Variable Selection

EIGENVECTOR RESEARCH INCORPORATED

Outline

- **Why select variables?**
- **What methods are there?**
- **So which to choose?**
- **In practice**
- **Some examples**

EIGENVECTOR RESEARCH INCORPORATED

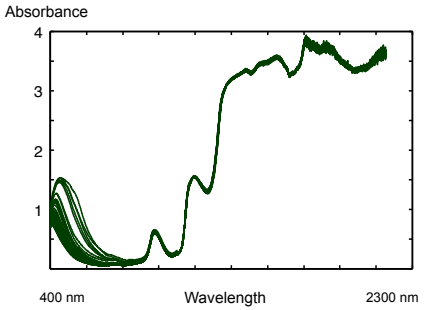
Course Materials

- These slides
- PLS_Toolbox or Solo 6.7 or later
- Data sets
  - From DEMS folder (distributed with software)
    - nir_data.mat (optional)
  - From EVRIHW folder (additional data sets)
    - beer.mat, nir_shootout_2002.mat (optional)

EIGENVECTOR RESEARCH INCORPORATED

VIS/NIR spectra of 61 beers

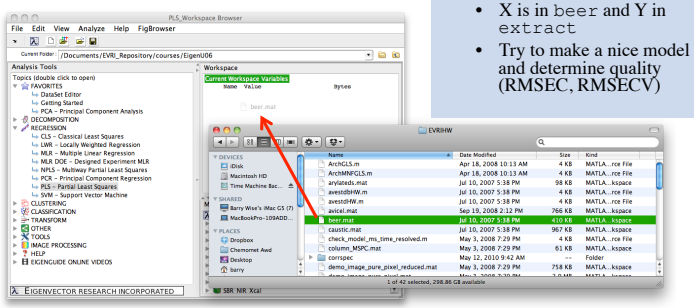
Purpose: prediction of real extract



EIGENVECTOR RESEARCH INCORPORATED

VIS/NIR spectra of 61 beers

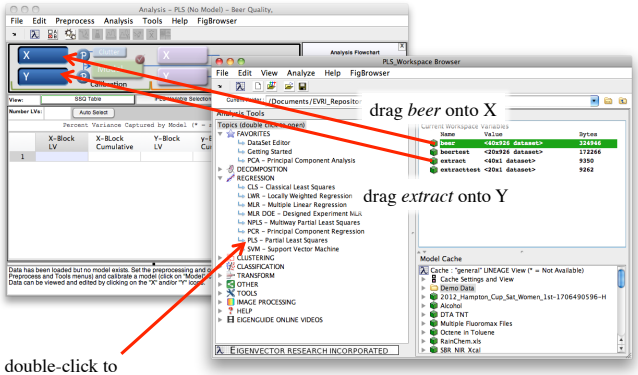
Try to make a PLS model for extract



- **Load beer.mat**
  - X is in `beer` and Y in `extract`
  - Try to make a nice model and determine quality (RMSEC, RMSECV)

EIGENVECTOR RESEARCH INCORPORATED

Start PLS, Load Beer



EIGENVECTOR RESEARCH INCORPORATED

Why select
What methods
Choose method
In practice
Some examples

Exercise Data

Determination of the amount of extract from NIR spectra of beers.

Dispersive visual & near-infrared data collected (at 25 C) NIRSystems Inc. (Model 6500) spectrophotometer. Split detector system – silicon detector 400-1100 nm & (PbS) detector 1100-2500 nm.

VIS-NIR transmission recorded directly on undiluted degassed beer in 30 mm quartz cell. Spectral data collected at 2 nm intervals 400-2250 nm & converted to absorbance units.

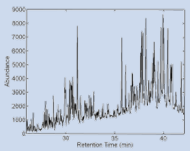
Original *extract* concentration is a quality parameter in the brewing industry, indicating the substrate potential for the yeast to ferment alcohol and serving as a taxation parameter. Original extract concentration determined by Carlsberg A/S in the range of 4.23-18.76% plato.

Data sorted by extract value, and a model independent test set was constructed by selecting every third sample of this full data set. There are thus two data sets: one for calibration (40 samples) and one for independent estimation of prediction error (20 samples).

