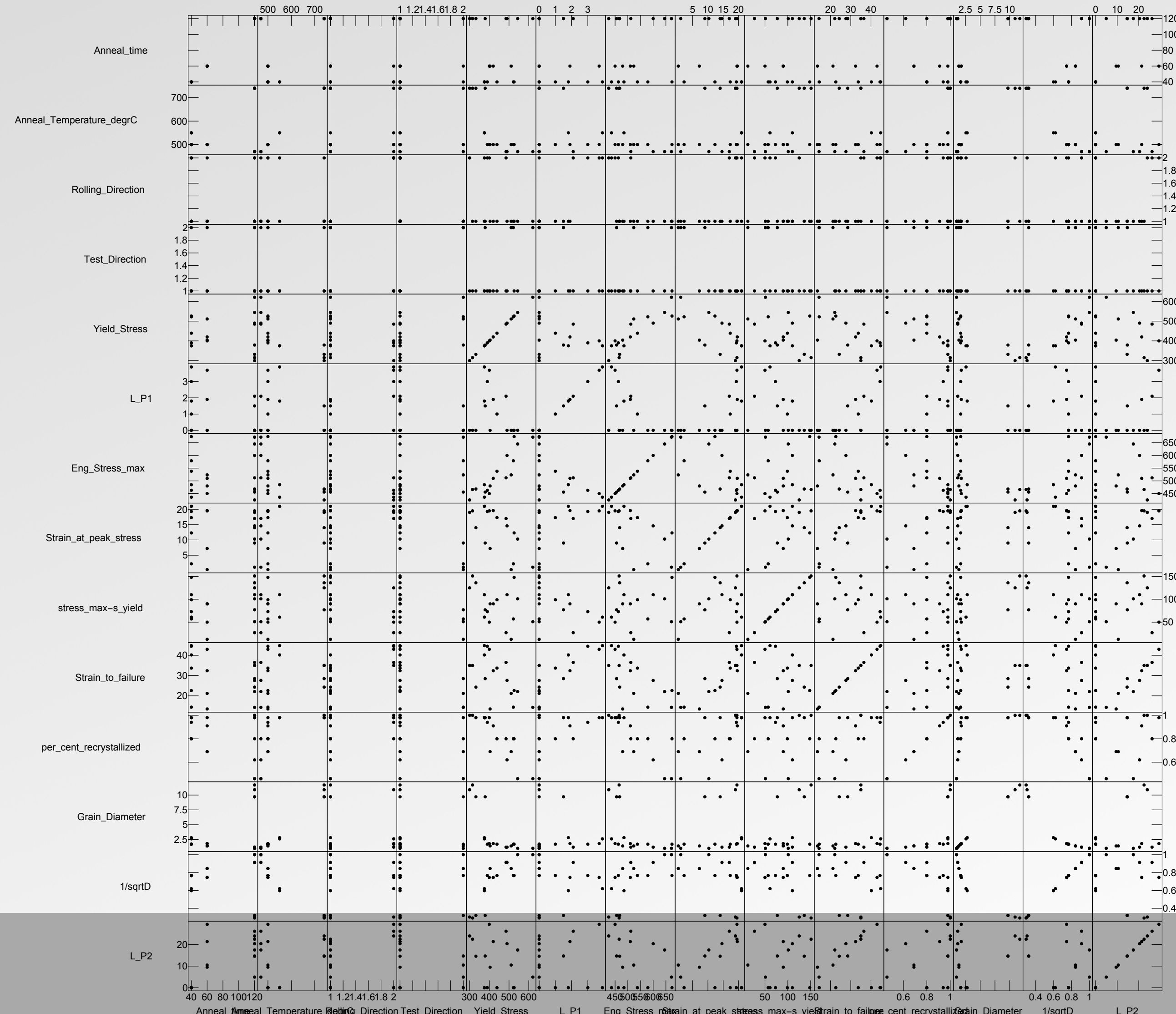
For CCA, we break the data types into two categories.

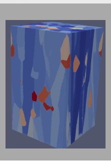
- Input are 4 *processing parameters*: (an 18×4 matrix)
  - $\mathbf{X}$ = Anneal time, Anneal Temperature (°C), Rolling Direction, Test Direction
- Output are 10 *results of tests*: (an 18×10 matrix)
  - $\mathbf{Y} =$ Yield Stress, L_P1, Eng. Stress (max)}, Strain at peak stress, stress at max yield, Strain to failure}, % recrystallized, Grain Diameter (D), $D^{-1/2}$, L_P2

Goal: find a representation for $\mathbf{X}$ and $\mathbf{Y}$ to capture the *maximum correlations between the inputs and outputs* based on a linear analysis.



Scatterplot matrix





Example 1:

Create autoscaled matrices:

$$\mathbf{X}_i' = \frac{\mathbf{X}_i - \overline{\mathbf{X}}_i}{\sigma_{\mathbf{X}_i}}$$

$$\mathbf{X} = \left\{\mathbf{X}_1', \mathbf{X}_2', \mathbf{X}_3', \mathbf{X}_4'\right\}$$

$$\mathbf{Y} = \left\{\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_{10}'\right\}$$

Tools for CCA available in R, MATLAB, SAS, …

Calculate correlation matrices:

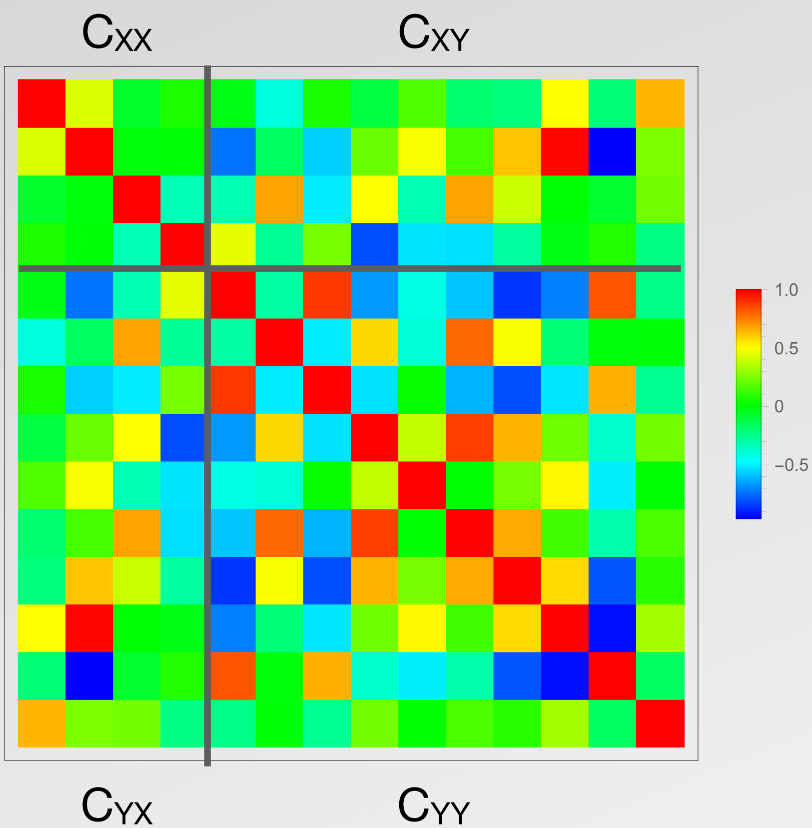
$$\mathbf{C_{XX}} = \frac{\mathbf{X}^T\mathbf{X}}{N-1} \qquad \mathbf{C_{YY}} = \frac{\mathbf{Y}^T\mathbf{Y}}{N-1}$$

$$\mathbf{C_{XY}} = \frac{\mathbf{X}^T\mathbf{Y}}{N-1} \qquad \mathbf{C_{YX}} = \frac{\mathbf{Y}^T\mathbf{X}}{N-1}$$

