Data Analytics for Materials Science

Canonical Correlation Analysis (CCA), part 2

Sudipto Mandal, Shivram Kashyap, Jacky Lao, Anthony Rollett

Materials Sci. & Eng., Carnegie Mellon University, Pittsburgh, PA.

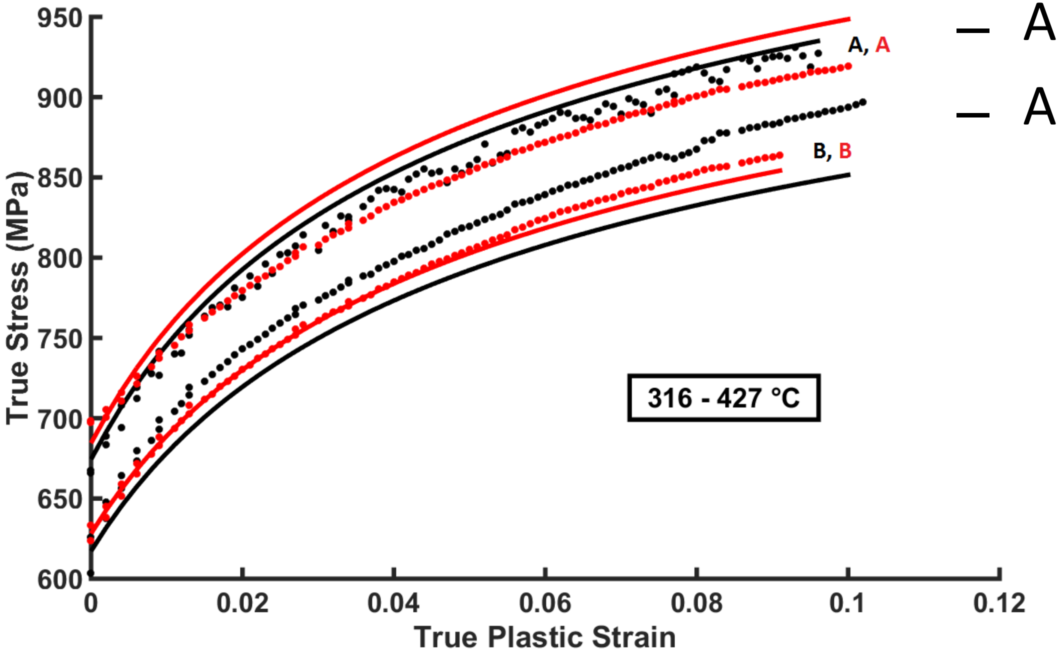
Revised by ADR:
Mar. 10th, 2021

Additions by RL:
6-7 September 2020

Constitutive Models

- Mechanical constitutive model: Relationship between stress and strain
- Predict deformation response to applied forces
- Mimic the actual material response subjected to different deformation conditions

Thesis work by Mandal (Ti5553) & Gockel (Ti6242)

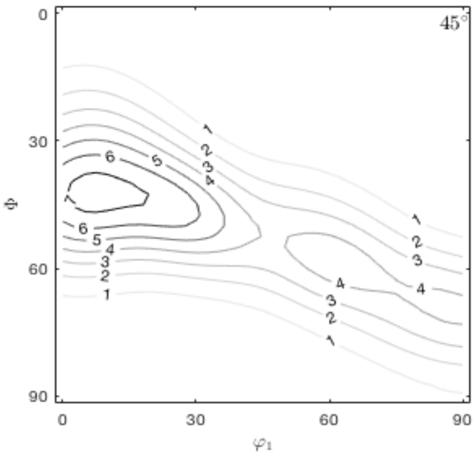


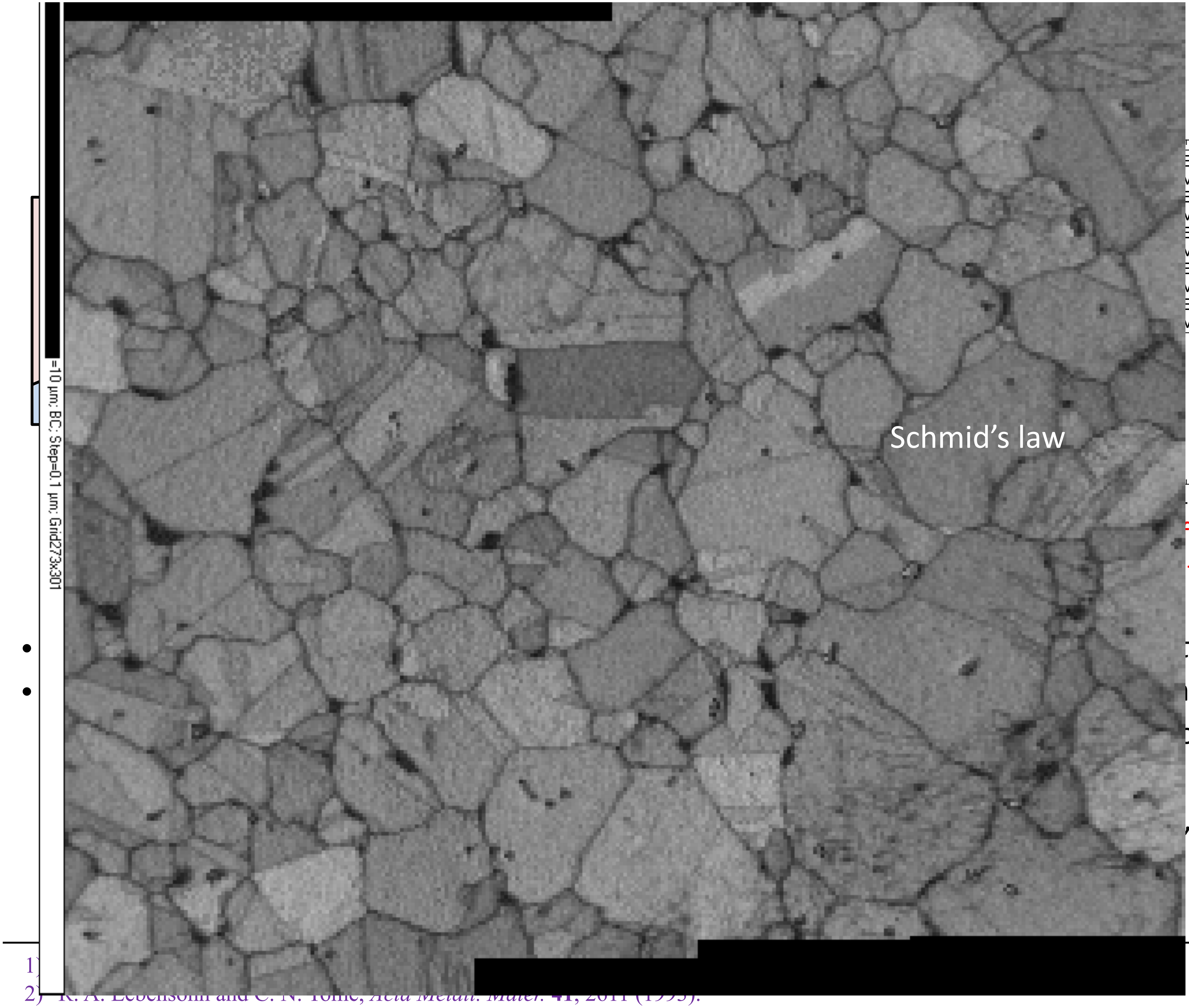
– Empirical and phenomenological models (Voce, JC)

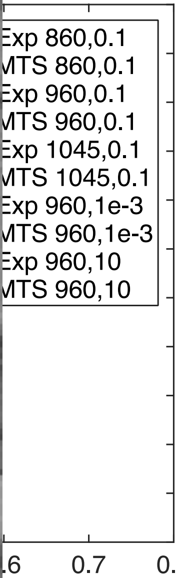
– Physics based models (MTS, ZA)

– Artificial Neural Networks

– Also, texture development







fferent

tes

- t evolves
- nperature:

ogical

, ZA)

1)

2) R. A. Lebensohn and C. N. Tomé, *Acta Metall. Mater.* **41**, 2611 (1993).

A problem with constitutive relations

Most work developing constitutive models is limited to finding the best fit parameters that capture a specific set of responses for a specific material under specific conditions — those models often (generally) lack transferability.

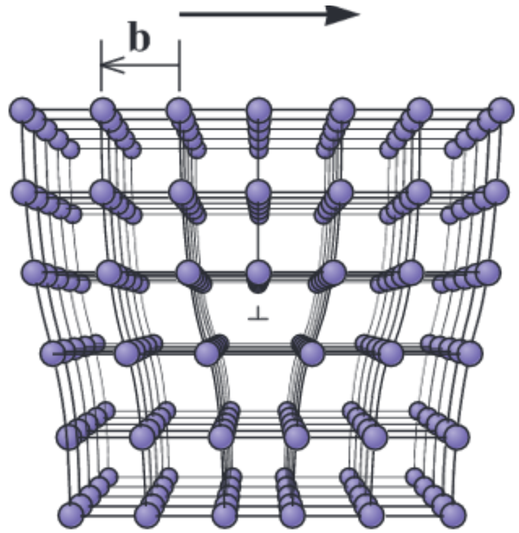
What is needed for efficient development of models is an understanding of how each parameter in the model affects the overall response of the model.

We used CCA to examine the *sensitivity* of the various parameters in a well-known, quite useful, but very complicated approach called the MTS model.

A Physics-based Model:
Mechanical Threshold Stress (MTS)

- Plastic deformation accommodated by dislocation motion
- Mechanical Threshold Stress (MTS):

  Flow stress at 0 K



$$\frac{\tau}{\mu} = \frac{\tau_a}{\mu} + S_i\left(\dot{\varepsilon},T\right)\frac{\hat{\tau}_i}{\mu_0} + S_\varepsilon\left(\dot{\varepsilon},T\right)\frac{\hat{\tau}_\varepsilon}{\mu_0}$$

Athermal stress: interaction of dislocations with grain boundaries. Grain size dependence:

$$\tau_a = k_y \big/ \sqrt{d_{gs}}$$

Rate dependent interaction of dislocations with obstacle populations that can be overcome with the assistance of thermal activation like interstitial atoms (S_i) & stored dislocations (S_ε)

Evolution

considered when there is increase of stored dislocation density with straining

$$\frac{d\hat{\tau}_\varepsilon}{d\varepsilon} = \theta_0 \frac{\mu}{\mu_0}\left[1 - \frac{\hat{\tau}_\varepsilon}{\hat{\tau}_{\varepsilon s}}\right]^\kappa$$

$$\ln\left(\frac{\hat{\tau}_{\varepsilon s}}{\hat{\tau}_{\varepsilon s0}}\right) = \frac{KT}{\mu b^3 g_{0\varepsilon s}} \ln\left(\frac{\varepsilon\boxed{?}}{\varepsilon\boxed{?}_{0\varepsilon s}}\right)$$

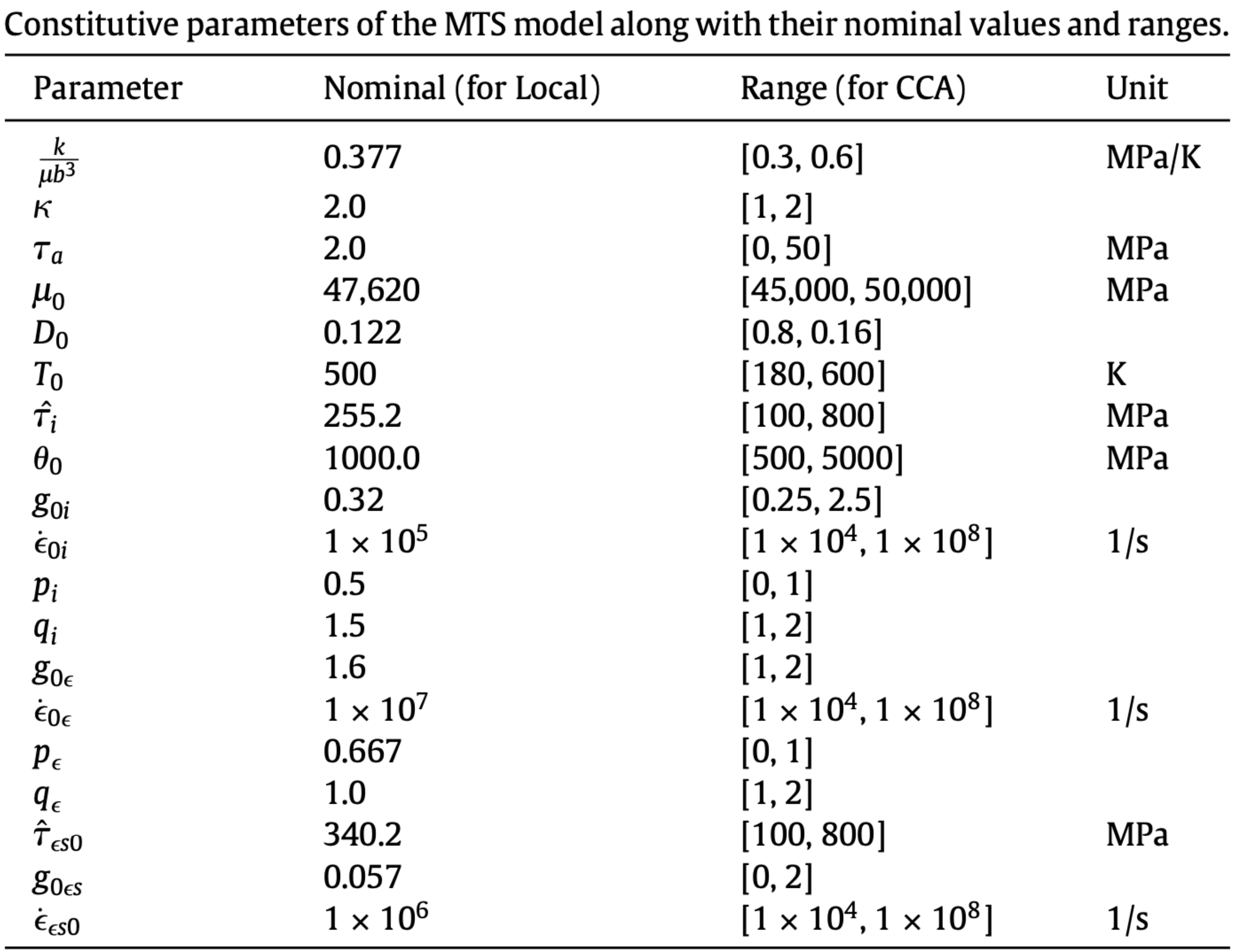
Scaling factor to reference term

$$S_i(\dot{\varepsilon},T) = \left\{1 - \left[\frac{kT}{\mu b^3 g_{0i}} \ln\left(\frac{\dot{\varepsilon}_{0i}}{\dot{\varepsilon}}\right)\right]^{1/q_i}\right\}^{1/p_i}$$

19 parameters, i.e., a large number

1) P. S. Follansbee, Fundamentals of Strength (John Wiley & Sons, Inc., 2014).

MTS parameters

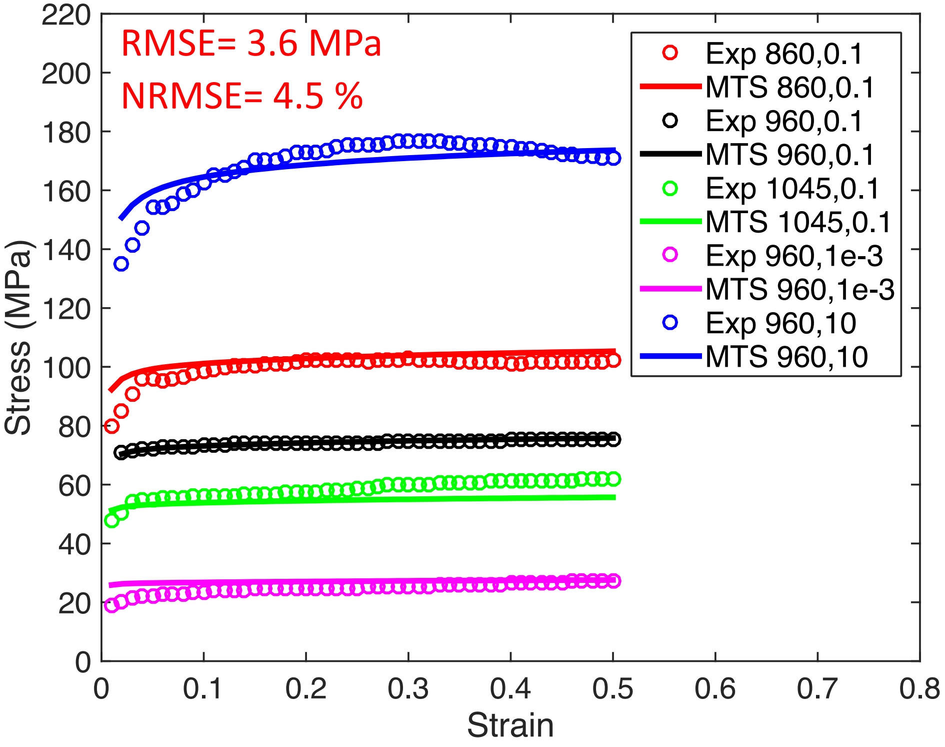


Constitutive parameters of the MTS model along with their nominal values and ranges.

