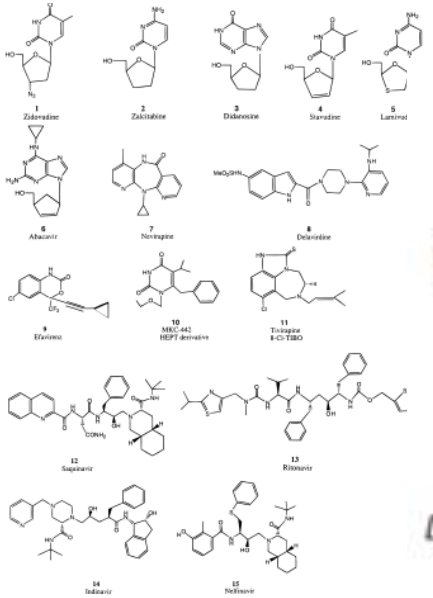
Linear Regression for QSAR

Prof. Em. Dr. ir. Mark J. Embrechts
Helmholtz Centrum for Health and Environment
Munich Germany

I know what I know	I know what I don’t know
I don’t know what I know	I don’t know what I don’t know

HELMHOLTZ
MUNICH

1860-1960: Origins of QSAR: Quantitative Structure Activity Relationship ➔ Drug Design



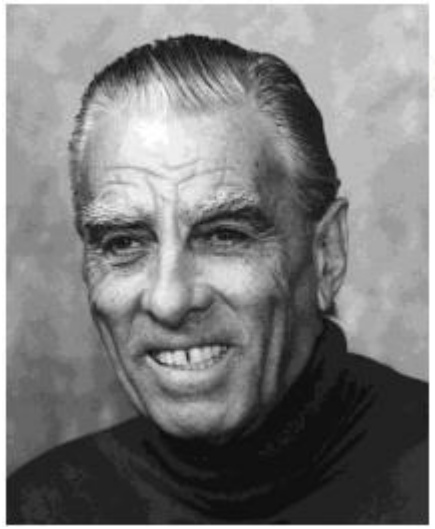
Chem. Rev. **1999**, *99*, 3525–3601

Comparative Quantitative Structure–Activity Relationship Studies on Anti-HIV Drugs

Rajni Garg,[†] Satya P. Gupta,[*,‡] Hua Gao,[§] Mekapati Suresh Babu,[||] Asim Kumar Debnath,[⊥] and Corwin Hansch[*,†]

Department of Chemistry, Pomona College, Claremont, California 91711, Departments of Chemistry and Pharmacy, Birla Institute of Technology and Science, Pilani 333031, India, Pharmacia & Upjohn, 301 Henrietta Street, Kalamazoo, Michigan 49007, and Biochemical Virology Laboratory, Lindsley F. Kimball Research Institute of The New York Blood Center, 310 E. 67th Street, New York, New York 10021

Received May 21, 1999



Contents

RNA virus is now called human immunodeficiency virus (HIV)[3,4] and two genetically distinct subtypes, HIV-1 and HIV-2, have been characterized,[5–7] of which the former has been found to be prevalent in causing the disease.

In the present review, the QSAR studies available or derivable on anti-HIV chemicals are discussed. We have compared the optimum Clog P values (log P_0) observed in correlation equations and then compared them with the Clog P values (calculated log P) of those anti-HIV chemicals which are in the market.

From Narcosis to Hyperspace: The History of QSAR

Hugo Kubinyi*

BASF AG, Ludwigshafen, Germany

The Early Days

QSAR history has no clear starting point. Its roots developed over about a century, from the 1860s to the 1960s [1–4]. The earliest report on a relationship between molecular and biological properties seems to be documented in a thesis by A. F. A. Cros, University of Strasbourg, in 1863. He observed an increase in the toxicity of alcohols to mammals, with decreasing water solubility, up to a maximum potency [5].