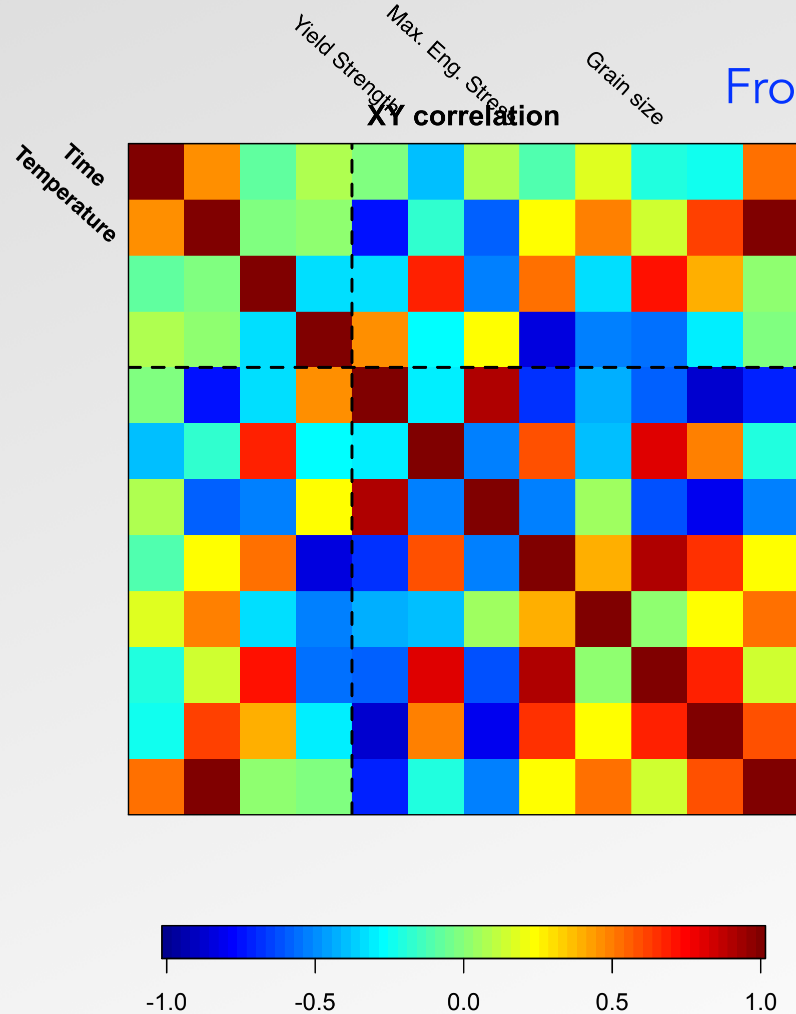
$$\mathbf{C_{XY}} = \mathbf{C}_{\mathbf{YX}}^T$$

C_{XX} *4x4*	C_{XY} *4x10*
C_{YX} *10x4*	C_{YY} *10x10*

CCA steps



From R



matcor tool in R

CCA steps: correlation matrix

Tools for CCA available in R, MATLAB, SAS, …

Derivation of CCA equations

Datasets: A_x and A_y

Directions: x and y

Projections (Canonical Variates): Z_x and Z_y

$$\begin{aligned} \mathbf{z}_x &= \mathbf{A}_x\mathbf{x} \\ \mathbf{z}_y &= \mathbf{A}_y\mathbf{y}. \end{aligned}$$

Correlation between the canonical variates

$$\rho = \frac{\mathbf{z}_y' \cdot \mathbf{z}_x}{\sqrt{\mathbf{z}_y' \cdot \mathbf{z}_y}\sqrt{\mathbf{z}_x' \cdot \mathbf{z}_x}}.$$

Choice of rescaling is arbitrary

$$\begin{aligned} \mathbf{z}_x' \cdot \mathbf{z}_x &= \mathbf{x}'\mathbf{A}_x'\mathbf{A}_x\mathbf{x} = \mathbf{x}'\Sigma_{xx}\mathbf{x} &= 1 \\ \mathbf{z}_y' \cdot \mathbf{z}_y &= \mathbf{y}'\mathbf{A}_y'\mathbf{A}_y\mathbf{y} = \mathbf{y}'\Sigma_{yy}\mathbf{y} &= 1. \end{aligned}$$

Maximization in the Lagrangian form

$$L(\rho_x, \rho_y, \mathbf{x}, \mathbf{y}) = \mathbf{y}'\Sigma_{yx}\mathbf{x} - \frac{\rho_x}{2}\left(\mathbf{x}'\Sigma_{xx}\mathbf{x} - 1\right) - \frac{\rho_y}{2}\left(\mathbf{y}'\Sigma_{yy}\mathbf{y} - 1\right),$$

Solve Lagrangian by taking partial derivatives

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{x}} &= \Sigma_{xy}\mathbf{y} - \rho_x\Sigma_{xx}\mathbf{x} &= \mathbf{0} \\ \frac{\partial L}{\partial \mathbf{y}} &= \Sigma_{yx}\mathbf{x} - \rho_y\Sigma_{yy}\mathbf{y} &= \mathbf{0}. \end{aligned}$$

multiply by x'

multiply by y'

$$\begin{aligned} \mathbf{0} &= \mathbf{y}'\Sigma_{yx}\mathbf{x} - \rho_y\mathbf{y}'\Sigma_{yy}\mathbf{y} - \mathbf{x}'\Sigma_{xy}\mathbf{y} + \rho_x\mathbf{x}'\Sigma_{xx}\mathbf{x} \\ &= \rho_x\mathbf{x}'\Sigma_{xx}\mathbf{x} - \rho_y\mathbf{y}'\Sigma_{yy}\mathbf{y}. \end{aligned}$$

From rescaling constraint & last equation, it can be concluded that: $\rho_x = \rho_y = \rho$

Derivation (contd.)

From the partial derivative

$$\mathbf{x} = \frac{\Sigma_{xx}^{-1}\Sigma_{xy}\mathbf{y}}{\rho}.$$

Rearranging terms

$$(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \rho^2\Sigma_{yy})\mathbf{y} = \mathbf{0}.$$

$$(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \rho^2\Sigma_{xx})\mathbf{x} = \mathbf{0}.$$

$$\boxed{\begin{aligned}(\Sigma'_{xy}\Sigma_{xx}^{-1}\Sigma_{xy} - \rho^2\Sigma_{yy})\mathbf{y} = \mathbf{0} \\ (\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma'_{xy} - \rho^2\Sigma_{xx})\mathbf{x} = \mathbf{0}.\end{aligned}}$$

