Principal Component Analysis (PCA)

- Skjult informasjon i data dekomponeres slik at ikke-observerbare trender og fenomener kan detekteres

Benyttes til å finne korrelasjoner mellom variabler og likheter og ulikheter i observasjoner (~prøver, batcher, etc).

- Variablene kan typisk være flere analytiske målinger fra prøver / batcher hvor man ønsker å få et raskt overblikk over korrelasjonen mellom dem (loadingsplot), samtidig som man prøvenes relative verdi på variablene (scoreplot)
- Metoden kan brukes på alle datamatriser
- Kreves i utgangspunktet ingen kunnskap om dataene for å gjøre analysen. (Men for at tolkningen skal bli pålitelig og nyttig bør man kjenne dataene for å finne forklaringer på de resultatene som PCA gir)

- Skiller på systematisk og tilfeldig variasjon (~noise, error) :

$$\boxed{X} = \boxed{\text{Model}} + \boxed{\text{Error/Noise}}$$

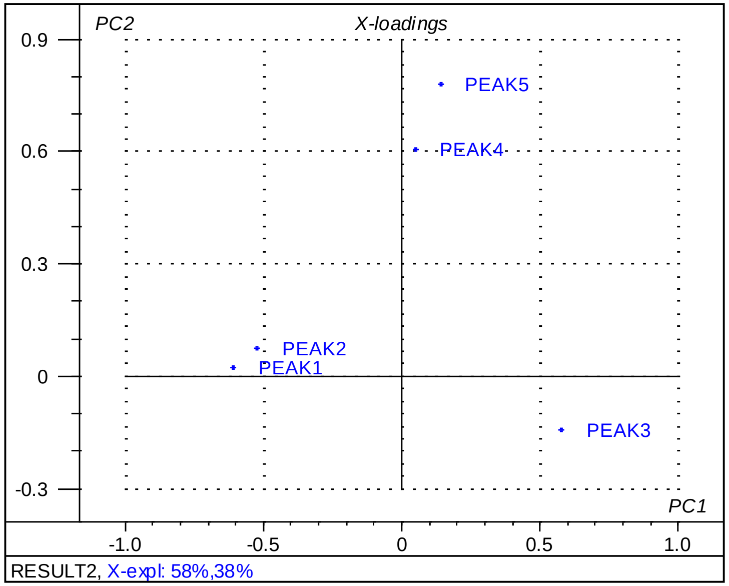
Principal Component Analysis (PCA)
Et eksempel med bruk av PCA først med excel, deretter i Unscrambler:

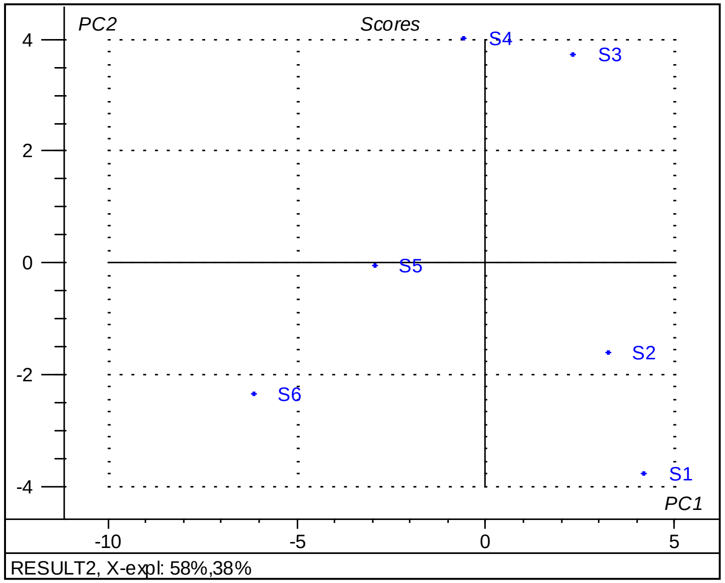
	PEAK1	PEAK2	PEAK3	PEAK4	PEAK5
S1	1	2	7	2	1
S2	2	2	6	4	2
S3	3	3	5	6	7
S4	4	5	3	7	6
S5	5	6	2	5	2
S6	8	7	1	2	1

