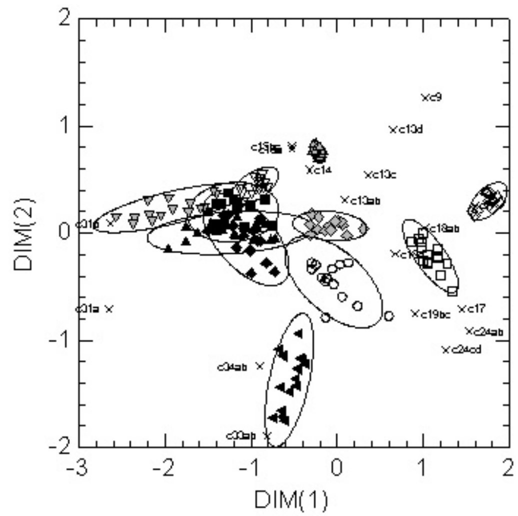
MULTIVARIATE STATISTICAL ANALYSIS FOR FOOD SCIENCE AND AGRICULTURE: AN INTRODUCTION 7. A MULTIVARIATE REGRESSION PROBLEM

Prof. Eugenio Parente
Scuola di Scienze Agrarie- Università della Basilicata



Outline

- (some) Inferential methods for multivariate data
  - Multiple Linear Regression (MLR)
  - Principal Component Regression (PCR)
  - Partial Least Square Regression (PLSR)
- a step by step approach to descriptive and inferential analysis for a dataset containing continuous and discrete variables
  - the initial exploratory phase
  - the need for data transformation
  - data treatment
    - Multiple Linear Regression
    - Principal Component Regression
    - Partial Least Square Regression (PLS1 and PLS2)

Multivariate data set

$$Y = \begin{bmatrix} y_{11} & \cdots & \cdots & y_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & \cdots & \cdots & y_{nm} \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & \cdots & \cdots & x_{1k} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & \cdots & x_{nk} \end{bmatrix}$$

n observations (cases) for which *k* x (independent) variables and *m* y (dependend variables have been measured)

Simple linear regression and Multiple Linear Regression

- One y, one x
  - **y**= b**x**+**e**

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = b \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \dots \\ e_n \end{bmatrix}$$

- One y, *k* x
  - **y**=**Xb**+**e**

$$\begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} b_1 \\ \dots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ \dots \\ e_n \end{bmatrix}$$

- *m* y, *k* x
  - **Y=XB+E**

$$\begin{bmatrix} y_{11} & \dots & y_{1m} \\ \dots & \dots & \dots \\ y_{n1} & \dots & y_{nm} \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} b_{11} & \dots & b_{1m} \\ \dots & \dots & \dots \\ b_{1k} & \dots & b_{mk} \end{bmatrix} + \begin{bmatrix} e_{11} & \dots & e_{1m} \\ \dots & \dots & \dots \\ e_{1n} & \dots & e_{nm} \end{bmatrix}$$

How many (*n, m, k*)?

- *k*>*n* : infinite number of solutions for b, which cannot be estimated
- *k*=*n* : unique solution for b, if X has full rank (the p variables are linearly independent) -> **e=y-Xb=0**
- *k*<*n* : no exact solution for **b**, but **b** can be estimated by least squares, i.e. the sum of squares of the residuals is minimized. This means solving the equation:

$$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

There might be no solution for this equation: there might be no inverse of **X'X** because of collinearity, 0 determinant, singularity

Frequent situations

- $k<n$: there are (hopefully far) more observations than variables but he **X** (and or **Y**) matrix is not full rank; the *p* (and/or *m*) independent variables are correlated, this results in high collinearity with very large standard errors for regression coefficients
  - Common examples: NIR spectrometry, RP-HPLC, etc.
  - One possibility is using stepwise regression to remove some of the variables, but this may be difficult because of lack of independence of regression coefficients; another possibility is using PCR
- $k>n$: there are less observations than variables; this may be combined with a collinearity problem
  - Reduce the number of variables by removing those which are less important and make them imdependent: PCR, PLS

Principal Component Regression

- Pretreatment of the data is done, as applicable (transformations, standardization, etc…)
- PCA is carried out on the **X** matrix to extract *a* principal components: principal components are new, orthogonal variables, which are (hopefully) few ($a<<p$) and summarize most of the variation in **X**
- MLR regression is carried out to estimate **y** from **T** (the principal components score matrix)
- Diagnostics (regression diagnostics, residuals, etc.) are used to evaluate the quality of the model, loadings can be used to interpret the model, PCA coefficients can be saved for validation or re-use in predictive mode (multivariate calibration models)

Principal Component Regression

$$t_{ia} = \sum_k w^*_{ka} x_{ik} \qquad \left(\mathbf{T} = \mathbf{X}\mathbf{W}^*\right)$$

$$x_{ik} = \sum_a t_{ia} p_{ak} + e_{ik} \qquad \left(\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}\right)$$

$$y_i = \sum_a b_a w^*_{ka} x_{ik} + f_i \qquad \left(\mathbf{y} = \mathbf{X}\mathbf{W}^*\mathbf{b} + \mathbf{F}\right)$$

- Here **T** is the *nxa* score matrix, **W** the *kxa* weights matrix, **P** the *axk* loadings, **E** is the *nxk* residuals (all for **X**) and **b** are the *a* coefficients and **F** are the *n* residulas (for the y=f(x) regression)
- Principal Component Regression may be very effective in several situations (and can be generalized for ANOVA problems)
- The main problem is that PCA may extract variation in the independent data set which is not related to the y data set and is therefore of little use in prediction

Partial Least Square Regression

- PLSR has been used for several applications in econometrics, chemometrics, biology, etc. since the ’80s to address problems with many collinear variables and with p<n
- PLS derives from the original algorithm for estimating the coefficients of the model: Nonlinear Iterative Partial Least Squares (NIPALS). The term refers to the fact that the **x** vector **u** (the y scores for each of the a components) is considered fixed in the estimation, so it is a partial regression

Advantages of PLSR

- It can usually find a parsimonious model (with a low nunber of predictors) which is also robust (when you estimate the parameters with different datasets, for example during validation, they usually change little) and has good predictive value
- It tolerates moderate amount of missing data
- In addition to find a model to predict **Y** from **X**, it also help exploring the data structure (i.e. the relationships among **X** and **Y** variables)

The PLSR model

- Two multivariate matrices are available: **Y** (*nxm*, dependend variable matrix; *m*=1 in PLS1 and *m*>1 in PLS2) and **X** (*nxk*) independent variable matrix;
- Both the *m* **Y** variables and the *k* **X** variables are not independent (i.e. they may have significant correlations) and are assumed to be realizations of *a* independent, orthogonal, latent variables, which model both **X** and **Y**
- Latent variables are extracted in such a way to maximize the correlation between **X** and **Y**
- The process of extraction is iterative and cross-validation is needed to identify the correct number of components

Geometric interpretation

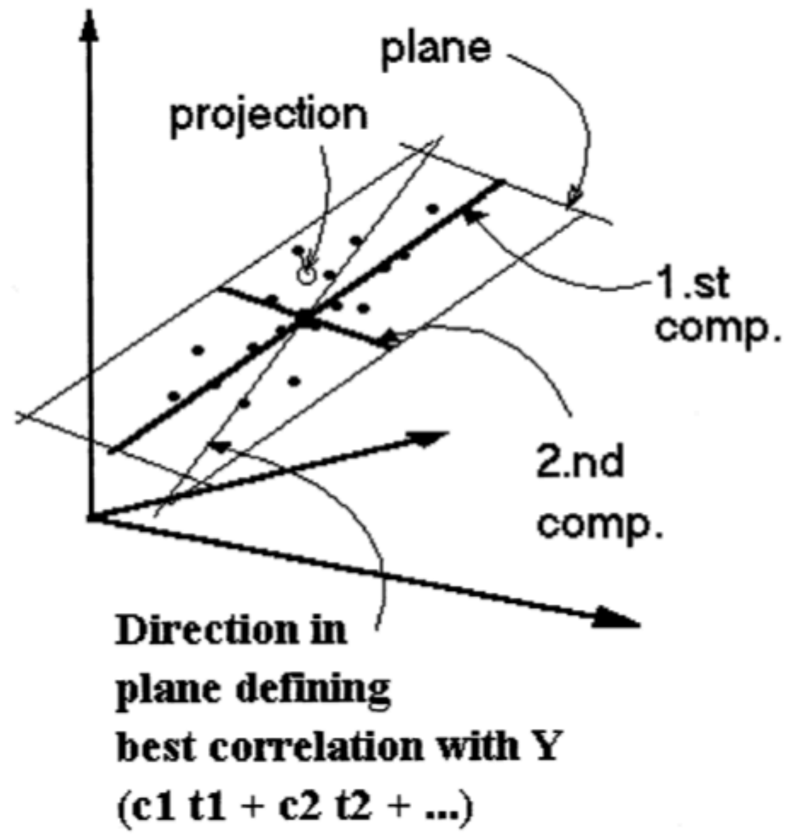


Fig. 2. The geometric representation of PLSR. The **X**-matrix can be represented as N points in the K dimensional space where each column of **X** ($\mathbf{x}_k$) defines one coordinate axis. The PLSR model defines an A-dimensional hyper-plane, which in turn, is defined by one line, one direction, per component. The direction coefficients of these lines are p_{ak}. The coordinates of each object, i, when its data (row i in **X**) are projected down on this plane are t_{ia}. These positions are related to the values of **Y**.

ELSEVIER
Chemometrics and Intelligent Laboratory Systems 58 (2001) 109–130
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemometrics

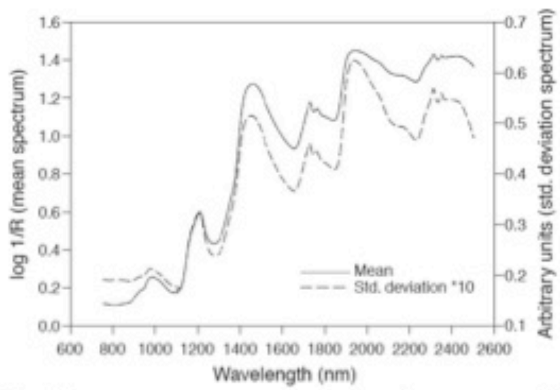
PLS-regression: a basic tool of chemometrics

Svante Wold [a,*], Michael Sjöström [a], Lennart Eriksson [b]
[a] *Research Group for Chemometrics, Institute of Chemistry, Umeå University, SE-901 87 Umeå, Sweden*
[b] *Umetrics AB, Box 7960, SE-907 19 Umeå, Sweden*

Steps in PLSR

- Pre-treat and standardize the data as needed
- Iteratively estimate x scores, weights, loadings and residuals, y scores, loadings and residuals, PLSR coefficients and residuals
- Use cross-validation to
  - Estimate the number of components
  - Calculate cross-validation statistics and indicators of goodness of fit
- Present the results
  - Number of components, amount of variance explained, R^2, Q^2 (crossvalidation R^2), PRESS (Predictive Residual Sum of Squares)
  - x scores vs y scores plots, x weights and y loadings plots, residuals plots, x scores and loading plots

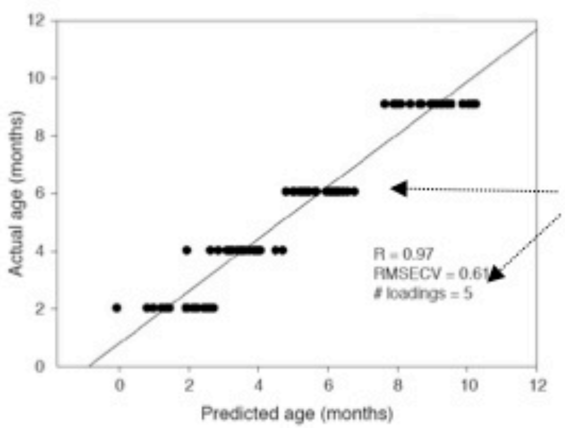
PLSR step by step - 1



A. Perform experiments, gather raw data

NIR reflectance spectra are measured on 100 Cheddar cheese samples, ripened for 2-9 months. Mean and standard deviation reflectance spectra for all samples are sown.