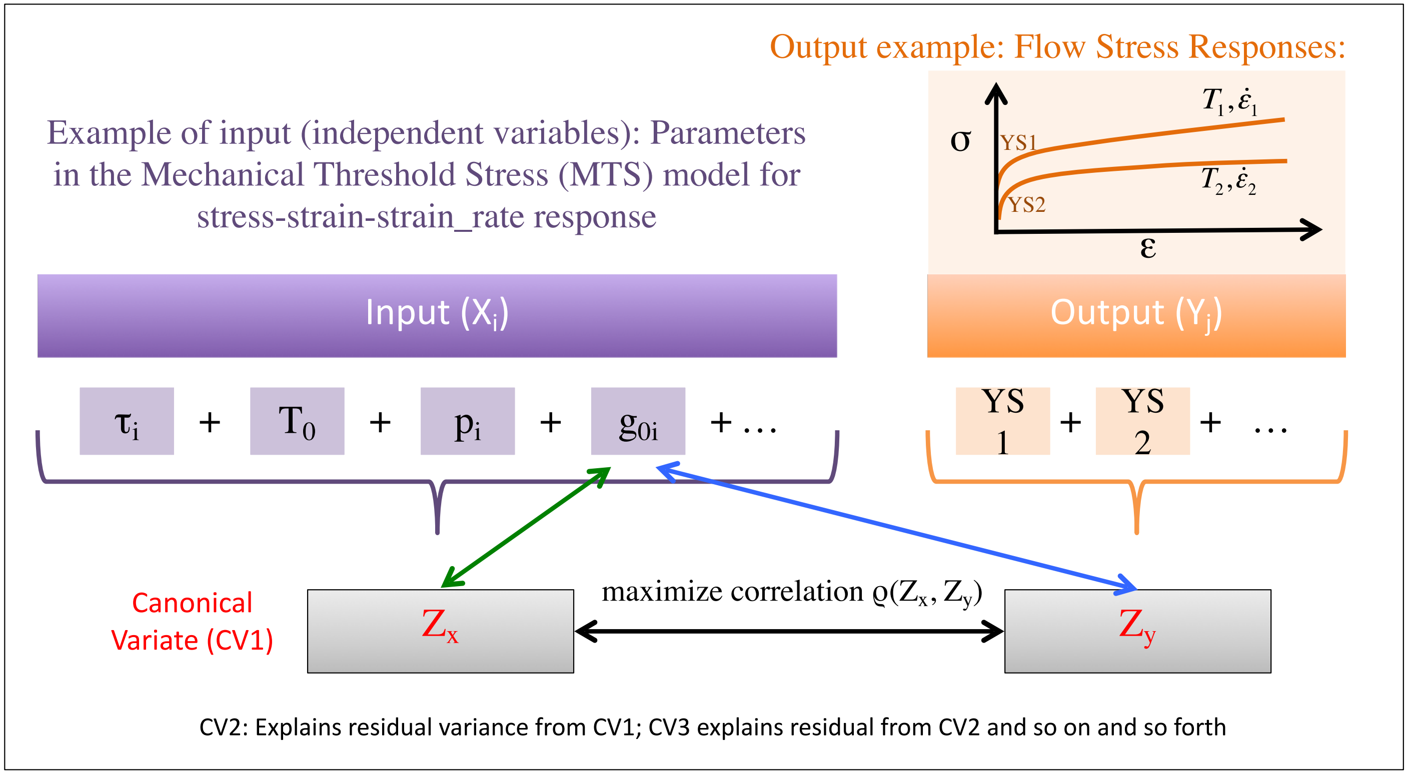
Equations for the Canonical Correlation Analysis

Generalized eigenvalue problems

Can be solved by:

- Given the correlation matrices, this eigenvalue problem can be written as a general singular value problem that can be solved by Cholesky factorization
- Given the data matrices, singular value decomposition (SVD) can be used

Application



Coefficients or Weights: Values that multiply each variable to make up a Canonical Variate (values that one sees in an equation)

Loadings: Bivariate correlations between canonical variate & real variable (relative importance)

Communality: Sum of squared loadings for all CVs (Overall usefulness)

Redundancy: Averaged cross-loadings across all CVs (Adequacy of prediction)

From Jobson Vol. 2

Note that the canonical correlations (square root of the eigenvalues) can be large, even when the proportion of variance of the underlying variables explained by the canonical variates is comparatively small. I.e., the canonical variate pairs may be very well correlated, even if the relationship to the actual variables is weaker.

The latter is quantified by the squared structure correlations (bottom right of the figure).

7.5 Multivariate Regression and Canonical Correlation 187

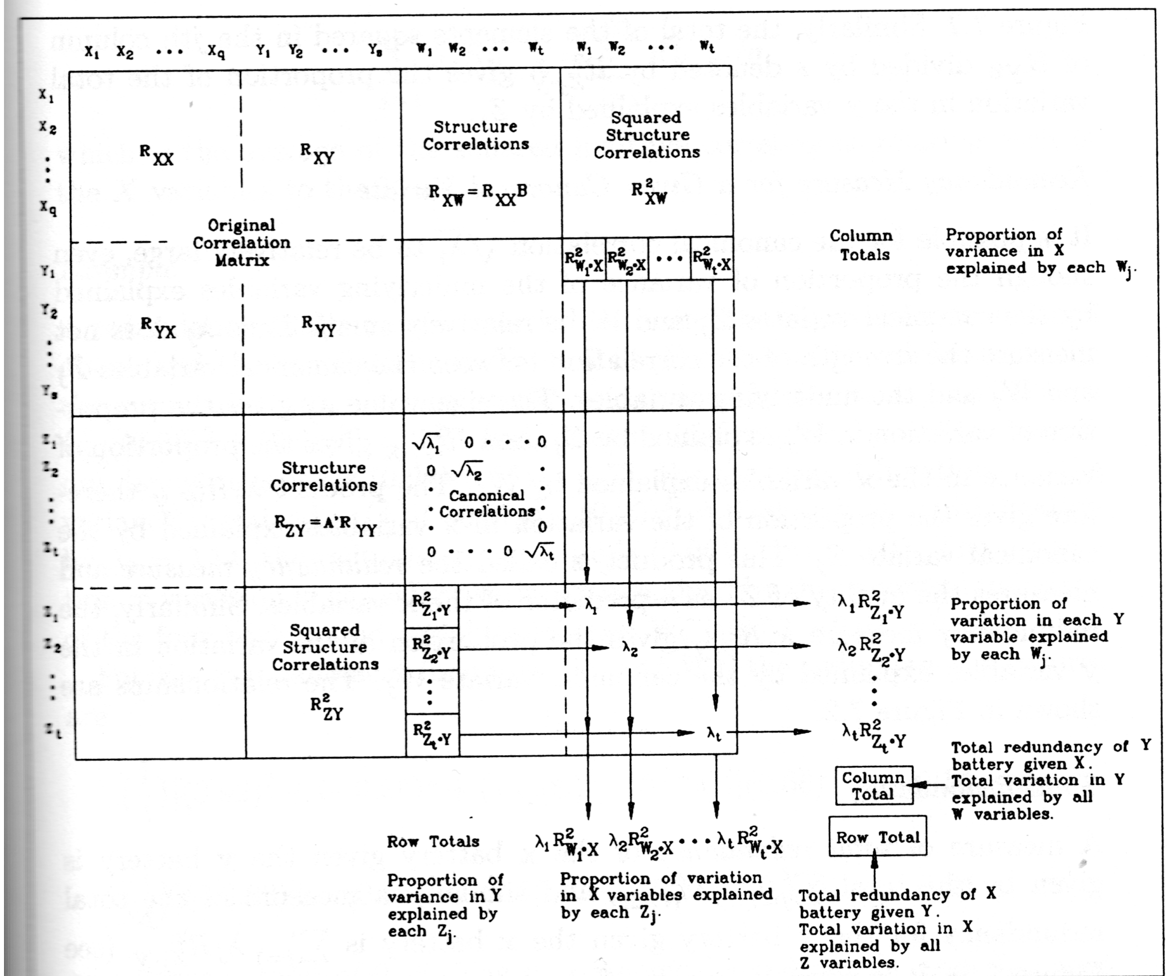


FIGURE 7.2. Summary of Canonical Correlation Terminology

There are a number of ways to solve for the linear combination of variables that maximizes the correlation.