In 1868, A. Crum Brown and T. Fraser studied the biological effects of certain alkaloids, prior to and after methylation of a basic nitrogen atom. They observed pronounced differences between the basic and the permanently charged quaternary compounds, which led them to the conclusion that "physiological activity" Φ should be a function of the chemical constitution C (Eq. 1) [6].

$$\Phi = f(C) \tag{1}$$

Of course, they had no chance to describe any specific example, using this relationship. There was no way to encode chemical structures in a quantitative manner. In addition, the chemical structures of most organic com-

relationships), or by the corresponding changes of molecular properties (Hansch-type analyses).

$$\Delta\Phi = f(\Delta C) \tag{2}$$

At about the same time as Crum Brown and Fraser formulated their general structure-activity relationship, B. J. Richardson showed that the narcotic activity of alcohols was proportional to their molecular weight [10]. In 1893, C. Richet observed that the toxicity of ethers, aldehydes, alcohols, ketones and other compounds was inversely related to their aqueous solubility: "*Plus ils sont solubles, moins ils sont toxiques*" [11].

More general theories to explain the mechanism of narcosis were independently formulated by H.H. Meyer [12, 13] and C. E. Overton [14, 15], at the turn of the 19th century. They proposed that the toxicity of neutral organic compounds is related to their ability to partition between water and a lipophilic biophase, where they exert their biological action. As a model system for partitioning they proposed the system olive oil/water.

Theories on Drug Action and Organic Reactivity

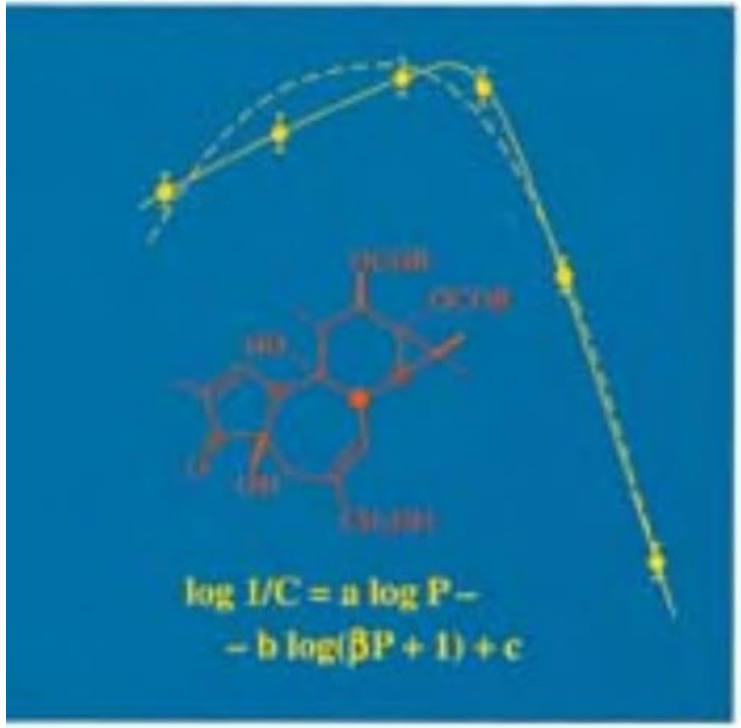
Hugo Kubinyi [2002] From narcosis to hyperspace: The history of QSAR. Quantitaive Structure-Activity Relationships, Vol. 21, pp. 348-356..