	SCORES	
	PC_01	**PC_02**
S1	4,18	-3,76
S2	3,23	-1,61
S3	2,31	3,74
S4	-0,59	4,03
S5	-2,96	-0,05
S6	-6,17	-2,35

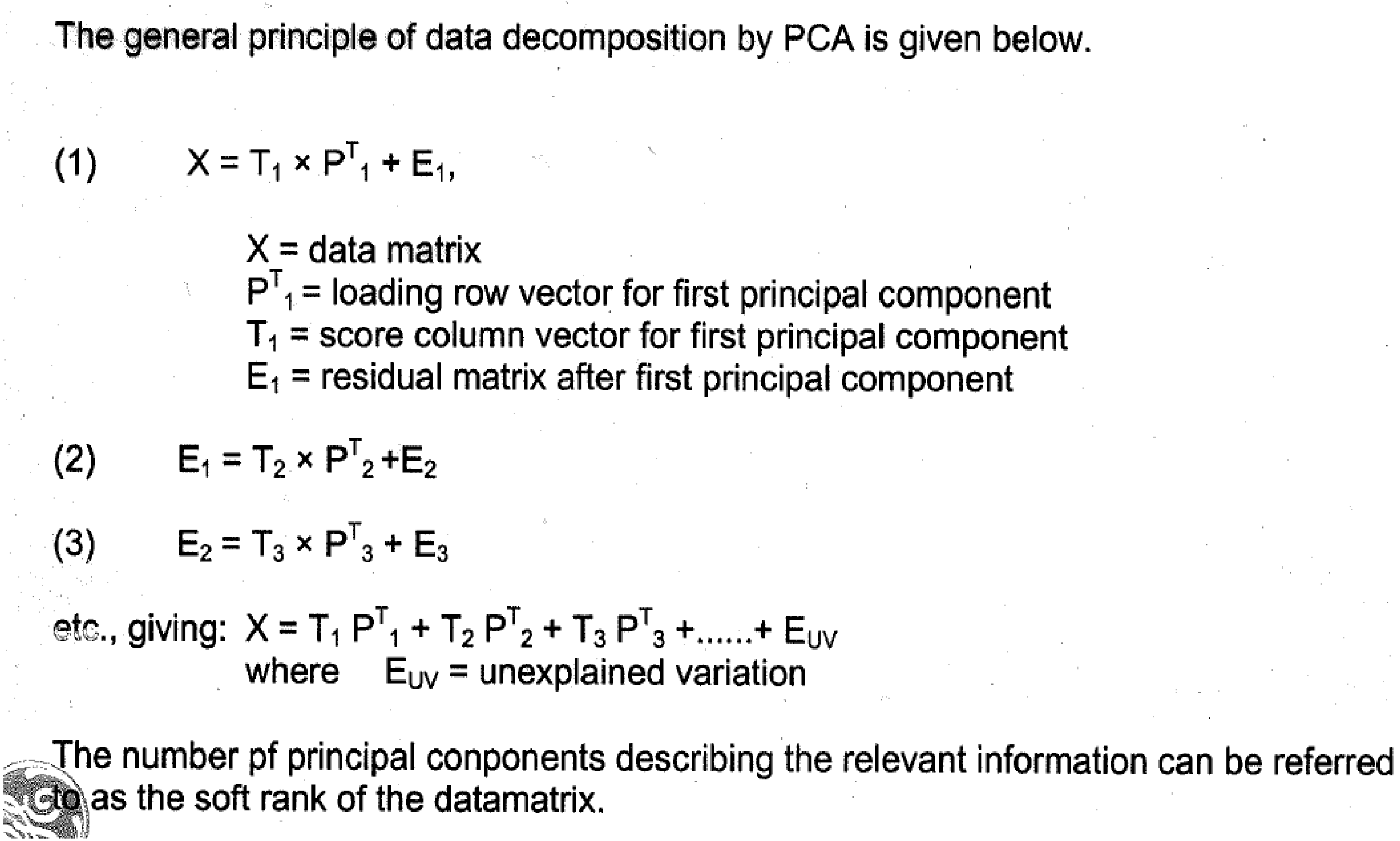
LOADINGS

	PEAK1	PEAK2	PEAK3	PEAK4	PEAK5
PC_01	-0,61	-0,52	0,58	0,05	0,14
PC_02	0,02	0,07	-0,15	0,60	0,78





Principal Component Analysis (PCA)



The general principle of data decomposition by PCA is given below.

