Direct Kernel Partial Least Squares (DK-PLS): Feature Selection with Sensitivity Analysis

Mark J. Embrechts (embrem@rpi.edu)
***Kristin Bennett**
www.drugmining.com

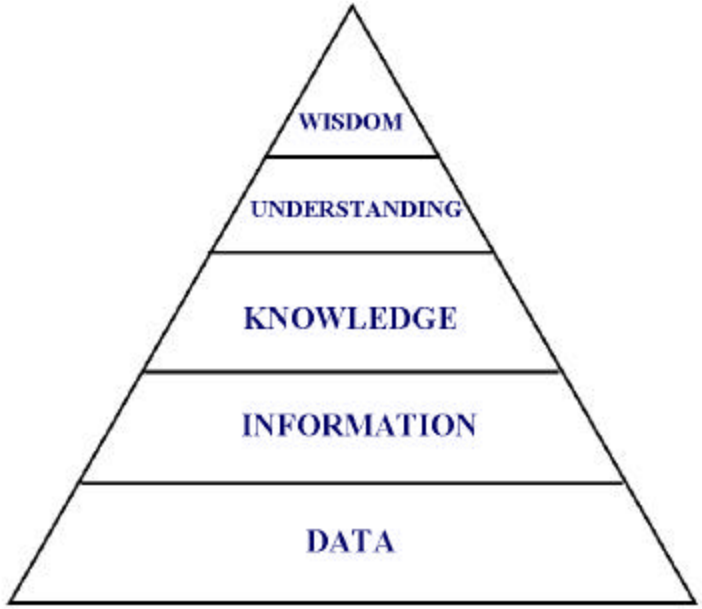
Department of Decision Sciences and Engineering Systems
****Department of Mathematics***
Rensselaer Polytechnic Institute, Troy, New York, 12180

Supported by NSF KDI Grant # 9979860

Presented at NIPS Feature Selection Workshop
November 12, 2003
Whistler, BC, Canada

Outline

- **PLS**
  - Please Listen to Svante Wold
  - Partial-Least Squares
  - Projection to Latent Structures
- **Kernel PLS (K-PLS)**
  - cfr Kernel PCA
  - Kernel makes PLS model nonlinear
  - Regularization by selecting small nu
- **Direct Kernel PLS**
  - Direct Kernel Methods
  - Centering the Kernel
- **Feature Selection with Analyze/StripMiner**
  - Filters: Naïve feature selection: drop “cousin features”
  - Wrappers: Based on sensitivity analysis
    - ➔ Iterative procedure
    - ➔ Training set for feature selection used in bootstrap mode



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*

Abstract

PLS-regression (PLSR) is the PLS approach in its simplest, and in chemistry and technology, most used form (two-block predictive PLS). PLSR is a method for relating two data matrices, **X** and **Y**, by a linear multivariate model, but goes beyond traditional regression in that it models also the structure of **X** and **Y**. PLSR derives its usefulness from its ability to analyze data with many, noisy, collinear, and even incomplete variables in both **X** and **Y**. PLSR has the desirable property that the precision of the model parameters improves with the increasing number of relevant variables and observations.