Edited by
R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman

QSAR: Hansch Analysis and Related Approaches

by Hugo Kubinyi

Volume 1



VCH

WILEY-VCH

Roberto Todeschini, Viviana Consonni

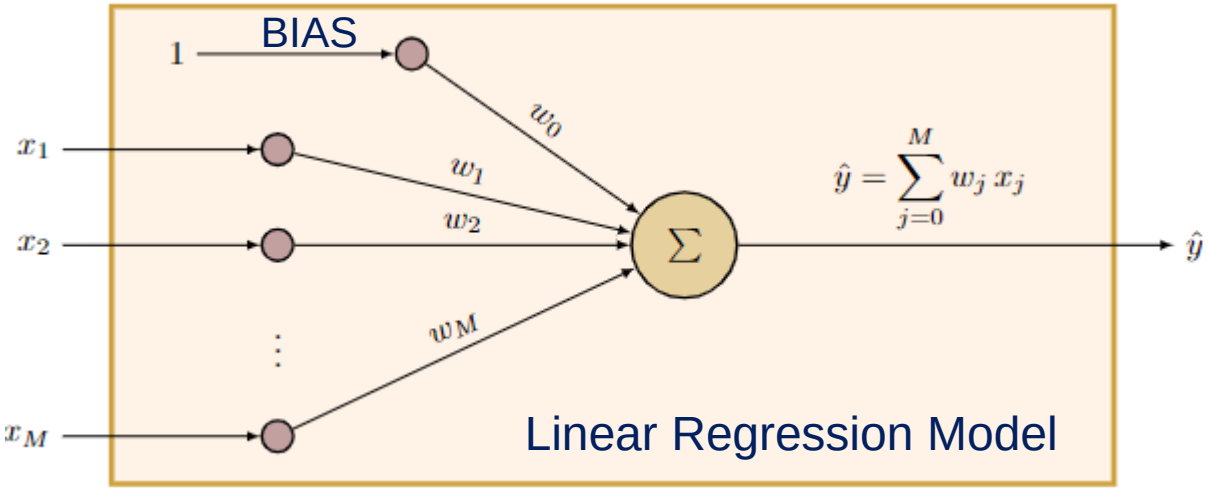
Handbook of Molecular Descriptors

Methods and Principles in Medicinal Chemistry

Volume 11

Edited by
R. Mannhold,
H. Kubinyi,
H. Timmerman

Linear Regression



Terminology:

x_i descriptors, features $(i = 1..M)$

y^μ responses $(i = 1..N)$

$\hat{y}^\mu$ predicted responses $(\mu = 1..N)$

$\vec{\boldsymbol{x}}$ data record

$\boldsymbol{X}_{NM}$ data matrix

N number of data patterns

M number of features

$\vec{w}$ weight vector

Linear Regression Model

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M$$

$$\hat{y} = \sum_{j=0}^{M} w_j x_j$$

How to make a linear model?

possible answer

$$\vec{\boldsymbol{y}} = \boldsymbol{X}_{NM} \vec{\boldsymbol{w}}$$

$$\vec{\boldsymbol{w}} = (\boldsymbol{X}_{NM}^T \boldsymbol{X}_{NM})^{-1} \boldsymbol{X}_{NM}^T \vec{\boldsymbol{y}}$$

How good is the model? ➔ Loss function

mean squared error loss function

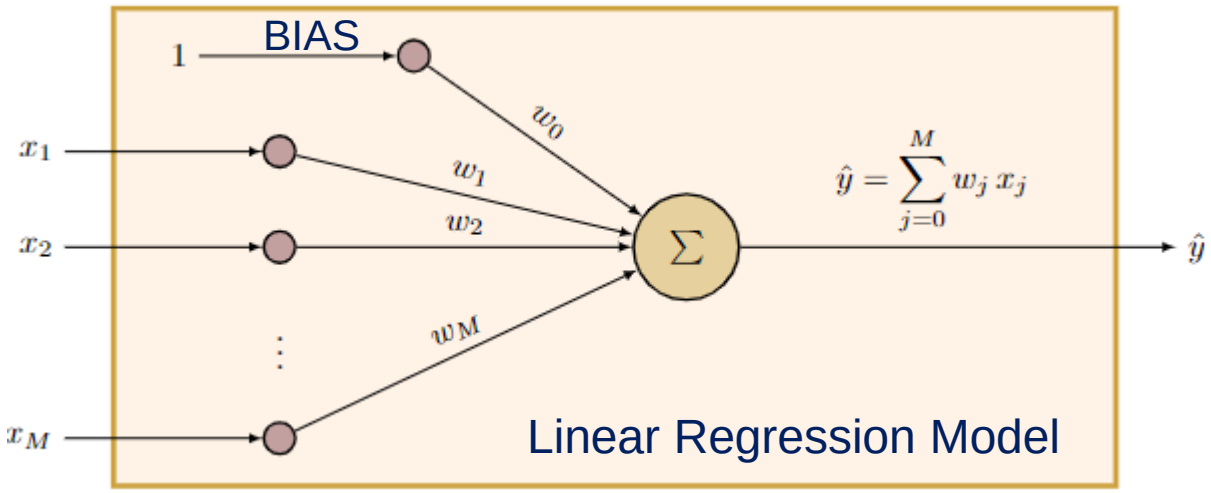
$$L_{MSE} = \frac{1}{N} \sum_{\mu=0}^{N} (y^\mu - \hat{y}^\mu)^2$$