B. Preprocess raw data: best results in terms of model robustness and ease of interpretation obtained with Savitzky-Golay 2nd derivative



C. Estimate parameters of a PLS1 model to predict cheese age from preprocessed spectral data

A PLS model with 5 components is used to predict age (**y**) from the transformed spectral data (**X**). The PLS model has a good predictive ability for age. The first 2 PLS components explain 62% and 82% of the **X** and **y** variance, respectively

ELSEVIER
Available online at www.sciencedirect.com
SCIENCE DIRECT
International Dairy Journal 15 (2005) 701–709
INTERNATIONAL DAIRY JOURNAL
www.elsevier.com/locate/idairyj

Prediction of maturity and sensory attributes of Cheddar cheese using near-infrared spectroscopy

G. Downey[a,*], E. Sheehan[b], C. Delahunty[b], D. O'Callaghan[c], T. Guinee[c], V. Howard[c]
[a]Teagasc, The National Food Centre, Ashtown, Dublin 15, Ireland
[b]Department of Nutritional Sciences, University College Cork, Cork, Ireland
[c]Teagasc, The Dairy Products Research Centre, Moorepark, Fermoy, Co. Cork, Ireland

Received 7 May 2004; accepted 8 June 2004

X scores, weights, loadings, residuals

- Extract x scores (**T**) in such a way that they are good predictors of **Y** and that they explain most of the X variation in a parsimonious way
- Use x scores to explore relationships among cases in the X matrix
- Use x loadings (**P**) to explore the relationships between **T** and the original variables
- Analyze residuals (**E**) to identify outliers and violations of assumptions

$$t_{ia} = \sum_k w^*_{ka} x_{ik} \qquad \left(\mathbf{T} = \mathbf{X}\mathbf{W}^*\right)$$

$$x_{ik} = \sum_a t_{ia} p_{ak} + e_{ik} \qquad \left(\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}\right)$$

Y scores, loadings, X weights and residuals

- Extract y scores (**U**) in such a way that they are good predictors of **Y** and that they explain most of the **Y** variation in a parsimonious way and that x scores (**T**) are good predictors of **Y**
- Analyze residuals (**G** and **F**) to identify outliers and violations of assumptions
- Use **Y** loadings (**C**) and X weights (**W***) to explore relationships among original variables and components
- Look at plots of x-scores vs y-scores for each component

$$y_{im} = \sum_a u_{ia} c_{am} + g_{im} \quad \left(\mathbf{Y} = \mathbf{UC}' + \mathbf{G}\right)$$

$$y_{im} = \sum_a t_{ia} c_{am} + f_{im} \quad \left(\mathbf{Y} = \mathbf{TC}' + \mathbf{F}\right)$$

$$y_{im} = \sum_a c_{am} \sum_a w^*_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im}$$

$$\left(\mathbf{Y} = \mathbf{XW}^*\mathbf{C}' + \mathbf{F} = \mathbf{XB} + \mathbf{F}\right)$$

The need for cross-validation

- PLSR will extract as many components as the rank of the **X'X** matrix; this model fits perfectly the data
- During extraction of components cross-validation is used to determine the number of components to extract to obtain a parsimonious model, with good predictive ability
- Cross-validation is carried out by estimating models on subsets of observations, and comparing the effect of adding one further component.
- There are two main cross-validation methods
  - Leave-one-out or jacknife: n models are calculated by leaving out each time one of the observations
  - Resampling: a random subset of observations (usually) is extracted without replacement and the predictive ability of the model developed on the remaining observations is evaluated; this cam be repeated over several subsamples

PLS1 or PLS2

- Before PLSR, PCA should be carried out on Y variables
- If there is **no structure**, carry out a PLS1 for each of the response variables
- If PCA explains a significant amount of variance opt for PLS2
  - Use a single model if variables are not strongly clustered
  - Use several models (one for each group for variables) if variables are strongly clustered

PLSR statistics

- Analysis of variance can be carried out on individual **Y** variables to evaluate if they are significantly affected by the **X** variables
- Standard errors and confidence limits of coefficients can be used to compare the coefficients
- The amount of **X** and **Y** variance explained can be used to evaluate how well the model explains **X** and **Y** variability
- Predictive Residual Sum of Squares (PRESS, must be as low as possible) and cross-validated R^2 (Q^2, 1-PRESS/SS, between 0 and 1, high values indicate better predictive ability) for each **Y** variable are used to evaluate the predictive ability of the model

How many components?

- Develop models for a-1 and a components; calculate the ration $PRESS_a/SS_{a-1}$; if smaller than 0.9 for at least one of the Y variables, extract another component and recalculate
- Calculate models with a, a+1, a+2, etc. components; choose the model with the lowest PRESS/(N-A-1)

AN EXAMPLE

Relationships between flour composition, kneading, dough properties and leavening

Please note

- These are unpublished data
- Data are courtesy of Dr. Pasquale Catzeddu, Porto Conte Ricerche
- Data and results should not be disclosed outside this classroom

The data set

- Qualitative (discrete) variables
  - 3 wheat varieties (L, G2, New)
  - 2 different milling (semola, semolato)
  - 2 hydration levels (optimal, 80% of oftimal)
  - 3 different kneading times (optimal, 450 sec, 7 min)
- Quantitative (continuous) variables
  - Composition of the flour (% ashes, % damaged starch, % gluten, % proteins, gluten index)
  - % moisture of the dough
  - Chopin alveograph variables (pressure P, extensibility L, strenght, P/L)
  - Pressure measured with a consistograph at different kneading times
  - Kneading time
  - Density after kneading
  - Stress relaxation test after kneading (Fmax, Elasticity)
  - Glutenin macropeptide after kneading
  - Volume after leavening

Chopin Alveograph

How does the Chopin Alveograph work?

What the graph means

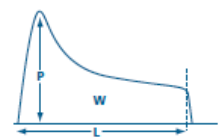
W =	L =	P/L =
baking strength of dough (the area under the curve)	extensibility of dough (time taken for bubble to burst)	dough strength and extensibility (ratio of curve height to length)
High W =	**P =**	**Low P/L =**
strong flour	maximum pressure required	very extensible and low strength

Typical characteristics of different wheat varieties

nabim Group 1

Suitable for bread flour – makes strong elastic dough and has excellent bread-making potential.

- high pressure (P)
- long time (L) to burst



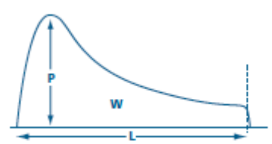
Typical range

Alveograph P/L	0.5 - 0.9
Alveograph W	>200

nabim Group 2

Suitable for bread and baking flours – most varieties having bread-making potential.

- low P/L ratio important



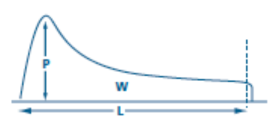
Typical range

Alveograph P/L	0.4 - 0.9
Alveograph W	170 - 310

nabim Group 3

Suitable for biscuit and blending flours – makes extensible dough, good for biscuits and blending with strong wheats.

- low pressure (P)
- long time (L)
- area under the curve (W) less critical



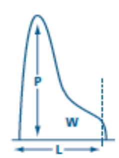
Typical range

Alveograph P/L	0.2 - 0.4
Alveograph W	70 - 100

nabim Group 4

Suitable for animal feed only – makes tough, inelastic dough

- high pressure (P)
- short time (L) to burst



Typical range

Alveograph P/L	0.3 - 1.5
Alveograph W	60 - 140

Results from the consistograph

ALVEOLINK NG CONSISTO AH CHOPIN

HEARTLAND MILL INC.
RT. 1 BOX 2
MARIENTHAL KS 67863

DATE: 07/07/2009
TIME: 6:49 pm

SAMPLE IDENTIFICATION: 9187.02UK
FILE NAME : 07070005A509

PARAMETERS

LAB.TEMP.:
FLOUR : 06239U9H9
MOISTURE : 12.40 %
PROTEIN : 11.00 %
S.D. :
ZELENY :
ASH CONT.:
GLUTEN :

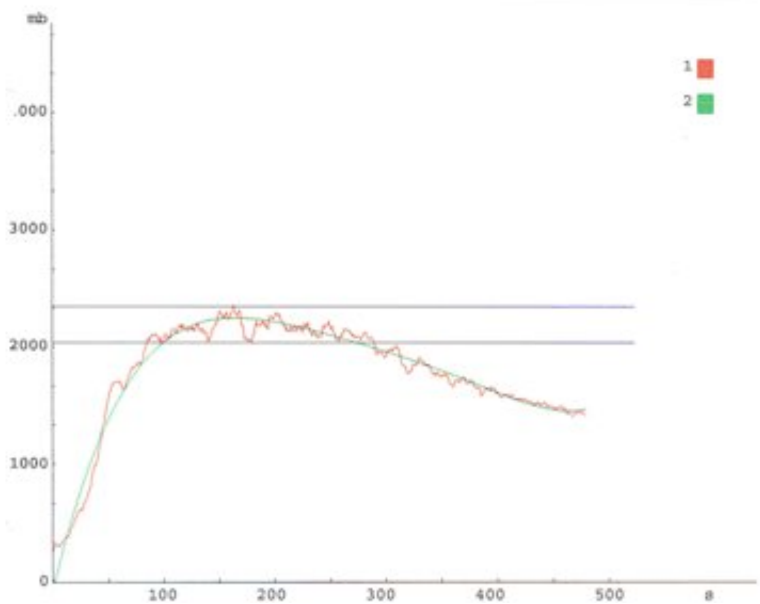
LAB.HYGROM.:
MILL :HMI
FN VALUE :
EXTRAC.R.:
WAC =58.0 % b 14

RESULTS

H2O = 12.40%
HYDHA = 53.4 % b 15
PrMax = 2257 mb
TPrMax = 165 s
Tol = 274 s
D250 = 146 mb
D450 = 782 mb
WAC = 56.2 % b 15

PROTOCOL : CHOPIN PrMax MINI : 0 PrMax TARGET : 2200

V:d2.8C +5.9



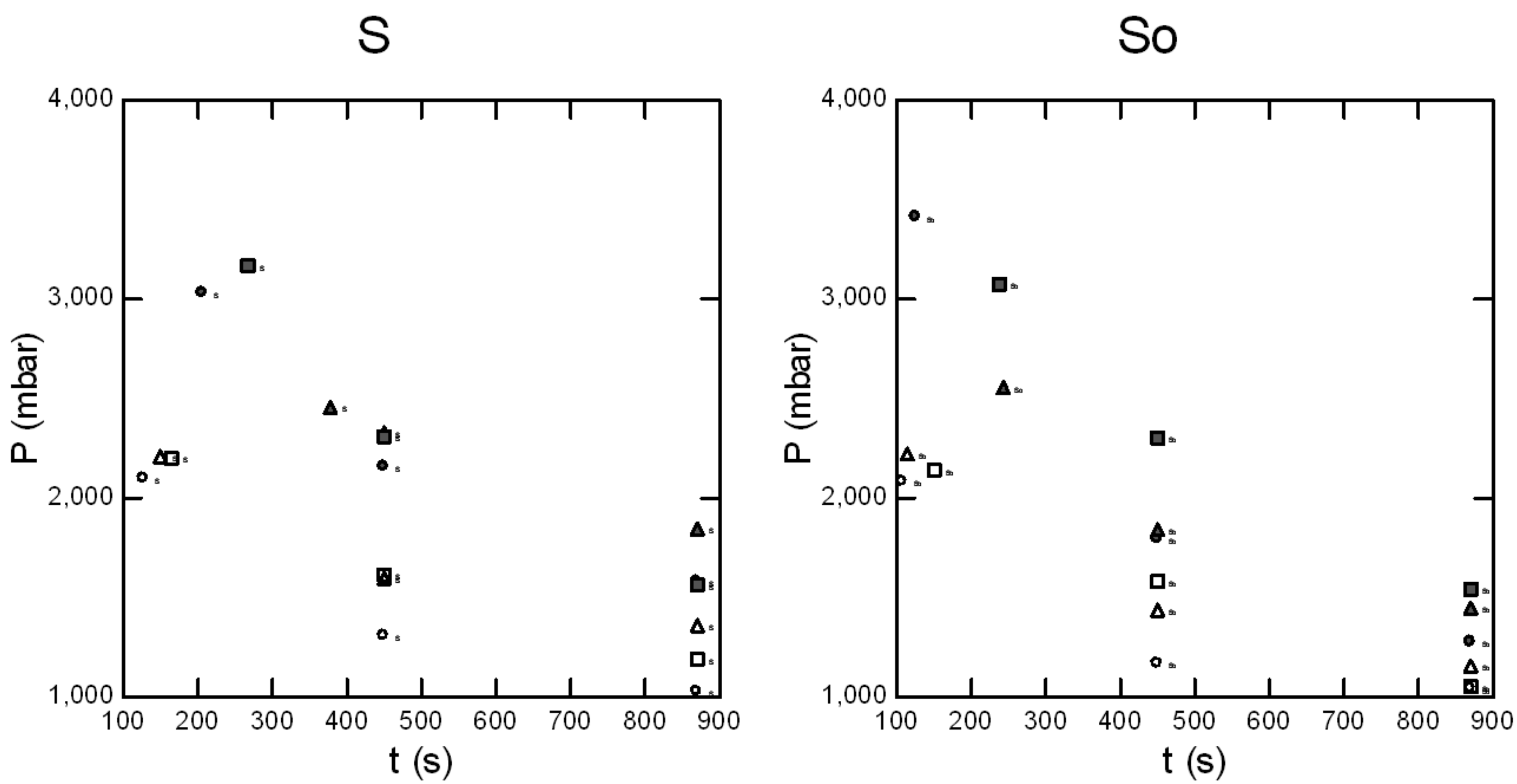
Objectives

- Can we predict the pressure recorded by consistograph from hydration, composition, alvograph variables and kneading time?
- Can we predict the dough properties (especially volume after leavening) from composition, pressure and kneading time?
- Can we explain what we predict?