EIGENVECTOR RESEARCH INCORPORATED

Why select
What methods
Choose method
In practice
Some examples

Why variable selection?

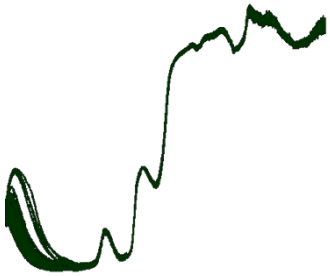


- Improvement of the model
  - Remove irrelevant, unreliable or noisy variables
  - Improve predictions
  - Improve statistical properties
- Interpretation
  - Obtain a model that is easier to understand
- Costs
  - Use less measurements to replace expensive or time-consuming one
- Development of fast instruments/routines for on-line control
  - Find wavelength ranges for a filter-based instrument

EIGENVECTOR RESEARCH INCORPORATED

What methods available?

- **A priori**
  - Choose measurements
- **A posteriori**
  - Use chemical/physical insight
- **Model based**
  - Look at loadings
- **"Random based"**
  - Genetic algorithms
  - Simulated annealing
- **Classical**
  - Forward, backward selection
  - Best subset selection
  - Significance tests
  - Significance based on Jack-knife
  - GOLPE
- **"Spectral"**
  - i-PLS
- **Other**
  - Pure variables
  - Principal variables
  - Iterative weighting with regression vector
  - …



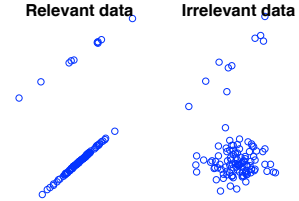
EIGENVECTOR RESEARCH INCORPORATED

What methods?

- Why not just choose the best variables?
  - Highly nonlinear problem
  - Exchaustive search not possible
  - How to validate what's good?

Best fitting variable of 1000 to ten calibration samples

New data (100 samples)



Method : A priori

Choose the right measurements

- The most important of all
- Beyond the scope of this course as we assume the data are already available/fixed

Important assumptions

- There may be indirect correlations that you did not anticipate
- Don't choose too few variables a priori

EIGENVECTOR RESEARCH INCORPORATED

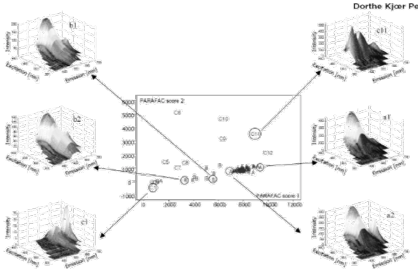
Method: A Priori

Example indirect relation

JOURNAL OF CHEMOMETRICS
J. Chemometrics 2002; 16: 421–460
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.735

Screening for dioxin contamination in fish oil by PARAFAC and N-PLSR analysis of fluorescence landscapes†

Dorthe Kjær Pedersen, Lars Munck and Søren Balling Engelsen*



EIGENVECTOR RESEARCH INCORPORATED

Method: A Priori

Example indirect relations

Even though not direct link – fluorescence works well!!!

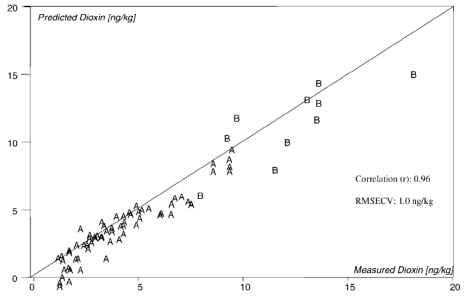
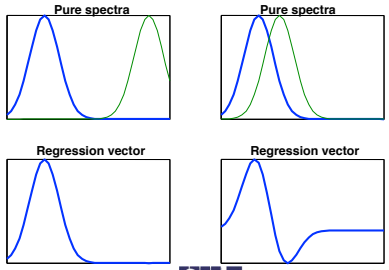


Figure 5. Predicted dioxin concentration versus measured dioxin concentration for N-PLSR model (four PCs) for sample set A + B fish oil samples ($n = 75$). The correlation coefficient (r) and prediction error (RMSECV) are reported.