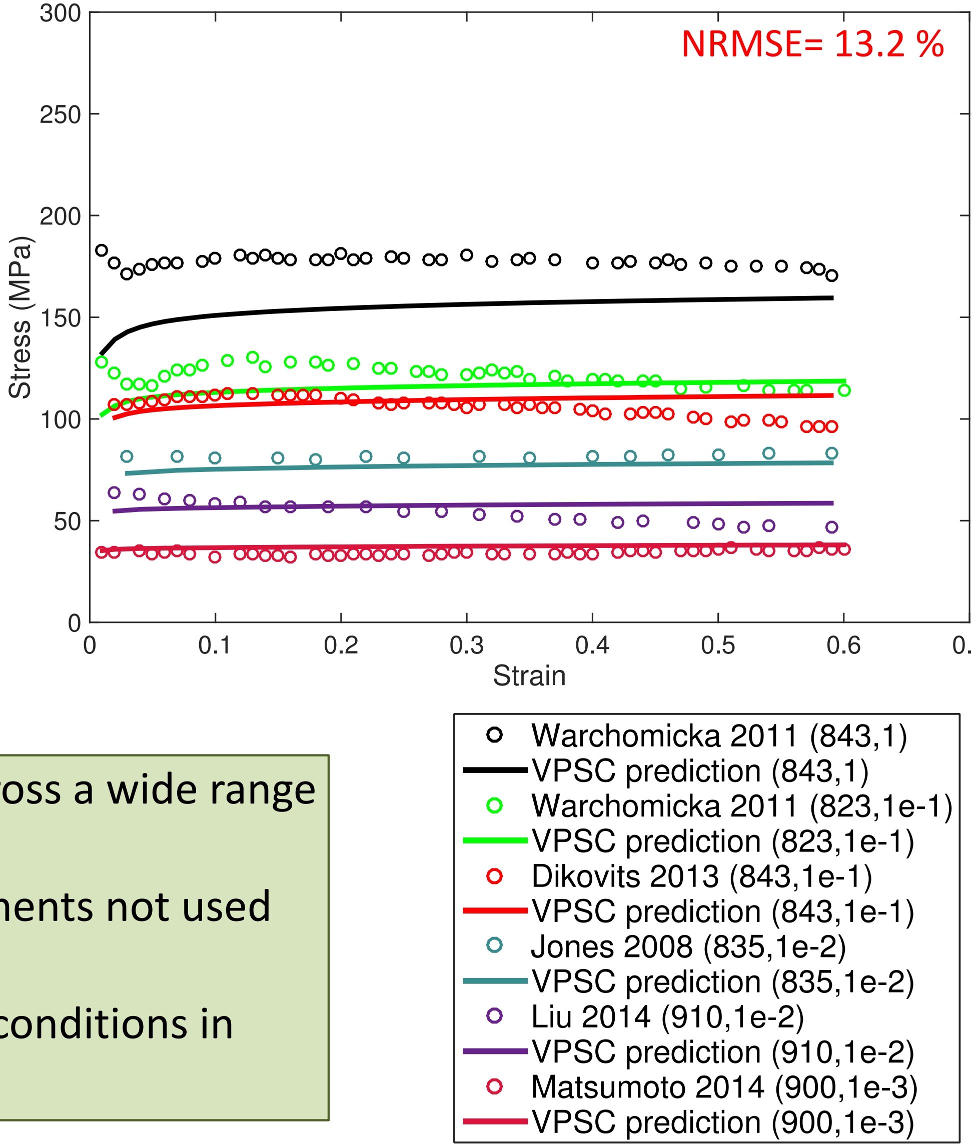
Parameter	Nominal (for Local)	Range (for CCA)	Unit
$\frac{k}{\mu b^3}$	0.377	[0.3, 0.6]	MPa/K
κ	2.0	[1, 2]	
τ_a	2.0	[0, 50]	MPa
μ_0	47,620	[45,000, 50,000]	MPa
D_0	0.122	[0.8, 0.16]	
T_0	500	[180, 600]	K
$\hat{\tau}_i$	255.2	[100, 800]	MPa
θ_0	1000.0	[500, 5000]	MPa
g_{0i}	0.32	[0.25, 2.5]	
$\dot{\epsilon}_{0i}$	1×10^5	$[1 \times 10^4, 1 \times 10^8]$	1/s
p_i	0.5	[0, 1]	
q_i	1.5	[1, 2]	
$g_{0\epsilon}$	1.6	[1, 2]	
$\dot{\epsilon}_{0\epsilon}$	1×10^7	$[1 \times 10^4, 1 \times 10^8]$	1/s
p_ϵ	0.667	[0, 1]	
q_ϵ	1.0	[1, 2]	
$\hat{\tau}_{\epsilon s0}$	340.2	[100, 800]	MPa
$g_{0\epsilon s}$	0.057	[0, 2]	
$\dot{\epsilon}_{\epsilon s0}$	1×10^6	$[1 \times 10^4, 1 \times 10^8]$	1/s

MTS Model for beta-Ti Alloys

Training



Testing



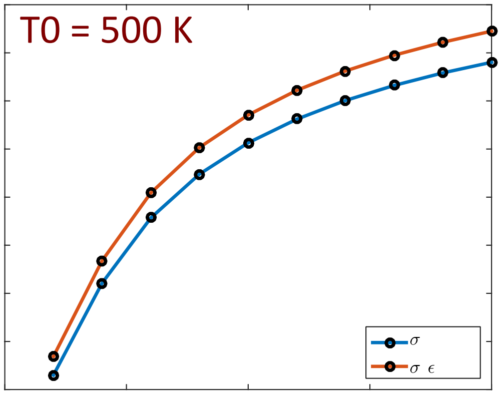
- **Single set** of parameters captures response across a wide range of conditions in single phase regime
- Validated with Ti-5553 experimental measurements not used to train the parameters
- Model **extrapolated** to corresponding loading conditions in **similar β-Ti alloys** from literature.

"Simulation of plastic deformation in Ti-5553 alloy using a self-consistent viscoplastic model", Mandal *et al*. (2017), *Int. J Plasticity,* **94** 54-73.

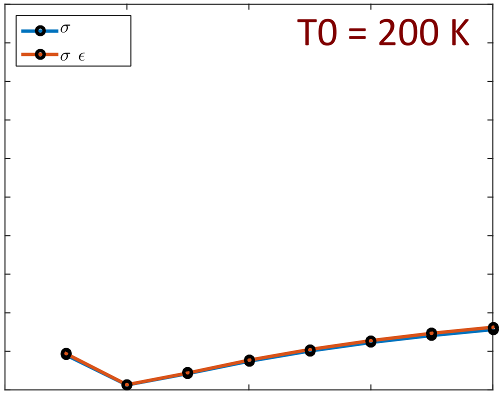
Why Sensitivity Analysis?

- Identification of important parameters is quite challenging, especially for complex models like MTS with many parameters
- Sensitivity analysis before calibration leads to simplification of the modeling, the spotting of errors, better extrapolation and understanding of the model
- **Local sensitivity analysis**: One-at-a-time (OAT) approach to obtain partial derivatives
- Local trend near the nominal values in parameter space may be different from global trends -> Motivates **global analysis**

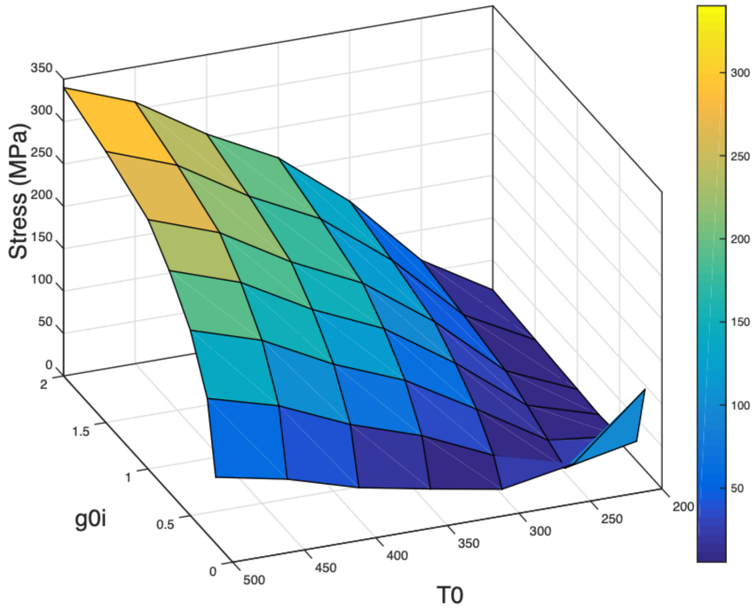
Example: How does yield stress change with the parameter g0i (activation barrier)?



All other parameters fixed to nominal values



Change one other parameter T0



Interaction between 2 parameters; Much more complicated for 19!

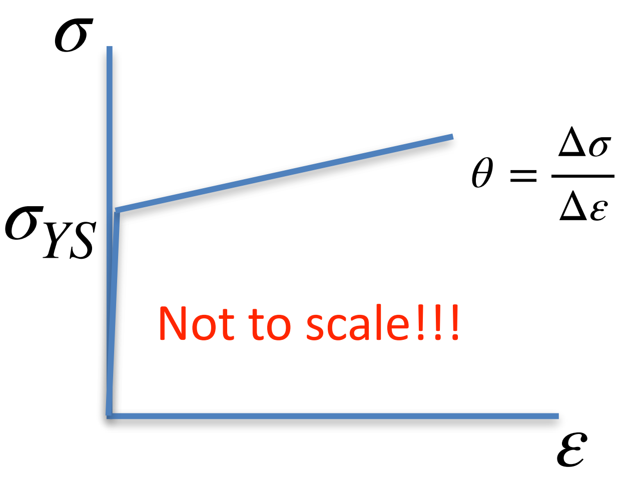
Mandal, Gockel, Rollett, *Materials Design* **132** 30 (2017)

The calculations

- Monte Carlo used to randomly choose parameters in their specified ranges (**input**)

- Do three polycrystal plasticity calculations for **each set of parameters** (**experiments**)
  1. Temperature = 1300 K, Strain rate= 1 s^{-1}
  2. Temperature = 1200 K, Strain rate= 1 s^{-1}
  3. Temperature = 1200 K, Strain rate= 10 s^{-1}

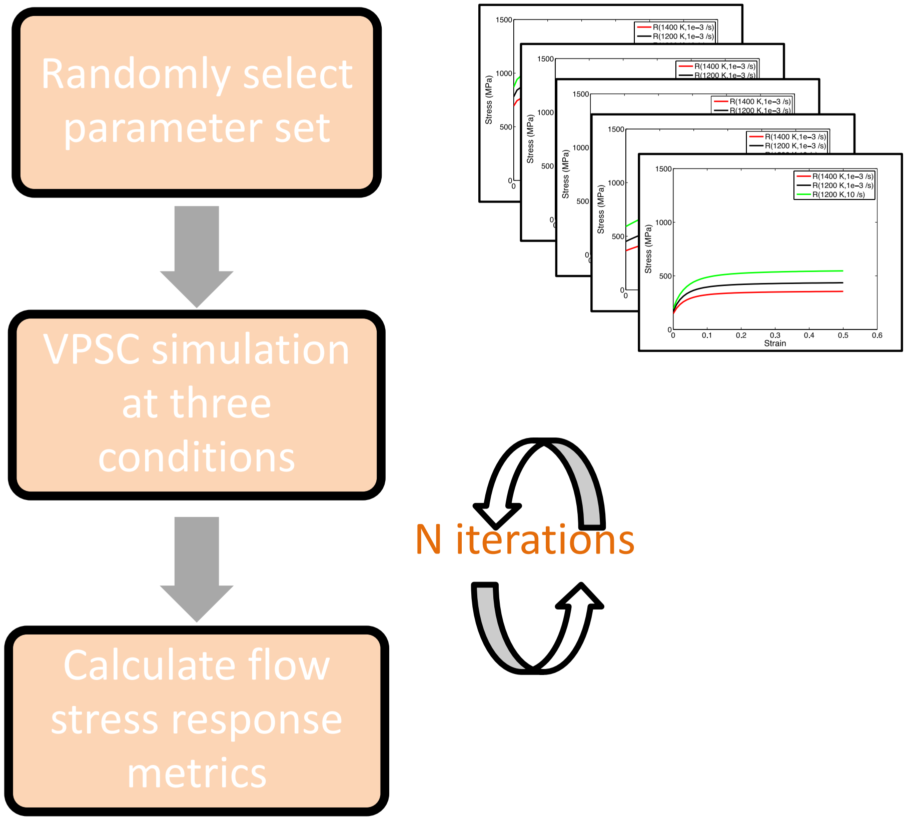
- Define response parameters: (**output**: 8 output parameters)
  – Yield Stresses for each calculation ($\sigma_{YS1}, \sigma_{YS2}, \sigma_{YS2}$)
  – Hardening: $\Delta\sigma_i = \left(\dfrac{\sigma_{i(\epsilon=0.5)} - \sigma_{i(\epsilon=0.1)}}{\sigma_{i(\epsilon=0.1)}} \right)$
  – Temperature sensitivity: $T_{sen} = \left\langle \dfrac{\sigma_{\epsilon 1} - \sigma_{\epsilon 2}}{\sigma_{\epsilon 2}} \right\rangle_{\epsilon}$
  – Rate sensitivity: $R_{sen} = \left\langle \dfrac{\sigma_{\epsilon 3} - \sigma_{\epsilon 2}}{\sigma_{\epsilon 2}} \right\rangle_{\epsilon}$



Not to scale!!!