Potential approaches

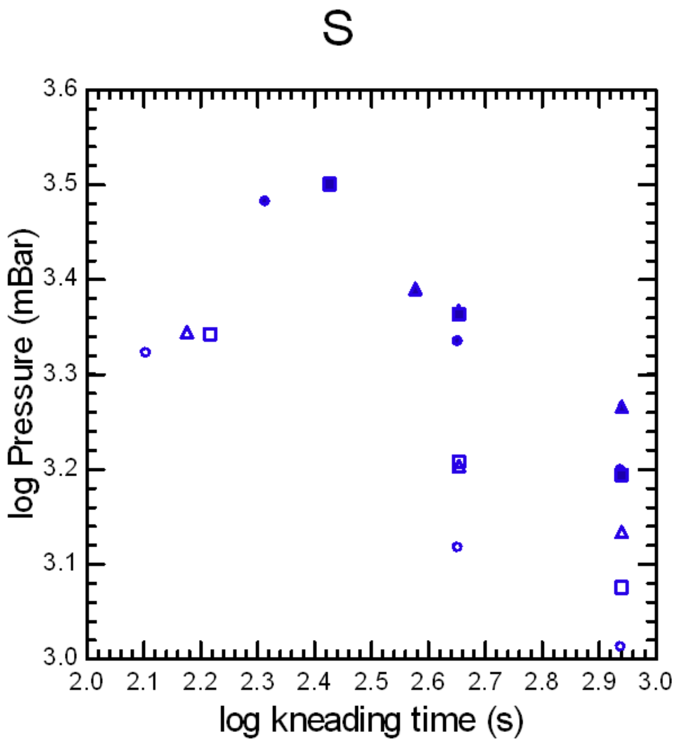
- Univariate or multivariate approach with ANOVA to test significance of effects (there is little to learn here)
- Descriptive approach using PCA to explore relationships among variables and observations
- Predictive approach:
  - Univariate regression (linear? Non linear?)
  - Multivariate regression
    - **Multiple linear regression** (MLR): high risk of collinearity, poor estimates of coefficients, poor predictive ability, risk of overfitting
    - **Principal Component Regression**: problems of collinearity reduced or cancelled but results contaminated by the part of X variance in which we are not interested in
    - Partial Least Squares Regression: can we prodict volume from everything else; can we predict a set (or subset) of independent variables from a set (or subse) of independent variables (PLS1 and PLS2 models)

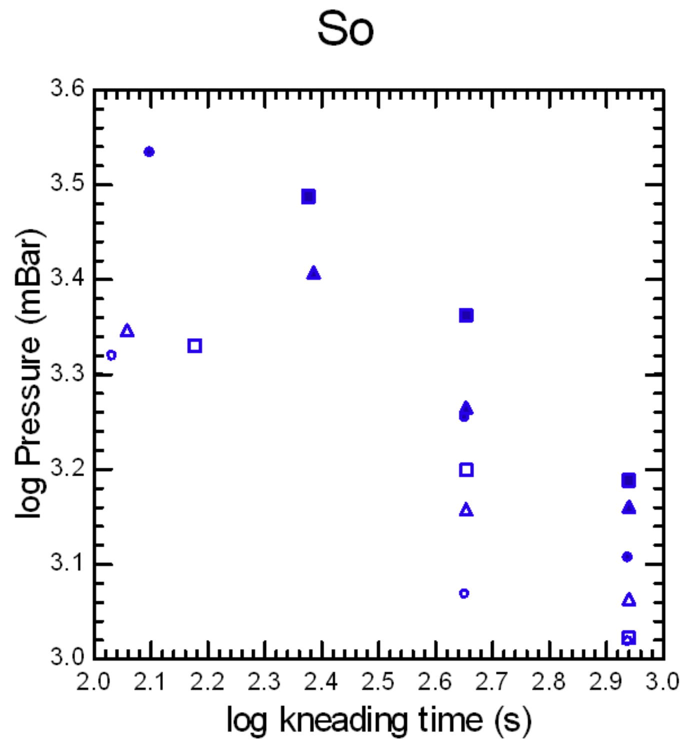
1. Explore the data file breadleavening.xls



Relationship between kneading time and pressure for three varieties (O L, △ New, □ G2), 2 milling sizes (S and So), optimal (empty symbols) and suboptimal (closed symbols) hydration

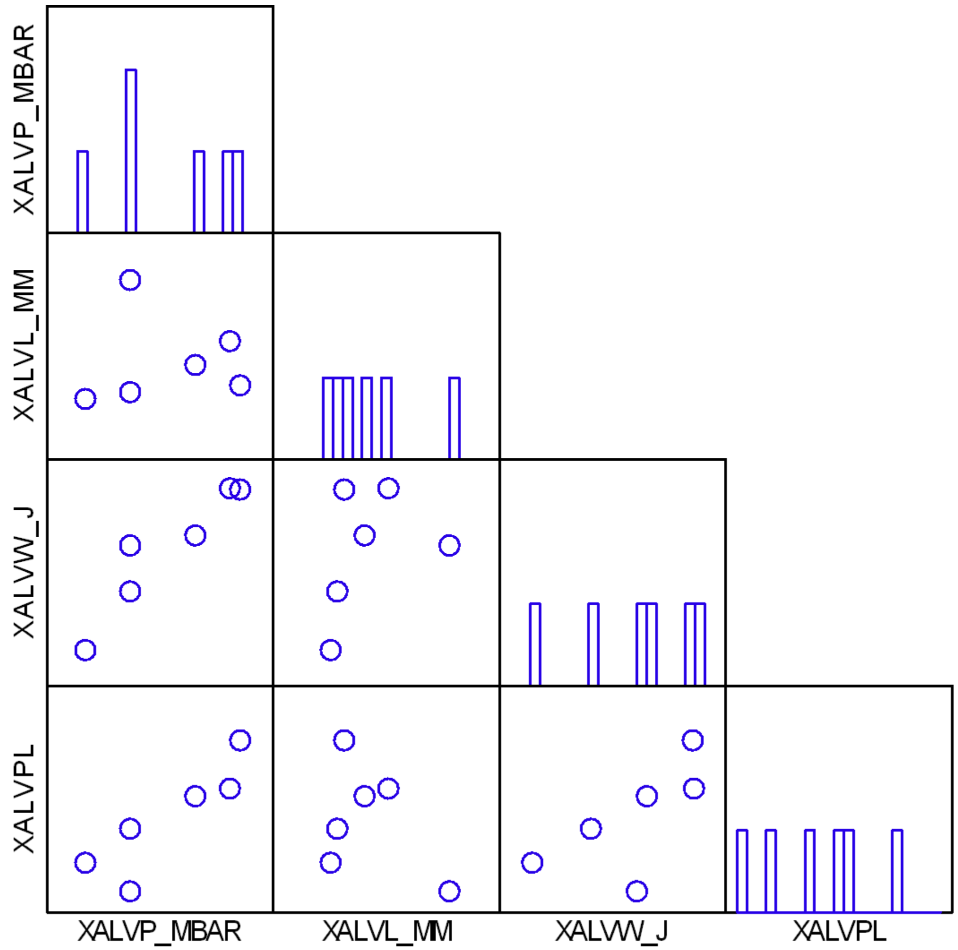
2. Transformation can linearize some relationships



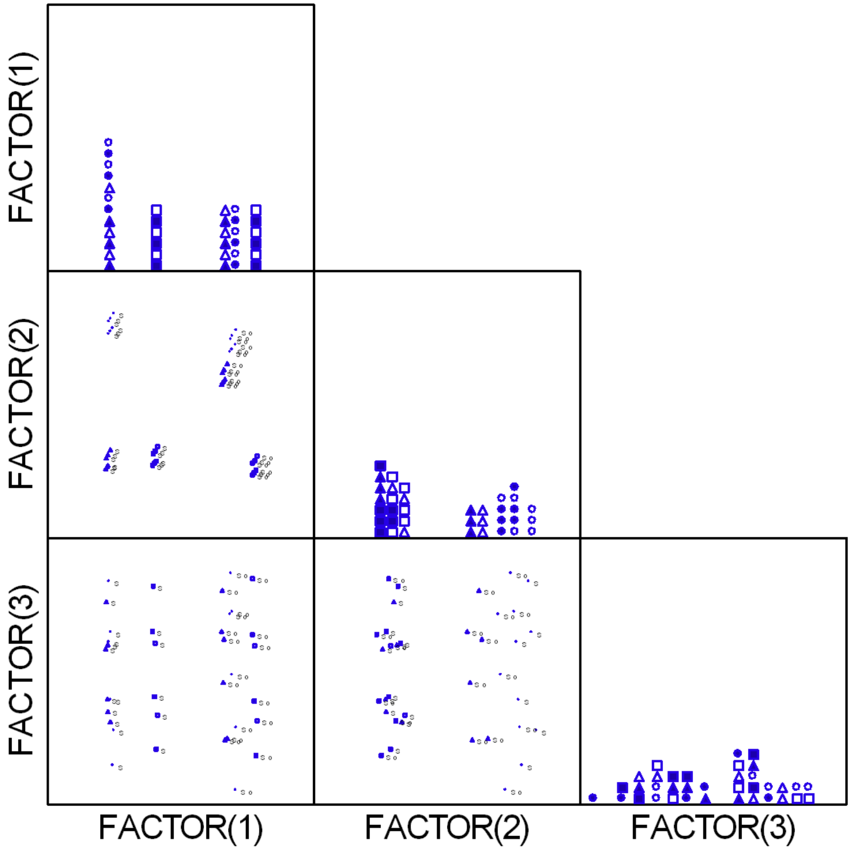


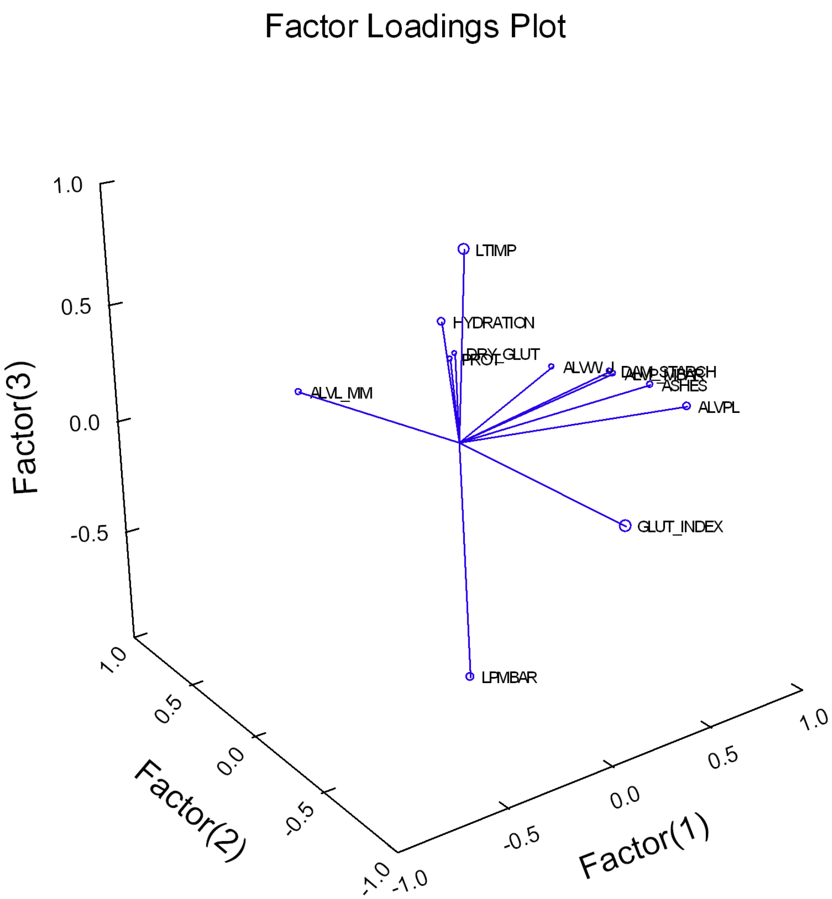
Log transformation of both kneading time and pressure results in determination coefficients close to 1 for most combinations