(1) $X = T_1 \times P^T{}_1 + E_1$,

X = data matrix
$P^T{}_1$ = loading row vector for first principal component
T_1 = score column vector for first principal component
E_1 = residual matrix after first principal component

(2) $E_1 = T_2 \times P^T{}_2 + E_2$

(3) $E_2 = T_3 \times P^T{}_3 + E_3$

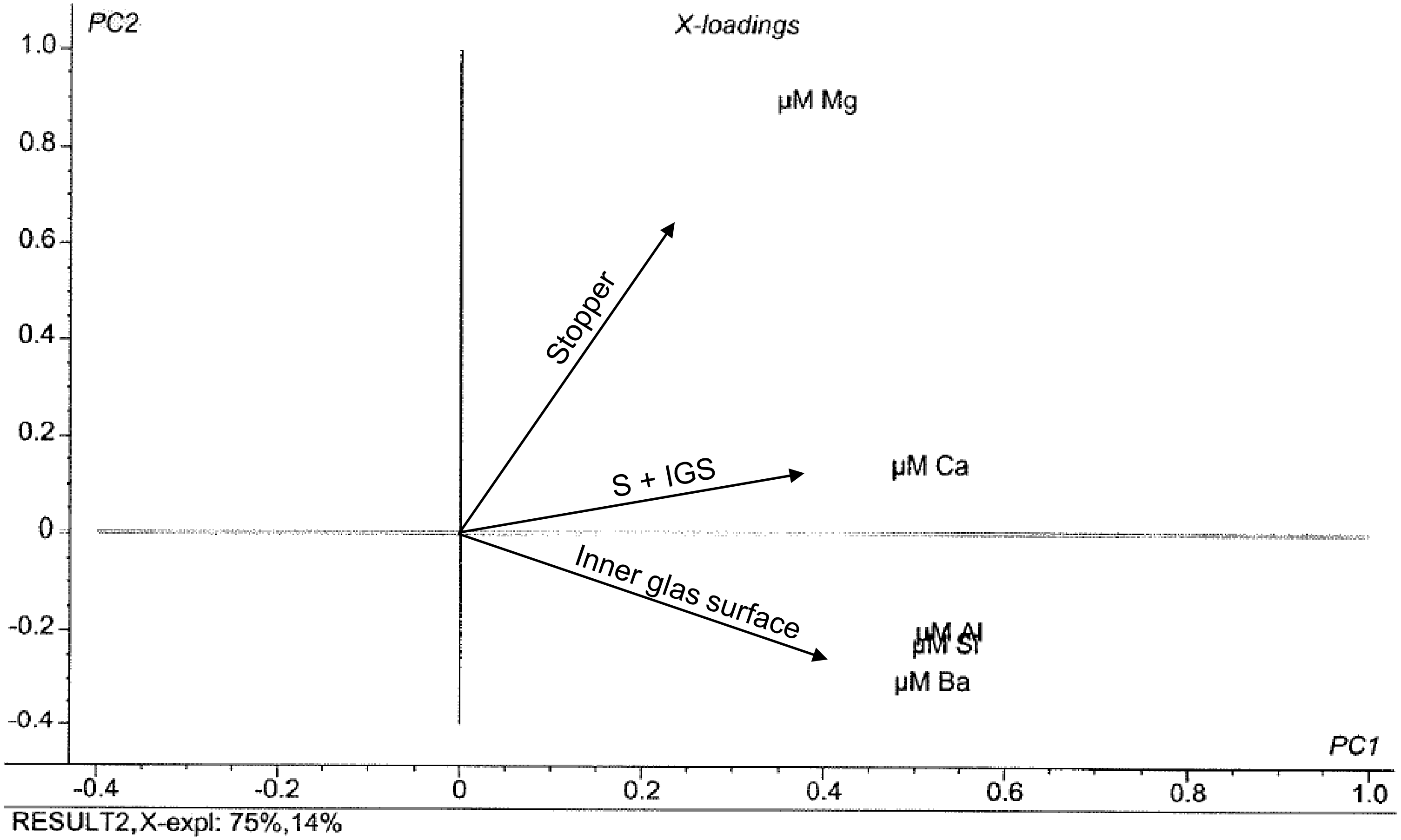
etc., giving: $X = T_1 P^T{}_1 + T_2 P^T{}_2 + T_3 P^T{}_3 + \ldots\ldots + E_{UV}$
where E_{UV} = unexplained variation

The number pf principal conponents describing the relevant information can be referred to as the soft rank of the datamatrix.