Mandal, Gockel, Rollett, *Materials Design* **132** 30 (2017)

Global Sensitivity Analysis



Randomly select parameter set

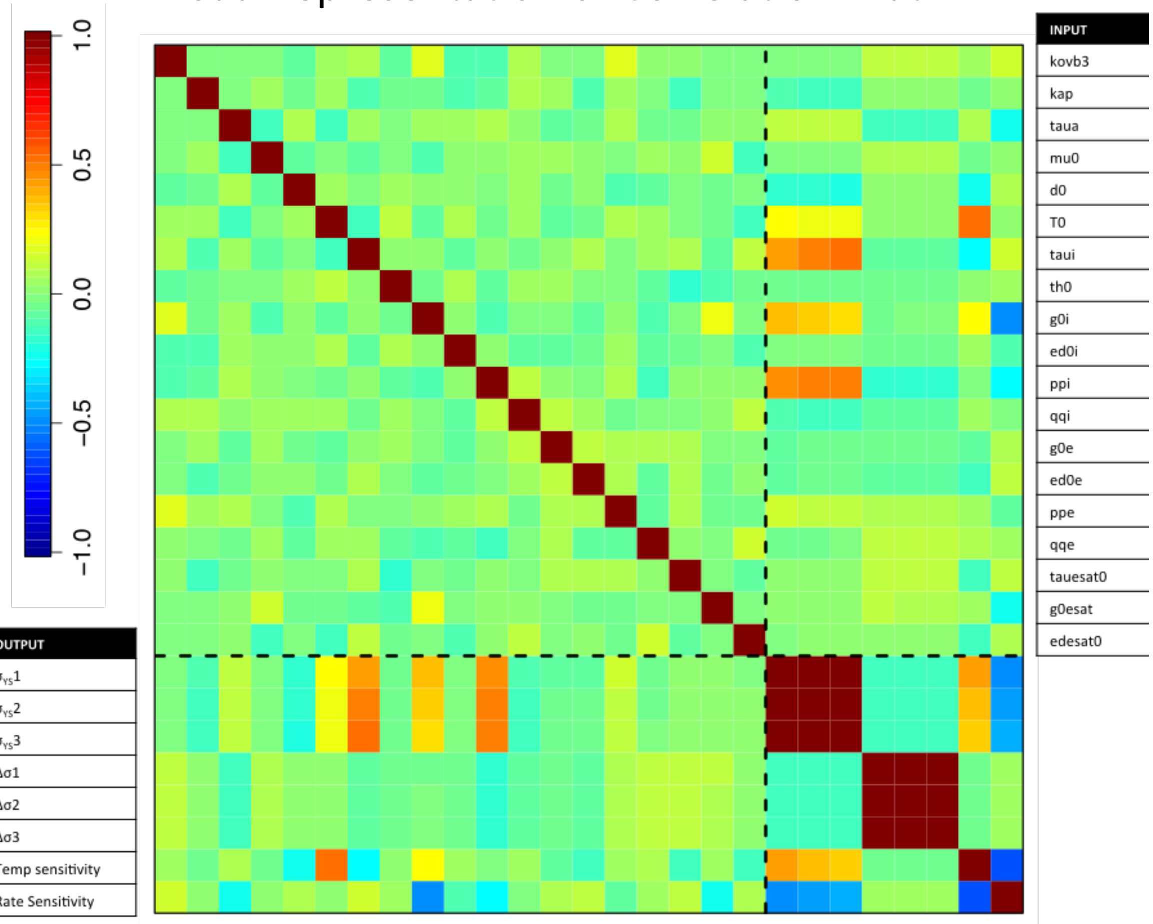
VPSC simulation at three conditions

Calculate flow stress response metrics

N iterations

$$Cor(x,y) = r_{XY} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}}$$

Visual representation of Correlation Matrix

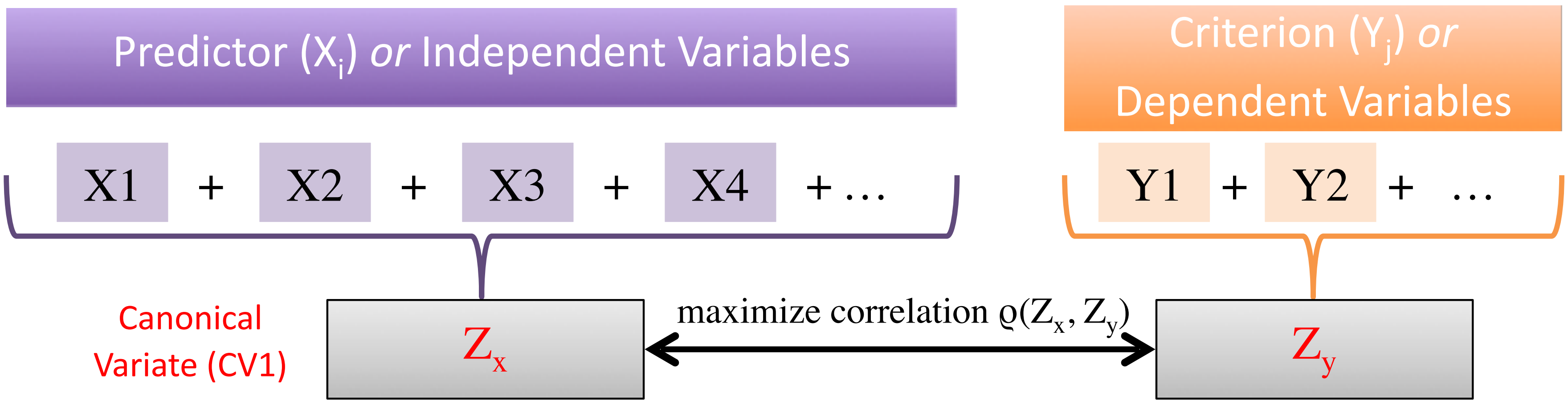


Flow stress response metrics

- Yield Stresses
- Hardening $(\sigma_{0.5}-\sigma_{0.1})/\sigma_{0.1}$
- Temperature sensitivity $(\sigma_2-\sigma_1)/\sigma_1$
- Rate sensitivity $(\sigma_3-\sigma_2)/\sigma_2$

Mandal, Gockel, Rollett, *Materials Design* **132** 30 (2017)

Canonical Correlation Analysis


Predictor (X_i) *or* Independent Variables
Criterion (Y_j) *or* Dependent Variables
X1 + X2 + X3 + X4 + …
Y1 + Y2 + …
Canonical Variate (CV1)
Z_x
maximize correlation $\varrho(Z_x, Z_y)$
Z_y

CV2: Explains residual variance from CV1; CV3 explains residual from CV2 and so on

Canonical Correlation Analysis (CCA): Multivariate statistical technique to study of relationships between multiple dependent & independent variables.

When & Why use CCA?

- Data can be logically split into two sets
- To check if two sets of variables are related
- To find (linear) interrelationships within and between set variables
- Presence of multicollinearity
- Dimensionality reduction; builds on Principal Component Analysis (PCA)
- To quantify variable importance in a model
- To check if a set of variables at one time step can be used to predicting variables at the next step, i.e., temporal prediction
- For extension to non-linear, see Rickman *et al. (Nature) Comp. Matls.* **3** 26 (2017)

Canonical Correlation Analysis (CCA)

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

- assume *p* variables in **X** and *q* variables in **Y**
- **X** is thus an *N* x *p* dimensional matrix and **Y** is an *N* x *q* dimensional matrix

CCA provides a way to find the maximum *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.

CCA: steps

Create autoscaled matrices:

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

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

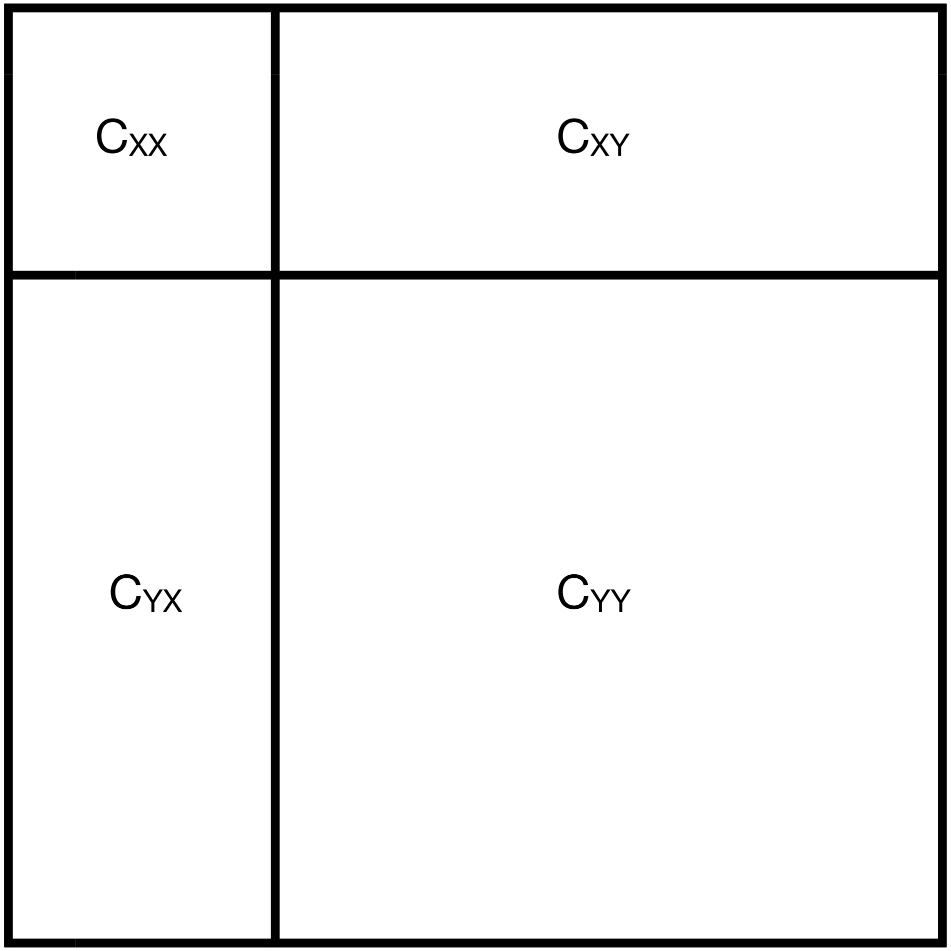
$$\mathbf{Y} = \{\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_{10}'\}$$

Calculate correlation matrices:

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

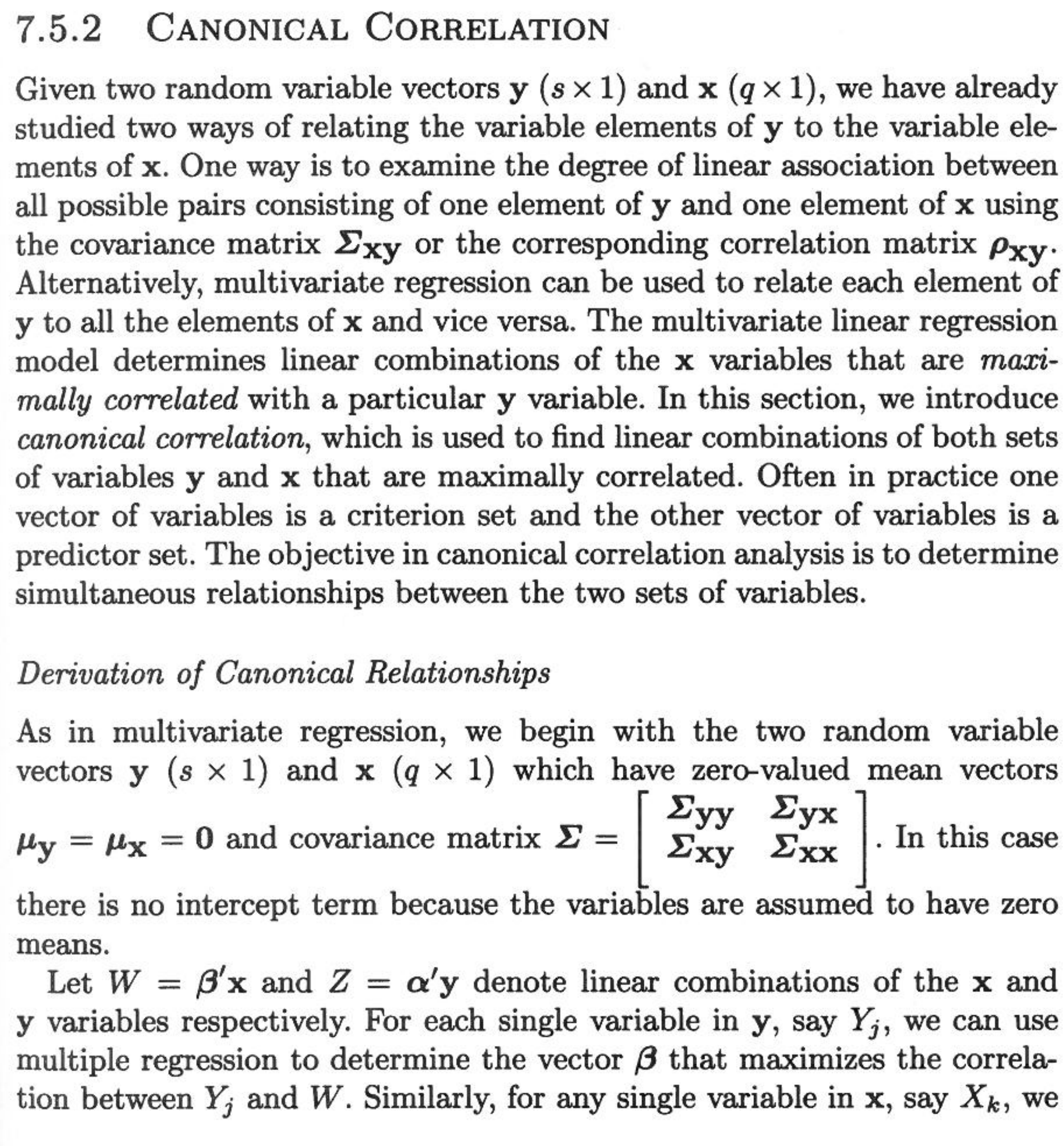
$$\mathbf{C}_{\mathbf{XY}} = \frac{\mathbf{X}^T\mathbf{Y}}{N-1} \qquad \mathbf{C}_{\mathbf{YX}} = \frac{\mathbf{Y}^T\mathbf{X}}{N-1}$$

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



C_{XX}	C_{XY}
C_{YX}	C_{YY}

From Jobson Vol. 2



7.5.2 Canonical Correlation

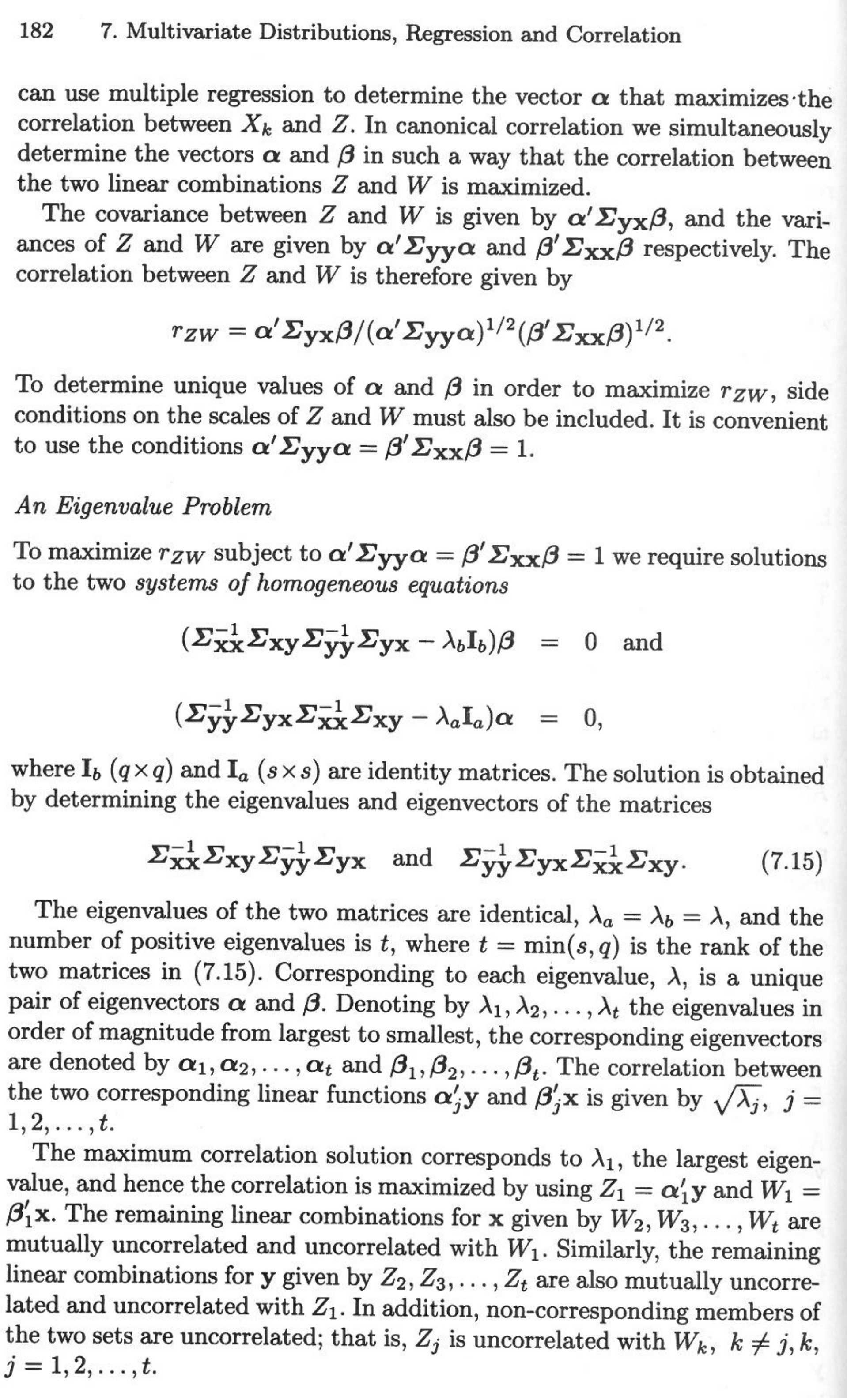
Given two random variable vectors $\mathbf{y}$ $(s \times 1)$ and $\mathbf{x}$ $(q \times 1)$, we have already studied two ways of relating the variable elements of $\mathbf{y}$ to the variable elements of $\mathbf{x}$. One way is to examine the degree of linear association between all possible pairs consisting of one element of $\mathbf{y}$ and one element of $\mathbf{x}$ using the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{xy}}$ or the corresponding correlation matrix $\boldsymbol{\rho}_{\mathbf{xy}}$. Alternatively, multivariate regression can be used to relate each element of $\mathbf{y}$ to all the elements of $\mathbf{x}$ and vice versa. The multivariate linear regression model determines linear combinations of the $\mathbf{x}$ variables that are *maximally correlated* with a particular $\mathbf{y}$ variable. In this section, we introduce *canonical correlation*, which is used to find linear combinations of both sets of variables $\mathbf{y}$ and $\mathbf{x}$ that are maximally correlated. Often in practice one vector of variables is a criterion set and the other vector of variables is a predictor set. The objective in canonical correlation analysis is to determine simultaneous relationships between the two sets of variables.