One approach is to define the vector: $\mathbf{K} = \mathbf{C}_{\mathbf{XX}}^{-1/2}\mathbf{C}_{\mathbf{XY}}\mathbf{C}_{\mathbf{YY}}^{-1/2}$

Suppose $\mathbf{C}^{-1} = \begin{bmatrix} 7 & 10 \\ 15 & 22 \end{bmatrix}$ $\left(\mathbf{C} = \begin{bmatrix} \frac{11}{2} & -\frac{5}{2} \\ -\frac{15}{4} & \frac{7}{4} \end{bmatrix} \right)$

$\mathbf{C}^{-1/2} = \begin{bmatrix} 3\sqrt{3/11} & 10/\sqrt{33} \\ 5\sqrt{3/11} & 8/\sqrt{3/11} \end{bmatrix}$ and $\mathbf{C}^{-1/2}\mathbf{C}^{-1/2} = \mathbf{C}^{-1}$

$$\mathbf{K} = \mathbf{C}_{\mathbf{XX}}^{-1/2}\mathbf{C}_{\mathbf{XY}}\mathbf{C}_{\mathbf{YY}}^{-1/2}$$

We then perform a *singular value decomposition* of $\mathbf{K}$:

$$\mathbf{K} = \mathbf{\Gamma\Lambda\Delta}$$

The **singular value decomposition** (**SVD**) is a factorization of a matrix that generalizes an eigenanalysis of a square normal matrix to any $m \times n$ matrix.

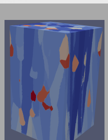
For the example that we chose of the annealed Ti, $\mathbf{\Gamma}$ is a 4×4 matrix, $\mathbf{\Delta}$ is a 4×10 matrix, and $\mathbf{\Lambda}$ is a 4×4 matrix with the diagonals being the eigenvalues of $\mathbf{K}$ (the correlations).

https://en.wikipedia.org/wiki/Singular_value_decomposition

$$\boldsymbol{\Lambda} = \begin{pmatrix} 0.999464 & 0. & 0. & 0. \\ 0. & 0.970684 & 0. & 0. \\ 0. & 0. & 0.95352 & 0. \\ 0. & 0. & 0. & 0.834025 \end{pmatrix}$$

"Thus, if $\mathbf{X}$ is the matrix containing the explanatory factors of $\mathbf{Y}$, the matrix containing the criterion measures (or criterion variables), it is possible to say that the explanatory factors would perfectly explain the criterion variables if $\lambda_1 = 1$. If $\lambda_1 = 0$, the explanatory factors have no influence on the criterion variables, and any value between 1 and 0 is merely an interpolation of these extreme cases."

[Canonical Correlation Analysis,](#) Malacarne 2014



CCA: Λ, the eigenvalues of $\mathbf{K}$

Define two new matrices, consisting of the canonical correlation vectors, which maximize the correlation between the **canonical variates** (new term!).

$$\mathbf{A} = \mathbf{C}_{\mathbf{XX}}^{-1/2}\boldsymbol{\Gamma} \quad \text{(a } 4 \times 4 \text{ matrix)}$$

$$\begin{array}{r c} & \begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \\ \begin{array}{r} \text{anneal. time} \\ \text{anneal. temp} \\ \text{roll. dir.} \\ \text{test dir.} \end{array} & \begin{pmatrix} 0.125047 & -0.161884 & -1.09568 & 0.0837215 \\ -1.03984 & 0.184989 & 0.352461 & -0.0171176 \\ -0.0133887 & -0.223218 & 0.0551519 & 1.03603 \\ -0.142058 & -1.02403 & 0.208802 & 0.122576 \end{pmatrix} \end{array}$$

$$\mathbf{B} = \mathbf{C}_{\mathbf{YY}}^{-1/2}\boldsymbol{\Delta} \quad \text{(a } 4 \times 10 \text{ matrix)}$$

$$\begin{array}{r c} & \begin{array}{cccc} b_1 & b_2 & b_3 & b_4 \end{array} \\ \begin{array}{r} \text{YS} \\ \text{LP1} \\ \\ \\ \\ \\ \\ \text{D} \\ \text{D}^{-1/2} \\ \\ \end{array} & \begin{pmatrix} 0.505018 & 0.168701 & 1.41821 & 2.96834 \\ -0.0883199 & -0.106505 & -0.210539 & 0.31741 \\ -0.418723 & -0.724131 & -1.48629 & -2.67865 \\ 0.230338 & 0.735017 & -1.05744 & 0.143887 \\ 0.243539 & 0.850836 & 1.02523 & 0.841286 \\ -0.0145254 & -0.279299 & 0.621457 & 0.353228 \\ 0.146178 & -0.0218443 & 0.909085 & 1.07527 \\ -0.351801 & -0.409257 & -1.54541 & 2.77108 \\ 0.815247 & 0.337397 & -0.75542 & 3.21709 \\ 0.0422002 & 0.0389418 & -0.288733 & -0.234883 \end{pmatrix} \end{array}$$

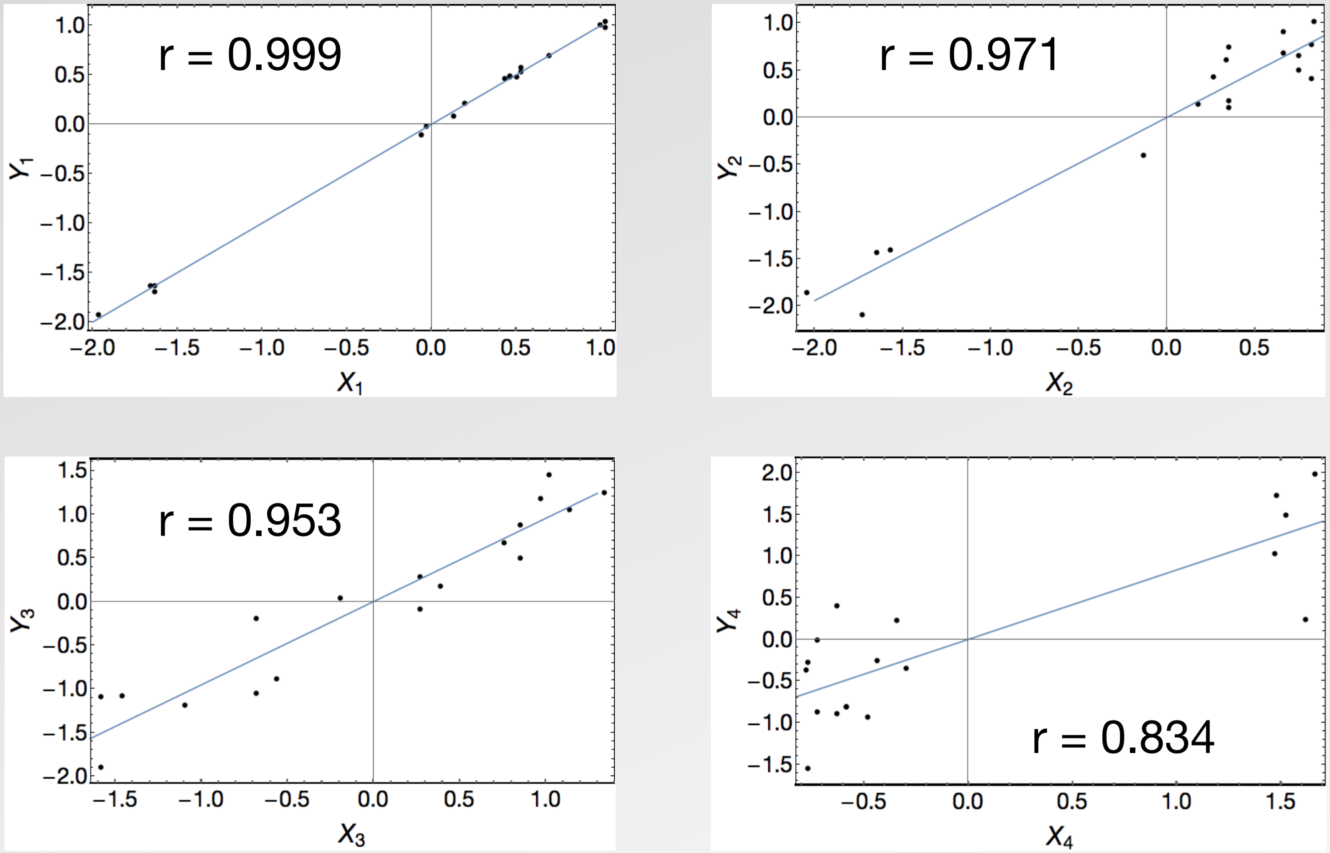
Project the data onto the $\mathbf{A}$ and $\mathbf{B}$ vectors (to get *scores*):