EIGENVECTOR RESEARCH INCORPORATED

Method : A posteriori

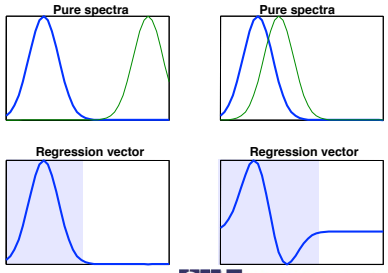
- Remember that a regression model is not only depending on the analyte directly
- Also has to adjust for overlapping signals



EIGENVECTOR RESEARCH INCORPORATED

Method : A posteriori

- Remember that a regression model is not only depending on the analyte directly
- Also has to adjust for overlapping signals



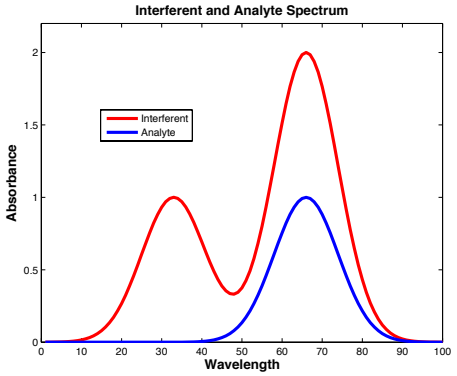
EIGENVECTOR RESEARCH INCORPORATED

Why is Variable Selection Hard?

- Is much easier when analyte of interest has "pure variables," *i.e.* variables where the analyte has a signal but no interferents do
- When interferents overlap with analyte, important variables can be positively correlated with concentration, negatively correlated, or uncorrelated!

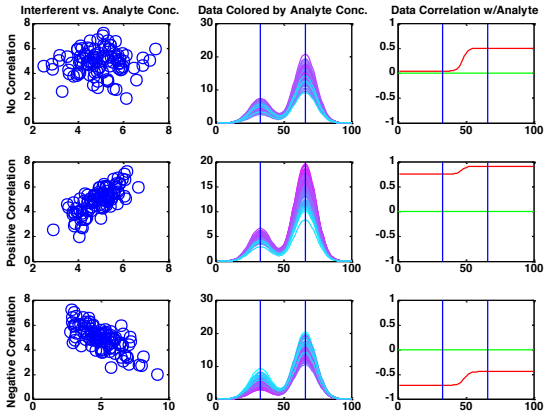
EIGENVECTOR RESEARCH INCORPORATED

Simple System Example



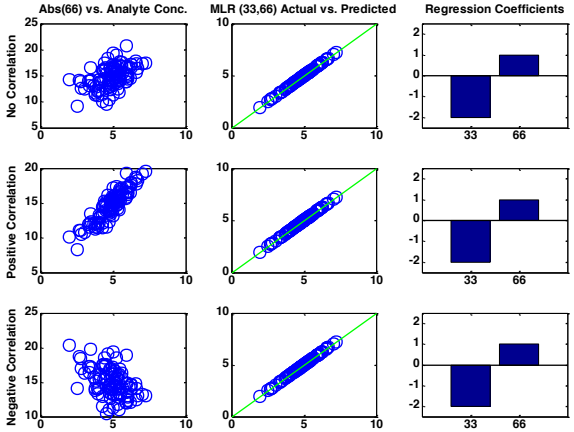
EIGENVECTOR RESEARCH INCORPORATED

Consider 3 Cases



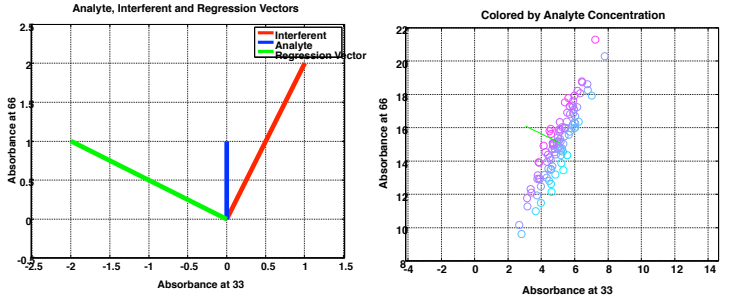
EIGENVECTOR RESEARCH INCORPORATED

Need Both Variables 33 & 66!