Principal Component Analysis (PCA)

- Viktige parametre / plott i PCA:
- Loading: Bidrag fra en variabel på prinsipalkomponenten
- Score: Vekten til en prøve på prinsipalkomponenten
- Forklart varians: Hvor mye av variasjonen i data matrisen som forklare ved det valgte antall prinsipalkomponenter
- Uteligger: Enprøve som er unormal sammenlignet med resten av dataene
- Leverage (vektstang): Et mål på distansert et punkt er fra resten av punktene (~modellen)
- Overtilpassing:En kosntruert model som ikke brae beskriver systematisk informasjon, men også støy
- Kryssvalidering: En metode for bestemme det antallet prinsipalkomponenter som beskriver viktig / sytematisk variasjon

Principal Component Analysis (PCA)



Principal Component Analysis (PCA)

Spektroskopi

- Bruk av PCA har blitt essentiell innenfor tolking av NIR, IR, RAMAN, UV of Fluorescens spektra. PCA brukes også til evaluering av impurities i kromatografiske systemer med DAD, MS og NMR deteksjon.
- Mange variabler (~ 1200 bølgelengder) i forhold til prøver gjøre denne teknikken viktig.
- Ofte brukes PCA i kombinasjon med spekterforbehandling som derivering, normalisering og “scattering” korreksjoner.

PLSR - Partial least squares regression (projection to latent squares regression)

I PLSR gjøres det en korrelasjon mellom X og Y som bestemmer regresjonskoeffisientene. Med andre ord, algoritmen i PLS forsøker vha variasjonen i Y å ekstrahere variasjonen i X som er korrelert til Y for å bestemme regresjonskoeffisientene. Modellen modifiseres for hver ny mengde variasjon som hentes fra X for å forklare variasjon i Y (=hver ny PC).

"In **PLSR**, Y is used to achieve a guided decomposition of X prior to the estimation of regression coefficients. The philosophy behind PLSR is to extract only the information in X necessary to give precise predictions of Y."

" In general, when colinear X variables exist, when the number of experiments are few, when the noise level is varying or/and when outliers may influence the model, PLSR is superior to MLR. Since these aspects are common, PLSR should have wider use than the classical MLR."