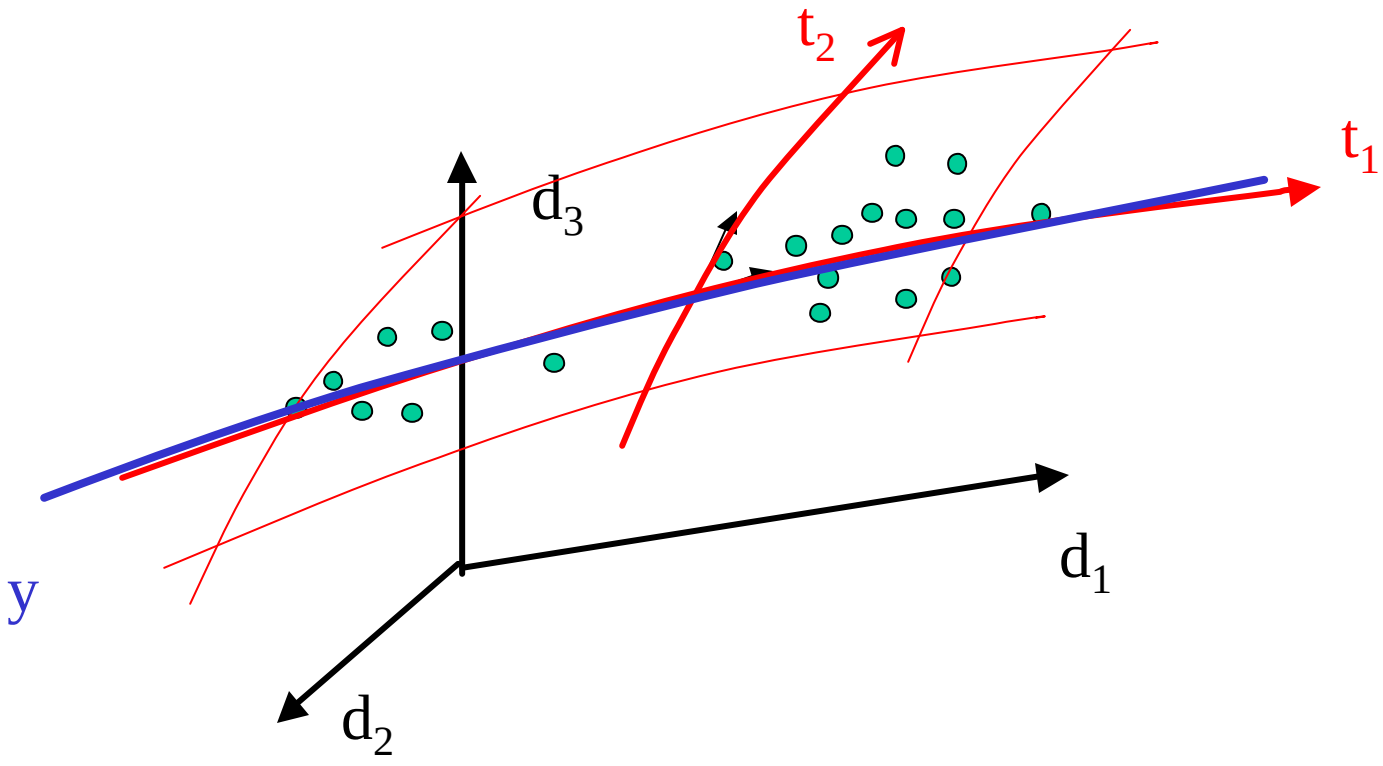
This article reviews PLSR as it has developed to become a standard tool in chemometrics and used in chemistry and engineering. The underlying model and its assumptions are discussed, and commonly used diagnostics are reviewed together with the interpretation of resulting parameters.

Two examples are used as illustrations: First, a Quantitative Structure–Activity Relationship (QSAR)/Quantitative Structure–Property Relationship (QSPR) data set of peptides is used to outline how to develop, interpret and refine a PLSR model. Second, a data set from the manufacturing of recycled paper is analyzed to illustrate time series modelling of process data by means of PLSR and time-lagged X-variables.

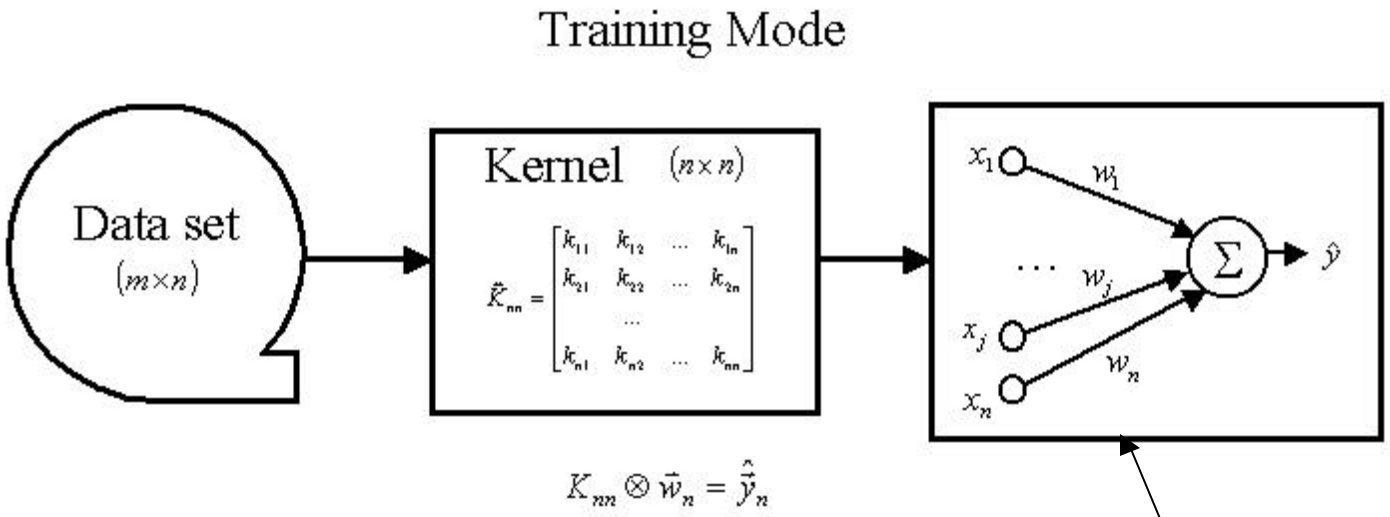
Keywords: PLS; PLSR; Two-block predictive PLS; Latent variables; Multivariate analysis