EIGENVECTOR RESEARCH INCORPORATED

Regression Coeffs and Data



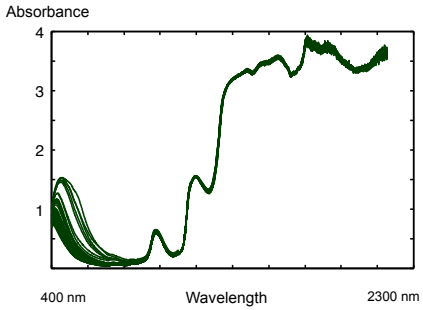
Model needs signal at 33 to be orthogonal to interferent!

EIGENVECTOR RESEARCH INCORPORATED

Example: A posteriori

VIS/NIR spectra of 61 beers

Purpose: prediction of real extract

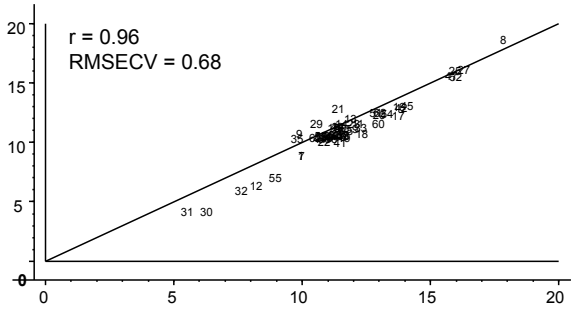


EIGENVECTOR RESEARCH INCORPORATED

Example: A posteriori

Full spectrum PLS-model

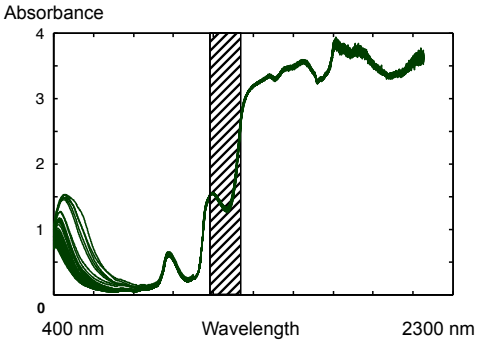
Cross validated prediction error (61 samples, 6 segments)
Five PLS factors



EIGENVECTOR RESEARCH INCORPORATED

Example: A posteriori

Selected interval: 1218-1300 nm

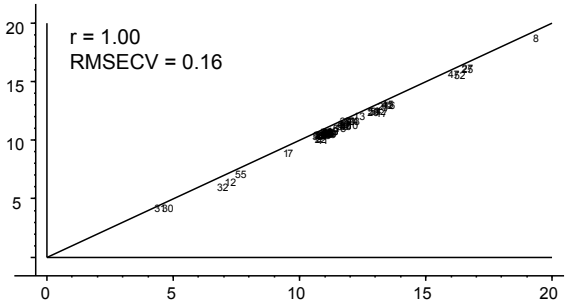


EIGENVECTOR RESEARCH INCORPORATED

Example: A posteriori

PLS-model based on 1218-1300 nm

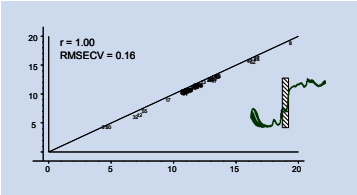
Cross validated prediction error (61 samples, 6 segments)
Three PLS factors



EIGENVECTOR RESEARCH INCORPORATED

Example: A posteriori
PLS-model based on 1218-1300 nm

This course is about how to achieve results as these even in situations where such detailed background knowledge is not accessible.



EIGENVECTOR RESEARCH INCORPORATED

Model based selection

- Simply use the visual appearance of the model
- E.g. small loadings, low regression coefficents etc.