Linear Regression Model

Prediction model for a single data pattern

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M$$

$$\hat{y} = \sum_{i=0}^{M} w_i x_i$$



Model for all data

$$\vec{\boldsymbol{y}} = \boldsymbol{X}_{NM}\vec{\boldsymbol{w}}$$

How to find the weight vector?: solution

$$\boldsymbol{X}_{NM}^T\vec{\boldsymbol{y}} = (\boldsymbol{X}_{NM}^T\boldsymbol{X}_{NM})\vec{\boldsymbol{w}}$$

$$\vec{\boldsymbol{w}} = (\boldsymbol{X}_{NM}^T\boldsymbol{X}_{NM})^{-1}\boldsymbol{X}_{NM}^T\vec{\boldsymbol{y}}$$

How good is the model? ➔ Loss function

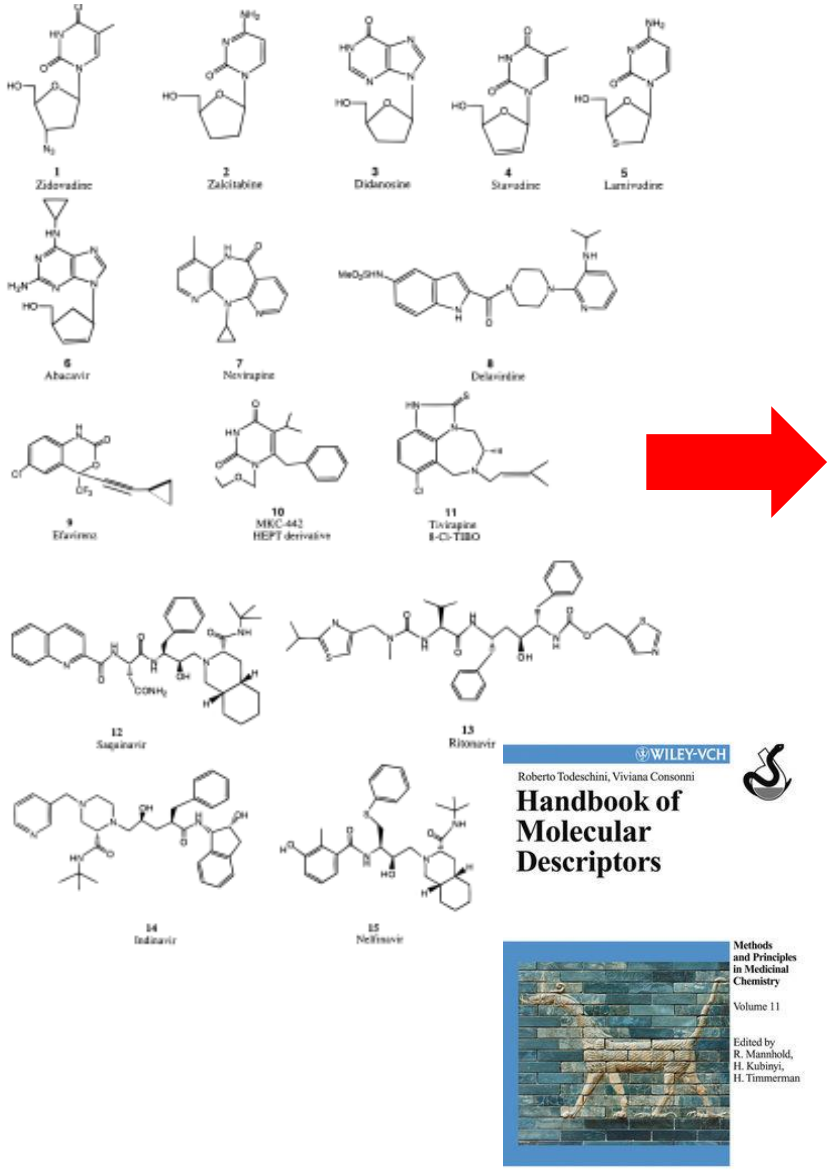
mean squared error loss function

$$L_{MSE} = \frac{1}{N}\sum_{\mu=0}^{N}(y^{\mu} - \hat{y}^{\mu})^2$$

Does the model predict well on new data?

How good is the model?
What can we learn from a model: XAI (explainable AI)?
Is this the best possible model?

Generating Descriptors (features/attributes) for Molecules in QSAR: unstructured data ➔ structured data