3. Some X (and Y) variables are not independent



PCA on X variables, (correlation matrix, varimax rotation): 42.1+29.2+15.0=86.3% expl. variance





A PCR on X variables, LPMBAR as y variable

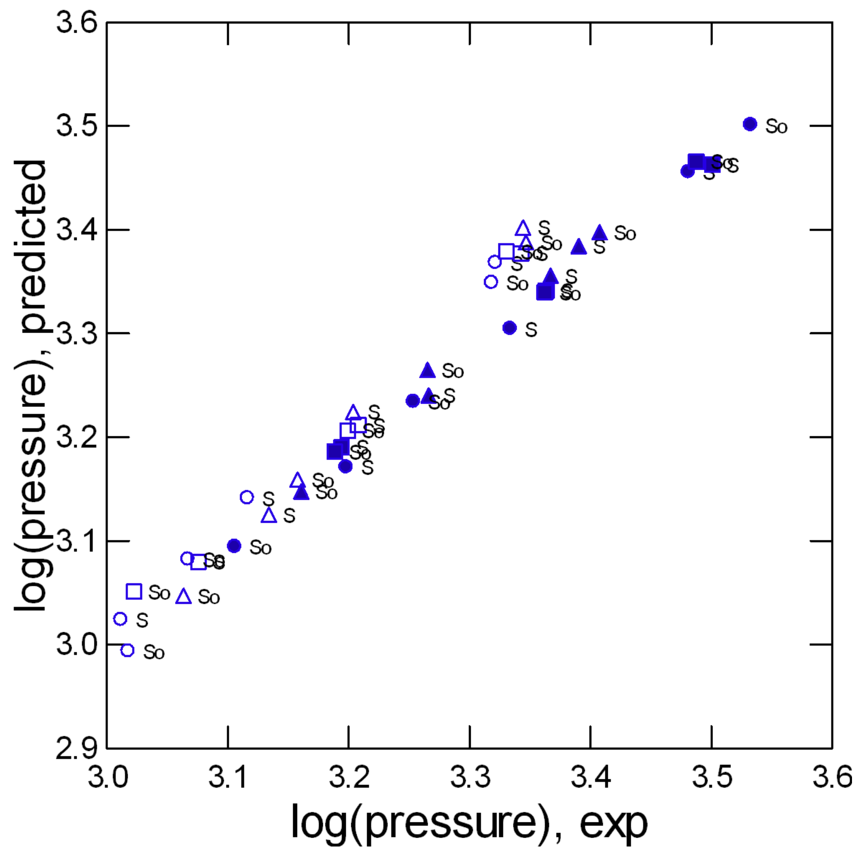
Dependent Variable	LPMBAR
N	36
Multiple R	0.9843
Squared Multiple R	0.9689
Adjusted Squared Multiple R	0.9660
Standard Error of Estimate	0.0267

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t	p-Value
CONSTANT	3.2539	0.0044	0.0000	.	732.1010	0.0000
FACTOR(1)	-0.0092	0.0045	-0.0636	1.0000	-2.0421	0.0495
FACTOR(2)	-0.0240	0.0045	-0.1661	1.0000	-5.3296	0.0000
FACTOR(3)	-0.1400	0.0045	-0.9681	1.0000	-31.0625	0.0000

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
CONSTANT	3.2539	3.2449	3.2630	.
FACTOR(1)	-0.0092	-0.0184	0.0000	1.0000
FACTOR(2)	-0.0240	-0.0332	-0.0148	1.0000
FACTOR(3)	-0.1400	-0.1492	-0.1308	1.0000



PLS1 on X variables (prediction of LPMBAR)

Dependent Variable(s): LPMBAR

Independent Variable(s): HYDRATION ASHES DAM_STARCH GLUT_INDEX DRY_GLUT PROT ALVP_MBAR ALVL_MM ALVW_J ALVPL LTIMP

Number of Observations : 36
Number of Factors Extracted : 4

The SIMPLS algorithm has been used to estimate the model.

Estimates of Regression Coefficients

	ESTIMATE	Standard Error
Constant	0.0000	0.0473
HYDRATION	-0.6575	0.0564
ASHES	0.0214	0.0393
DAM_STARCH	-0.0189	0.0329
GLUT_INDEX	-0.0137	0.0540
DRY_GLUT	-0.0078	0.0183
PROT	-0.0110	0.0395
ALVP_MBAR	0.0021	0.0245
ALVL_MM	-0.0159	0.0546
ALVW_J	0.0032	0.0372
ALVPL	0.0037	0.0174
LTIMP	-0.8266	0.0608

Analysis of Variance for LPMBAR

Source	SS	df	Mean Squares	F-Ratio	p-Value
Regression	33.0000	4	8.2500	127.8746	0.0000
Error	2.0000	31	0.0645		

Percent Variation Explained by Factors for Predictors and Responses

Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum. Percentage	Percentage	Cum. Percentage
1	31.7379	31.7379	56.0970	56.0970
2	27.7978	59.5357	35.3439	91.4408
3	24.5615	84.0972	2.2337	93.6746
4	9.9202	94.0175	0.6111	94.2857

PLS1 on X variables (prediction of LPMBAR)

X-Loadings

	FACTOR1	FACTOR2	FACTOR3	FACTOR4
HYDRATION	-3.3350	-0.2530	-0.6417	-4.8018
ASHES	-3.0104	3.6058	-3.3651	0.8652
DAM_STARCH	-3.6356	3.9754	-2.1024	0.8374
GLUT_INDEX	3.1521	-1.2836	-4.6449	0.8302
DRY_GLUT	-4.3751	3.3343	1.9407	0.0476
PROT	-4.0169	2.9841	1.9183	0.1632
ALVP_MBAR	-3.5668	3.9141	-2.2761	0.6320
ALVL_MM	-2.1904	0.7034	4.8281	-0.8689
ALVW_J	-3.8758	3.7688	-0.6748	0.2718
ALVPL	-2.0018	3.1379	-4.4684	1.0204
LTIMP	-2.6509	-4.0763	-0.6869	3.2708

Y-Loadings

	FACTOR1	FACTOR2	FACTOR3	FACTOR4
LPMBAR	4.4310	3.5172	0.8842	0.4625

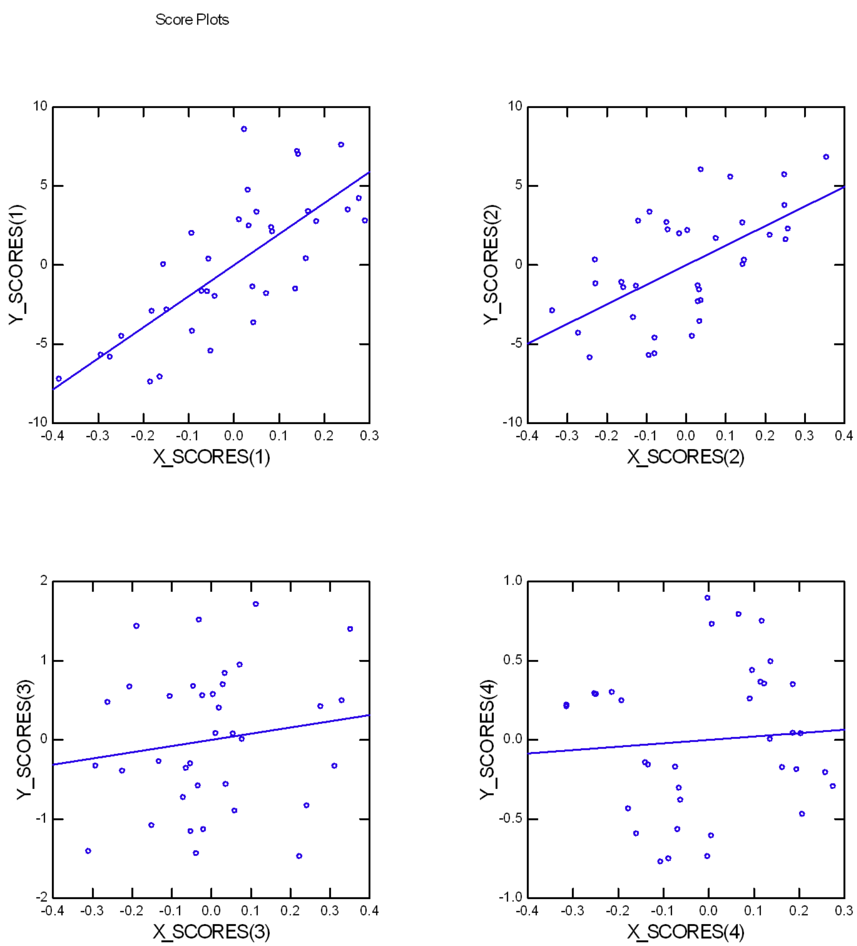
The "Leave One Out" method has been used for cross-validation.

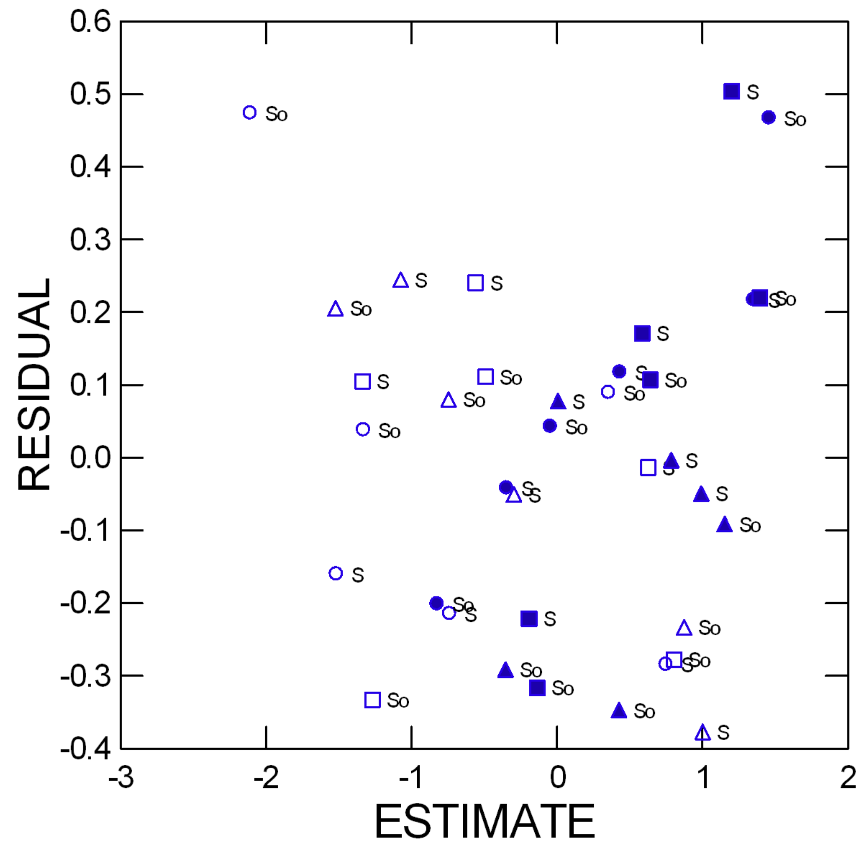
Number of Factors Extracted after Cross-Validation : 4

