

Relating potato flesh colour to a metabolomics data set.

Animesh Acharjee, Chris Maliepaard

Group: Quantitative aspects of plant breeding

The Netherlands

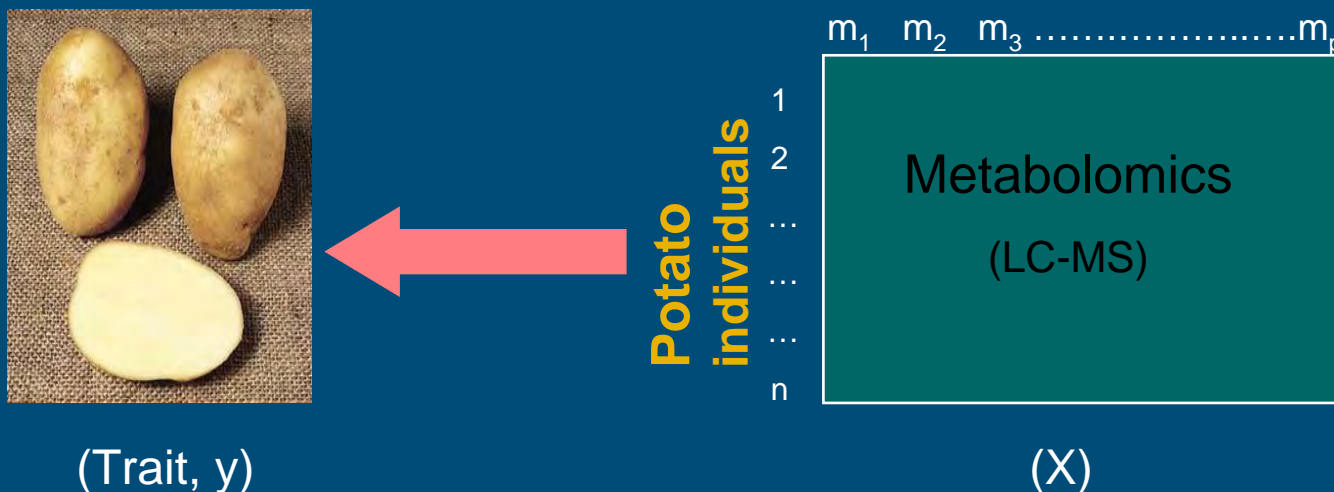


Outline

- Objective
- Materials
- Methods
- Comparison
- Conclusions

Objective

- Relate a phenotypic trait to a large ~omics data set
- ~omics data: metabolomics/transcriptomics data of potato.



- Comparison of regression approaches suitable for ~omics
- 'pipeline' for these methods