	A	B	C	D	E	F	G	H	I
1	0.23	0.31	-0.55	254.2	2.126	-0.02	82.2	8.5	1
2	-0.48	-0.6	0.51	303.6	2.994	-1.24	112.3	8.2	2
3	-0.61	-0.77	1.2	287.9	2.994	-1.08	103.7	8.5	3
4	0.45	1.54	-1.4	282.9	2.933	-0.11	99.1	11	4
5	-0.11	-0.22	0.29	335	3.458	-1.19	127.5	6.3	5
6	-0.51	-0.64	0.76	311.6	3.243	-1.43	120.5	8.8	6
7	0	0	0	224.9	1.662	0.03	65	7.1	7
8	0.15	0.13	-0.25	337.2	3.856	-1.06	140.6	10.1	8
9	1.2	1.8	-2.1	322.6	3.35	0.04	131.7	16.8	9
10	1.28	1.7	-2	324	3.518	0.12	131.5	15	10
11	-0.77	-0.99	0.78	336.6	2.933	-2.26	144.3	7.9	11
12	0.9	1.23	-1.6	336.3	3.86	-0.33	132.3	13.3	12
13	1.56	1.79	-2.6	366.1	4.638	-0.05	155.8	11.2	13
14	0.38	0.49	-1.5	288.5	2.876	-0.32	106.7	8.2	14
15	0	-0.04	0.09	266.7	2.279	-0.4	88.5	7.4	15
16	0.17	0.26	-0.58	283.9	2.743	-0.53	105.3	8.8	16
17	1.85	2.25	-2.7	401.8	5.755	-0.31	185.9	9.9	17
18	0.89	0.96	-1.7	377.8	4.791	-0.84	162.7	8.8	18
19	0.71	1.22	-1.6	295.1	3.054	-0.13	115.6	12	19