Cross-Validation Statistics

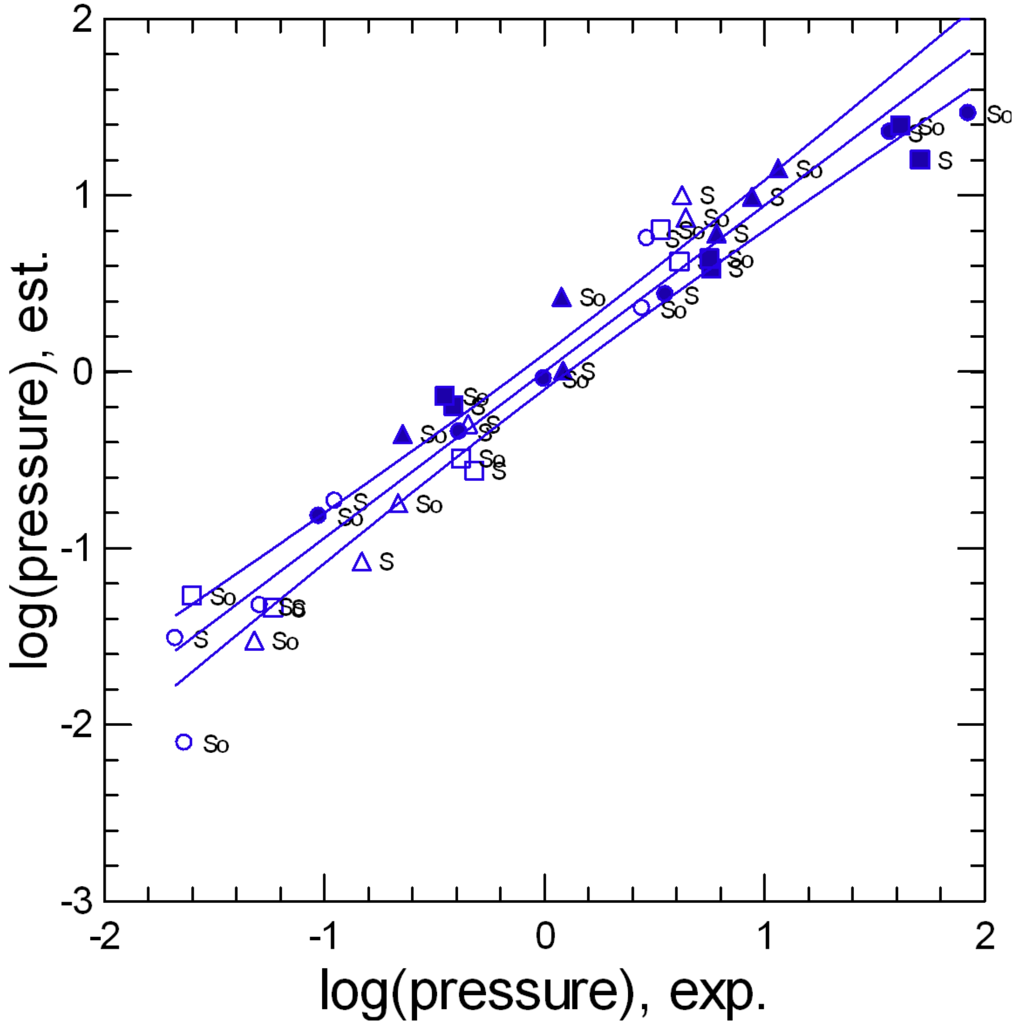
	LPMBAR
PRESS	2.9776
R-Square(Prediction)	0.9149

PLS1 on X variables (prediction of LPMBAR)

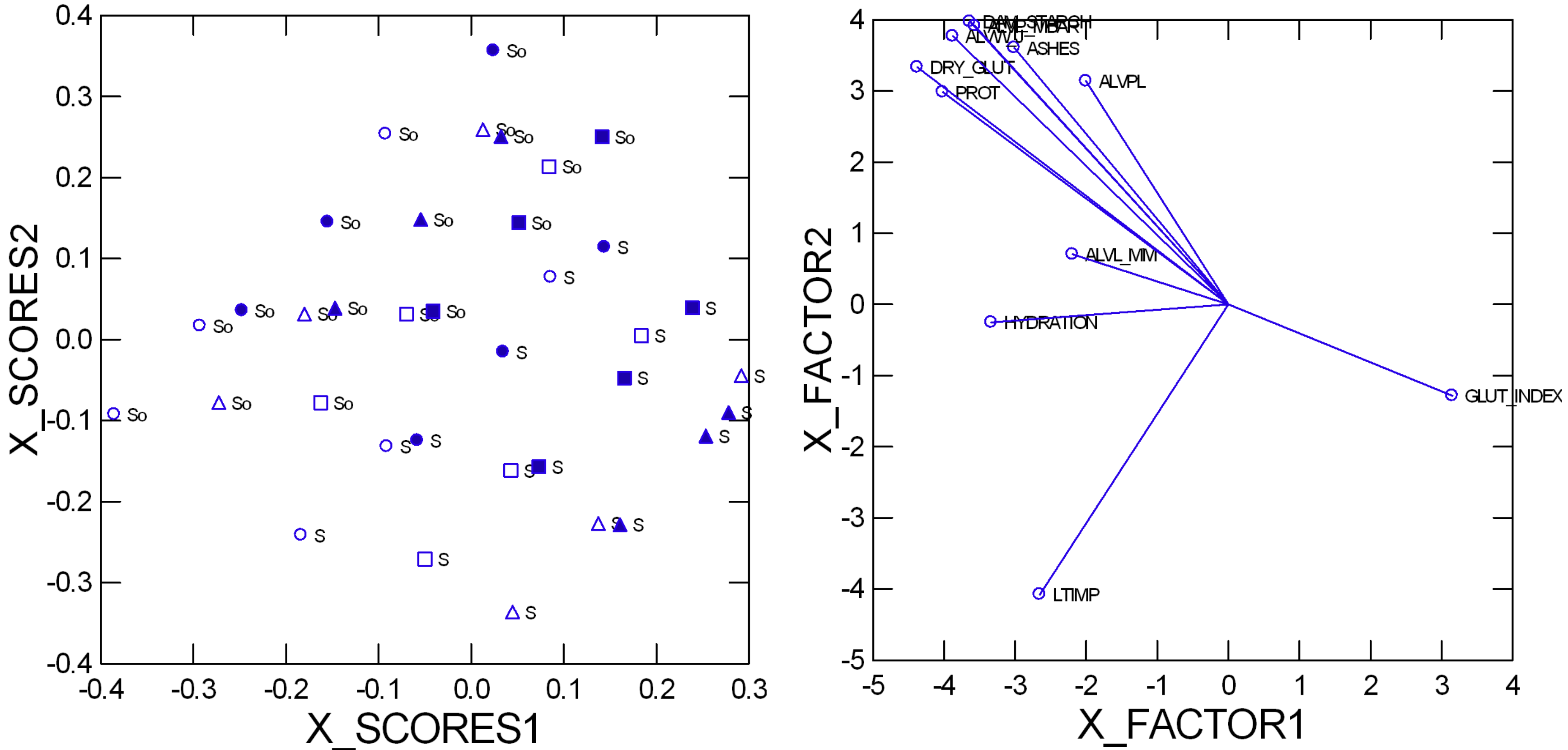




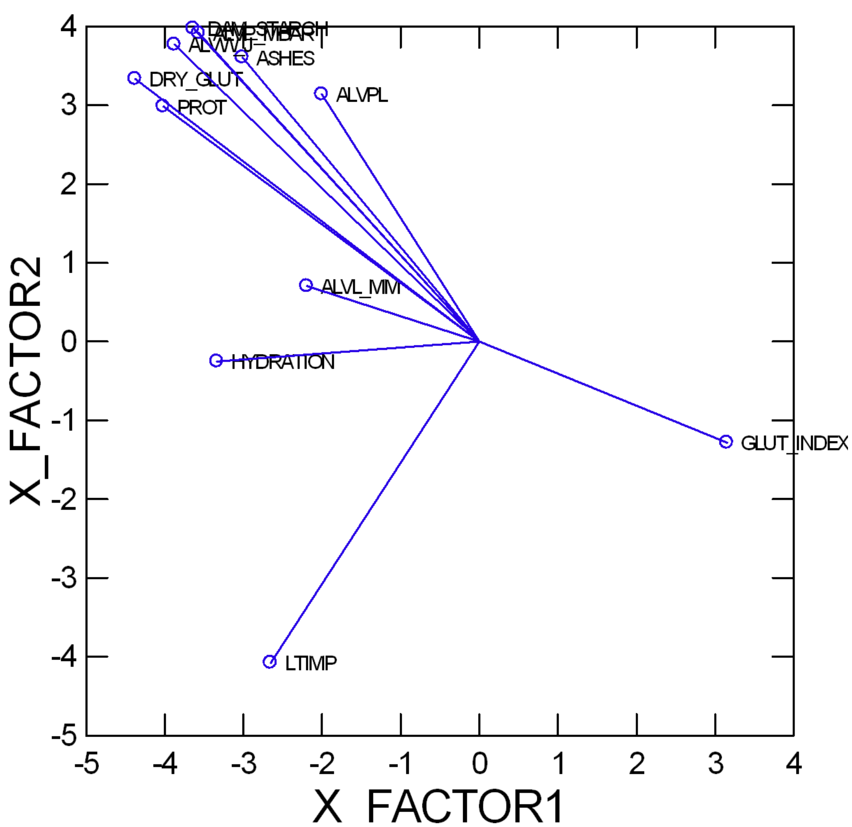
PLS1 on X variables (prediction of LPMBAR)

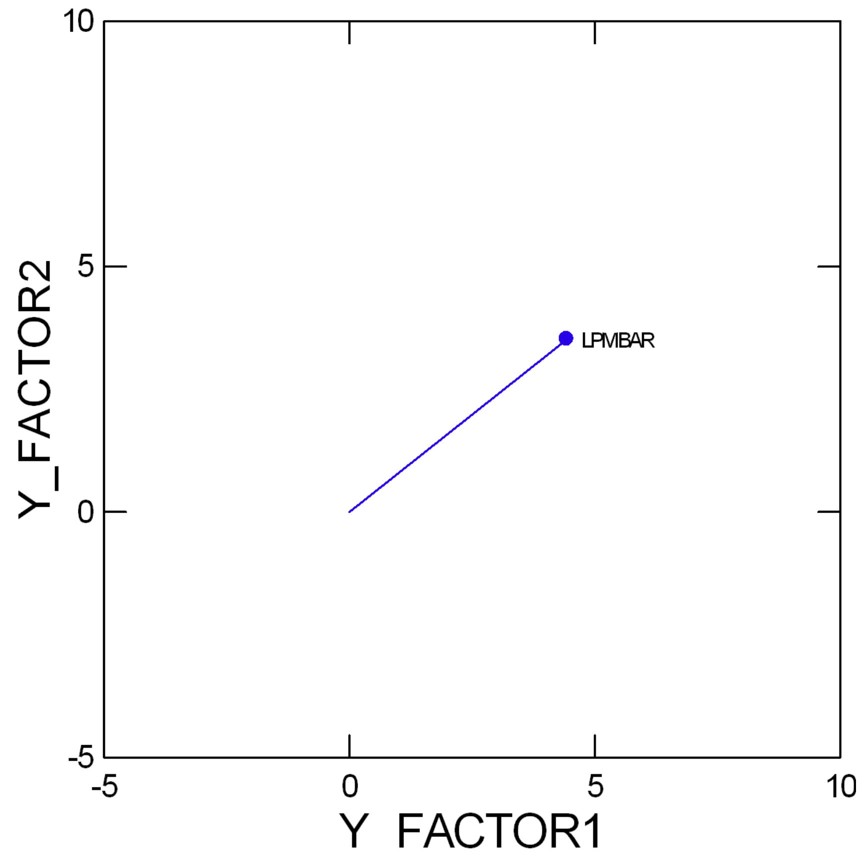


PLS1 on X variables (prediction of LPMBAR)



PLS1 on X variables (prediction of LPMBAR)