EIGENVECTOR RESEARCH INCORPORATED

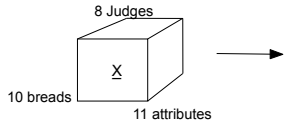
Model based selection
Important assumptions

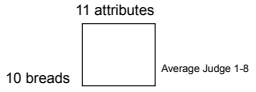
- **Model is reasonable**
  - if 900 out of 1000 variables irrelevant, the model may be reflecting those and hence the relevant ones look insignificant
- **Model is certain**
  - Few samples or noisy measurements =>
  - Statistical uncertainty high =>
  - Do not trust the model parameters too much

EIGENVECTOR RESEARCH INCORPORATED

Model based selection
Example – Sensory analysis

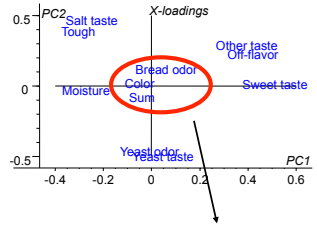
- Sensory profiling of bread
  - 10 breads (replicates) × 11 attributes × 8 judges
  - Average over judges: 10 × 11 attributes
  - Data from Magni Martens



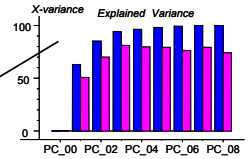


EIGENVECTOR RESEARCH INCORPORATED

Model based selection
Example



Loadings indicate some variables not important



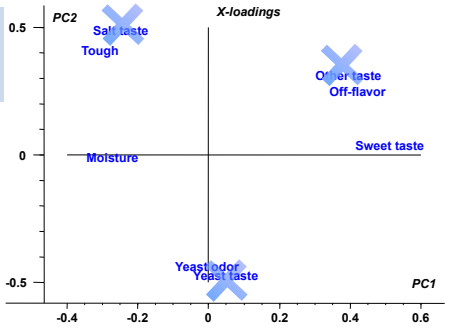
Seems trustworthy as the model is otherwise well-behaved

EIGENVECTOR RESEARCH INCORPORATED

Model based selection
Example - Many ways to skin a cat

Want to get rid of more?

Remove redundant variables



EIGENVECTOR RESEARCH INCORPORATED

Model based selection
Automating it a bit

Variable Importance for Projection (VIP)

Relative weighted sum of squares of PLS-weights, weighted by components importance for predicting

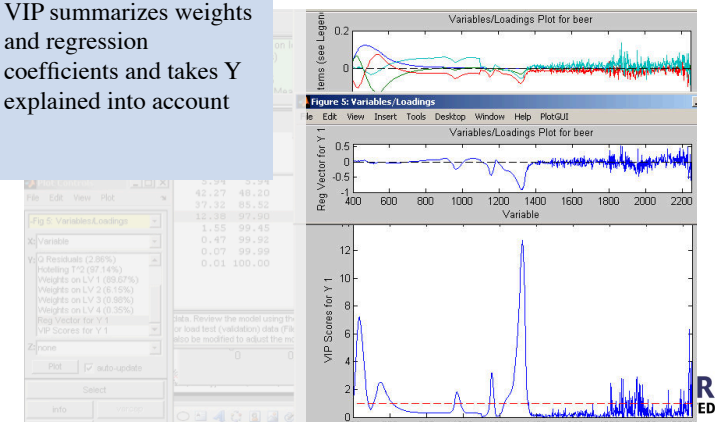
- Assumes valid model
- Hence only remove few variables at a time
- And check for outliers etc. along the way
- VIP smaller than one indicates low importance

S. Wold, E. Johansson, M. Cocchi, 3D QSAR in Drug Design; Theory, Methods, and Applications, ESCOM, Leiden, Holland, 1993, pp. 523– 550.

EIGENVECTOR RESEARCH INCORPORATED

Model based selection
Automating it a bit

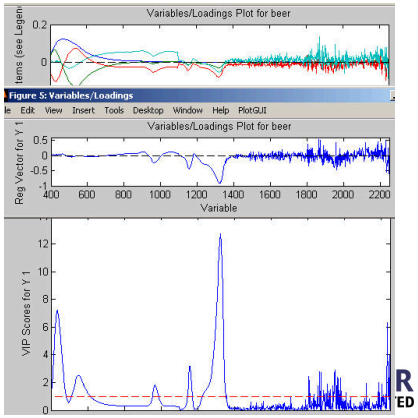
VIP summarizes weights and regression coefficients and takes Y explained into account



Model based selection
Automating it a bit

Try yourself on the beer data.