Kernel PLS (K-PLS)

- **Direct Kernel PLS is PLS with the kernel transform as a pre-processing step**
  - K-PLS ➔ "better" nonlinear PLS
  - PLS ➔ "better" Principal Component Analysis (PCA) for regression
- **K-PLS gives almost identical (but more stable) results as SVMs**
  - Easy to tune (5 latent variables)
  - Unlike SVMs there is no patent on K-PLS
- **K-PLS transforms data from a descriptor space to a t-score space**



Implementing Direct Kernel Methods

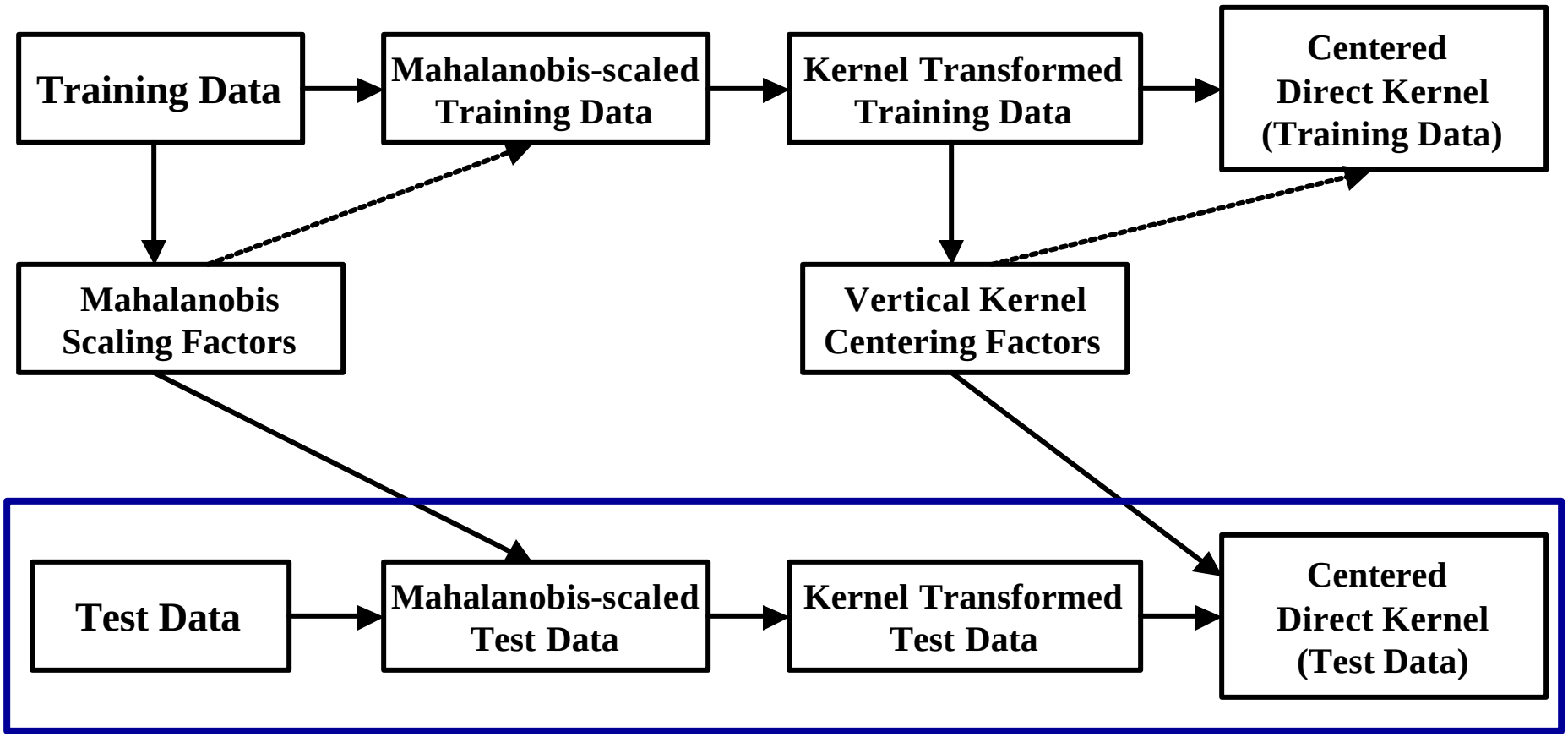


Linear Model:

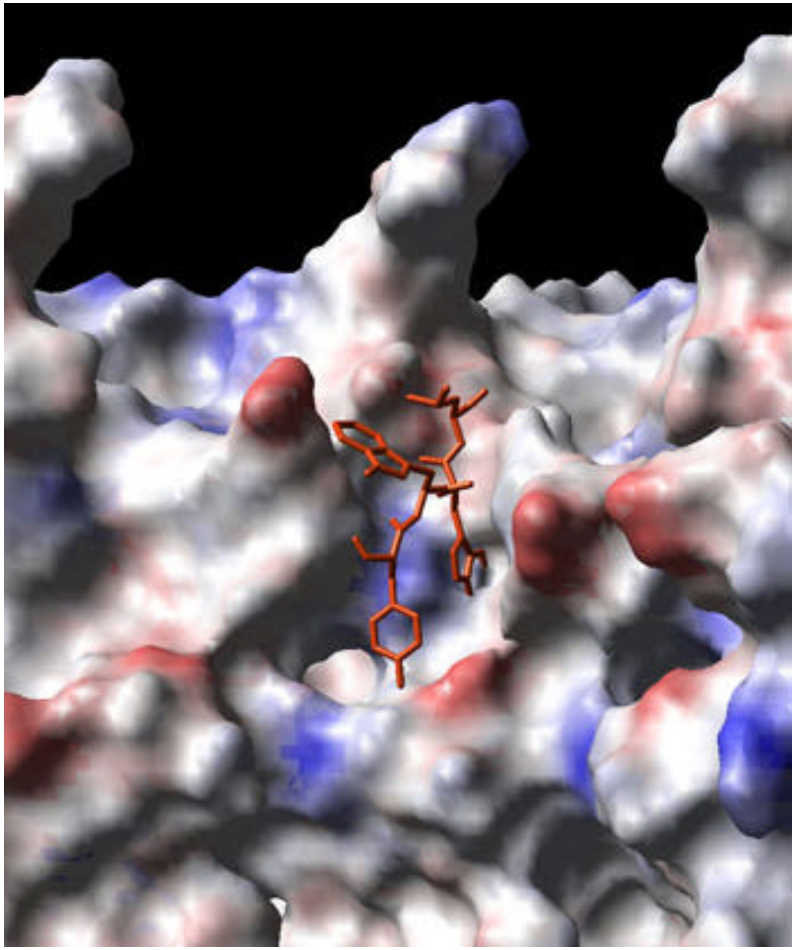
- **PCA model**
- **PLS model**
- **Ridge Regression**
- **Self-Organizing Map**

. . .

Scaling, centering & making the test kernel centering consistent



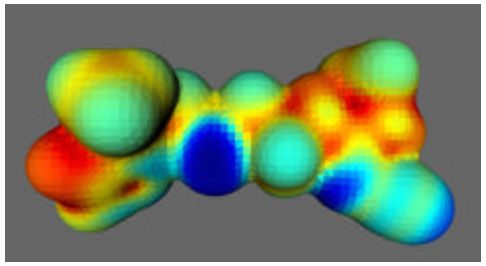
Docking Ligands is a Nonlinear Problem



Electron Density-Derived TAE-Wavelet Descriptors

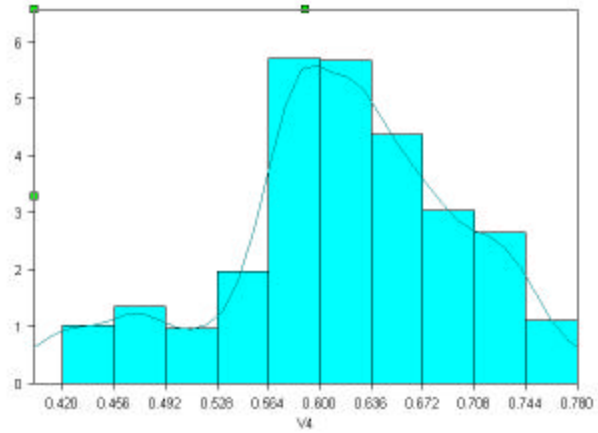
- Surface properties are encoded on 0.002 e/au^3 surface