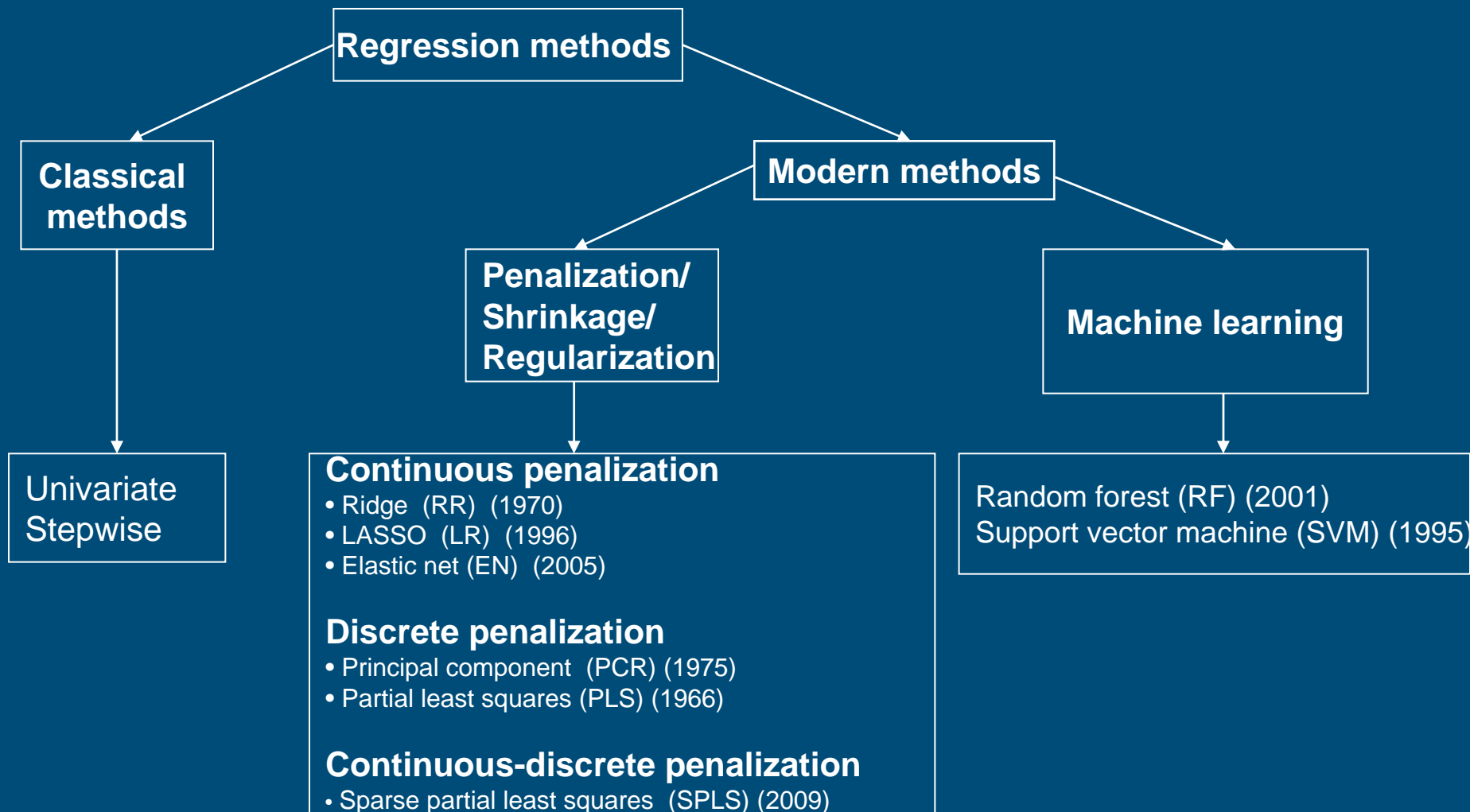
Materials

- Parent C: hybrid between *S. phureja* x *S. tuberosum*
- Parent E: cross between clone C and *S. verneii*
- Potato diploid backcross population (C x E)
- Progeny: 91 individuals (n)
- Phenotypic traits: flesh colour and enzymatic discoloration of potato
- ~omics data: 163 metabolite (p) peaks from untargeted LC-MS analysis
- Predictors are \log_{10} transformed and autoscaled and response is mean centered.

Problems

- Nr. of variables (**metabolites**: p) \gg nr. of objects (individuals, samples, n)
- Variable selection : reduce dimensionality
- Significance of the variables and models
- Computational time
- Statistical power

Methods used



Penalization regression methods

- Ordinary least squares :

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

- Ridge regression :

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda_2 \sum_{j=1}^p \beta_j^2$$

- LASSO regression :

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j|$$

- Elastic net regression :

$$\min_{\beta} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda_2 \sum_{j=1}^p \beta_j^2 + \lambda_1 \sum_{j=1}^p |\beta_j|$$

Penalization regression methods

| | Variable selection | Grouping | Variables selected | Number of penalty (s) |
|-----------------------------|--------------------|----------|--------------------|-----------------------|
| Ridge regression (RR) | No | Yes | "p" | 1 |
| LASSO regression (LR) | Yes | No | 0 to "n" | 1 |
| Elastic net regression (EN) | Yes | Yes | 0 to "p" | 2 |

Optimization of shrinkage parameter (s)

- 10-fold cross-validation (CV)
 - Use, a number of times, different subsets of the data
 - Each time use only 90% of the samples to make a regression model
 - Try different values for the shrinkage parameter
 - Calculate the error of prediction in the 10% left out

Discrete penalization methods

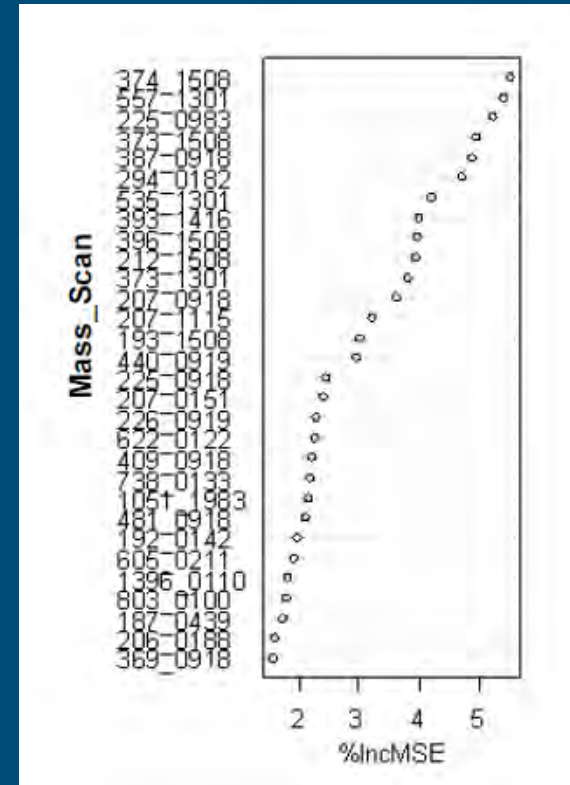
- Principal component regression (PCR)
 - Criterion: maximization of variance in X matrix.
 - Select first few principal components
 - Use these in regression
 - Optimum nr. of components has to be chosen (by CV)
- Partial least squares regression (PLS)
 - Criterion: maximize covariance of y with latent variables
 - Fewer components compared to PCR