Comparing MLR and PLS

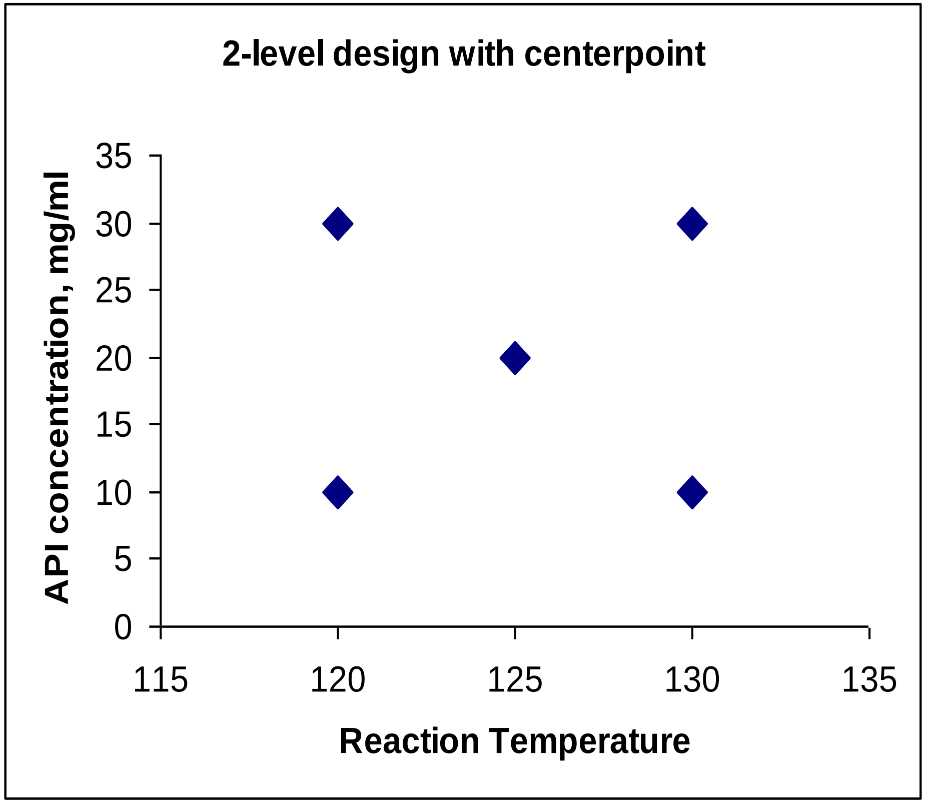
MLR

- Orthogonal X-variables
- #samples > #reg.coef.
- No missing data
- Only model one response at a time
- Simpel
- Established statistical procedures

PLS

- Handles correlations within the X-matrix
- No limitations with respect to dimensionality
- Estimates missing data
- PLS2 allows several Y-variables
- Needs proper validation
- Several methods for statistical evaluations and meaningful plots

Example screening design



A full factorial design includes:

- All variable combinations to determine main and interaction effects
- Three or more center points to determine variability and linearity
- Balanced design: no confounding of variables

- Calculation of Number of experiments:

-L^V, L=Levels, V=No. Variables,

-Plus centre points

Regression model/transfer function calculated by statistical software

Y = β1xT + β2xC + β3xTxC + β0,

βs = effects of the variables, β0 = only constant value

Screening design

The screening design contains the 4 corners and 3 centrepoints.

We begin with the screening design and develop a PLS- regression model with Unscrambler.

After performing a PLS regression a infinite number of analytical plots are available.

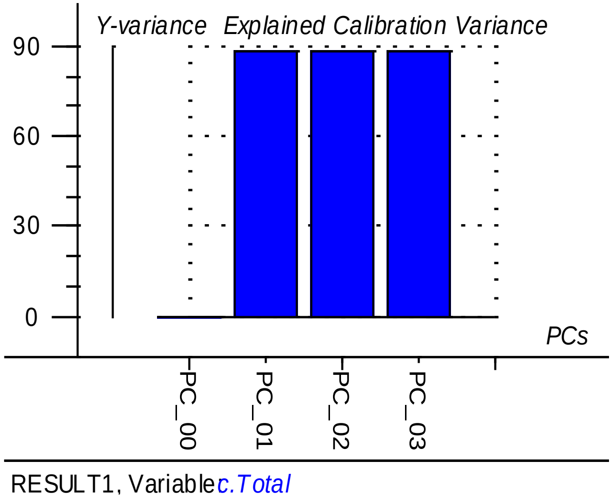
The 5 most important Unscrambler-plots are:

- •Explained variance to determine the number of principal components
- •Weighed regression coefficients to rank the effect of variables
- •Prediction versus measured to understand the quality of model and the predictive power
- •Response surface to get the quantitative influence of the Xs and to evaluate Design Space
- •Influence plot to detect outliers

Purpose	**Temperature, °C**	**API Concentration, mg/ml**	**Yield, %**
Screening	**120**	**10**	**60**
Screening	**130**	**10**	**82**
Screening	**120**	**30**	**65**
Screening	**130**	**30**	**91**
Screening	**125**	**20**	**80**
Screening	**125**	**20**	**81**
Screening	**125**	**20**	**83**
Optimization	125	10	73
Optimization	125	30	87
Optimization	120	20	65
Optimization	130	20	89

Example from PET-Alzheimer

PLSR plot – Explained calibration variance



The explained variance plot shows the variation in Y that is explained by the regression model.

A low explained variance may indicate:

1. The variation in Xs cannot explain the variation in Y for various reasons
2. The model is inappropriate
3. The variation in Y is small (~noise)