Is a one-shot selection optimal?



"Random" variable selection

Random selection means choosing without considering improvement in fit directly

Increases chance of obtaining unforeseen and complex interactions

Genetic algorithms is a good example

www.rdecom.army.mil

Genetic algorithm

- **Method**
  - Survival of the fittest (best fit)
- **Principle**
  - Every combination of variables (a model) is defined by an index of which variables to use *and* the goodness of this model
    - *Example: One species is given by the calibration model with variable 1, 3, 14 & 27 yields and RMSEP of 1.23.*
  - Find better models by "mating" species so that the good ones mate more than bad ones (survival of the fittest).
    - *Example: Calibration models with variable 8 are generally better and therefore increasingly part of new calibration models.*

Genetic algorithm

- **Terminlogy**
  - Population = Set of individuals
  - One individual = Model with a given set of variables
  - One gene = Codes for one variable (in/out)
- **Algorithm**
  - Make start population (e.g. 50 different models with different variables included)
  - Evaluate each model (RMSEP or similar)
  - Have a party and let the best ones have most fun
  - Fun: two models mate and make a child which has similar variables
  - Arrange a new party for the 50 children and continue

1 0 0 1 0 0 0 0 0 0 0 0 0

Gene for one model defines which variables are in. RMSEP defines the quality of the individual

$$\begin{bmatrix} y \end{bmatrix} \quad \begin{bmatrix} X \end{bmatrix}$$

Riccardo Leardi has written many papers on how to make genetic algorithms work

Random methods
Assumptions

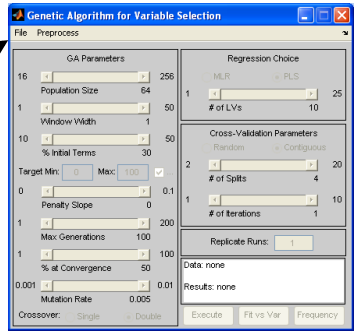
- "Random" methods are based on combining a random search (individuals) with a guided search (mating)
  - Mostly used because exhaustive search is too expensive
  - Good and sound principle
  - Excellent for getting ideas
  - Not good, though, for refining

EIGENVECTOR RESEARCH INCORPORATED

Genetic algorithms
Exercise

```
In MATLAB
>> load beer
>> genalg
```

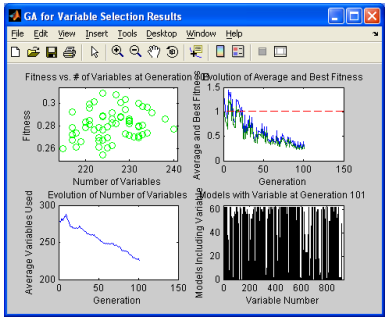
- Load calibration data (beer and extract)
- Execute



EIGENVECTOR RESEARCH INCORPORATED

Genetic algorithms
Exercise

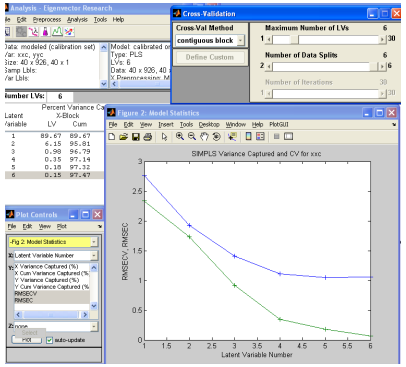
- Does the result look nice?
- Maybe try increasing iterations?