Sparse partial least squares regression (SPLS)

- Dimension reduction via PLS
- PLS: no variable selection
- SPLS does dimension reduction and variable selection simultaneously
- Variable selection by LASSO
- Grouping of highly correlated variables

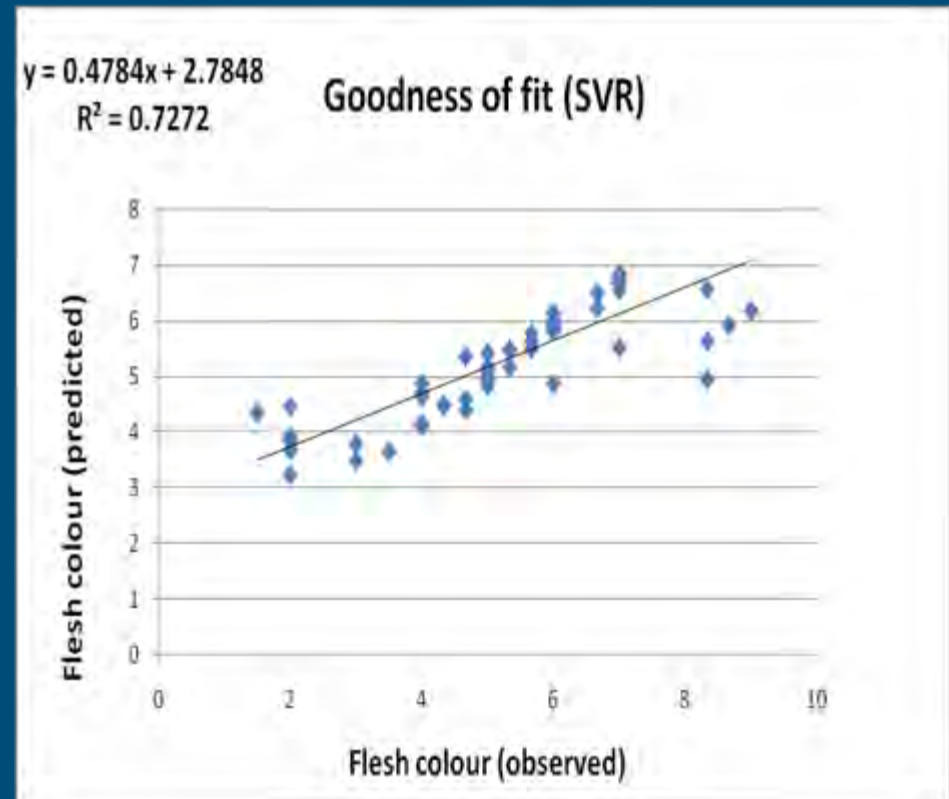
Random forest regression (RF)

- Based on classification and regression trees
- Handles high numbers of variables ($p \gg n$)
- Internal cross validation
- Variable importance is estimated
- Fast algorithm



Support vector regression (SVR)

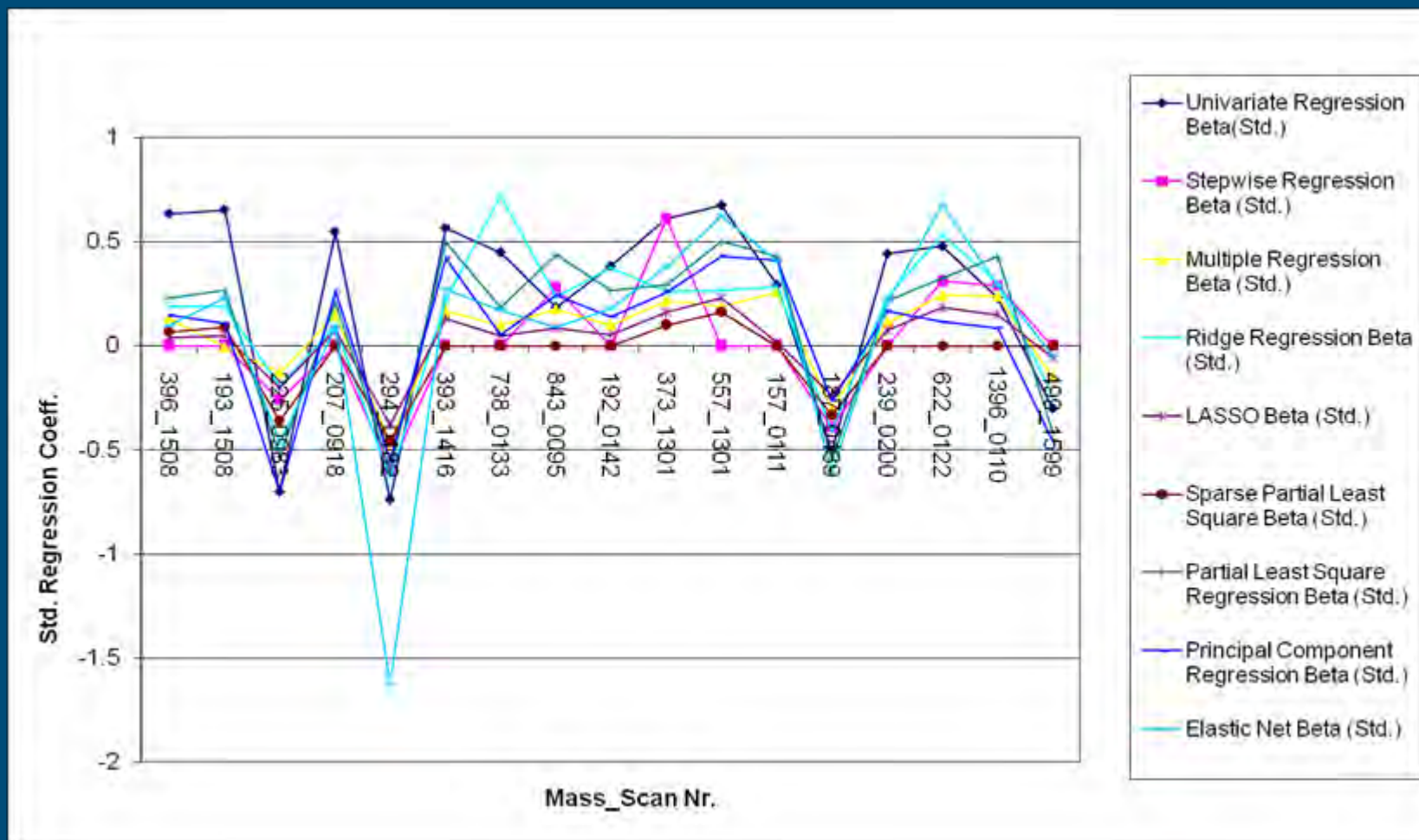
- Based on SVM classification
- Objective of SVR: find a robust function $f(x)$ with some constraint
- Data is transformed to a different space