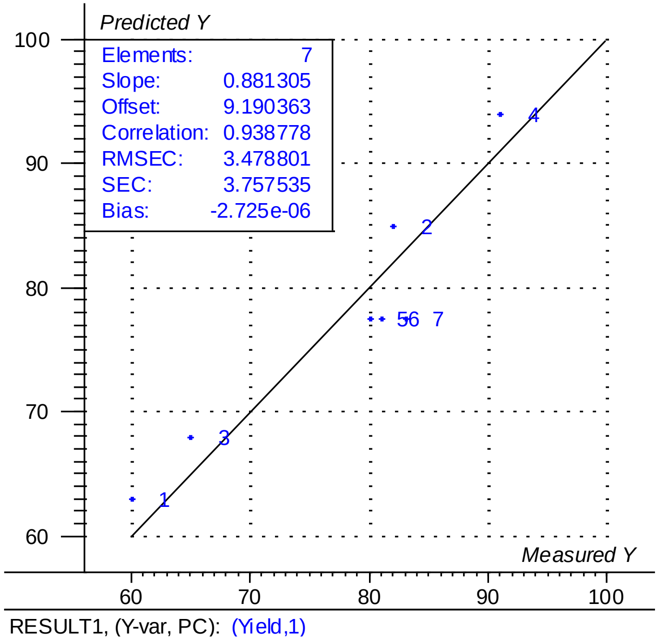
Always remember to choose the same number of Principal Components when looking at the different plots.

PLSR plot – Predicted vs Measured

The predicted versus measured plot indicates the accuracy and precision of the developed model:

Important elements of the plot:

1. RMSEC ~ the standard deviation of the model
2. The spread and positioning of center points
3. Any non-linear curvature aligned by the spots
4. Estimating analytical method precision



Without a reasonable RMSEC the model is not very useful

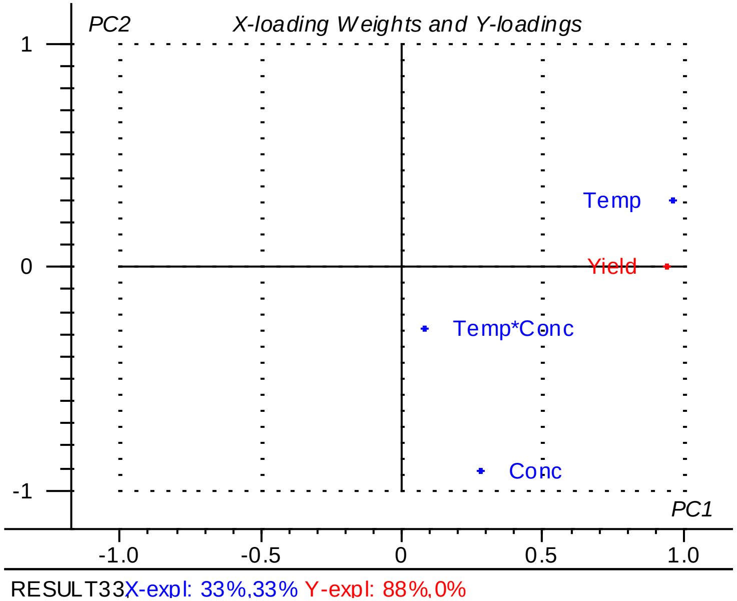
PLSR plot – X-Y loading

The loadings plot gives an impression bout the most important Xs regarding Ys.

Important elements of the plot:

1. Xs closely placed may indicate confounding (~covariance)

2. The real effect of Interaction effects cannot be understood in this plot



Loadings is often used to evaluate confounded Xs

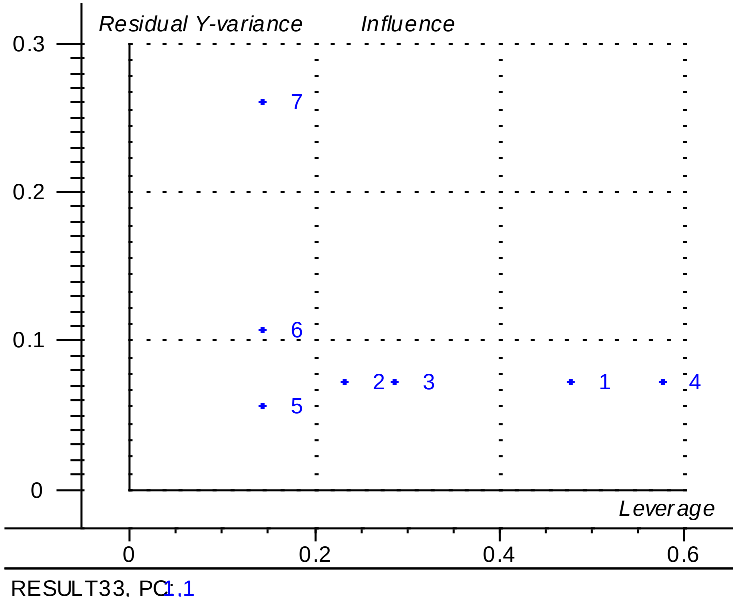
PLSR plot – Influence

The Influence plot is used to identify possible outliers

Important elements of the plot:

1. Samples placed in upper right corner an be outliers an destroy the model

2. Inhomogeneous spread of samples in the plot indicate different impact on the model by the experiments



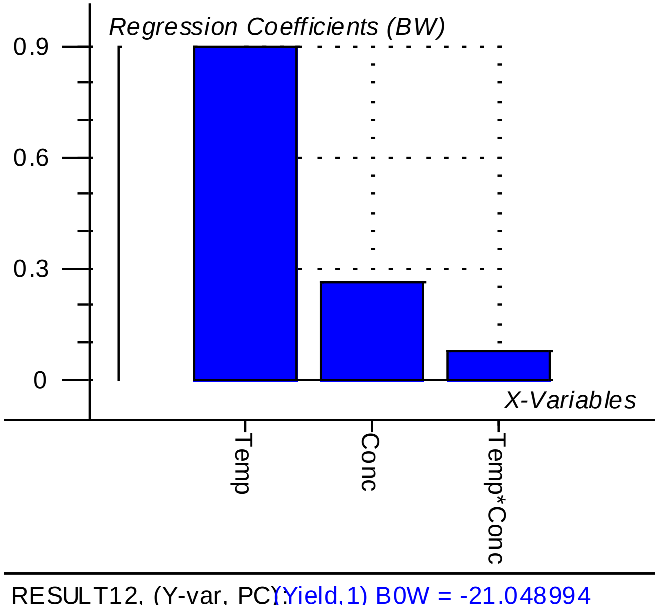
Experiments having high leverage compared to the other experiments may give an unreliable model

PLSR plot - Weighed regression coefficients

The weighed regression coefficient plot from PLSR is used to rank the effect of variables on yield:

1. Temp (β1)
2. Conc (β2)
3. T*C (β3)

The real effect of Interaction effects cannot be understood in this plot

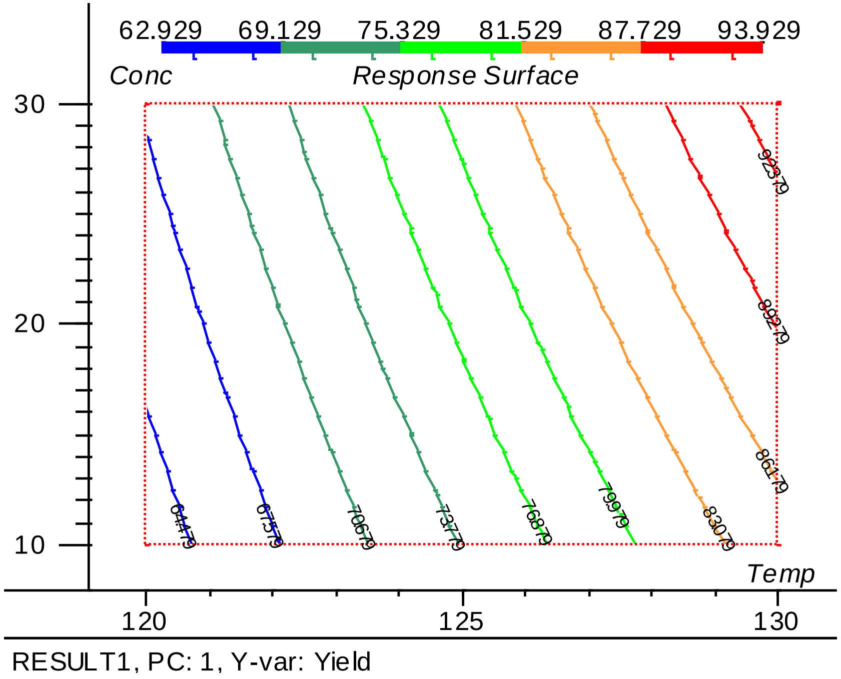


With few experiments we cannot use the PLSR-significance test. We can instead recalculate without T*C and evaluate the prediction plot.

Or one can use MLR which is the most straight forward regression analysis.

The test for determination of significant regression coefficients (~variables) with PLSR will be discussed in the Principal Component training

PLSR plot – response surface



The response surface is a target plot used to finally understand the exact effects of the variables. This is essential in the design space determination.