  Breneman, C.M. and Rhem, M. [1997] *J. Comp. Chem.*, Vol. *18* (2), p. 182-197
- Histograms or wavelet encoded of surface properties give Breneman's TAE property descriptors
- 10x16 wavelet descriptore

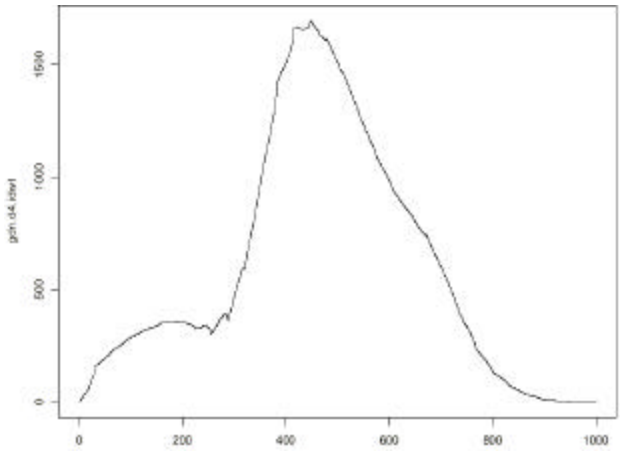


PIP (Local Ionization Potential)

Histograms



Wavelet Coefficients



Data Preprocessing

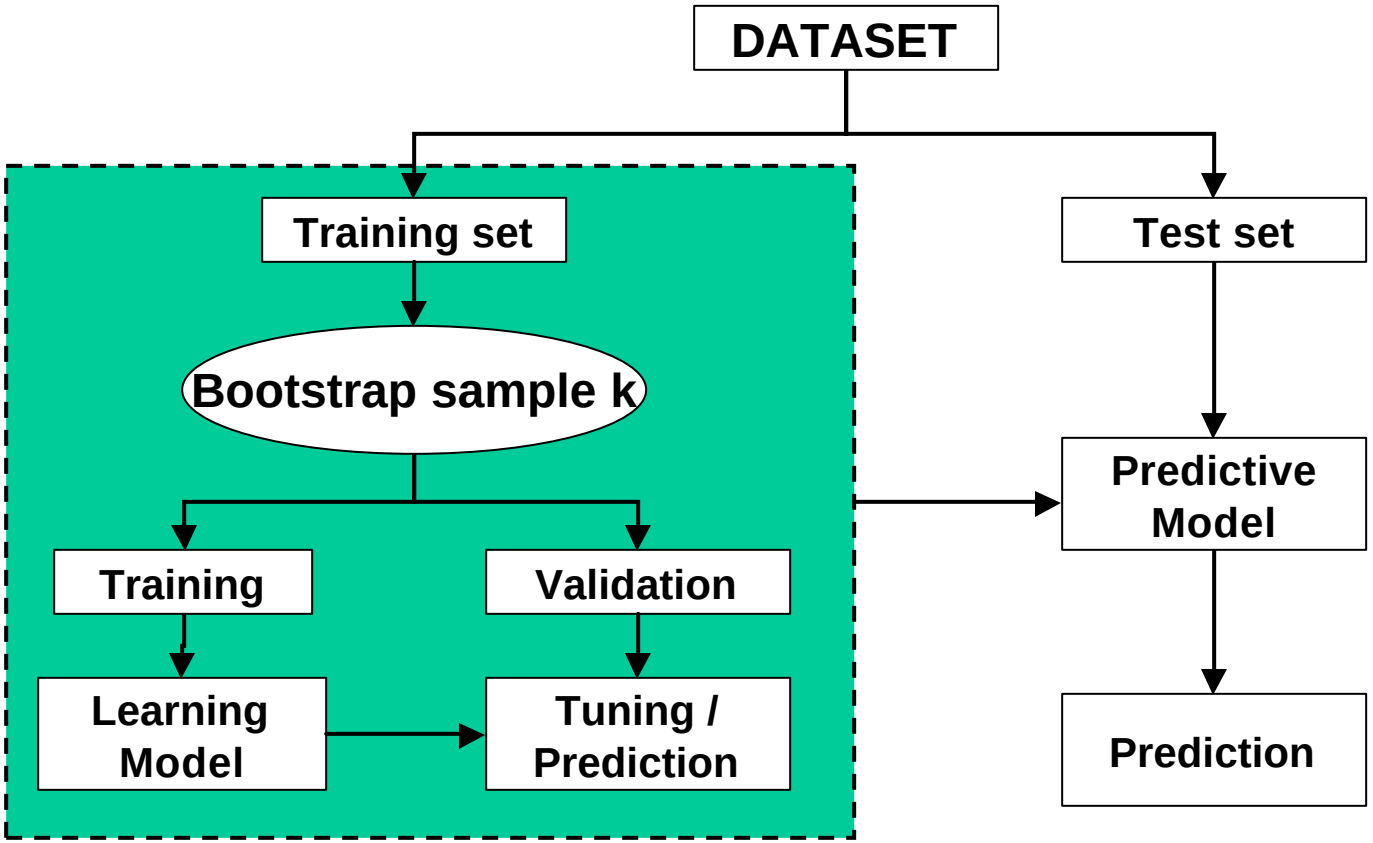
- Data Preprocessing for Competition
  - data centering
  - to normalize or not? (no)
- General Data Preprocessing Issues:
  - extremely important for the success of an application
  - if you know what the data are you can do smarter preprocessing
  - drop features with extremely low correlation coefficient and sparsity
  - outlier detection and cherry picking?