Comparative study: p=163

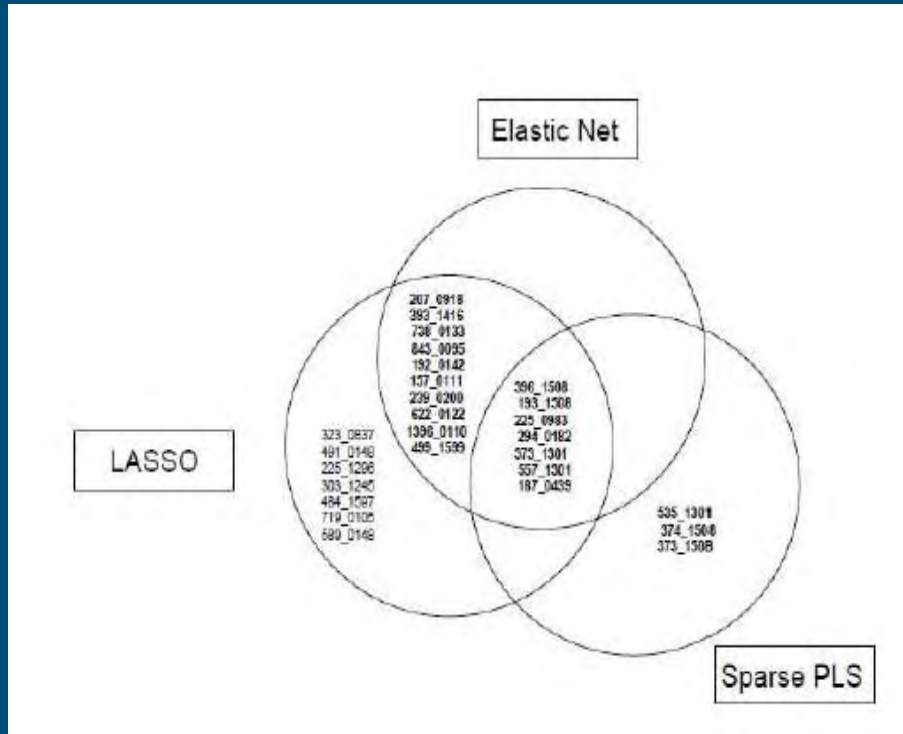
| | Regularization | Obs. Vs. fitted (R2) | Variables Selected | Grouping | MSEP |
|-------|-------------------------|----------------------|--------------------|----------|--------|
| RR | Continuous | 69% | "p" | Yes | 1.2731 |
| LASSO | Continuous | 65% | 24 | No | 1.2016 |
| EN | Continuous | 58% | 17 | Yes | 1.1852 |
| PCR | Discrete | 60% | "p" | No | 1.2596 |
| PLS | Discrete | 70% | "p" | Yes | 1.3134 |
| SPLS | Continuous and Discrete | 50% | 10 | Yes | 1.2941 |
| RF | Discrete | 24% (Test) | "p" | No | 1.2690 |
| SVM | Continuous | 72 % | "p" | No | 1.3898 |