Derivation of Canonical Relationships

As in multivariate regression, we begin with the two random variable vectors $\mathbf{y}$ $(s \times 1)$ and $\mathbf{x}$ $(q \times 1)$ which have zero-valued mean vectors $\boldsymbol{\mu}_{\mathbf{y}} = \boldsymbol{\mu}_{\mathbf{x}} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{yy}} & \boldsymbol{\Sigma}_{\mathbf{yx}} \\ \boldsymbol{\Sigma}_{\mathbf{xy}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{bmatrix}$. In this case there is no intercept term because the variables are assumed to have zero means.

Let $W = \boldsymbol{\beta}'\mathbf{x}$ and $Z = \boldsymbol{\alpha}'\mathbf{y}$ denote linear combinations of the $\mathbf{x}$ and $\mathbf{y}$ variables respectively. For each single variable in $\mathbf{y}$, say Y_j, we can use multiple regression to determine the vector $\boldsymbol{\beta}$ that maximizes the correlation between Y_j and W. Similarly, for any single variable in $\mathbf{x}$, say X_k, we

From Jobson Vol. 2



can use multiple regression to determine the vector $\boldsymbol{\alpha}$ that maximizes the correlation between X_k and Z. In canonical correlation we simultaneously determine the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in such a way that the correlation between the two linear combinations Z and W is maximized.

The covariance between Z and W is given by $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\beta}$, and the variances of Z and W are given by $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathbf{yy}}\boldsymbol{\alpha}$ and $\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta}$ respectively. The correlation between Z and W is therefore given by

$$r_{ZW} = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\beta}/(\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathbf{yy}}\boldsymbol{\alpha})^{1/2}(\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta})^{1/2}.$$

To determine unique values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in order to maximize r_{ZW}, side conditions on the scales of Z and W must also be included. It is convenient to use the conditions $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathbf{yy}}\boldsymbol{\alpha} = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta} = 1$.

An Eigenvalue Problem

To maximize r_{ZW} subject to $\boldsymbol{\alpha}'\boldsymbol{\Sigma}_{\mathbf{yy}}\boldsymbol{\alpha} = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{xx}}\boldsymbol{\beta} = 1$ we require solutions to the two *systems of homogeneous equations*

$$\begin{aligned}(\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}} - \lambda_b\mathbf{I}_b)\boldsymbol{\beta} &= 0 \quad \text{and} \\ (\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}} - \lambda_a\mathbf{I}_a)\boldsymbol{\alpha} &= 0,\end{aligned}$$

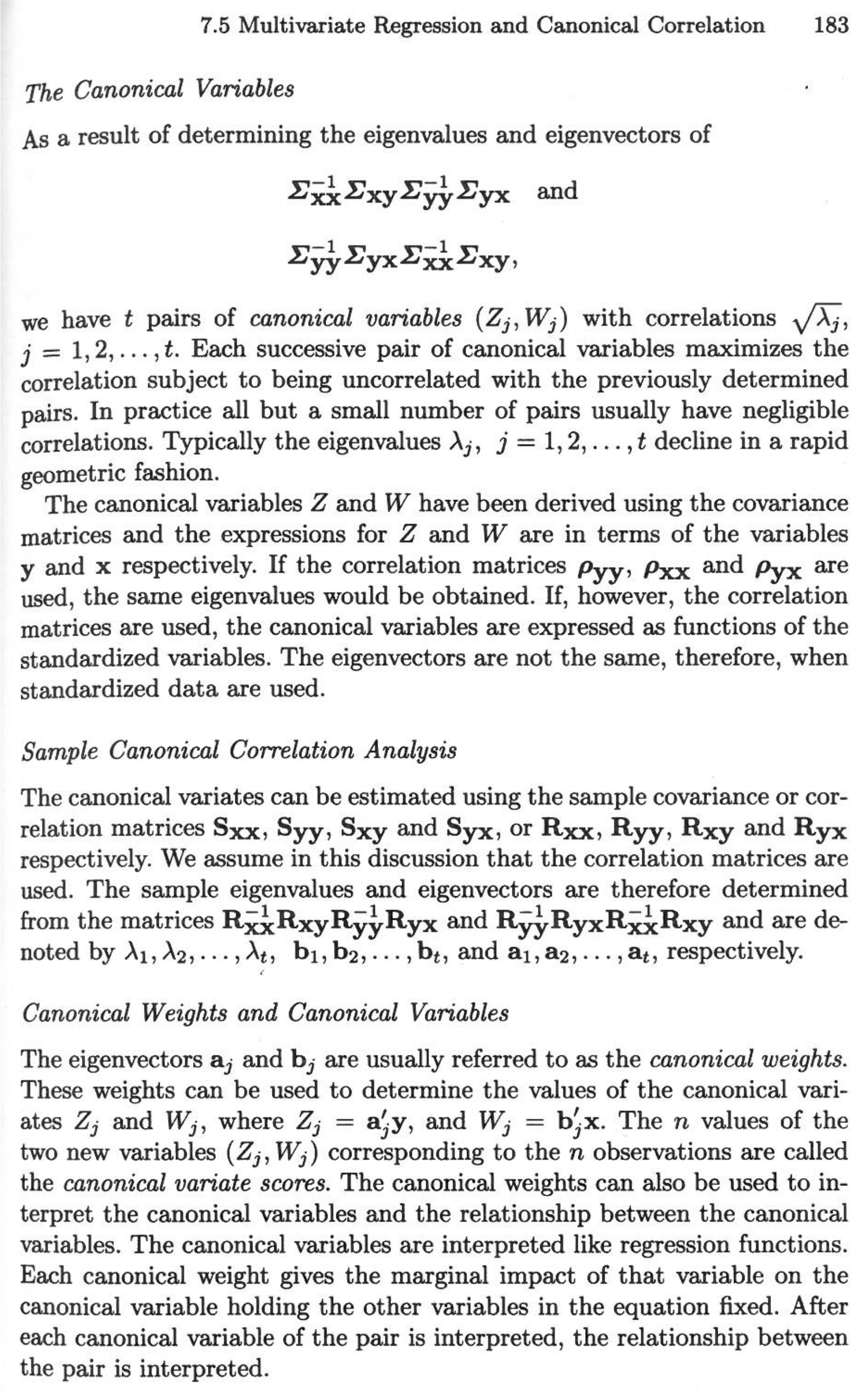
where $\mathbf{I}_b$ $(q \times q)$ and $\mathbf{I}_a$ $(s \times s)$ are identity matrices. The solution is obtained by determining the eigenvalues and eigenvectors of the matrices

$$\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}. \tag{7.15}$$

The eigenvalues of the two matrices are identical, $\lambda_a = \lambda_b = \lambda$, and the number of positive eigenvalues is t, where $t = \min(s, q)$ is the rank of the two matrices in (7.15). Corresponding to each eigenvalue, λ, is a unique pair of eigenvectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Denoting by $\lambda_1, \lambda_2, \ldots, \lambda_t$ the eigenvalues in order of magnitude from largest to smallest, the corresponding eigenvectors are denoted by $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_t$. The correlation between the two corresponding linear functions $\boldsymbol{\alpha}_j'\mathbf{y}$ and $\boldsymbol{\beta}_j'\mathbf{x}$ is given by $\sqrt{\lambda_j}$, $j = 1, 2, \ldots, t$.

The maximum correlation solution corresponds to λ_1, the largest eigenvalue, and hence the correlation is maximized by using $Z_1 = \boldsymbol{\alpha}_1'\mathbf{y}$ and $W_1 = \boldsymbol{\beta}_1'\mathbf{x}$. The remaining linear combinations for $\mathbf{x}$ given by $W_2, W_3, \ldots, W_t$ are mutually uncorrelated and uncorrelated with W_1. Similarly, the remaining linear combinations for $\mathbf{y}$ given by $Z_2, Z_3, \ldots, Z_t$ are also mutually uncorrelated and uncorrelated with Z_1. In addition, non-corresponding members of the two sets are uncorrelated; that is, Z_j is uncorrelated with W_k, $k \neq j, k$, $j = 1, 2, \ldots, t$.

From Jobson Vol. 2



The Canonical Variables

As a result of determining the eigenvalues and eigenvectors of

$$\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}} \quad \text{and}$$

$$\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\Sigma}_{\mathbf{xy}},$$

we have t pairs of *canonical variables* (Z_j, W_j) with correlations $\sqrt{\lambda_j}$, $j = 1, 2, \ldots, t$. Each successive pair of canonical variables maximizes the correlation subject to being uncorrelated with the previously determined pairs. In practice all but a small number of pairs usually have negligible correlations. Typically the eigenvalues λ_j, $j = 1, 2, \ldots, t$ decline in a rapid geometric fashion.

The canonical variables Z and W have been derived using the covariance matrices and the expressions for Z and W are in terms of the variables $\mathbf{y}$ and $\mathbf{x}$ respectively. If the correlation matrices $\boldsymbol{\rho}_{\mathbf{yy}}$, $\boldsymbol{\rho}_{\mathbf{xx}}$ and $\boldsymbol{\rho}_{\mathbf{yx}}$ are used, the same eigenvalues would be obtained. If, however, the correlation matrices are used, the canonical variables are expressed as functions of the standardized variables. The eigenvectors are not the same, therefore, when standardized data are used.

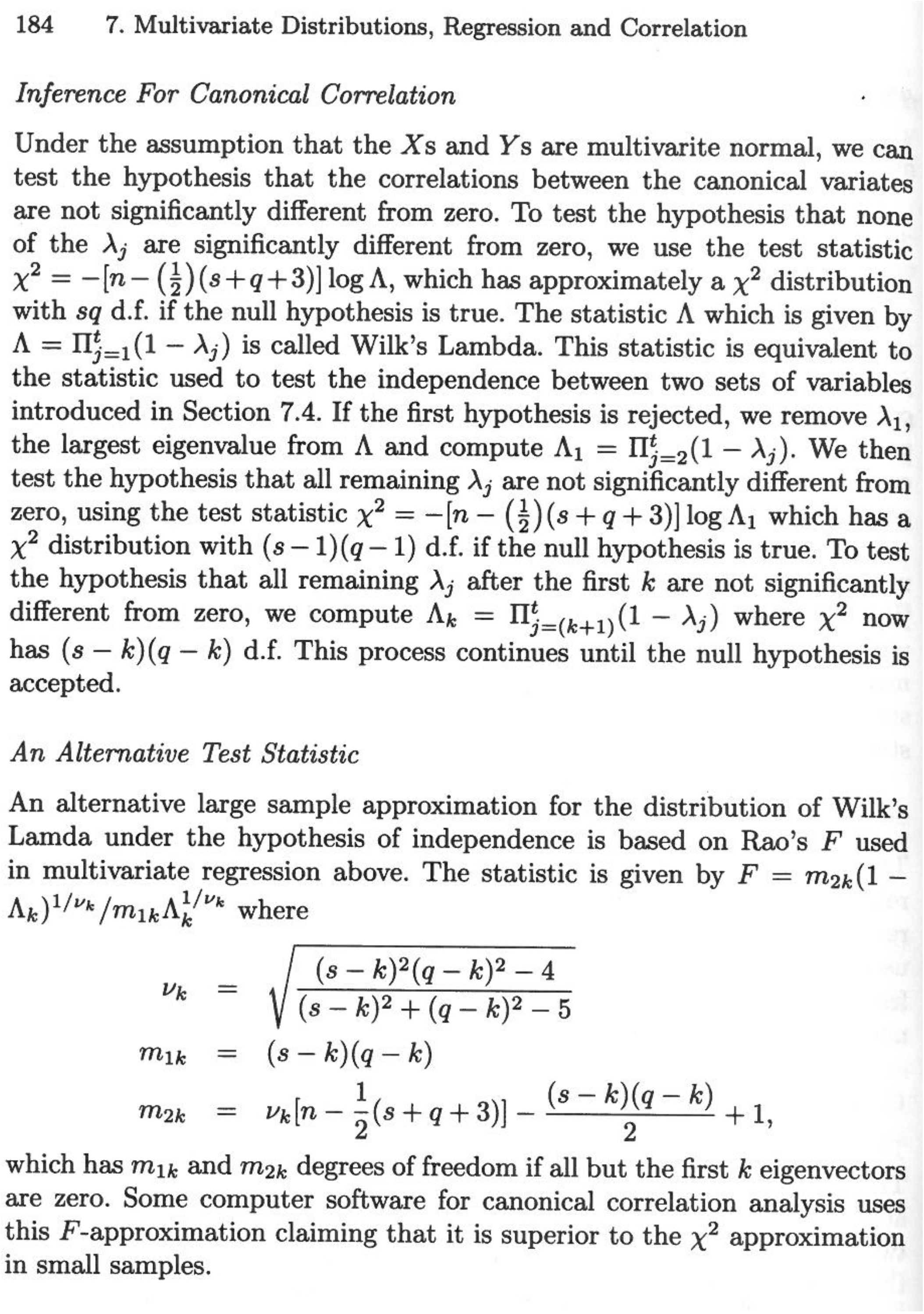
Sample Canonical Correlation Analysis

The canonical variates can be estimated using the sample covariance or correlation matrices $\mathbf{S_{xx}}$, $\mathbf{S_{yy}}$, $\mathbf{S_{xy}}$ and $\mathbf{S_{yx}}$, or $\mathbf{R_{xx}}$, $\mathbf{R_{yy}}$, $\mathbf{R_{xy}}$ and $\mathbf{R_{yx}}$ respectively. We assume in this discussion that the correlation matrices are used. The sample eigenvalues and eigenvectors are therefore determined from the matrices $\mathbf{R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}}$ and $\mathbf{R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}}$ and are denoted by $\lambda_1, \lambda_2, \ldots, \lambda_t$, $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_t$, and $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_t$, respectively.

Canonical Weights and Canonical Variables

The eigenvectors $\mathbf{a}_j$ and $\mathbf{b}_j$ are usually referred to as the *canonical weights*. These weights can be used to determine the values of the canonical variates Z_j and W_j, where $Z_j = \mathbf{a}_j'\mathbf{y}$, and $W_j = \mathbf{b}_j'\mathbf{x}$. The n values of the two new variables (Z_j, W_j) corresponding to the n observations are called the *canonical variate scores*. The canonical weights can also be used to interpret the canonical variables and the relationship between the canonical variables. The canonical variables are interpreted like regression functions. Each canonical weight gives the marginal impact of that variable on the canonical variable holding the other variables in the equation fixed. After each canonical variable of the pair is interpreted, the relationship between the pair is interpreted.