Acknowledgment: C. Breneman

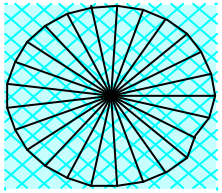
Feature Selection

- Why feature selection
  - explanation of models
  - simplifying models
  - improving models
- Naïve feature selection (filters):
  - drop all features that are more than 95% correlated but one
  - drop features with less than 1% sparsity (binary features)
  - drop features with extremely low correlation coefficient
- Sensitivity analysis for feature selection (wrappers)
  - make model (e.g., SVM, K-PLS, neural network)
  - keep features frozen at average
  - tweak all features and drop 10% of the least sensoitive features
    - ➔ boostrap mode
    - ➔ random gauge parameter
- Note: For most competition datasets we could find an extremely small feature set that works perfect on training date, but did not generalize to validation data.

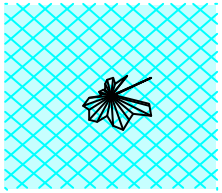
Bootstrapping: Model Validation



Caco-2 – 14 Features (SVM)



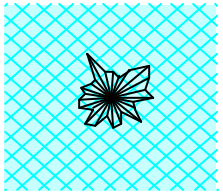
a.don



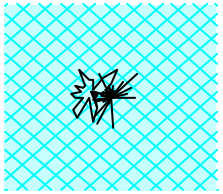
DRNB10

PEOE.VSA.FNEG

BNPB31



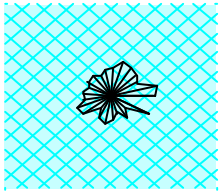
KB54



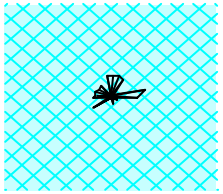
ABSDRN6

ABSKMIN

FUKB14

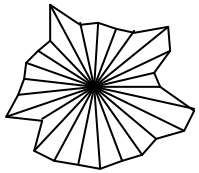


SMR.VSA2

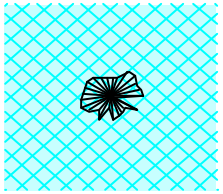


PEOE.VSA.FPPOS

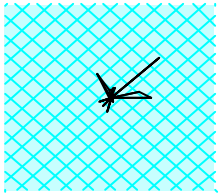
SIKIA



SlogP.VSA0



ANGLEB45



DRNB00

- Each star represents a descriptor
- Each ray is a separate bootstrap
- The area of a star represents the relative importance of that descriptor
- Descriptors shaded cyan have a negative effect
- Unshaded ones have a positive effect

- **Hydrophobicity** - a.don
- **Size and Shape** - ABSDRN6, SMR.VSA2, ANGLEB45 Large is bad. Flat is bad. Globular is good.
- **Polarity** – PEOE.VSA...: negative partial charge good.

Conclusions

- Thanks to competition organizers for a challenging and fair competition
- Congratulations to the winners
- Congratualtions to those who ranked in front of me