descriptors response ID#

Working with standardized (scaled) data in Linear Regression

	A	B	C	D	E	F	G	H	I
1	0.23	0.31	-0.55	254.2	2.126	-0.02	82.2	8.5	1
2	-0.48	-0.6	0.51	303.6	2.994	-1.24	112.3	8.2	2
3	-0.61	-0.77	1.2	287.9	2.994	-1.08	103.7	8.5	3
4	0.45	1.54	-1.4	282.9	2.933	-0.11	99.1	11	4
5	-0.11	-0.22	0.29	335	3.458	-1.19	127.5	6.3	5
6	-0.51	-0.64	0.76	311.6	3.243	-1.43	120.5	8.8	6
7	0	0	0	224.9	1.662	0.03	65	7.1	7
8	0.15	0.13	-0.25	337.2	3.856	-1.06	140.6	10.1	8
9	1.2	1.8	-2.1	322.6	3.35	0.04	131.7	16.8	9
10	1.28	1.7	-2	324	3.518	0.12	131.5	15	10
11	-0.77	-0.99	0.78	336.6	2.933	-2.26	144.3	7.9	11
12	0.9	1.23	-1.6	336.3	3.86	-0.33	132.3	13.3	12
13	1.56	1.79	-2.6	366.1	4.638	-0.05	155.8	11.2	13
14	0.38	0.49	-1.5	288.5	2.876	-0.32	106.7	8.2	14
15	0	-0.04	0.09	266.7	2.279	-0.4	88.5	7.4	15
16	0.17	0.26	-0.58	283.9	2.743	-0.53	105.3	8.8	16
17	1.85	2.25	-2.7	401.8	5.755	-0.31	185.9	9.9	17
18	0.89	0.96	-1.7	377.8	4.791	-0.84	162.7	8.8	18
19	0.71	1.22	-1.6	295.1	3.054	-0.13	115.6	12	19

(structured) raw data

scaling

	A	B	C	D	E	F	G	H	I
1	-0.21	-0.25	0.20	-1.38	-1.27	0.90	-1.38	-0.52	1
2	-1.18	-1.19	1.10	-0.21	-0.35	-1.05	-0.33	-0.63	2
3	-1.36	-1.36	1.68	-0.58	-0.35	-0.79	-0.63	-0.52	3
4	0.09	1.02	-0.52	-0.70	-0.41	0.76	-0.79	0.42	4
5	-0.68	-0.79	0.91	0.53	0.15	-0.97	0.21	-1.34	5
6	-1.22	-1.23	1.31	-0.02	-0.08	-1.35	-0.04	-0.40	6
7	-0.52	-0.57	0.67	-2.07	-1.76	0.98	-1.98	-1.04	7
8	-0.32	-0.43	0.45	0.58	0.57	-0.76	0.66	0.08	8
9	1.12	1.29	-1.11	0.24	0.03	1.00	0.35	2.58	9
10	1.23	1.19	-1.03	0.27	0.21	1.13	0.35	1.91	10
11	-1.58	-1.59	1.32	0.57	-0.41	-2.68	0.79	-0.74	11
12	0.71	0.70	-0.69	0.56	0.57	0.41	0.37	1.27	12
13	1.61	1.28	-1.53	1.27	1.40	0.86	1.20	0.49	13
14	-0.01	-0.06	-0.60	-0.57	-0.47	0.42	-0.52	-0.63	14
15	-0.52	-0.61	0.74	-1.08	-1.10	0.30	-1.16	-0.93	15
16	-0.29	-0.30	0.17	-0.67	-0.61	0.09	-0.57	-0.40	16
17	2.00	1.76	-1.62	2.11	2.59	0.44	2.25	0.01	17
18	0.69	0.43	-0.77	1.54	1.56	-0.41	1.44	-0.40	18
19	0.45	0.69	-0.69	-0.41	-0.28	0.73	-0.21	0.79	19

(structured) scaled data

standardizing inputs and outputs

$$y_i' = \frac{y_i - \bar{y}}{\sigma_y}$$

$$x_j' = \frac{x_j - \bar{x}}{\sigma_x}$$

D. W. Marquardt [1948] You should standardize the predictor variables in your regression Models. Journal of the American Statistical Association, Vol. 75, pp.87-91.