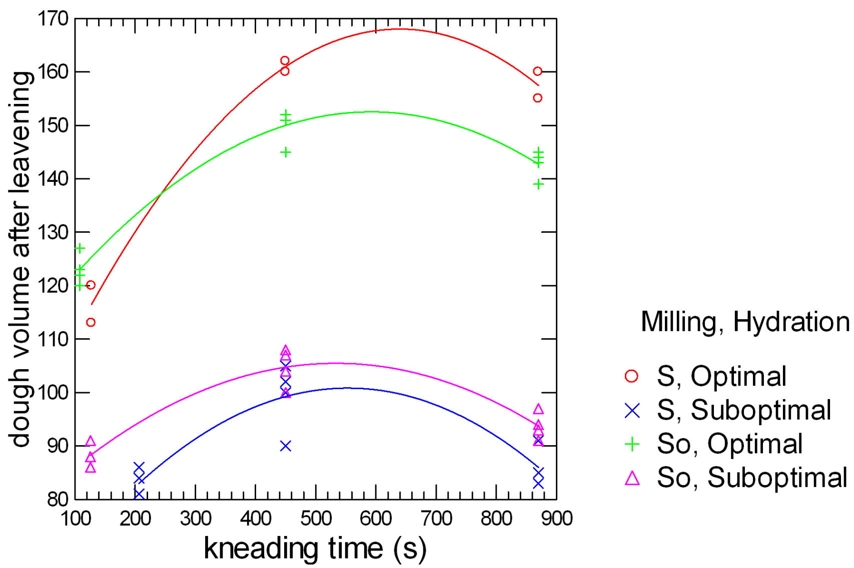




PCR, PLS and pills

- Blue pill: PCR fits the model better than PLS but there might be a problem with overfitting (red pill); crossvalidation may show this
- PLS1 gives a worse fit (but not too bad) but does a better job in relating **X** and y and in revealing the structure; it is crossvalidated

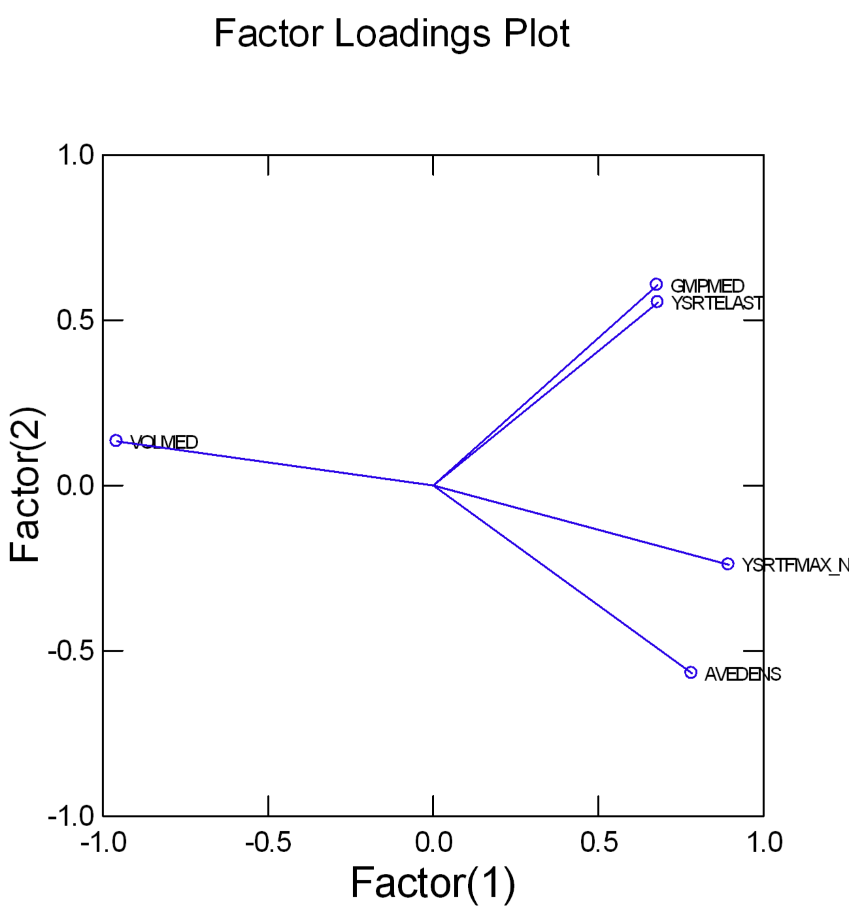
Problems with the Y data set



CV=new, a quadratic smoother is shown

- Here we are primarily interested in predicting volume after leavening (this is what the baker needs to know)
- Some relationships are not only nonlinear, but also non-monotonic
- It is also interesting to predict the y variables in a single model
- An empirical model based on quadratic transformations may help (it is a blue pill)

PCA on Y variables



- A PCA on the correlation matrix of the Y variables explains 86% of the variance with 2 components
- A PLS2 model may be justified

PLS1: volume

▼Partial Least Squares Regression

Dependent Variable(s): LEAVDOUGHVOLUME

Independent Variable(s): HYDRATION KNEADTIME_SEC XASHES XDAM_STARCH XGLUT_INDEX XDRY_GLUTEN XPROT XALVP_MBAR XALVL_MM XALVW_J XALVPL KNEADTIME_SEC2 XPR_MBAR

Number of Observations : 138
Number of Factors Extracted : 6

The SIMPLS algorithm has been used to estimate the model.

Estimates of Regression Coefficients

	ESTIMATE	Standard Error
Constant	0.0000	0.0276
HYDRATION	0.8877	0.0451
KNEADTIME_SEC	0.2789	0.1099
XASHES	0.4262	0.1533
XDAM_STARCH	-0.5074	0.1150
XGLUT_INDEX	-0.1720	0.0880
XDRY_GLUTEN	-0.1084	0.0152
XPROT	-0.1684	0.0444
XALVP_MBAR	-0.0647	0.0500
XALVL_MM	0.1138	0.0610
XALVW_J	0.2030	0.0401
XALVPL	-0.1300	0.0430
KNEADTIME_SEC2	-0.2844	0.1338
XPR_MBAR	-0.0838	0.0562

Analysis of Variance for LEAVDOUGHVOLUME

Source	SS	df	Mean Squares	F-Ratio	p-Value
Regression	124.1579	6	20.6930	211.0853	0.0000
Error	12.8421	131	0.0980		

Percent Variation Explained by Factors for Predictors and Responses

Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum. Percentage	Percentage	Cum. Percentage
1	27.7015	27.7015	54.9939	54.9939
2	28.0670	55.7685	24.2202	79.2140
3	17.5326	73.3011	7.2139	86.4280
4	4.7976	78.0988	1.6135	88.0415
5	18.8129	96.9116	0.2995	88.3409
6	2.0377	98.9493	2.2852	90.6262

PLS1: volume

X-Loadings

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
HYDRATION	9.6968	3.5707	5.0467	-1.2011	0.4526	0.9482
KNEADTIME_SEC	1.8219	0.1401	-6.2736	-0.1845	9.5116	-1.8012
XASHES	3.8098	-8.6707	5.8833	-1.4660	2.9817	0.9363
XDAM_STARCH	5.2084	-9.3773	3.9777	-1.6909	0.8749	-1.4469
XGLUT_INDEX	-8.0574	2.1987	5.4331	-1.2465	5.1488	-2.5703
XDRY_GLUTEN	8.4046	-7.2654	-1.5047	-1.1146	-3.0293	0.9220
XPROT	7.8135	-6.5446	-2.2014	-3.8931	-3.2490	1.4279
XALVP_MBAR	5.0295	-9.0174	4.9155	1.5101	1.6764	-0.8969
XALVL_MM	6.2927	-1.2004	-5.5329	6.0991	-5.2033	-0.2468
XALVW_J	6.1583	-8.4326	3.2224	4.1087	0.3792	-0.6552
XALVPL	1.3844	-7.8575	7.3489	-0.7833	4.1173	-1.2633
KNEADTIME_SEC2	2.1777	0.0517	-6.1116	-0.0562	9.4225	-1.5844
XPR_MBAR	-6.9490	-2.8162	1.9613	2.0079	-7.2608	3.7791

Y-Loadings

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6
LEAVDOUGHVOLUME	8.6800	5.7604	3.1437	1.4868	0.6405	1.7694

PLS1: volume

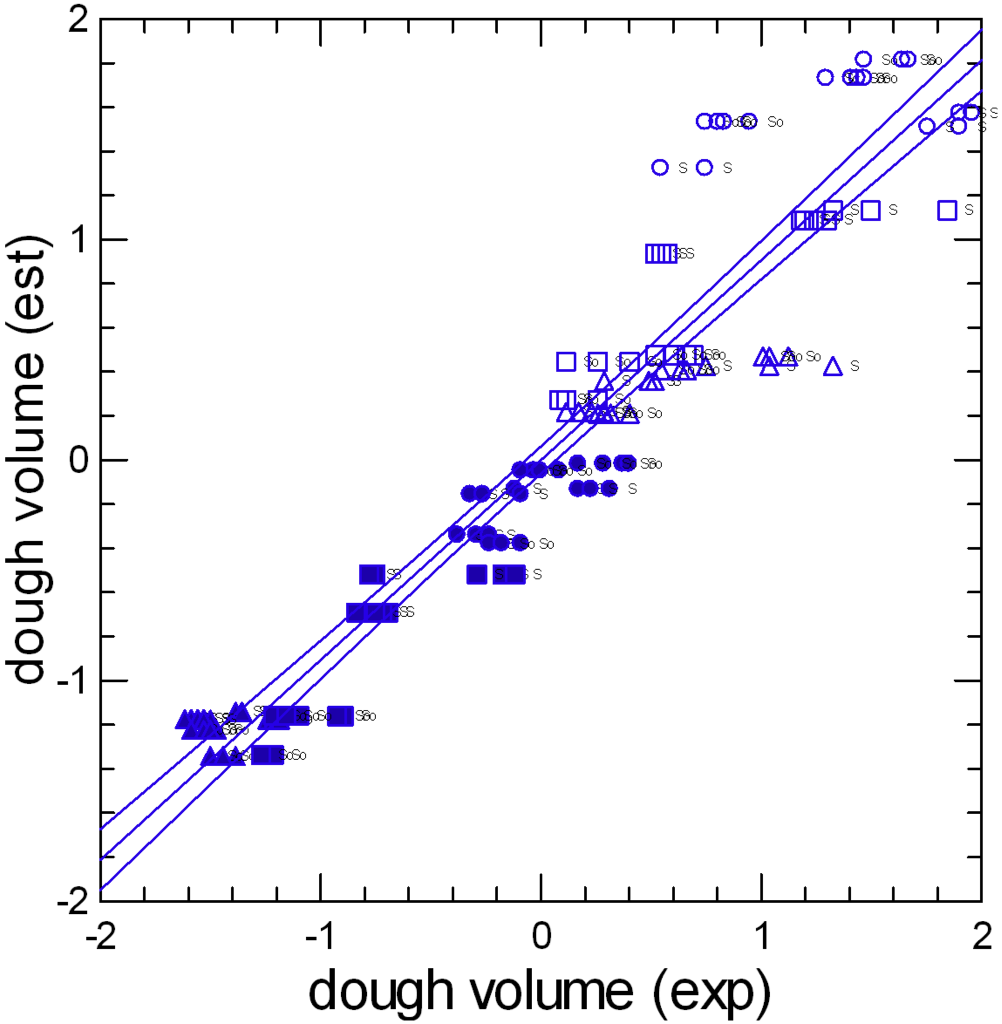
The "Random Exclusion" method has been used for cross-validation.

Number of Repetitions	:	1
Test Set Size	:	69
Number of Factors Extracted after Cross-Validation	:	6

Cross-Validation Statistics

	LEAVDOUGHVOLUME
Average PRESS	6.9359
R-Square(Prediction)	0.9494

PLS1: volume



PLS2, all Y variables

Dependent Variable(s): YSRTFMAX_N YSRTELAST AVEVOL AVEGMP AVEDENS

Independent Variable(s): PR_MBAR XASHES XDAM_STARCH XGLUT_INDEX XDRY_GLUTEN XPROT XALVP_MBAR XALVL_MM XALVW_J XALVPL KNEADTIME_SEC KNEADTIME_SEC2

Number of Observations : 216
Number of Factors Extracted : 7

The SIMPLS algorithm has been used to estimate the model.