From Jobson Vol. 2



Inference For Canonical Correlation

Under the assumption that the Xs and Ys are multivarite normal, we can test the hypothesis that the correlations between the canonical variates are not significantly different from zero. To test the hypothesis that none of the λ_j are significantly different from zero, we use the test statistic $\chi^2 = -[n - (\frac{1}{2})(s+q+3)] \log \Lambda$, which has approximately a χ^2 distribution with sq d.f. if the null hypothesis is true. The statistic Λ which is given by $\Lambda = \Pi_{j=1}^{t}(1-\lambda_j)$ is called Wilk's Lambda. This statistic is equivalent to the statistic used to test the independence between two sets of variables introduced in Section 7.4. If the first hypothesis is rejected, we remove λ_1, the largest eigenvalue from Λ and compute $\Lambda_1 = \Pi_{j=2}^{t}(1-\lambda_j)$. We then test the hypothesis that all remaining λ_j are not significantly different from zero, using the test statistic $\chi^2 = -[n - (\frac{1}{2})(s+q+3)] \log \Lambda_1$ which has a χ^2 distribution with $(s-1)(q-1)$ d.f. if the null hypothesis is true. To test the hypothesis that all remaining λ_j after the first k are not significantly different from zero, we compute $\Lambda_k = \Pi_{j=(k+1)}^{t}(1-\lambda_j)$ where χ^2 now has $(s-k)(q-k)$ d.f. This process continues until the null hypothesis is accepted.

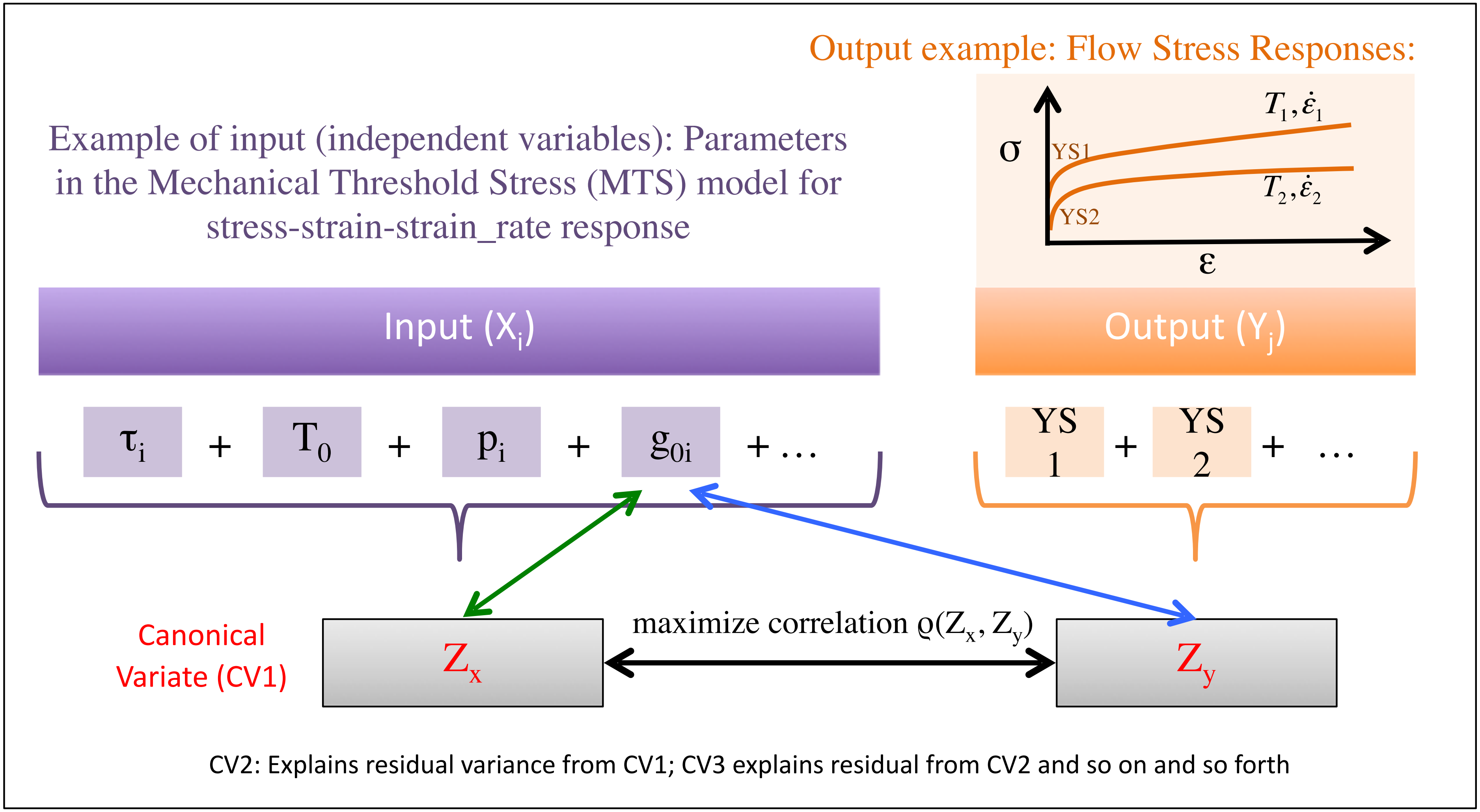
An Alternative Test Statistic

An alternative large sample approximation for the distribution of Wilk's Lamda under the hypothesis of independence is based on Rao's F used in multivariate regression above. The statistic is given by $F = m_{2k}(1 - \Lambda_k)^{1/\nu_k}/m_{1k}\Lambda_k^{1/\nu_k}$ where

$$\begin{aligned}
\nu_k &= \sqrt{\frac{(s-k)^2(q-k)^2-4}{(s-k)^2+(q-k)^2-5}} \\
m_{1k} &= (s-k)(q-k) \\
m_{2k} &= \nu_k[n - \frac{1}{2}(s+q+3)] - \frac{(s-k)(q-k)}{2} + 1,
\end{aligned}$$

which has m_{1k} and m_{2k} degrees of freedom if all but the first k eigenvectors are zero. Some computer software for canonical correlation analysis uses this F-approximation claiming that it is superior to the χ^2 approximation in small samples.

Application



Output example: Flow Stress Responses:

Example of input (independent variables): Parameters in the Mechanical Threshold Stress (MTS) model for stress-strain-strain_rate response

Input (X_i)

$\tau_i + T_0 + p_i + g_{0i} + \ldots$

Output (Y_j)

YS 1 + YS 2 + …

Canonical Variate (CV1)

Z_x

maximize correlation $\varrho(Z_x, Z_y)$

Z_y

CV2: Explains residual variance from CV1; CV3 explains residual from CV2 and so on and so forth

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)

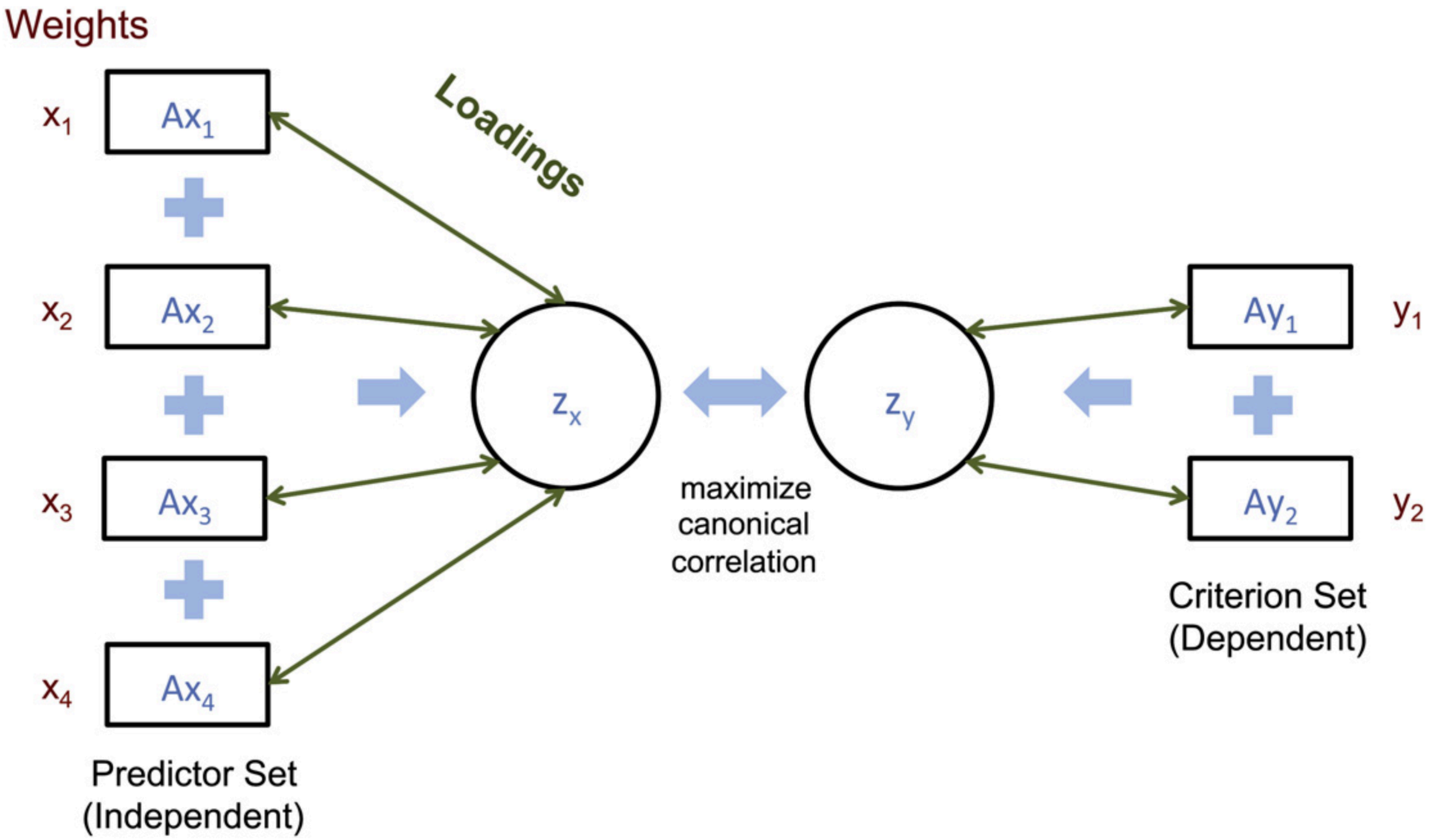
Communality: Sum of squared loadings for all CVs (Overall usefulness)

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

Canonical Correlation Analysis (CCA)

- The idea of CCA is to evaluate how much of the variance in one set of variables (“output”, or “dependent”) can be explained by an associated set of variables (“input”, or “independent”). Any relationship is assumed to be linear and combinations are sought that maximize the variance explained. The mathematics used is very similar to that of Principal Component Analysis (CPA).
- Here we aim to make connections between the calculated plasticity response (“output”) and a set of random model parameters (“input”).

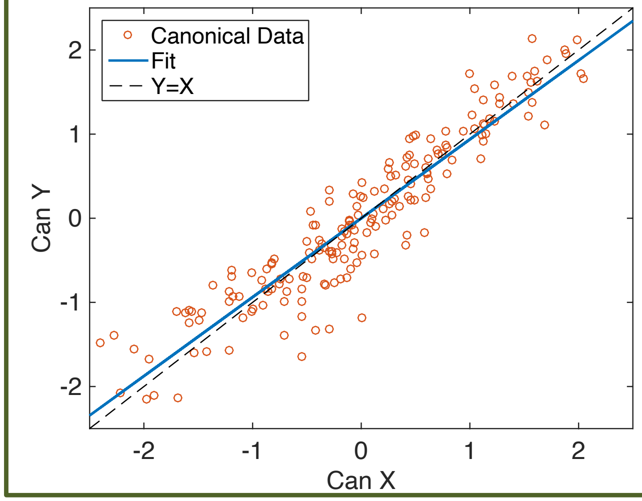
Construction of Canonical Functions



Mandal, Gockel, Rollett, *Matls. Design* **132** 30 (2017)

CCA Results for MTS model

How much variance is explained? **Rc²**

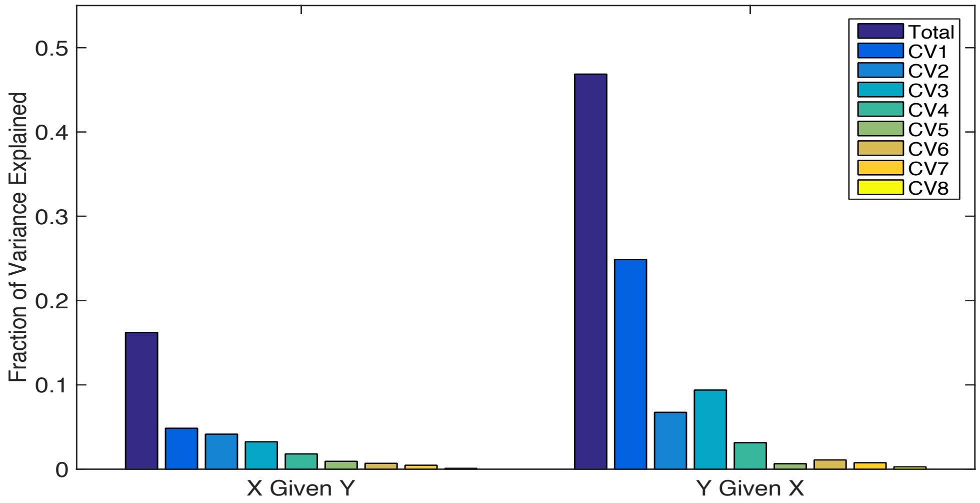


For CV1

$R_c=0.94$

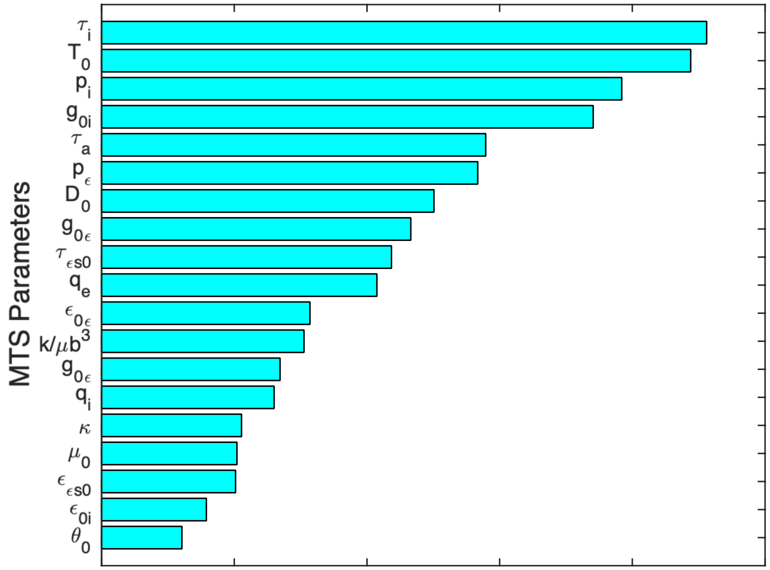
$R_c^2=0.88$

How adequate is the prediction? **Redundancy**

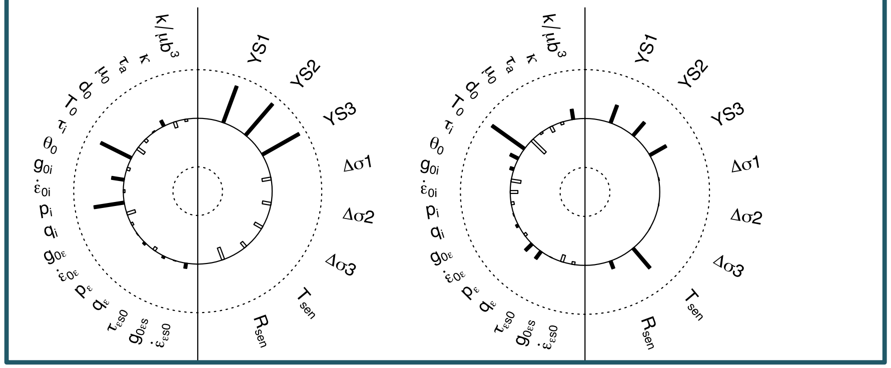


Mandal, Gockel, Rollett, *Matls. Design* **132** 30 (2017)