Comparison of the standardized regression coefficients

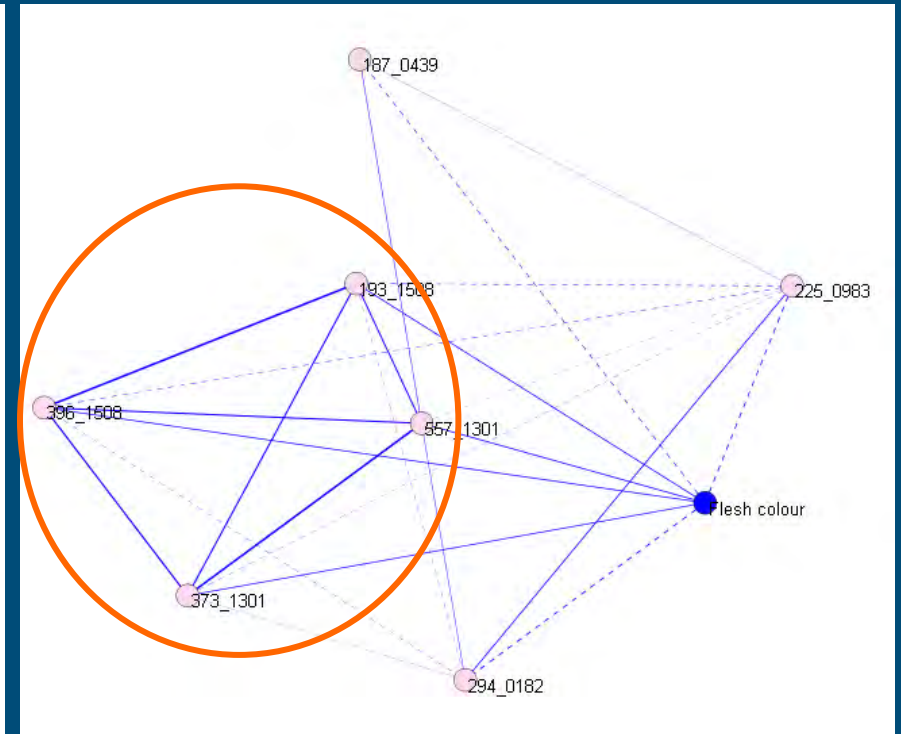


Variable selection (p=163)

Venn diagram



Network analysis



Dotted line: negative correlation

Solid line: positive correlation

Conclusions

- $p \gg n$ problem can be tackled using these methods.
- Variable selection
 - E.g. EN selects 17 variables, LASSO selects 24 variables, SPLS selects 10 variables
 - 7 variables in common between EN, LASSO and SPLS
- Lowest MSE: elastic net
- Variables putatively annotated as glycosides of carotenoid related compounds
- Enzymatic discoloration: Caffeoylquinic acid, methyl ester linked to the chlorogenic acid
- Developed a pipeline in “R”
 - Prediction of phenotype from ~omics data
 - Variable selection

Acknowledgements

Dr. Chris Maliepaard, WUR Plant Breeding
(Supervisor)

Prof. Richard Visser, WUR Plant Breeding (Promoter)