- $\mathbf{X_A} = \mathbf{A}^T\mathbf{X}^T$
  - $\mathbf{A}^T$: (a 4×4 matrix) times $\mathbf{X}^T$ (a 4×18 matrix): $\mathbf{X_A}$ (a 4×18 matrix)
  - $\mathbf{X}_{\mathbf{A}1i} = 0.125047X_{1i} - 1.03984X_{2i} - 0.0133887X_{3i} - 0.142058X_{4i}$
- $\mathbf{Y_B} = \mathbf{B}^T\mathbf{Y}^T$
  - $\mathbf{B}^T$: (a 4×10 matrix) times $\mathbf{Y}^T$ (a 10×18 matrix): $\mathbf{Y_A}$ (a 4×18 matrix)
  - $\mathbf{Y}_{\mathbf{B}1i} = 0.505018Y_{1i} - 0.0883199Y_{2i} - 0.418723Y_{3i} + \ldots + 0.0422002Y_{10i}$

$\mathbf{X_A}$ and $\mathbf{Y_B}$ correspond to the data projected onto the four eigenvectors of the covariance.

We plot them as pairs of data, i.e., $\left\{X_{A1}, Y_{B1}\right\}$ for all the data. The first plot generally should have the best correlation between the two *canonical variates pair.*

CCA: the canonical variates



The eigenvalues of **K** give *r* for each plot.

CCA: the loadings

$$
\begin{array}{rc}
 & \begin{array}{cccc} \text{CV 1} & \text{CV2} & \text{CV3} & \text{CV4} \end{array} \\
\begin{array}{r} \text{annealing time} \\ \text{temperature} \\ \text{RD} \\ \text{TD} \end{array} &
\begin{pmatrix}
-0.340349 & -0.140564 & -0.929615 & 0.0148034 \\
-0.985202 & 0.112832 & -0.127704 & 0.0183764 \\
0.0266028 & 0.126994 & 0.059899 & 0.989736 \\
-0.12997 & -0.961614 & 0.110832 & -0.214763
\end{pmatrix}
\end{array}
$$

$$
\begin{array}{rc}
\begin{array}{r} \text{yield stress} \\ \ldots \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} &
\begin{pmatrix}
0.708172 & -0.546803 & -0.16923 & -0.323792 \\
0.147977 & 0.161124 & 0.383525 & 0.776901 \\
0.565525 & -0.264445 & -0.261142 & -0.57743 \\
-0.13621 & 0.802808 & 0.0551283 & 0.495478 \\
-0.415386 & 0.6918 & -0.147229 & -0.468224 \\
-0.11684 & 0.468428 & 0.202036 & 0.752874 \\
-0.641999 & 0.365102 & 0.445681 & 0.425161 \\
-0.962596 & 0.127422 & -0.220168 & 0.0330277 \\
0.962728 & -0.222035 & -0.0949568 & -0.0772196 \\
-0.156153 & 0.130911 & -0.689754 & 0.322648
\end{pmatrix}
\end{array}
$$

The *loadings* are the projection of the variables onto the canonical coefficients (just as in PCA):

$$\mathbf{L_A = C_{XX}A}$$

$$\mathbf{L_B = C_{YY}B}$$

$$Var_i = a_{i1}CV_1 + a_{i2}CV_2 + a_{i3}CV_3 + a_{i4}CV_4 \qquad Corr_{ij} = Var_i \cdot Var_j$$

Applying the CCA technique is almost as simple as PCA. One main difference is to decide which set of variables (columns) should be regarded as *input variables* and which set as *output variables*.

```
> invars = allvars[,1:4]
> outvars = allvars[,5:12]
```

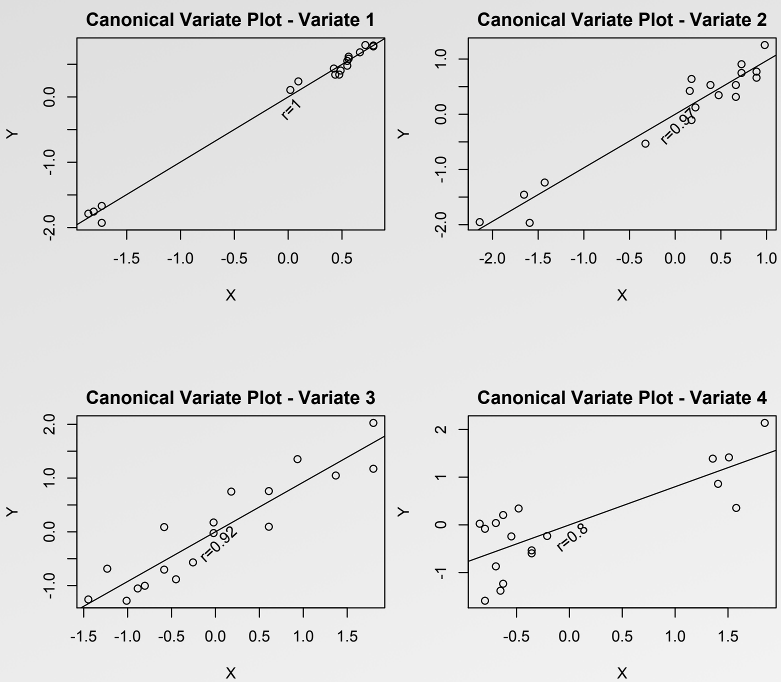
Then we apply the CCA itself (NB. you can find options in the *yacca* page).