The SIMPLS algorithm has been used to estimate the model.

Estimates of Regression Coefficients

	YSRTFMAX_N	YSRTELAST	AVEVOL	AVEGMP	AVEDENS
Constant	0.0000	0.0000	0.0000	0.0000	0.0000
PR_MBAR	0.7617	0.9083	-1.1007	1.0815	0.6347
XASHES	-0.9327	0.7176	0.6861	0.2374	-1.1422
XDAM_STARCH	1.0884	-0.5514	-0.7943	-0.2299	1.1847
XGLUT_INDEX	0.9276	-0.0434	-0.6821	0.1784	0.7835
XDRY_GLUTEN	0.2576	-0.0225	-0.1535	-0.0152	0.2062
XPROT	0.5614	0.0505	-0.3213	0.0495	0.4081
XALVP_MBAR	-0.1168	-0.1404	0.0293	-0.1387	-0.0095
XALVL_MM	0.0843	0.0377	-0.1103	0.1805	0.0228
XALVW_J	-0.4918	0.1975	0.2814	0.0900	-0.5082
XALVPL	0.0758	-0.1340	-0.0887	-0.1343	0.1509
KNEADTIME_SEC	1.2292	3.3685	-0.8869	2.8063	-0.8081
KNEADTIME_SEC2	-0.5843	-2.4497	0.1033	-2.0738	1.1383

Standard Error of the Estimated Coefficients

	YSRTFMAX_N	YSRTELAST	AVEVOL	AVEGMP	AVEDENS
Constant	0.0382	0.0426	0.0354	0.0367	0.0412
PR_MBAR	0.0589	0.0658	0.0577	0.0528	0.0587
XASHES	0.1820	0.2805	0.1222	0.2485	0.1670
XDAM_STARCH	0.1304	0.1655	0.0945	0.1619	0.1180
XGLUT_INDEX	0.1480	0.3345	0.1210	0.2514	0.1223
XDRY_GLUTEN	0.0285	0.0439	0.0270	0.0279	0.0228
XPROT	0.0822	0.1720	0.0678	0.1240	0.0635
XALVP_MBAR	0.0757	0.1589	0.0502	0.1278	0.0641
XALVL_MM	0.0794	0.1877	0.0649	0.1432	0.0789
XALVW_J	0.0457	0.0774	0.0363	0.0660	0.0576
XALVPL	0.0559	0.1175	0.0355	0.0958	0.0514
KNEADTIME_SEC	0.2669	0.2133	0.1810	0.3074	0.2336
KNEADTIME_SEC2	0.2650	0.2131	0.1849	0.3007	0.2205

PLS2, all Y variables

Analysis of Variance for YSRTFMAX_N

Source	SS	df	Mean Squares	F-Ratio	p-Value
Regression	152.2146	7	21.7449	72.0382	0.0000
Error	62.7854	208	0.3019		

Analysis of Variance for YSRTELAST

Source	SS	df	Mean Squares	F-Ratio	p-Value
Regression	136.7719	7	19.5388	51.9516	0.0000
Error	78.2281	208	0.3761		

Analysis of Variance for AVEVOL

Source	SS	df	Mean Squares	F-Ratio	p-Value
Regression	160.8945	7	22.9849	88.3619	0.0000
Error	54.1055	208	0.2601		

Analysis of Variance for AVEGMP

Source	SS	df	Mean Squares	F-Ratio	p-Value
Regression	156.8806	7	22.4115	80.2072	0.0000
Error	58.1194	208	0.2794		

Analysis of Variance for AVEDENS

Source	SS	df	Mean Squares	F-Ratio	p-Value
Regression	141.0423	7	20.1489	56.6671	0.0000
Error	73.9577	208	0.3556		

PLS2, all Y variables

Percent Variation Explained by Factors for Predictors and Responses

Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum. Percentage	Percentage	Cum. Percentage
1	35.4660	35.4660	14.0336	14.0336
2	31.9123	67.3783	7.6049	21.6385
3	4.6729	72.0512	22.7700	44.4085
4	22.1707	94.2219	1.4715	45.8800
5	4.8168	99.0387	4.9453	50.8252
6	0.5463	99.5850	7.7675	58.5928
7	0.1598	99.7448	10.9704	69.5632

PLS2, all Y variables

X-Loadings

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7
PR_MBAR	6.5414	-2.8809	5.1622	11.3163	3.0164	0.3071	-0.1046
XASHES	-5.5301	-13.2250	-2.0289	-0.4402	0.5795	1.9321	0.1276
XDAM_STARCH	-8.4027	-11.8285	-0.8987	0.6954	-0.1683	-1.7049	0.0398
XGLUT_INDEX	11.9122	-5.2205	-3.8490	-4.6626	1.3372	-2.1774	0.3724
XDRY_GLUTEN	-13.2694	-3.8696	2.5084	3.6003	-2.1554	0.0095	0.1305
XPROT	-11.9634	-3.9490	3.5727	3.2455	-5.6795	0.1181	0.3433
XALVP_MBAR	-8.3098	-11.2034	-2.2724	0.6974	3.7651	-0.1926	-0.2014
XALVL_MM	-10.2415	7.9912	2.1523	4.4757	4.2622	-1.5741	0.1287
XALVW_J	-10.6207	-7.6043	-1.8753	1.9820	6.0628	-0.2566	-0.0753
XALVPL	-2.1917	-13.8104	-3.4172	-1.6378	2.1857	-0.1375	-0.1249
KNEADTIME_SEC	-2.9079	1.6071	3.5576	-13.7460	0.9565	-0.0481	1.1700
KNEADTIME_SEC2	-3.1544	1.3766	4.1799	-13.4890	1.1520	-0.0577	-1.5403

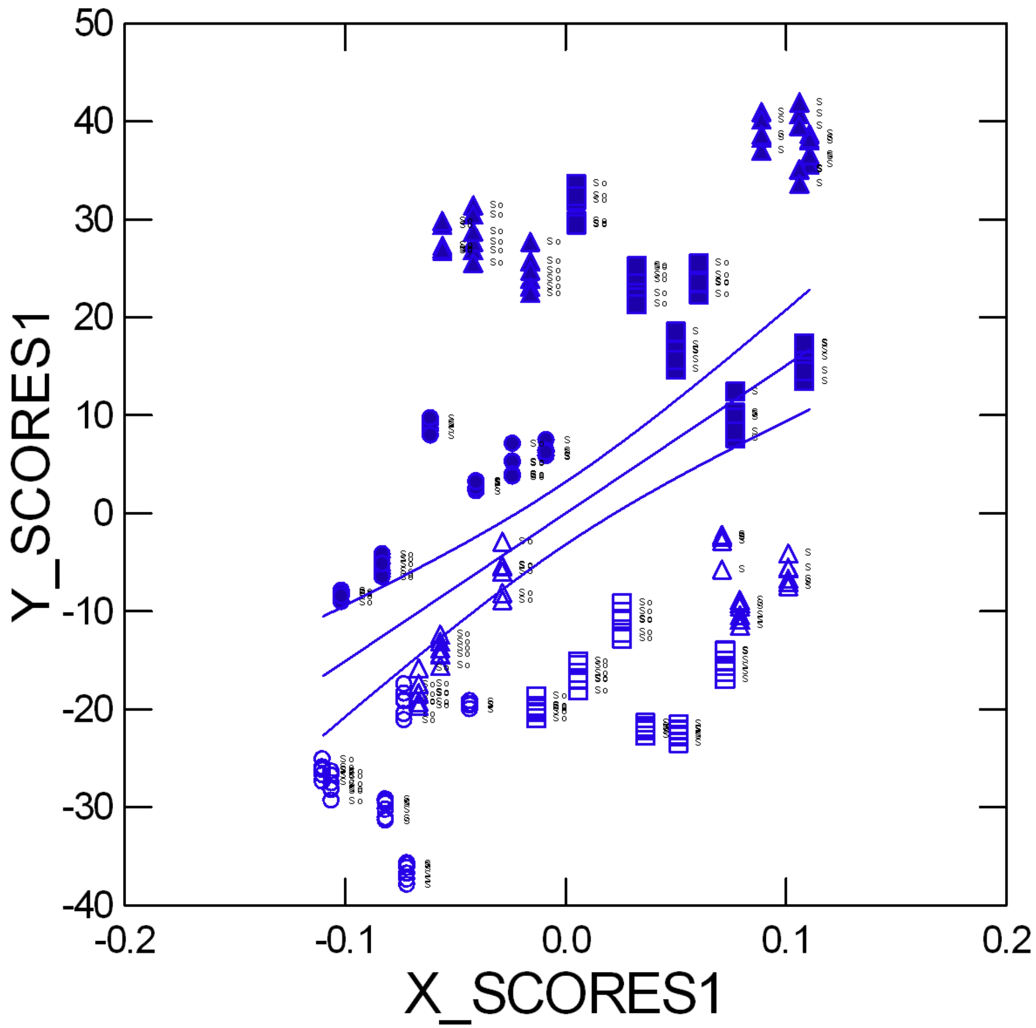
Y-Loadings

	FACTOR1	FACTOR2	FACTOR3	FACTOR4	FACTOR5	FACTOR6	FACTOR7
YSRTFMAX_N	5.3430	-4.9452	6.9681	-1.6061	-3.3233	-5.3946	2.8166
YSRTELAST	2.6927	-1.2053	6.2517	-2.7140	3.8878	2.6205	7.7225
AVEVOL	-6.2186	5.5092	-8.5199	-0.6971	-1.1225	3.9323	-1.4403
AVEGMP	6.4502	1.2184	7.0698	1.7566	4.2227	0.5226	6.5283
AVEDENS	5.8981	-4.8999	5.8795	1.5171	-2.8127	-5.6389	-2.3808