Prof. Cajo Ter Braak, Biometris

Dr. Bjorn Kloosterman, WUR Plant Breeding

Dr. Ric de Vos, Plant Research International

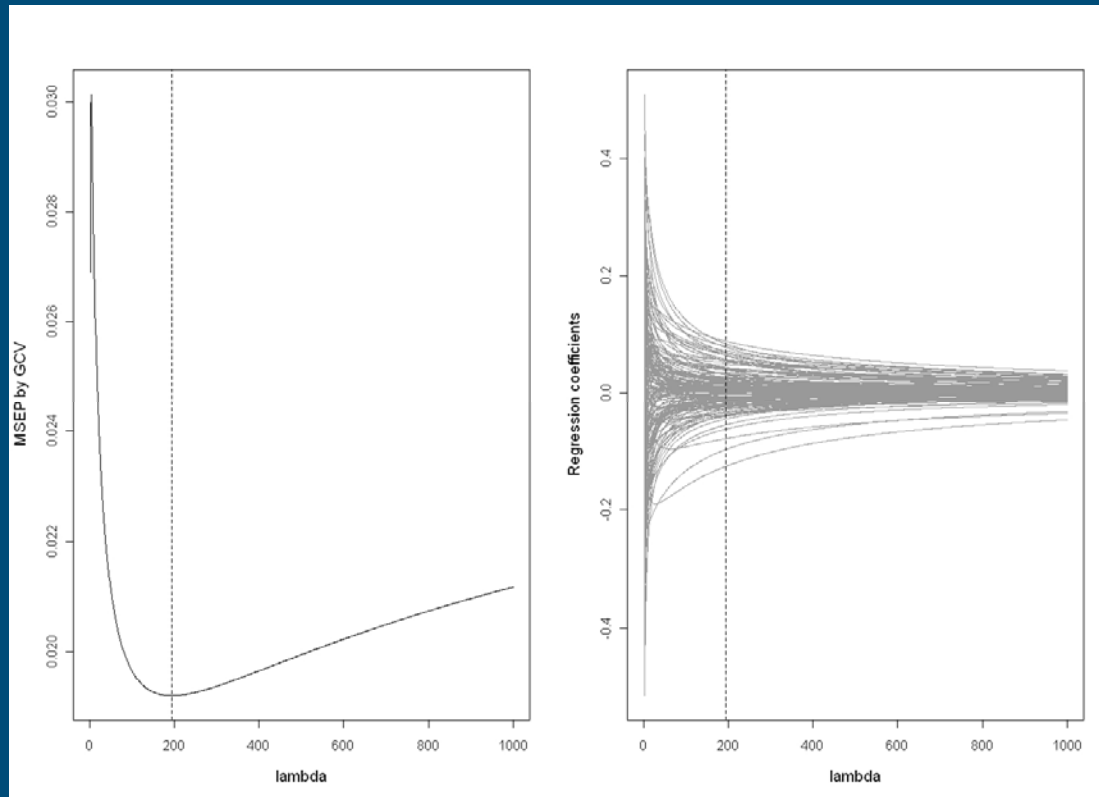
© Wageningen UR



Thanks

Optimization of the hyper parameter

- Parameter optimized (lambda) by 10 fold cross validation
- The value of the hyperparameter (λ) = 193



LC-MS data set

16,000 individual mass peaks

A subtraction of the back ground signal

10,000 mass peaks

skewness of data

1,100 mass peaks

correlations were calculated with metabolites linked to the quality traits using the student t-test. ($p < 0.0005$)

163 mass peaks (p)

Basic Idea of SV Regression

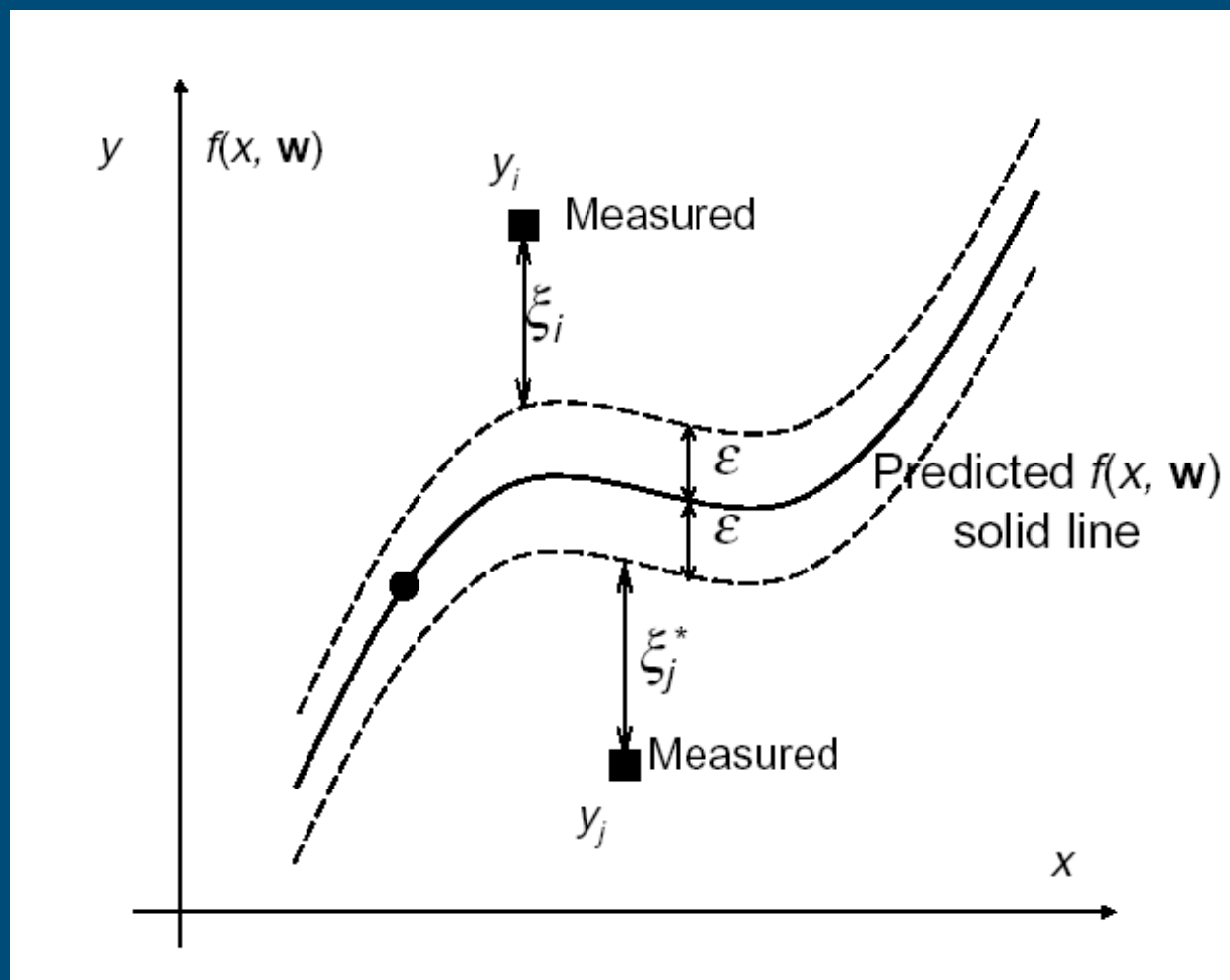
- Objective : To find a robust function $f(\underline{x})$ that has at most ε deviation from the targets y , while at the same time being as *flat* as possible.
- Idea : Simple regression problem + optimization + kernel trick