How important is each parameter? **Communality**



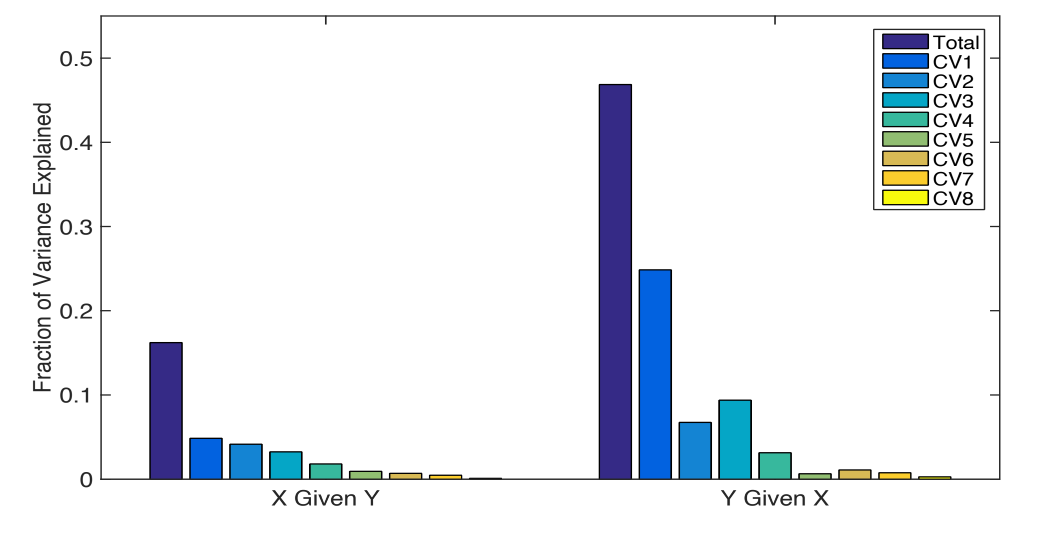
Which parameter affects which output responses? **Weights & Loadings**



Some explanations

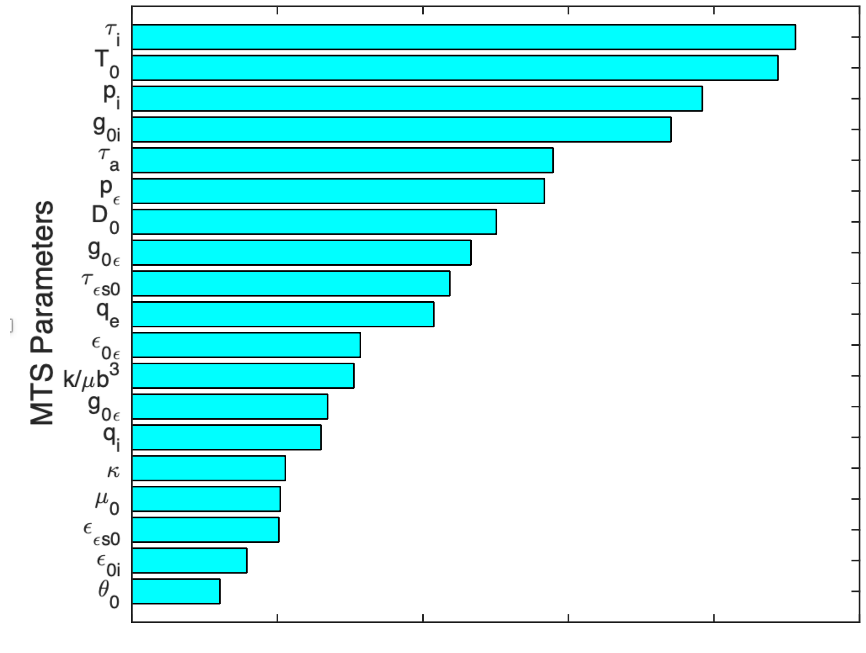
Redundancy coefficients show the proportion of variance in the variables in one set that is reproducible from the variables in the other set. This figure shows that, given the MTS parameters, the flow stress response can be modeled with a higher confidence than the other way around. In other words, the latter variables cannot be used to infer the model parameters.

How adequate is the prediction? **Redundancy**



Mandal, Gockel, Rollett, *Matls. Design* **132** 30 (2017)

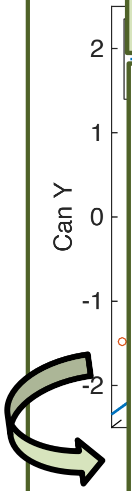
How important is each parameter?
Communality



Communality coefficients refer to the sum of squared loadings across all canonical functions. The communality coefficient of a variable quantifies what proportion of that variable's variance is reproducible from the total canonical results and hence informs about the usefulness of an observed variable in the entire analysis.

CCA Results for MTS model

How

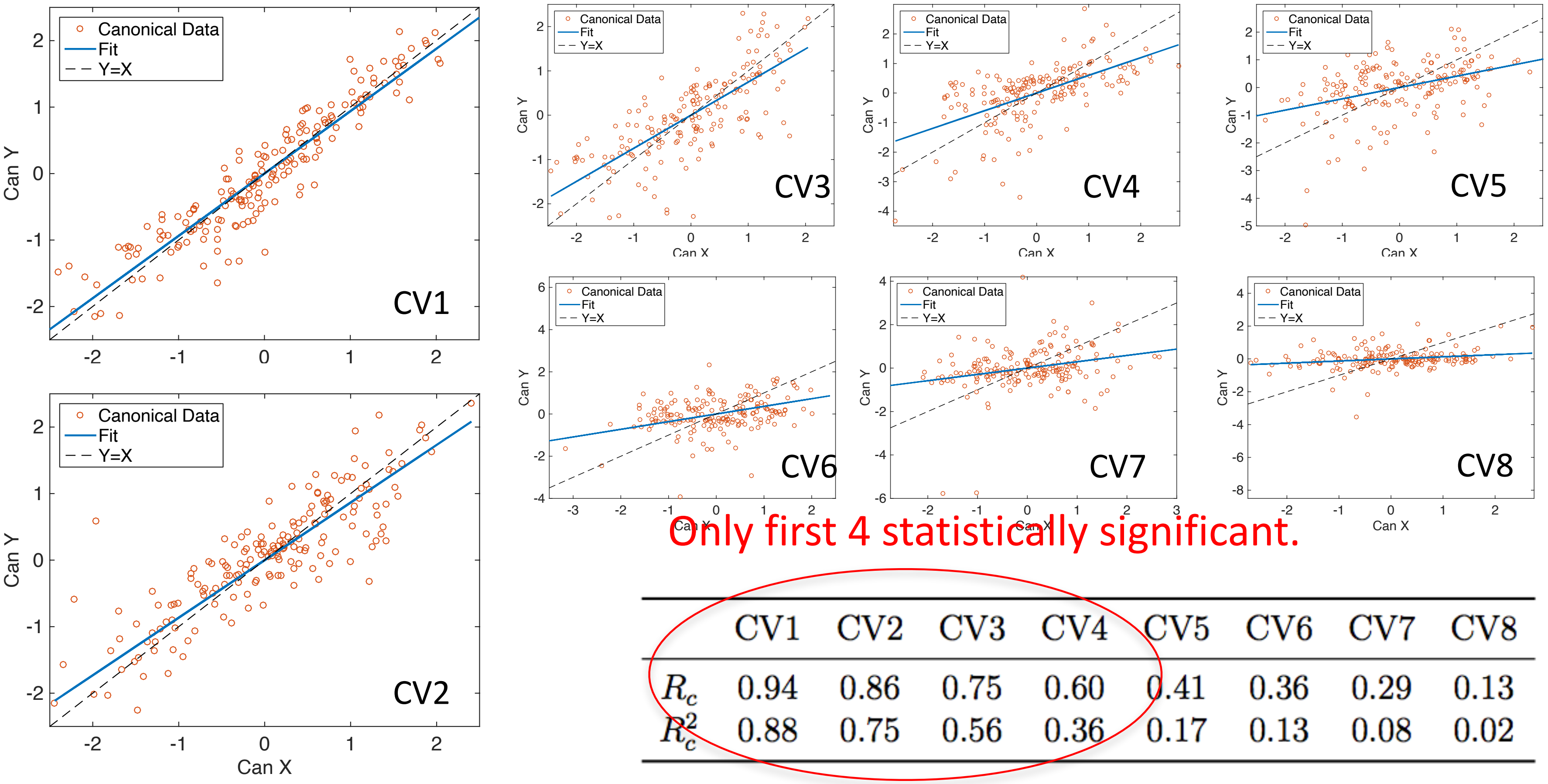


How

How much variance is explained?

Squared correlation

1st Canonical Variate captures 88% (R_c^2) of the total variance in the dataset.



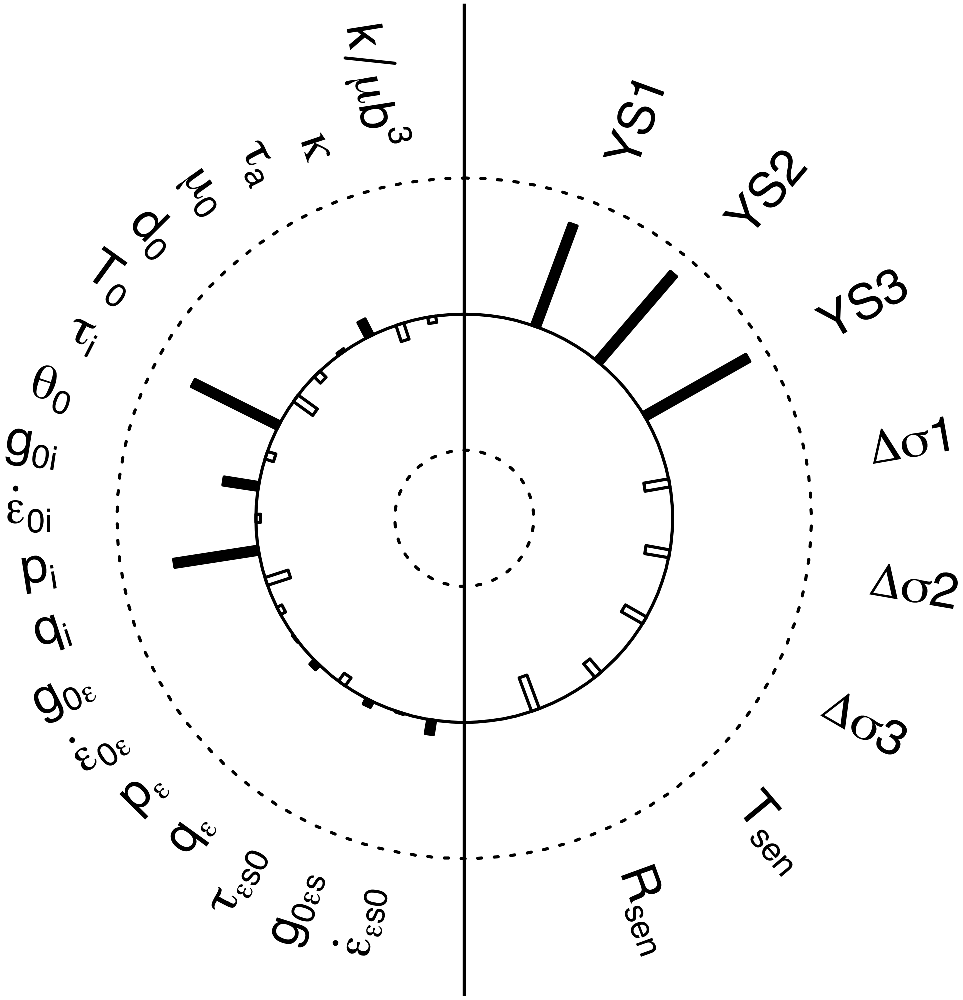
Only first 4 statistically significant.

	CV1	CV2	CV3	CV4	CV5	CV6	CV7	CV8
R_c	0.94	0.86	0.75	0.60	0.41	0.36	0.29	0.13
R_c^2	0.88	0.75	0.56	0.36	0.17	0.13	0.08	0.02

CCA Results

p_i , τ_i , g_{0i} ↑

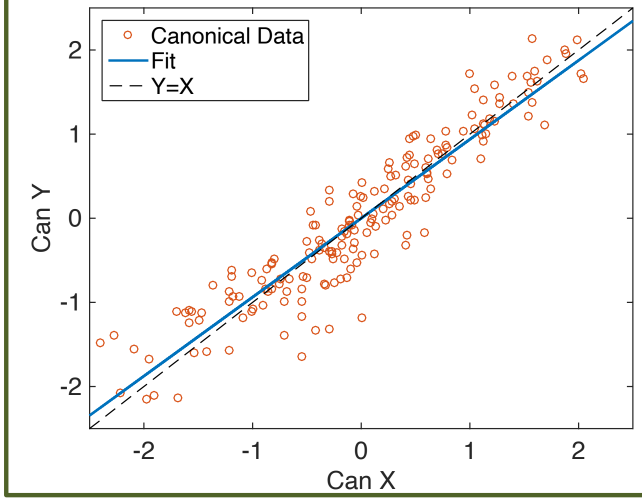
YS ↑



CV1

CCA Results

How much variance is explained? **Rc²**

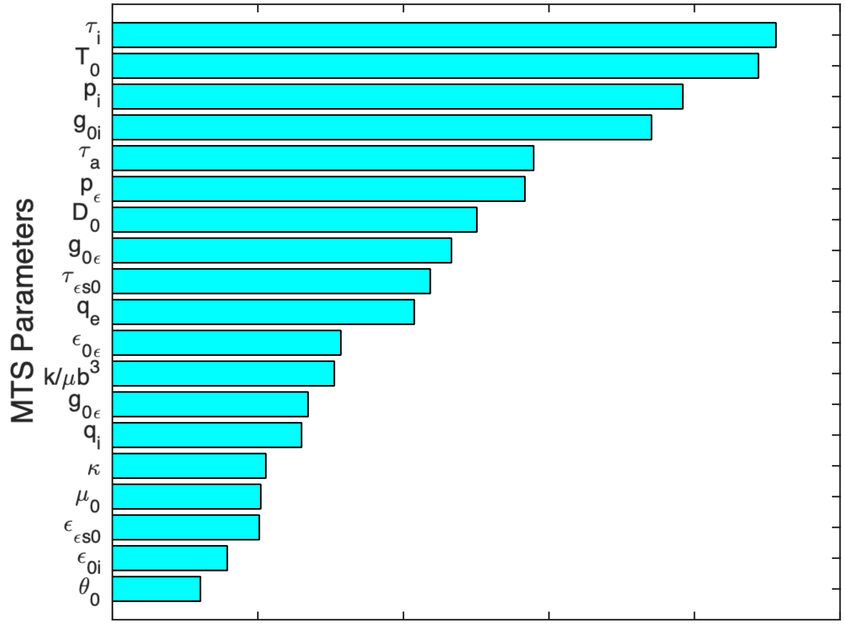


For CV1

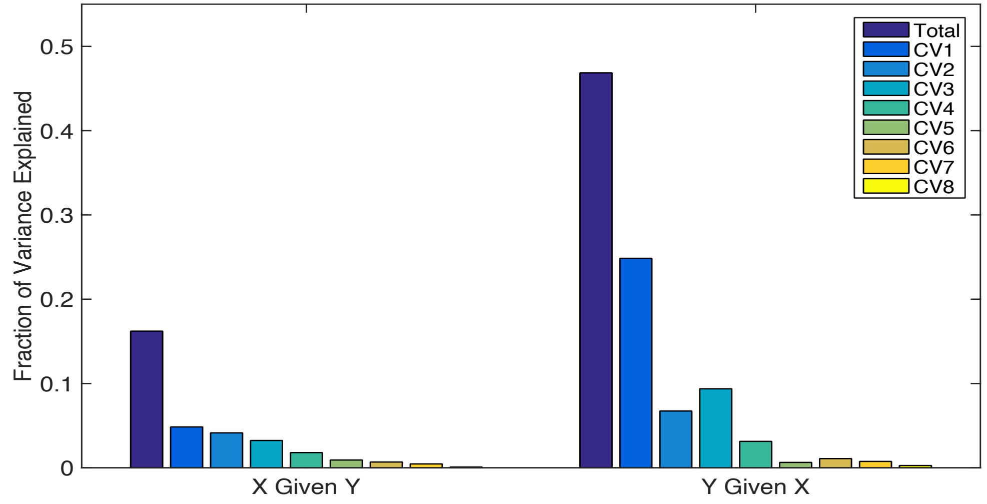
R_c=0.94

R_c^2=0.88

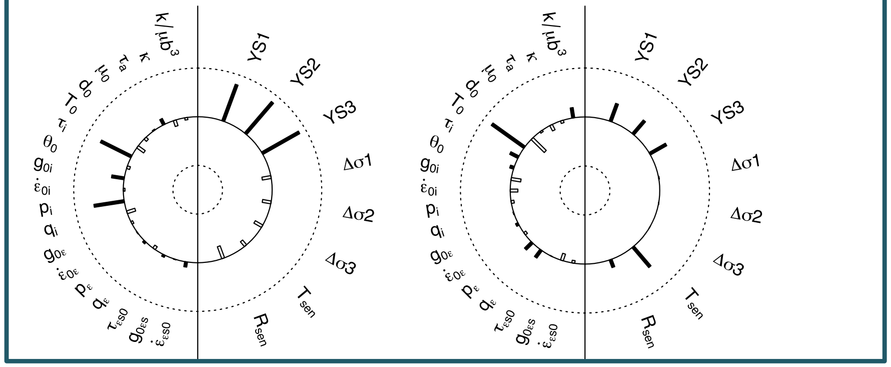
How important is each parameter? **Communality**