EIGENVECTOR RESEARCH INCORPORATED

Genetic algorithms
Exercise

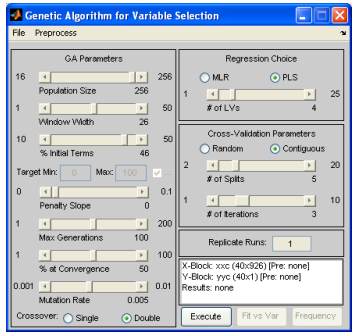
- **Generally**
- Lower the number of components based on initial PLS analyses but use a little more
- Keep ending criteria sensitive. The more iterations which occur, the more feedback from the cross-validation information, thus more likely over-fitting,
- Use random cross-validation and multiple iterations if practical,
- Repeat the GA run multiple times and watch for general trends
- For data with many variables and fewer samples, increase the window width



EIGENVECTOR RESEARCH INCORPORATED

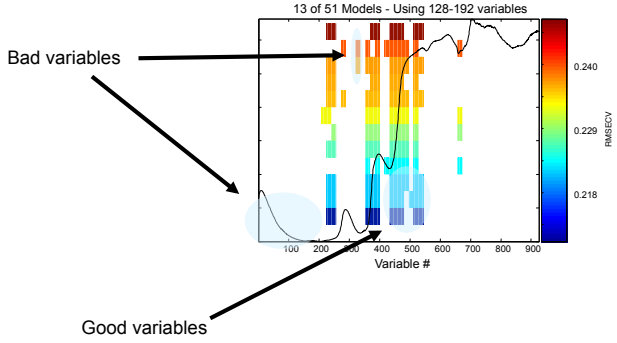
Genetic algorithms
Exercise

- **In PLSToolbox**
- **Size of Population** Larger populations provide a better representation of different variable combinations.
- **Window Width** When adjacent variables contain correlated information the original variables can be included or excluded in "blocks".
- **% Initial Terms** Appr. # variables included in the initial subsets. Few initial terms will make identification of useful variables more difficult, but will bias the end solution towards models with fewer variables.
- **• Target Min/Max & Penalty slope** For guiding towards a specific number of variables



EIGENVECTOR RESEARCH INCORPORATED

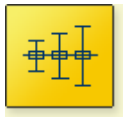
Genetic algorithms
Exercise



EIGENVECTOR RESEARCH INCORPORATED

Classical methods

- Subset selection
  - Statistical significance tests
  - Forward selection
  - Backward selection
  - Best subset selection



EIGENVECTOR RESEARCH INCORPORATED

Classical methods
Statistical significance tests

Principle
- Do regression
- Eliminate variables with non-significant regression coefficients

Properties
- Very fast
- Assumes a statistically valid model
- Assumes statistically valid significance tests
- Hence only works for models with insignificant amount of irrelevant variables

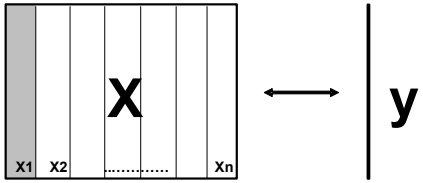
The function `Calibsel` selects variables based on significance

EIGENVECTOR RESEARCH INCORPORATED

Intermezzo

iPLS: Interval PLS

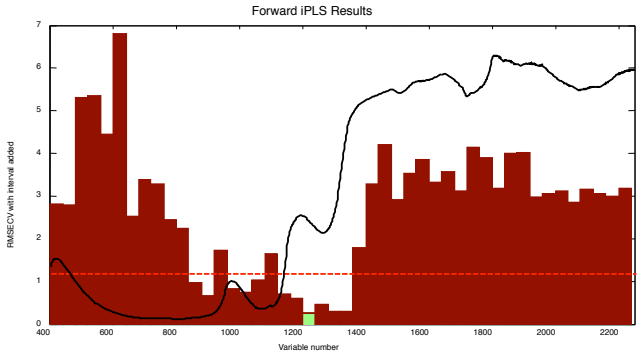
Local models in n intervals. Very intuitive and useful approach that can be easily combined with variable selection



L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. B. Engelsen. Interval partial least-squares regression (iPLS). *Appl.Spectrosc.* 54 (3):413-419, 2000.

iPod introduced 2001
iPLS introduced 2000
Lawsuit?!

iPLS result on beer data



EIGENVECTOR RESEARCH INCORPORATED

Classical methods

Forward selection

Principle

- Select best fitting variable (or better cross-validated)
- Regress y on this and select variable fitting best on residual
- Regress y on these two and select best fitting on the residual
- Etc.