Linear ε - insensitive loss regression

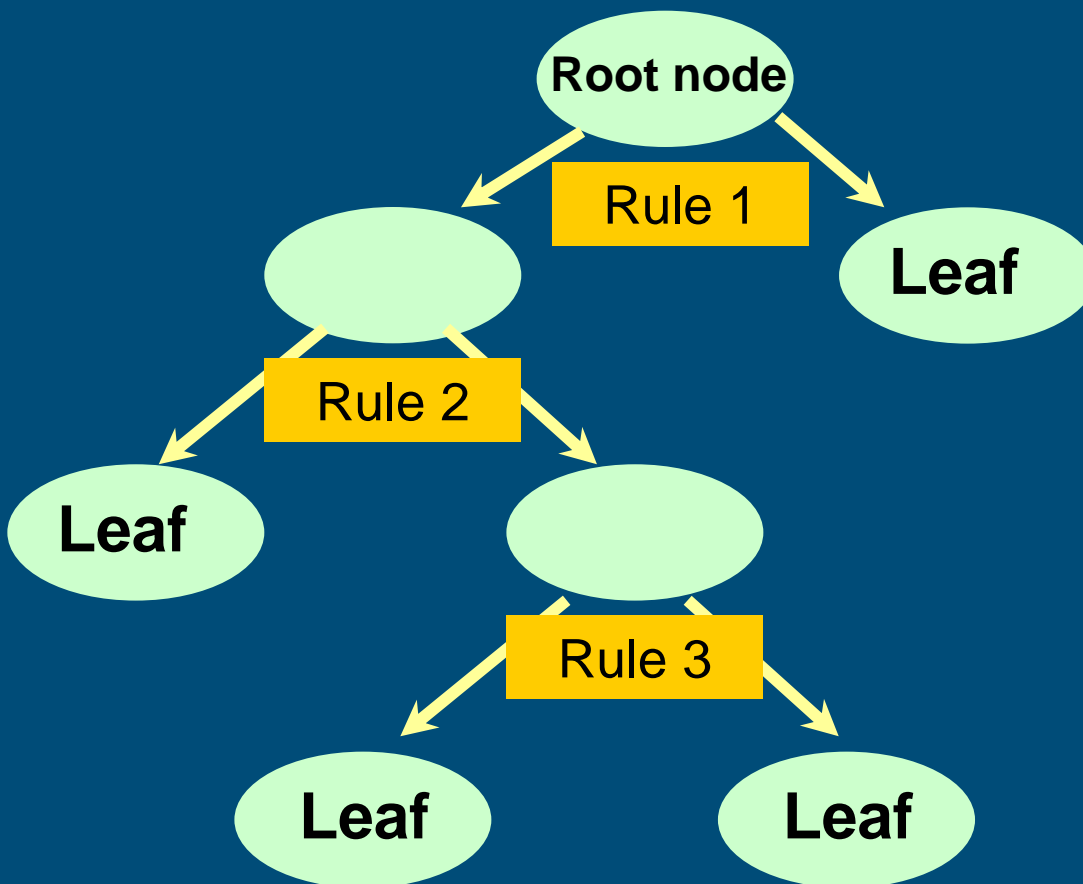
$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ \text{subject to} \quad & \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b & \leq \varepsilon + \xi_i \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \end{aligned}$$

- ε : decide insensitive zone
- C : a trade-off between error and $\|\mathbf{w}\|$
- ε and C must be tuned simultaneously