Working with training and validation (test) sets in QSAR

- Split data in training set and validation (test) set
- Training set (80% of data patterns): to build the model
- Validation (test) set (20% of data patterns): to see how well model works on new data
- Error metrics for a regression model: L_{MSE}, r^2 and R^2, q^2 and Q^2
- How well does the model generalize? r^2 and R^2 on validation data set

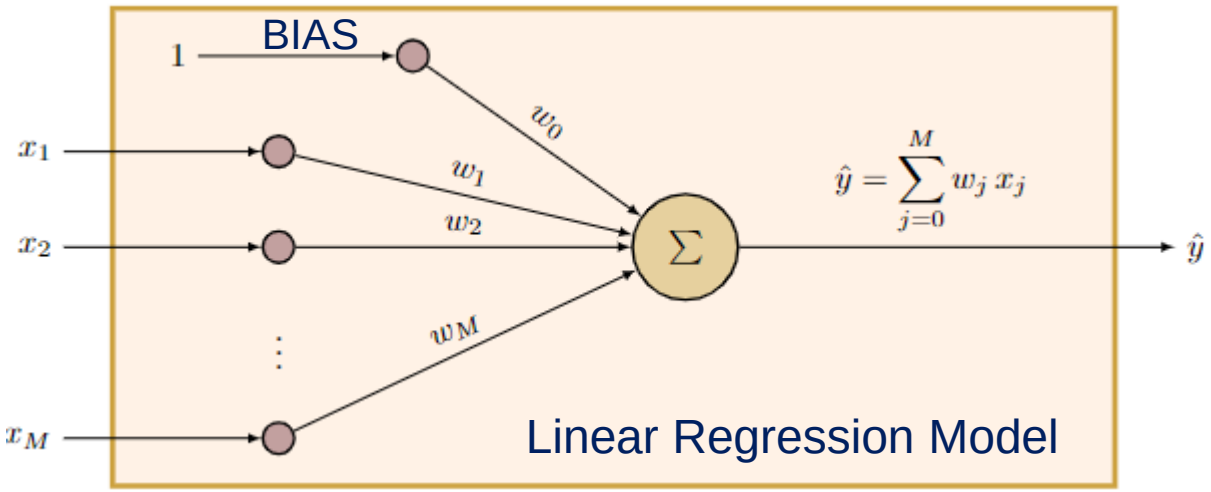
$$L_{MSE} = \frac{1}{N}\sum_{\mu=0}^{N}(y^{\mu} - \hat{y}^{\mu})^2$$

$$r^2 = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \qquad q^2 = 1 - r^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \qquad Q^2 = 1 - q^2$$

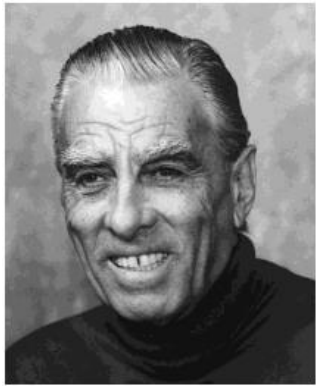
QSAR Model for Hansch's HIV Data

- Split data in training set and validation set
- Training set (80% of data patterns): to build the model
- Validation (test) set (20% of data patterns): to see how well model works on new data
- Error metrics for a regression model: L_{MSE}, r^2 and R^2, q^2 and Q^2
- How well does model generalize? Evaluate r^2 and R^2 on validation data set



R. Garg, S. P. Gupta, H. Gao, M. S. Babu, A. K., Debnath, and C. Hansch [1999] Comparative Quantitative Structure-Activity Relationship Studies on Anti-HIV Drugs. Chemical Review, Vol. 99, pp. 3525 - 3601.

Applying linear regression for QSAR on Hansch HIV data



- Structure data: Make descriptors for Hansch data (e.g., MOE descriptors)
  - 64 data records (N)
  - 184 descriptors (M)
- Standardize (scale) data
- Split data in training data and validation (test) data
- Apply linear model on training data
- See how well linear model works on validation (test) data
- What can we learn from a model?
  ➔Factor Analysis: To determine most important descriptors

R. Garg, S. P. Gupta, H. Gao, M. S. Babu, A. K., Debnath, and C. Hansch [1999] Comparative Quantitative Structure-Activity Relationship Studies on Anti-HIV Drugs. Chemical Review, Vol. 99, pp. 3525 - 3601.