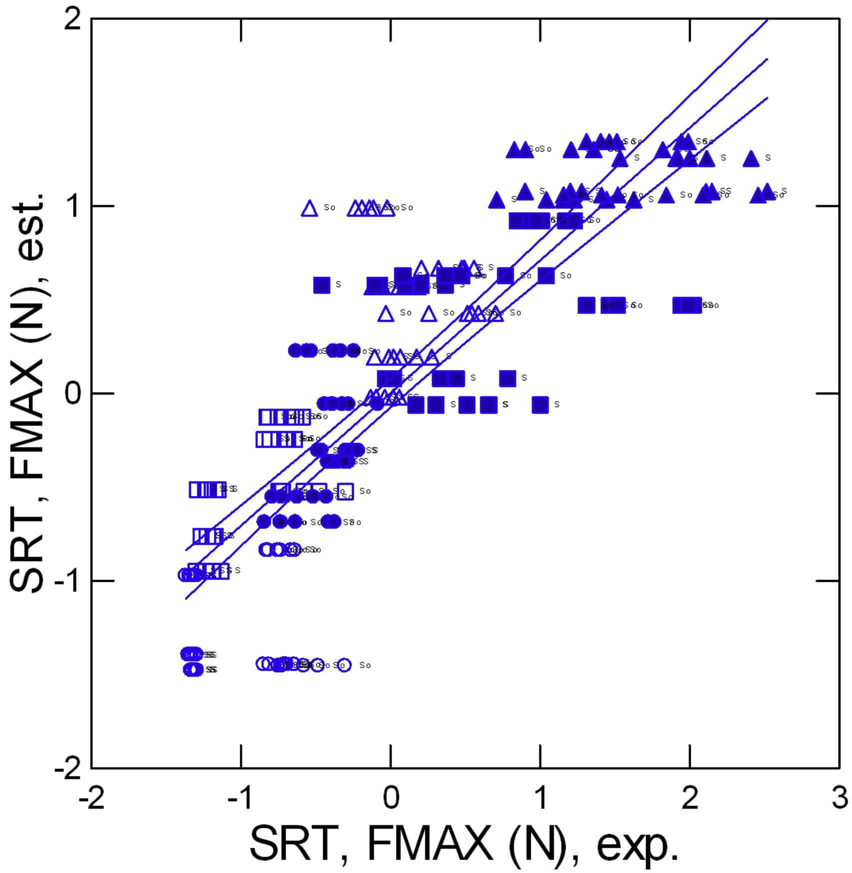
Cross-Validation Statistics

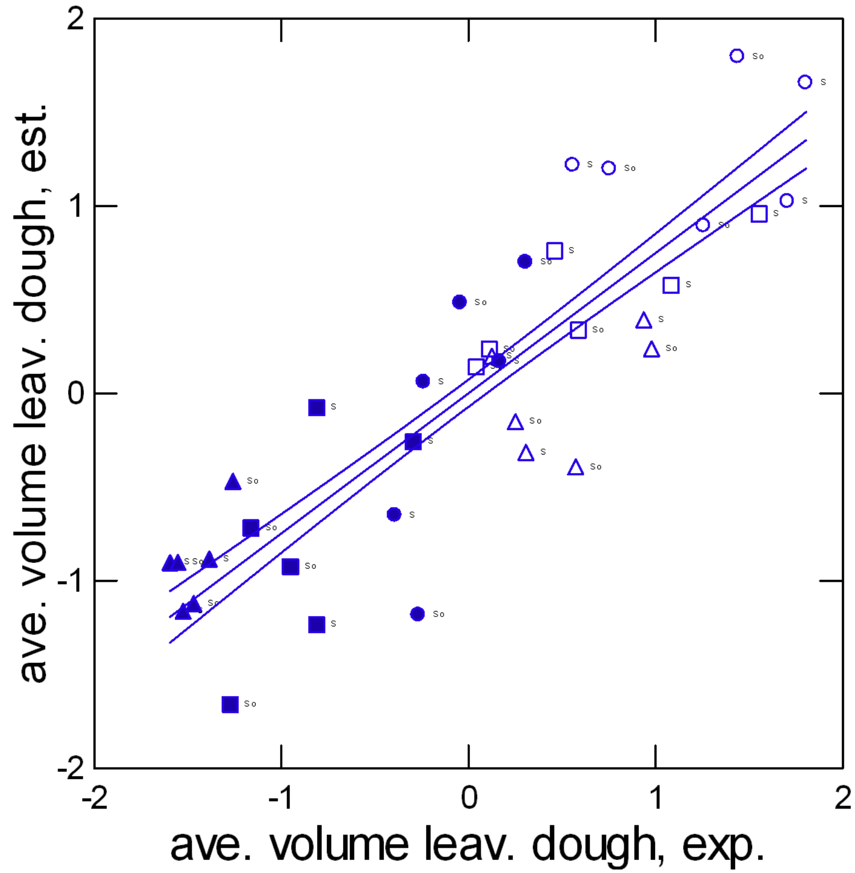
	YSRTFMAX_N	YSRTELAST	AVEVOL	AVEGMP	AVEDENS
Average PRESS	31.1593	42.7524	29.4652	28.9869	40.8482
R-Square(Prediction)	0.8551	0.8012	0.8630	0.8652	0.8100

PLS2, all Y variables

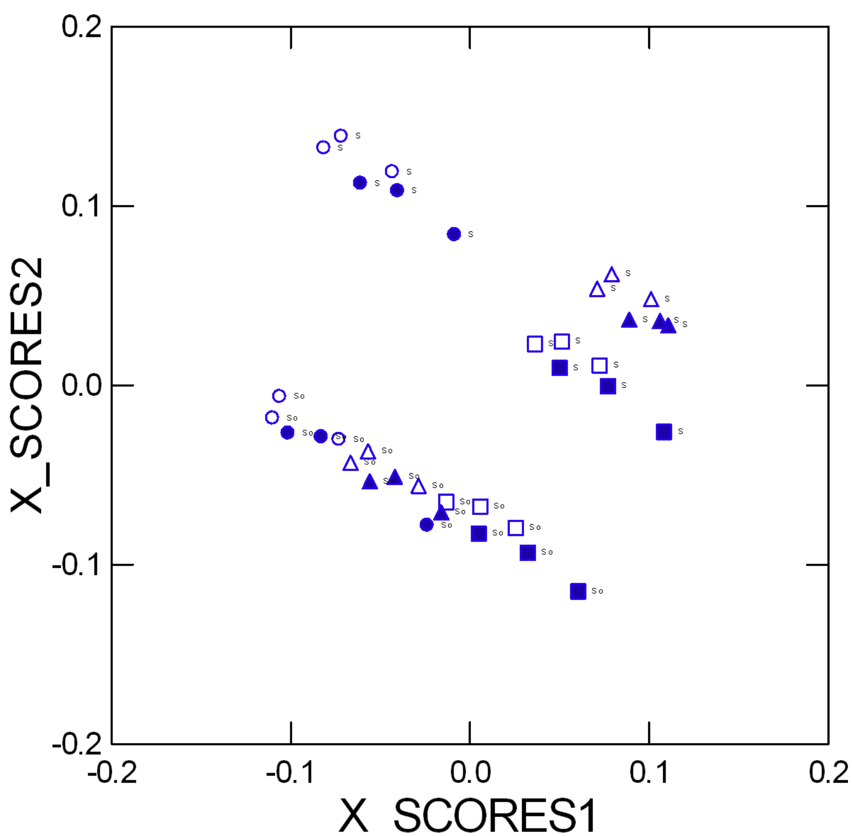


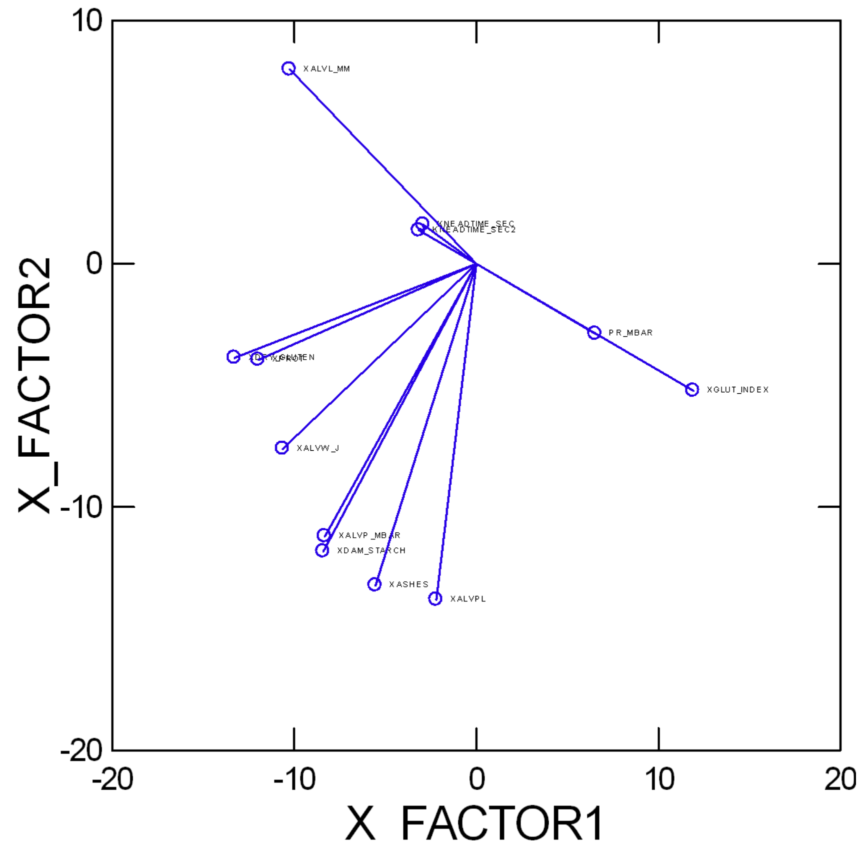
PLS2, all Y variables



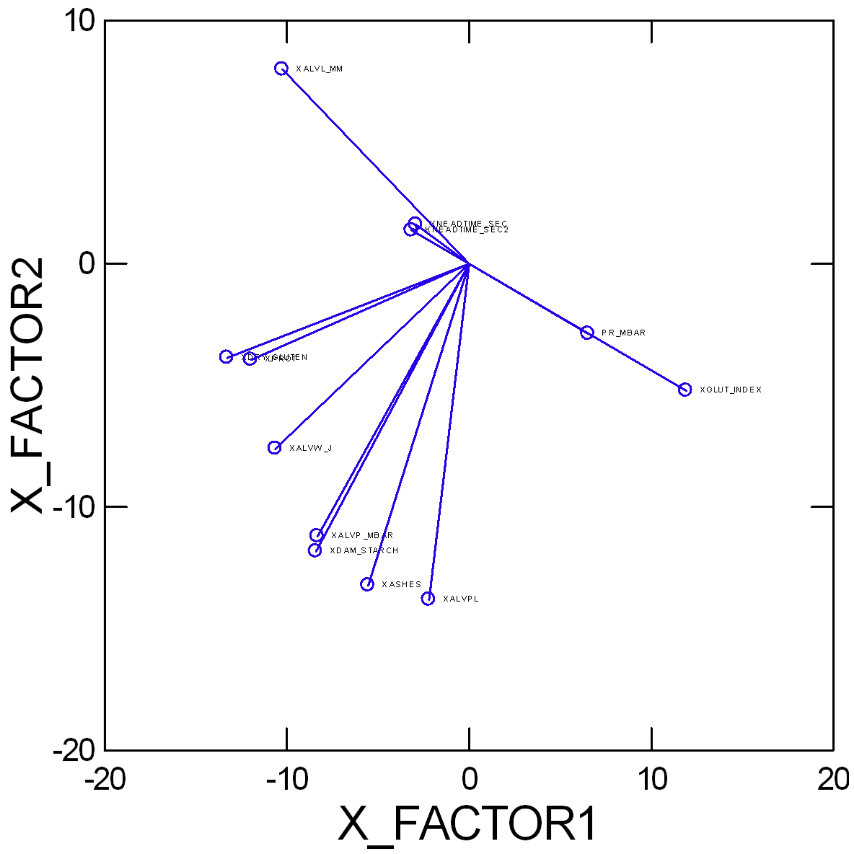


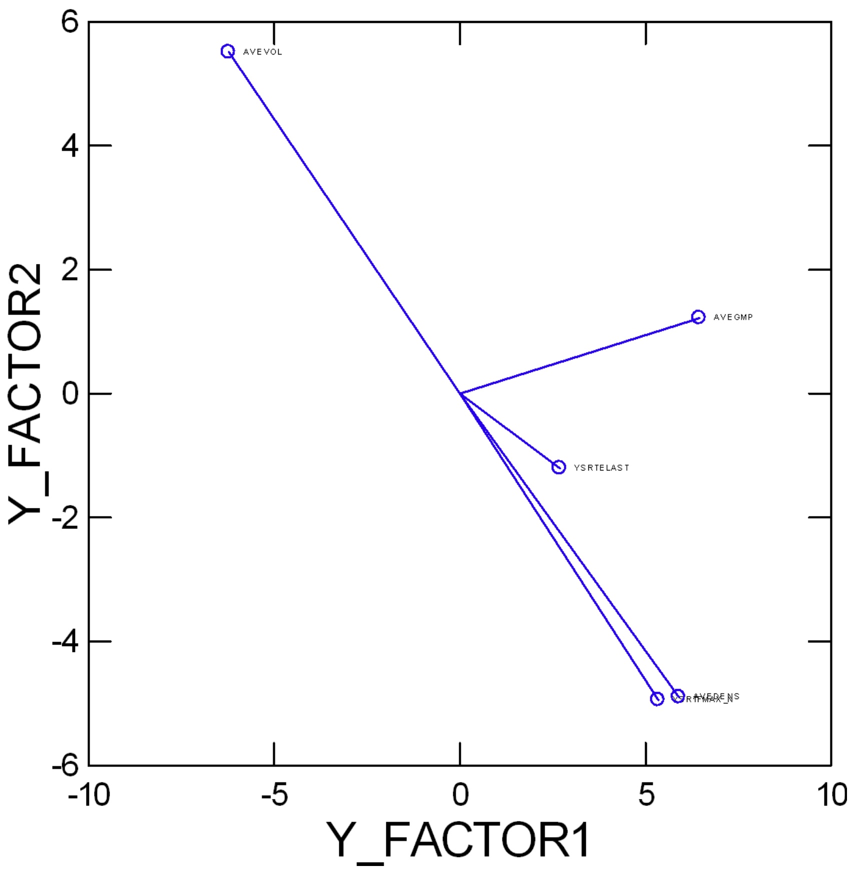
PLS2, all Y variables





PLS2, all Y variables





Oops, too late, gotta go