Parameters used in SV Regression



Example binary regression tree

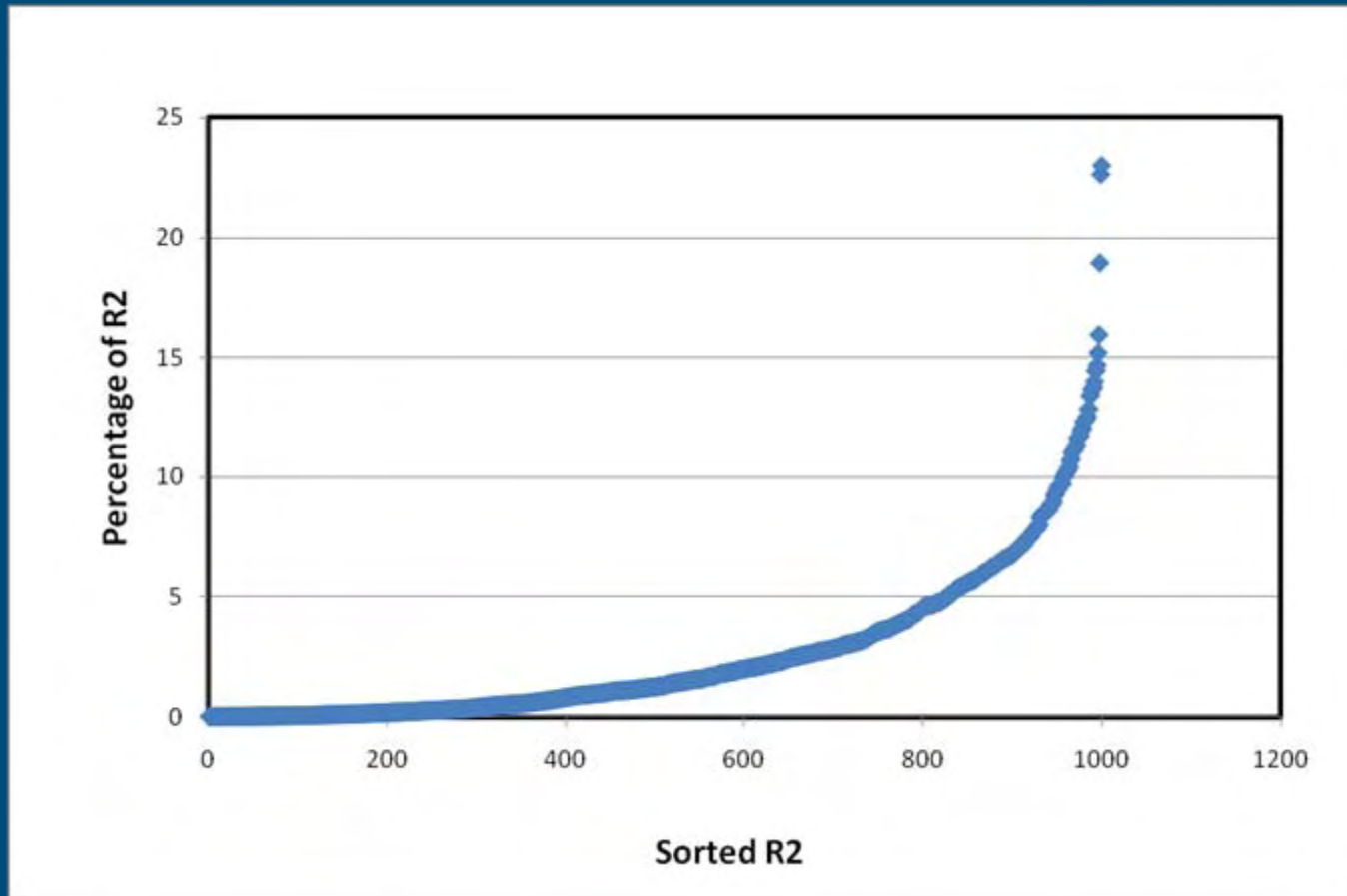


- The root contains all samples
- Each subsequent node contains a fraction of samples
- Each rule splits up the samples into two groups
- Every rule is of the form
 - $x > t$ for continuous x
 - $x \in A$ for categorical xOnly one variable per rule
- Each leaf more or less 'pure', similar values for y
- A new sample is run through the tree and one looks for the leaf it ends up. Prediction is the average.

Summary : Enzymatic discoloration

| | Auto Scaled | Optimization Of hyper parameters | Goodness of fit (R^2) | Variable Selected | Variable selected with highest coeff. |
|-------------------|-------------|----------------------------------|---------------------------|-------------------|---------------------------------------|
| Ridge | Yes | $\lambda = 210$ | 73 % | ALL | 1076_396_1508 (-0.08570240) |
| EN | Yes | $\lambda = 0.1$ | 52 % | 26 | 818_795_918 (-0.25444) |
| LASSO | Yes | $\lambda = 0.1$ | 54 % | 24 | 818_795_918 (-0.25909) |
| PCR | Yes | ncomp=27 | 54 % | All | 733_631_758 (0.1168027) |
| PLS | Yes | ncomp= 3 | 56 % | All | 1076_396_1508 (-0.1309039) |
| SVM | Yes | gamma= 0.00390625 cost=1 | 87 % | All | NA |
| RF | Yes | mtry=27 | 7% (test set) | NA | 818_795_918 (top rank) |
| SPLS | Yes | eta = 0.2, K =3 | 66 % | 215 | 1076_396_1508 (-0.12799564) |
| Stepwise | Yes | NA | 63 % | 12 | 2810_1140_2182 (-0.8425) |
| Univariate | Yes | NA | 17 % | 1 | 818_795_918 (-0.4170099) |

Permutation test (RF)



Permutation test (PLS)