How adequate is the prediction? **Redundancy**

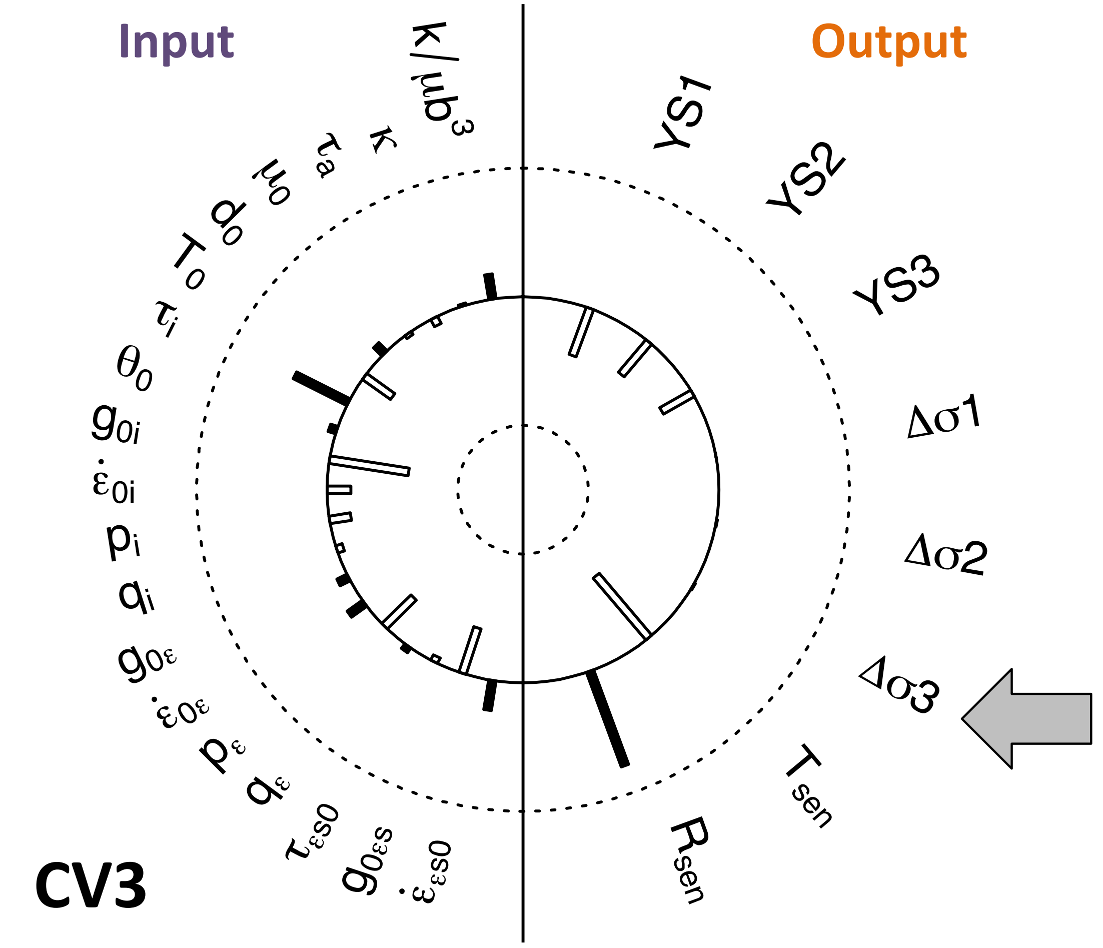


Which parameter affects which output responses? **Weights & Loadings**



Which parameter affects which output responses?

Loadings



CV1 Predictions

p_i , τ_i , g_{0i} ↑

YS ↑

CV3 Predictions

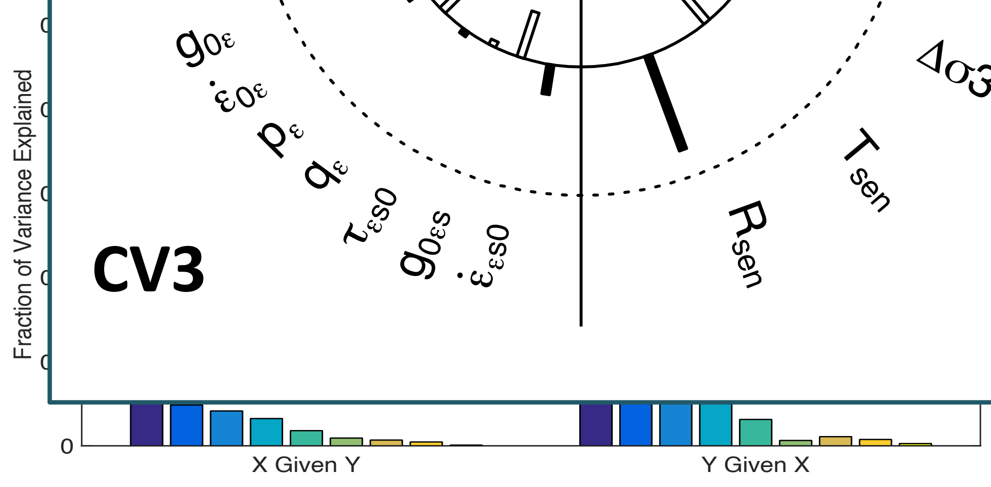
g_{0i} , $g_{0\varepsilon s}$ ↓ τ_i ↑

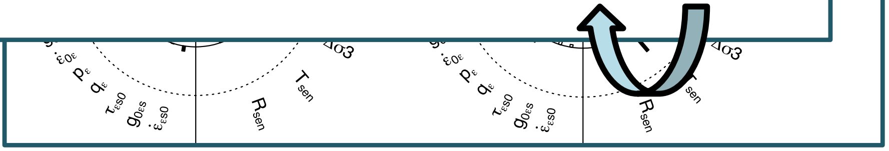
R_{sen} ↑

Hov

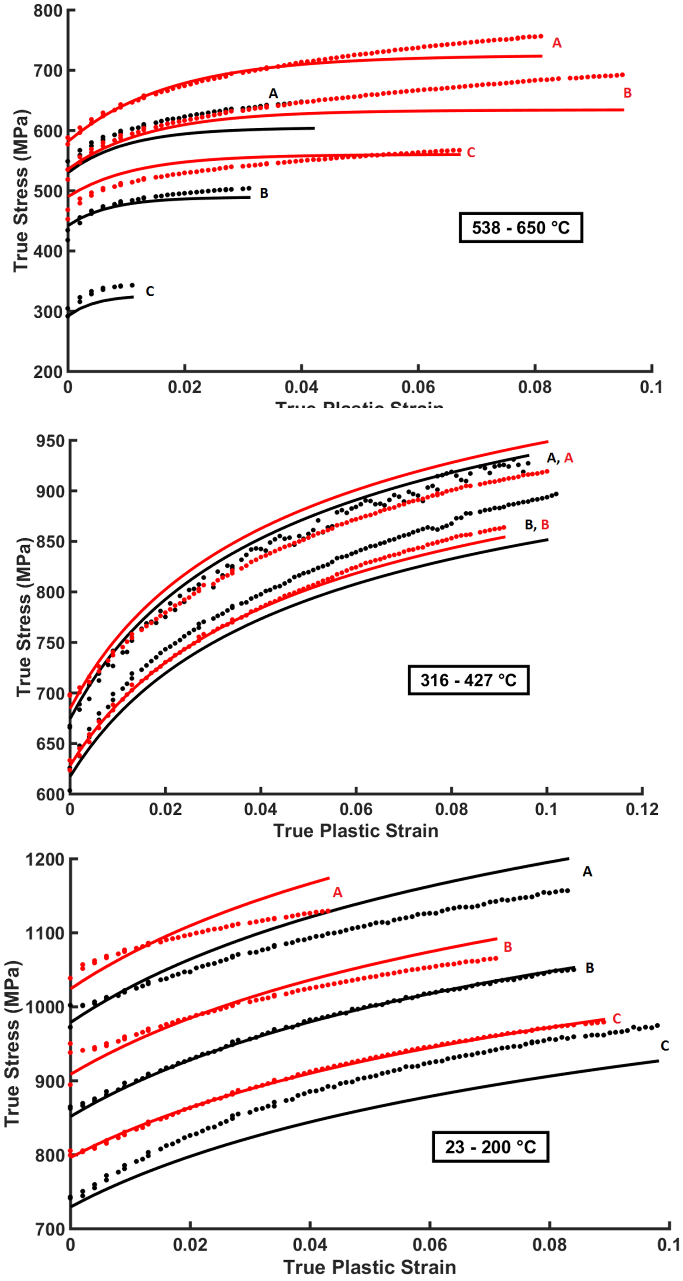
Can Y

How





Validation of CCA Predictions: Experiment



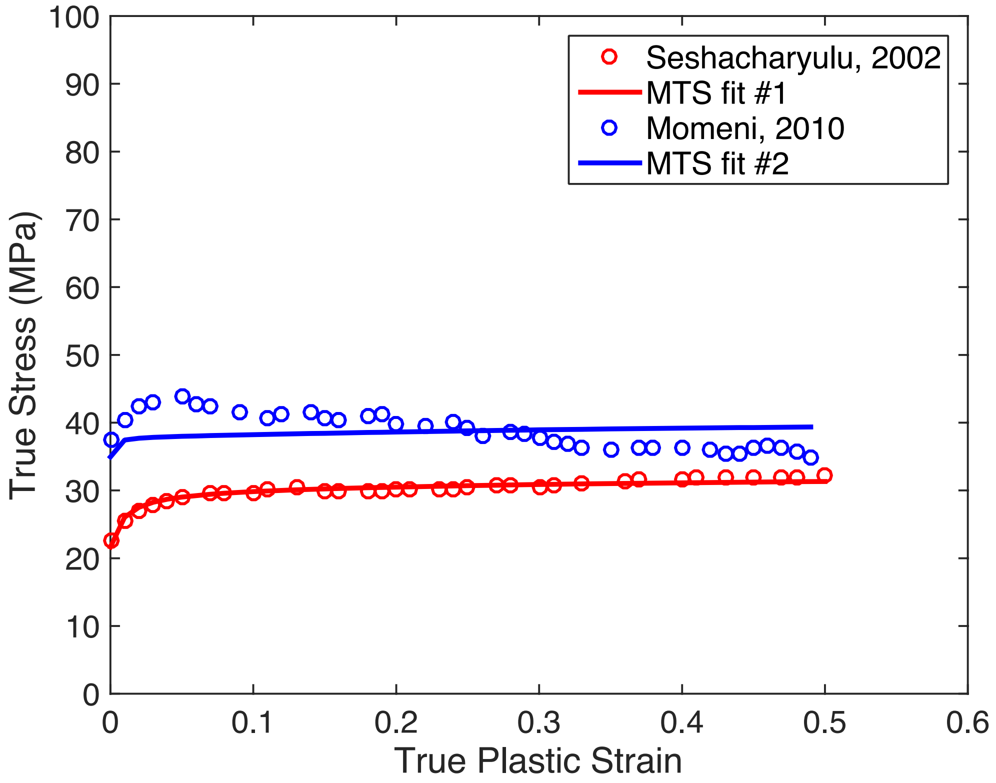
Exp: points, MTS: line; Black: Quasi-static strain rate, Red: High strain rate (0.01/s)

- Uniaxial compression of Ti-6242 (**Courtesy: Dr. Brian Gockel, AFRL**)
- MTS model fitted separately for the three different temperature regimes with varying deformation mechanisms

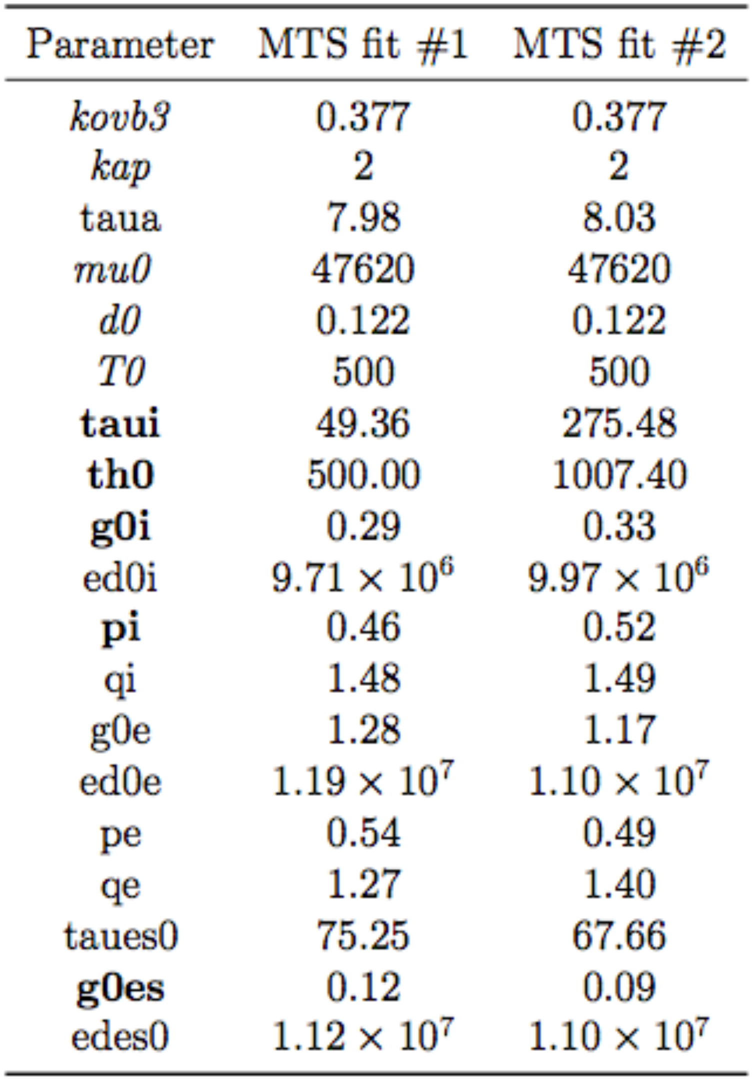
low R_{sen}

Parameter	RT–200 °C	316–427 °C	538–650 °C
$(\frac{k}{\mu b^3})$	0.5721	0.5721	0.5721
κ	3.47	2.55	0.95
(τ_a)	33	33	33
(μ_0)	48516.3	48516.3	48516.3
(D_0)	0.057	0.057	0.057
(T_0)	98.7	98.7	98.7
$\hat{\tau}_{\mathbf{i}}$	760	519	487
θ_0	2080	5015	9545
$\mathbf{g_{0i}}$	1.14	4.79	0.61
$(\dot{\epsilon}_{0i})$	1×10^7	1×10^7	1×10^7
$\mathbf{p_i}$	0.48	1.78	2.4
q_i	1.08	1.89	0.30
$g_{0\epsilon}$	1.6	1.6	1.6
$(\dot{\epsilon}_{0\epsilon})$	1×10^7	1×10^7	1×10^7
p_ϵ	4.1	3.69	0.66
q_ϵ	0.13	0.05	1.0
$\hat{\tau}_{\epsilon s0}$	499	416	1328
$\mathbf{g_{0\epsilon s}}$	38.5	83	0.20
$(\dot{\epsilon}_{\epsilon s0})$	1×10^7	1×10^7	1×10^7

Validation of CCA predictions: Literature



[1] Momen et al. (2010), *Mater. Des*. **31**, 3599–3604.
[2] Seshacharyulu *et al*. (2002), *Mater. Sci. Eng. A* **325**, 112–125