Good

- Fast
- Handles many irrelevant variables
  - If many samples or test set evaluation
- Often works well

Bad

- Disregards interactions to some extent

EIGENVECTOR RESEARCH INCORPORATED

Forward selection

Exercise

Try forward selection on NIR data

Many variables – computationally expensive

- If correlation between neighbors use windows instead of individual variables.
- E.g. use every 10 neighbors as one set and ex/ include them all together

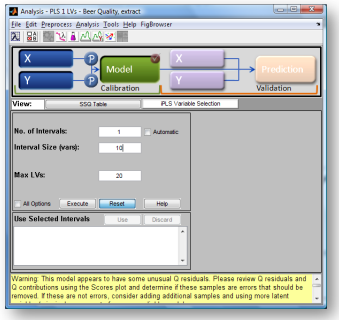
EIGENVECTOR RESEARCH INCORPORATED

Forward selection
Exercise

Maybe more than one interval could be useful?

Maybe bigger or smaller intervals are better?

Investigate ..



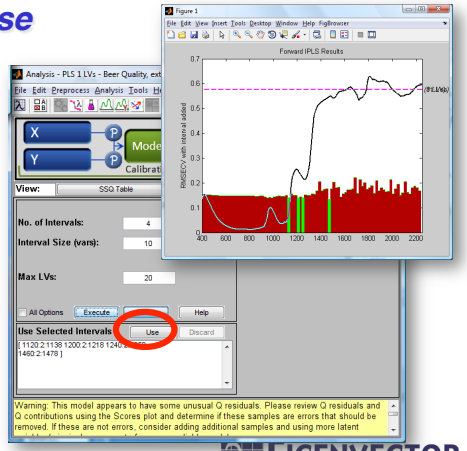
EIGENVECTOR RESEARCH INCORPORATED

Forward selection
Exercise

Once you have a good model, either

1) Quick and dirty: use the good intervals

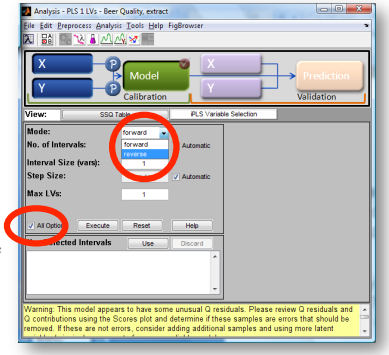
2) Better: Remove the clearly bad intervals



EIGENVECTOR RESEARCH INCORPORATED

Backward selection
Exercise

Rather than forward selection, you can choose advanced options and rather than *selecting* variables, you can *de-select* variables by choosing the reverse mode



Principle
- Make full model and remove the variable contributing the least to the fit
- Repeat that

Good
- Takes interactions into account

Bad
- Often less efficient than forward selection

EIGENVECTOR RESEARCH INCORPORATED

Choosing method
If there are many irrelevant variables

Why select
What methods
Choose method
In practice
Some examples

- Do not use statistically based tests
- The fewer samples, the more important the independence of the optimization criterion is
  - Do not use fit values but rather cross-validation or test set
- Genetic algorithms and forward selection are good choices
- Refine solution using background knowledge and other variable selection tools

EIGENVECTOR RESEARCH INCORPORATED

Choosing method
If refining an already ok model

- The more samples (to variables), the less important optimization criterion is
- Statistics makes sense
  - Significance is very helpful, model parameters can be interpreted meaningfully
- Use *i*PLS, backward selection, model statistics such as VIP etc.

How to do it in practice

By now it is clear that most methods do not like outliers:

Remove outliers!!

Even if they are not finally removed, meaningful variable selection is not possible with even slightly significant outliers (unless very many samples).

So get rid of them and add them again afterwards

Alternative – robust methods

How to do it in practice

- **Never trust results**
  - Use several methods as inspiration
  - Evaluate selected variables
    - Add some, remove some, check with a priori knowledge
- **The better the initial model, the better results from variable selection**
  - Remove obviously irrelevant variables
  - Remove outliers
  - Bin similar variables