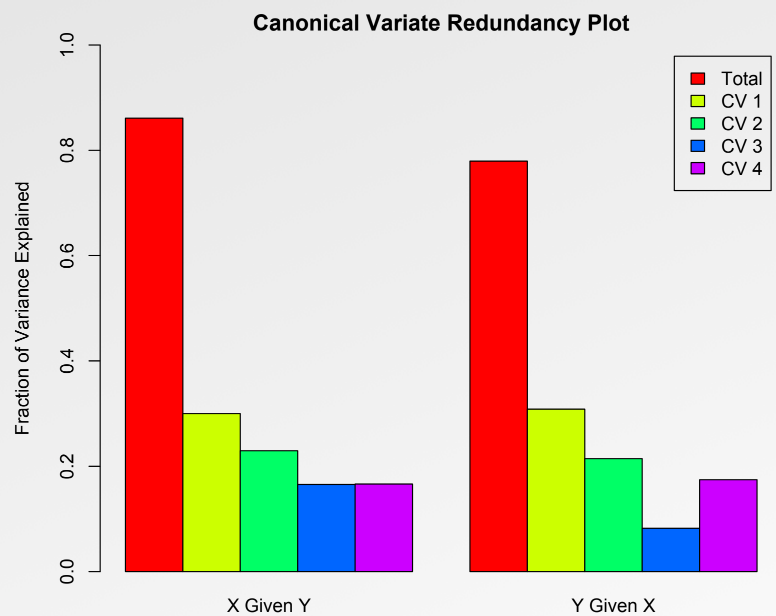
```
ccares=cca(invars,outvars,standardize.scores=T)
```

To learn about the results of the analysis, the quickest thing is to do this:

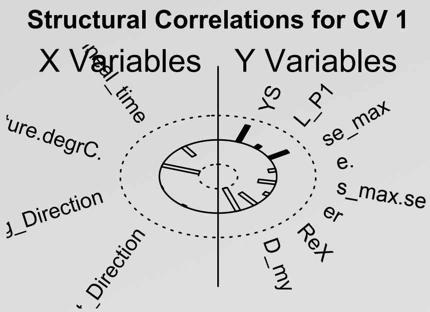
```
plot(ccares)
```

This gives 4 different plots, of which, the first shows that we can get a good fit.

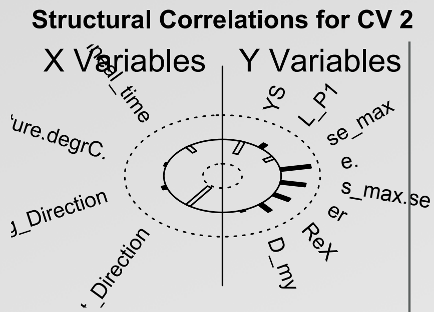




CCA in R: output

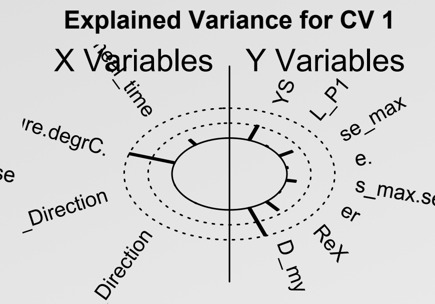


Canonical Variate1

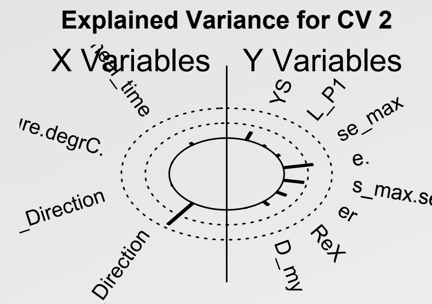


Canonical Variate2

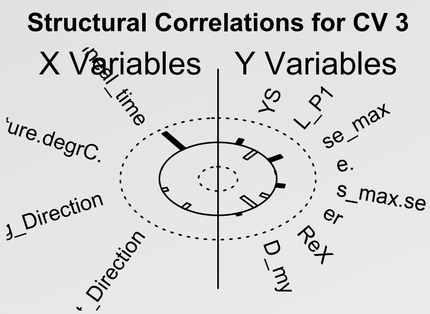
Most important *input* variable = Anneal Temp.
Most sensitive *output* variable = Grain size



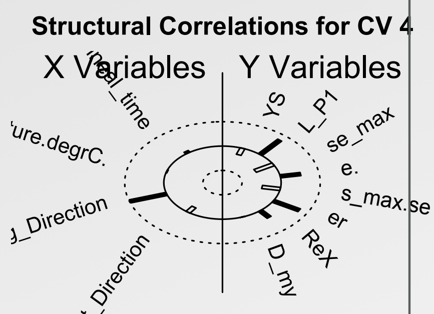
Canonical Variate1



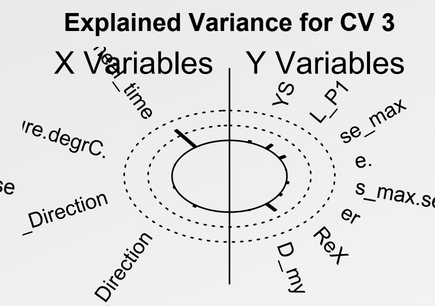
Canonical Variate2



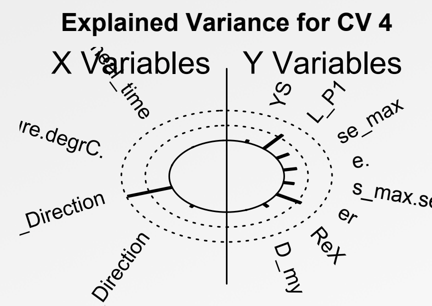
Canonical Variate3



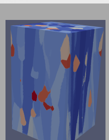
Canonical Variate4



Canonical Variate3



Canonical Variate4



CCA in R: circle or “helio” plots

CCA in R: the numbers

You can also type the name of the output dataset, here “ccares”, to get all this output:

```
Canonical Correlation Analysis

Canonical Correlations:
     CV 1      CV 2      CV 3      CV 4
0.9962597 0.9700839 0.9236971 0.7996017

X Coefficients:
                            CV 1          CV 2         CV 3         CV 4
Anneal_time         -0.0006349686 -0.008221356  0.028115903  0.005145620
Temperature.degrC.  -0.0093461287  0.001797585 -0.004406831 -0.000999307
Rolling_Direction   -0.0746291915 -0.504465534 -0.427833074  2.204475609
Test_Direction      -0.1248844862 -2.318961588 -0.863723232  0.146430906

Y Coefficients:
                 CV 1        CV 2         CV 3        CV 4
YS         0.028881576  0.01215613  0.009389696  0.16569611
L_P1      -0.140398526 -0.12340927 -0.008045269 -0.02628963
se_max    -0.026242062 -0.01873527 -0.004615684 -0.15381921
e.         0.031916944  0.09428803  0.298802031  0.12324637
s_max.se   0.024755042  0.03107842 -0.013281423  0.13067277
er         0.009540447 -0.01841757 -0.104140854  0.07363285
ReX        0.169052999  0.09136824 -7.559605691 -0.47708781
D_my      -0.236509432 -0.20259176  0.317418586  0.26674964
```

```
Structural Correlations (Loadings) - X Vars:
                          CV 1        CV 2         CV 3        CV 4
Anneal_time        -0.45853671 -0.28241766  0.838548257  0.08259038
Temperature.degrC. -0.99832200  0.05343568 -0.002871607 -0.02212593
Rolling_Direction  -0.01373942  0.11708491 -0.145841273  0.98225897
Test_Direction     -0.04519674 -0.93729575 -0.227480651 -0.26016625

Structural Correlations (Loadings) - Y Vars:
              CV 1        CV 2        CV 3         CV 4
YS       0.7224250 -0.52965623  0.23681354 -0.284652405
L_P1     0.1617876  0.20599013 -0.41777901  0.776076784
se_max   0.5606940 -0.26473525  0.37066197 -0.536592353
e.      -0.1923460  0.78878927 -0.01607383  0.525962537
s_max.se -0.4563148  0.65457757  0.22427023 -0.472436524
er      -0.1423697  0.47465970 -0.23168154  0.764658673
ReX     -0.6124121  0.38876123 -0.53960296  0.355374438
D_my    -0.9917832  0.05516687  0.08771473 -0.004104244

Aggregate Redundancy Coefficients (Total Variance Explained):
      X | Y: 0.861146
      Y | X: 0.7796086
```

This provides *some* of the numbers that you may wish to have. Note, e.g., the aggregated Redundancy Coefficients (bottom right), as well as the *coefficients* (LHS) and the *loadings* (RHS).