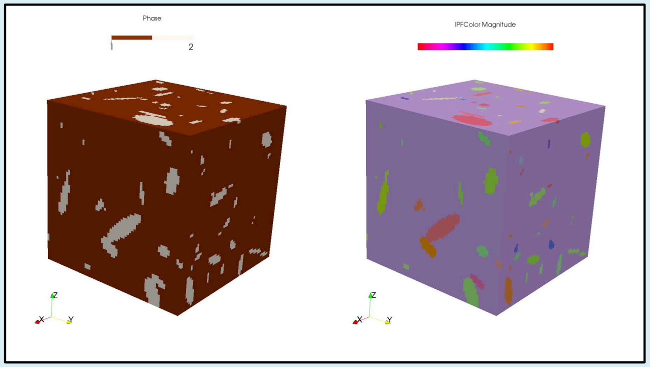
Parameter	MTS fit #1	MTS fit #2
kovb3	0.377	0.377
kap	2	2
taua	7.98	8.03
mu0	47620	47620
d0	0.122	0.122
T0	500	500
taui	49.36	275.48
th0	500.00	1007.40
g0i	0.29	0.33
ed0i	9.71×10^6	9.97×10^6
pi	0.46	0.52
qi	1.48	1.49
g0e	1.28	1.17
ed0e	1.19×10^7	1.10×10^7
pe	0.54	0.49
qe	1.27	1.40
taues0	75.25	67.66
g0es	0.12	0.09
edes0	1.12×10^7	1.10×10^7

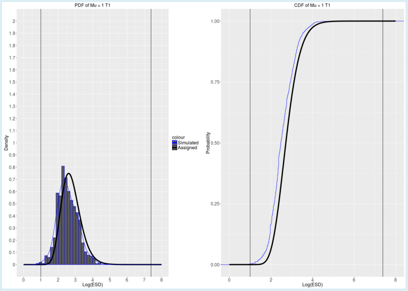
- Two different Ti-6Al-4V plots from literature showing **different YS**
- MTS parameters fitted separately
- Most important based on fitting:

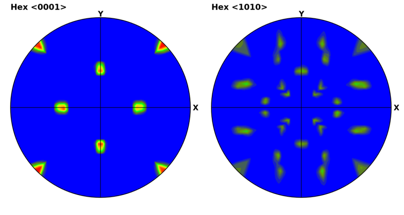
$$p_i\ ,\ \tau_i\ ,\ g_{0i}\ ,\theta_0\ ,\ g_{0\varepsilon}$$

- Matches with CV1 predictions for 3 parameters

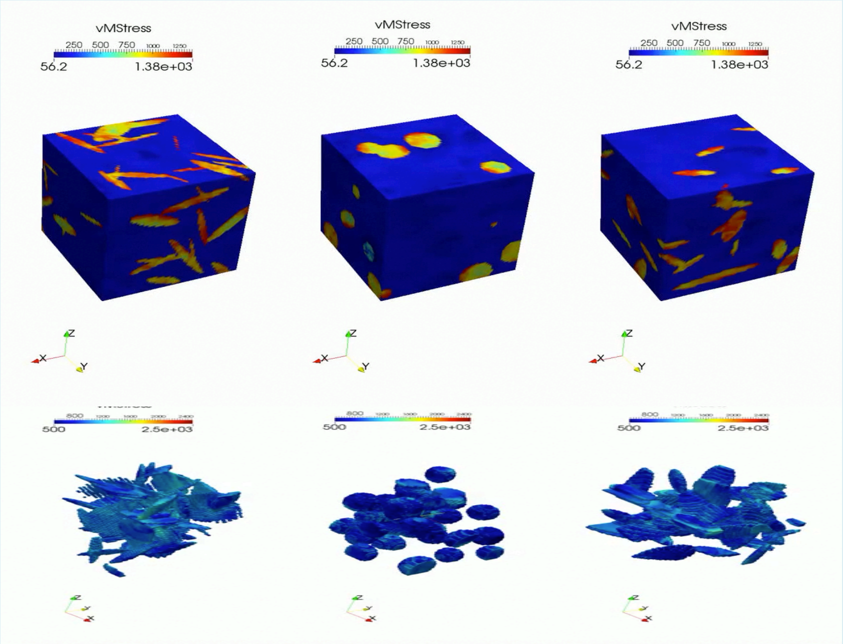
Other Applications of CCA

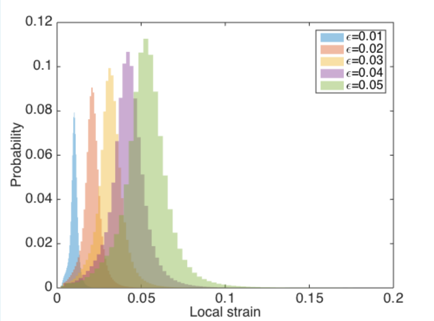


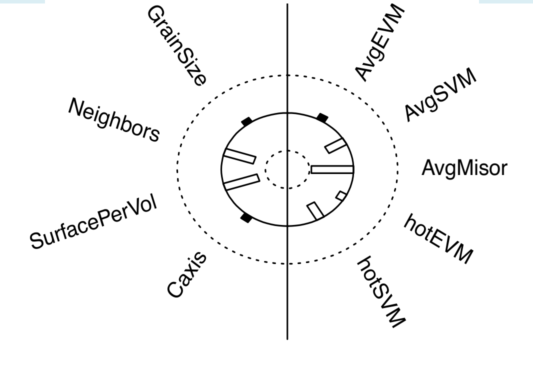


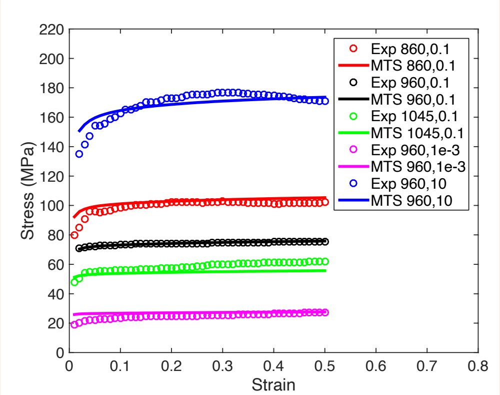


Microstructure-Property Relationships Quantified

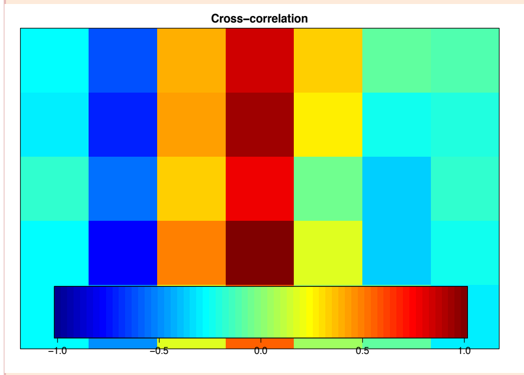








Error Signals



Inverse uncertainty in parameter calibration

Non-linear relationships

- One possibility for non-linear relationships is to use Kolmogorov-Gabor polynomials, also used in non-linear neural nets:

$$f\left(\vec{x}\right) = c_0 + \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} c_{ijk} x_i x_j x_k + \cdots ,$$

The same Kolmogorov-Gabor polynomials can be used but as ratios of variable sets:

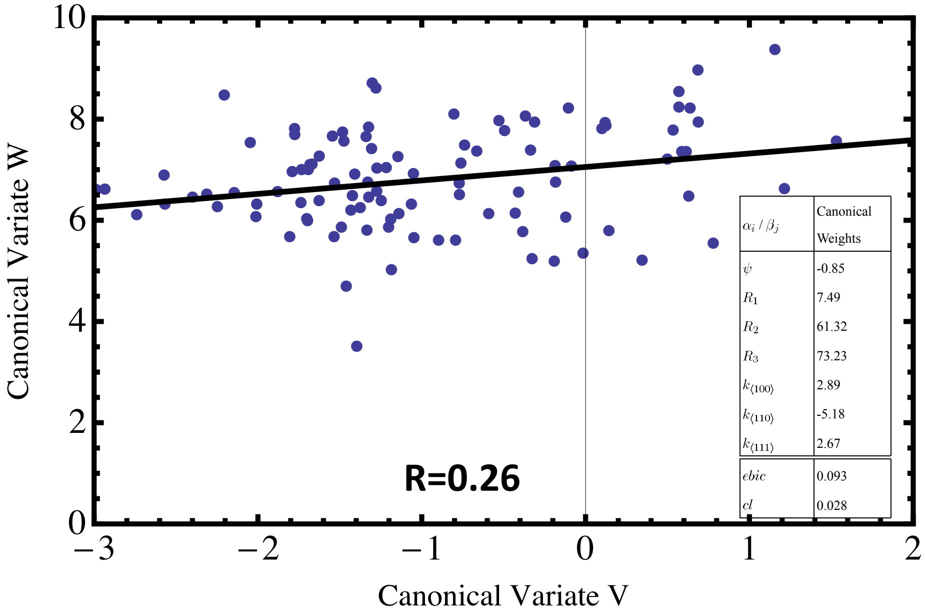
$$f\left(\vec{x}\right) = \frac{\sum_{i=0}^{m} \sum_{j=0}^{m} \sum_{k=0}^{m} c_{ijk} x_1^i x_2^j x_3^k}{\sum_{i=0}^{m} \sum_{j=0}^{m} \sum_{k=0}^{m} d_{ijk} x_1^i x_2^j x_3^k}$$

Such a set of variables with trial values of the coefficients can then be used in a simulated annealing procedure to maximize, say, an eigenvalue.

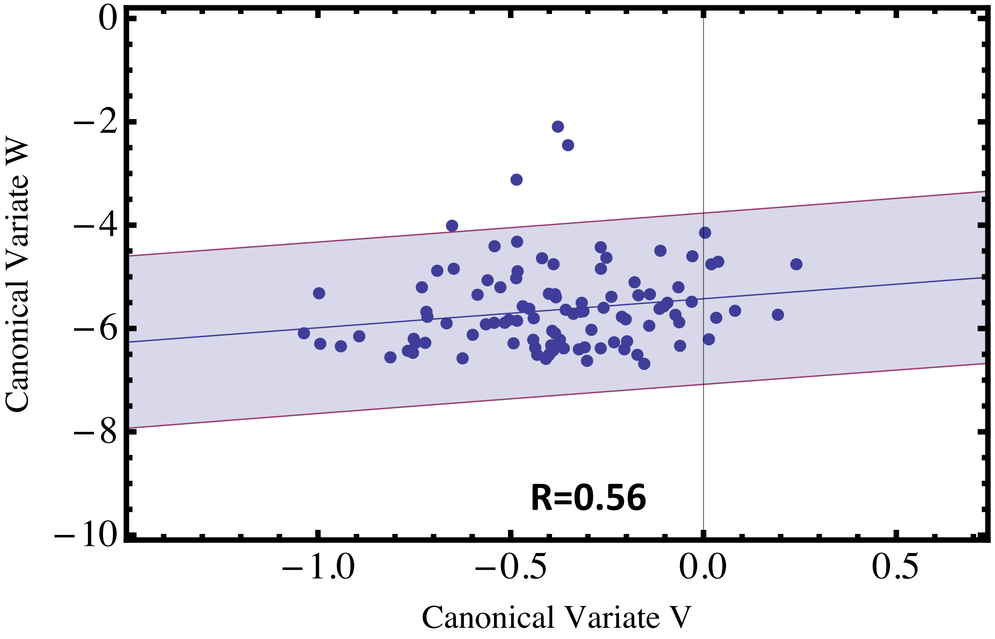
"Data Analytics using Canonical Correlation Analysis and Monte Carlo Simulation", J. Rickman, Y. Wang, A.D. Rollett, M.P. Harmer and C. Compson, *(Nature) Computational Materials* **3** 26 (2017);

Example: non-linear CCA

- Re-analysis of the EBIC and CL signals versus grain boundary data.
- Here, the canonical variate emphasized the EBIC signal, which displayed no significant correlation in the linear form with R=0.26.
- Re-analysis allowing for non-linear relationships yielded R=0.56 and a high likelihood of correlation.



α_i / β_j	Canonical Weights
ψ	-0.85
R_1	7.49
R_2	61.32
R_3	73.23
$k_{\langle 100 \rangle}$	2.89
$k_{\langle 110 \rangle}$	-5.18
$k_{\langle 111 \rangle}$	2.67
$ebic$	0.093
cl	0.028



"Data Analytics using Canonical Correlation Analysis and Monte Carlo Simulation", J. Rickman, Y. Wang, A.D. Rollett, M.P. Harmer and C. Compson, *(Nature) Computational Materials* **3** 26 (2017)

Why is CCA not popular?

“virtually all of the commonly encountered parametric tests of significance can be treated as special cases of canonical-correlation analysis, which is the general procedure for investigating the relationships between two sets of variables.”
- Knapp (1978)

“One reason why the technique is [somewhat] rarely used involves the difficulties which can be encountered in trying to interpret canonical results... **The neophyte student of CCA may be overwhelmed by the myriad coefficients which the procedure produces... [But] CCA produces results which can be theoretically rich, and if properly implemented, the procedure can adequately capture some of the complex dynamics involved in reality.”**
-Thompson (1980)

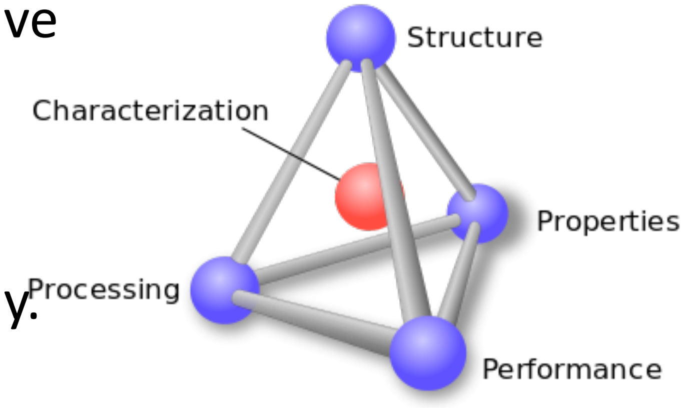
“CCA is only as complex as reality itself”
-Thompson (1991)

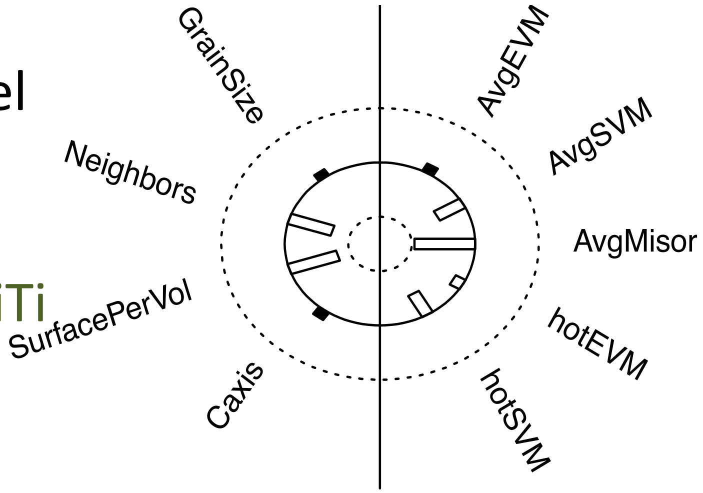
When to use CCA?

- When the data can be split into two sets and one predicts another
- As an exploratory tool to see if two sets of variables are related
- To find the interrelationships between different variables in both the sets
- To quantify the importance of variables in the overall model/ dataset
- To check if a set of variables at one time step can be used to predicting variables at the next step i.e., temporal prediction
- Don’t be afraid to try CCA – given a spreadsheet of values, it only takes a few minutes to calculate. Interpretation is straightforward once familiar.

Summary

- Local vs. global sensitivity analysis
- CCA, a multivariate technique presented as an alternative
- CCA shows relative impact & interrelationships (can be used for calibration & extrapolation)
- The parameters **taui, T0, pi and g0i** in the MTS model were found to be the most influential overall statistically.
- Validation tests based on experiments and literature supported the CCA predictions.
- This technique can be potentially used to quantify structure-property correlations at microstructural level thereby aiding in better integration with continuum models.
- Current work: sensitivity of texture development in NiTi to CRSS, hardening on different systems
- Future work: Effect of texture on micromechanical properties, Analysis of full-field MASSIF simulations





"Application of canonical correlation analysis to a sensitivity study of constitutive model parameter fitting", Mandal, S., Gockel, B.T., Rollett, A.D, Materials & Design **132**, 30-43 (2017).