Remember that *linear* combinations of the input and output variables are what we get. The *Loadings* provide the coefficients.

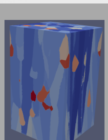
	CV 1
Anneal_time	-0.45853671
Temperature.degrC.	-0.99832200
Rolling_Direction	-0.01373942
Test_Direction	-0.04519674

	CV 1
YS	0.7224250
L_P1	0.1617876
se_max	0.5606940
e.	-0.1923460
s_max.se	-0.4563148
er	-0.1423697
ReX	-0.6124121
D_my	-0.9917832

The 1st block is for the input variables; the 2nd block is for the output variables.

In decreasing order of importance of the *input* variables, we have: Annealing temp., Anneal time, then Test direction, then Rolling direction.

The Annealing temperature is dominant – for the CV1 pair (but not for the other 3 CVs).



CCA in R: the first combination, i.e. CV1

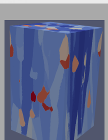
Remember that *linear* combinations of the input and output variables are what we get. The *Loadings* provide the coefficients.

	CV 1
Anneal_time	-0.45853671
Temperature.degrC.	-0.99832200
Rolling_Direction	-0.01373942
Test_Direction	-0.04519674

	CV 1
YS	0.7224250
L_P1	0.1617876
se_max	0.5606940
e.	-0.1923460
s_max.se	-0.4563148
er	-0.1423697
ReX	-0.6124121
D_my	-0.9917832

For the *output* variables, we have:
Grain size, fraction recrystallized, (negative) Yield strength, then (negative) Max. Eng. stress, then (negative) hardening etc.

The magnitudes offer useful clues as to which variables have the most influence (input) and which are the most sensitive (output).



CCA in R: the first combination, i.e. CV1

For additional numbers, values that were calculated, type summary(ccares)

```
Canonical Correlation Analysis - Summary


Canonical Correlations:

     CV 1      CV 2      CV 3      CV 4
0.9856295 0.9557033 0.8503614 0.6064473

Shared Variance on Each Canonical Variate:

     CV 1      CV 2      CV 3      CV 4
0.9714656 0.9133689 0.7231146 0.3677784

Bartlett's Chi-Squared Test:

        rho^2    Chisq df    Pr(>X)
CV 1  0.97147 85.19947 28 1.095e-07 ***
CV 2  0.91337 46.07639 18 0.0002892 ***
CV 3  0.72311 19.16933 10 0.0381639 *
CV 4  0.36778  5.04367  4 0.2828462
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Canonical Variate Coefficients:

      X Vars:
                                  CV 1         CV 2          CV 3          CV 4
Anneal_time                0.009040850 -0.008839731  0.0160612775 -0.0216132554
Anneal_Temperature_degrC -0.009698037  0.004081959 -0.0002298943  0.0004827524
Rolling_Direction          0.211199897  0.456271447 -1.6711080432 -1.5024705383
Test_Direction             0.876743388  2.288704128 -0.0328560195 -0.3907860553

      Y Vars:
                                 CV 1        CV 2        CV 3        CV 4
Yield_Stress               0.050951361 -0.07140689 -0.21501563  0.31787003
L_P1                      -0.051359196  0.13307434  0.03610953  0.12277281
Eng_Stress_max            -0.042068365  0.06922974  0.20256132 -0.29377099
Strain_at_peak_stress      0.049197647 -0.26235931  0.04351561 -0.10975734
stress_max-s_yield         0.030094283 -0.06659755 -0.19965679  0.32261589
Strain_to_failure          0.004225387  0.05582825 -0.12737839  0.06374365
per_cent_recrystallized   -1.124900635  3.61154239 -5.39739964  9.26162832
```

Structural Correlations (Loadings):

X Vars:

	CV 1	CV 2	CV 3	CV 4
Anneal_time	-0.08887783	-0.08560315	0.6441199	-0.754905508
Anneal_Temperature_degrC	-0.87567709	0.28694873	0.2413038	-0.304339509
Rolling_Direction	-0.04869550	-0.09228957	-0.8065009	-0.581951604
Test_Direction	0.36617810	0.88569375	0.2853964	0.003011926

Y Vars:

	CV 1	CV 2	CV 3	CV 4
Yield_Stress	0.906547517	0.07039372	0.2940504	0.21006405
L_P1	-0.003648265	-0.05721319	-0.8989282	-0.18498015
Eng_Stress_max	0.653377457	-0.13447373	0.5213911	0.37415434
Strain_at_peak_stress	-0.528216052	-0.58539486	-0.5277438	-0.19428978
stress_max-s_yield	-0.687062911	-0.44278984	0.4067414	0.32232930
Strain_to_failure	-0.362154467	-0.26075605	-0.7559954	-0.36492856
per_cent_recrystallized	-0.794671036	0.15426887	-0.5363385	-0.03395401

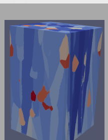
Fractional Variance Deposition on Canonical Variates:

X Vars:

	CV 1	CV 2	CV 3	CV 4
Anneal_time	0.007899269	0.007327899	0.41489051	5.698823e-01
Anneal_Temperature_degrC	0.766810358	0.082339573	0.05822753	9.262254e-02
Rolling_Direction	0.002371252	0.008517365	0.65044371	3.386677e-01
Test_Direction	0.134086398	0.784453426	0.08145110	9.071697e-06

Y Vars:

	CV 1	CV 2	CV 3	CV 4
Yield_Stress	8.218284e-01	0.004955275	0.08646565	0.044126904
L_P1	1.330984e-05	0.003273349	0.80807197	0.034217655
Eng_Stress_max	4.269021e-01	0.018083183	0.27184866	0.139991472
Strain_at_peak_stress	2.790122e-01	0.342687143	0.27851355	0.037748519
stress_max-s_yield	4.720554e-01	0.196062846	0.16543857	0.103896181
Strain_to_failure	1.311559e-01	0.067993719	0.57152900	0.133172855
per_cent_recrystallized	6.315021e-01	0.023798883	0.28765898	0.001152875



… Loadings & Fractional Variance on CVs …

```
Canonical Communalities (Fraction of Total Variance
Explained for Each Variable, Within Sets):

      X Vars:
             Anneal_time Anneal_Temperature_degrC        Rolling_Direction
Test_Direction
                       1                        1                        1
1

      Y Vars:
           Yield_Stress                    L_P1          Eng_Stress_max
Strain_at_peak_stress      stress_max-s_yield       Strain_to_failure
              0.9573762               0.8455763               0.8568254
0.9379614                0.9374530               0.9038514
per_cent_recrystallized
              0.9441128


Canonical Variate Adequacies (Fraction of Total Variance
Explained by Each CV, Within Sets):


      X Vars:
     CV 1      CV 2      CV 3      CV 4
0.2277918 0.2206596 0.3012532 0.2502954

      Y Vars:
      CV 1       CV 2       CV 3       CV 4
0.39463848 0.09383634 0.35278948 0.07061521
```

```
Redundancy Coefficients (Fraction of Total Variance
Explained by Each CV, Across Sets):


      X | Y:
      CV 1       CV 2       CV 3       CV 4
0.22129191 0.20154358 0.21784059 0.09205323

      Y | X:
      CV 1       CV 2       CV 3       CV 4
0.38337769 0.08570719 0.25510722 0.02597074


Aggregate Redundancy Coefficients (Total Variance
Explained by All CVs, Across Sets):

      X | Y: 0.7327293
      Y | X: 0.7501629
```

The redundancy coefficients (RHS) are used to make the CCA screeplots. In this case, at least, they are very similar to the Canonical Variate Adequacies (LHS)

… Communalities & Redundancies.

CCA is mathematically elegant but difficult to interpret because solutions are not unique.

A variate is interpreted by considering the pattern of variables that are highly correlated (loaded) with it. Variables in one set of the solution can be very sensitive to the identity of the variables in the other set; solutions are based upon correlation within and between sets, so a change in a variable in one set will likely alter the composition of the other set.

There is no implication of causation in solutions. The pairings of canonical variates must be independent of all other pairs.

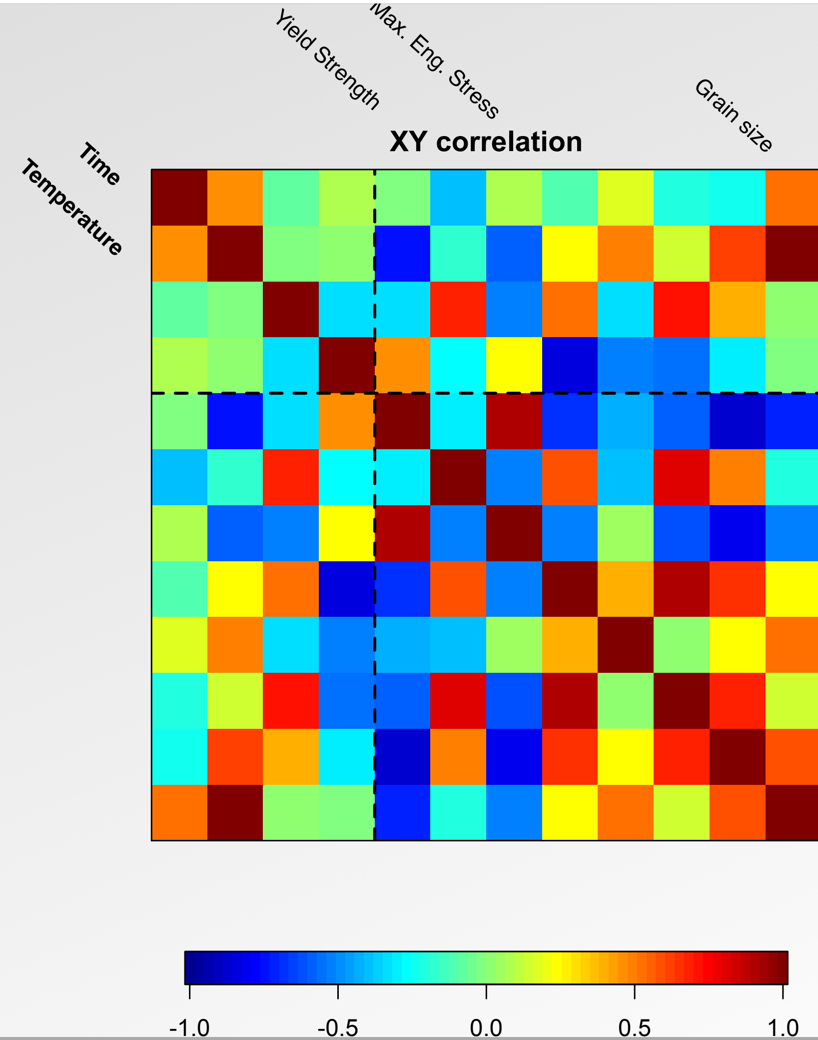
Only linear relationships are appropriate.

http://userwww.sfsu.edu/efc/classes/biol710/pca/CCandPCA2.htm

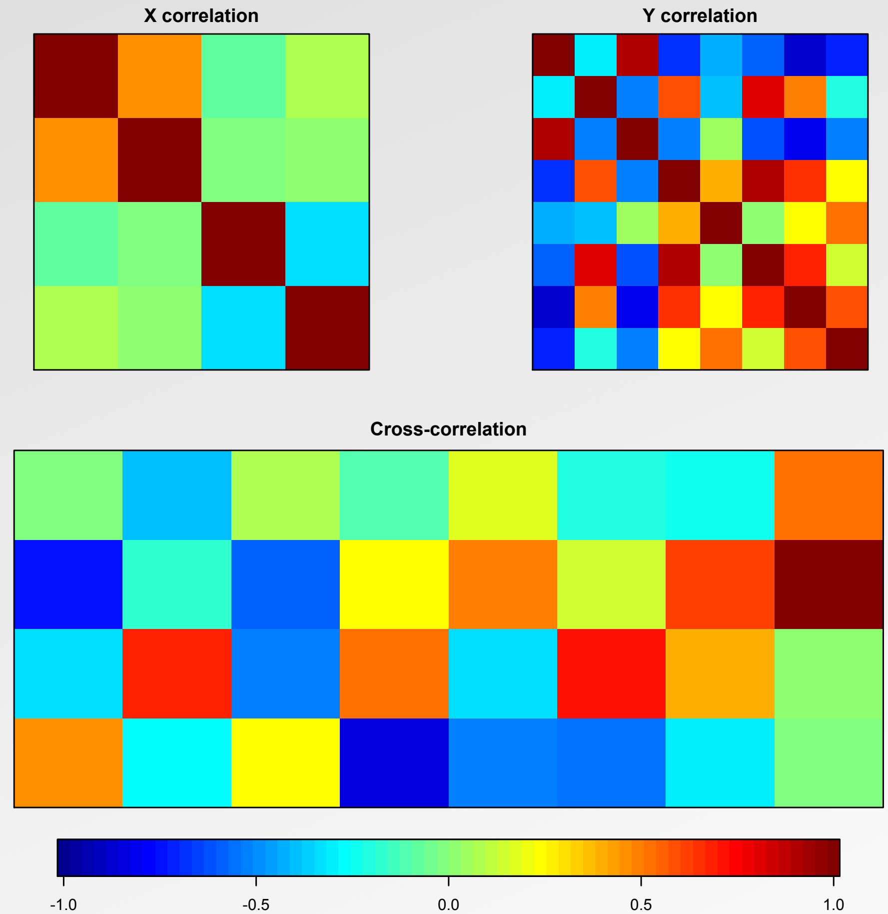
For a very straightforward, visual way to visualize correlations and cross-correlations, one can use *matcor*. The top right and bottom left quadrants are mirror images and show the cross-correlation between input & output.

```
> simpleCorr=matcor(invars,\
     outvars)
> img.matcor(simpleCorr)
```

You can also try using ggcorrplot



Here we show the 2nd type of *img.matcor* plot, which separates out the input and output variables, and shows the cross-correlation separately.




```
> simpleCorr=matcor(invars, outvars)
> img.matcor(simpleCorr, type=2)
```

R: matcor

```
scatterplotMatrix(allvars)
```

needs the "car" package

… illustrates why a general correlation plot is confusing!

R: scatterplotMatrix

There are two main packages for performing CCA:

One is in the *CCA* package and the procedure is called “cc”.

The other, illustrated in some detail here, is from the *yacca* package and the procedure is called “cca”.

All of which is somewhat confusing … but we recommend using “cca” from the *yacca* package because reading its output is more straightforward. For example all of the major items of interest such as *coefficients, loadings, commonalities*, and *redundancies* are easily found from using cca.

CCA package versus *yacca* package