Problem with Hansch HIV data

Problem: Ill-conditioned matrix because $M > N$ (i.e., Matrix inverse does not exist)

$$\vec{\boldsymbol{w}} = (\boldsymbol{X}_{NM}^T \boldsymbol{X}_{NM})^{-1} \boldsymbol{X}_{NM}^T \vec{\boldsymbol{y}}$$

Solution: 1. Use less descriptors (make $M < N$)
2. Get more data
3. Mathematical tricks

Math Solution #1: Tikhanov regularization

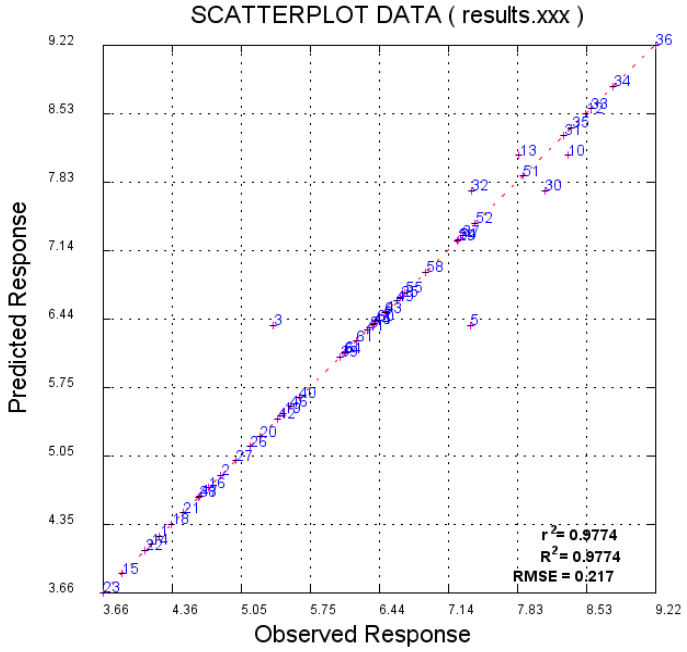
$$\vec{\boldsymbol{w}} = (\boldsymbol{X}_{NM}^T \boldsymbol{X}_{NM} + \lambda \boldsymbol{I}_{MM})^{-1} \boldsymbol{X}_{NM}^T \vec{\boldsymbol{y}}$$ (λ is a small regularization or ridge parameter)

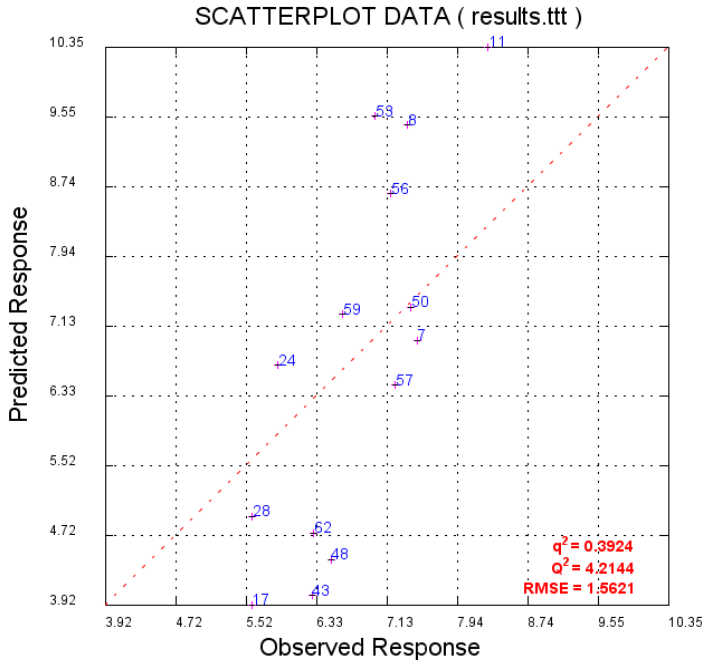
Math Solution #2: Partial Least Squares (PLS) methods
Math Solution #3: Principal Component regression (PCR)
Math Solution #4: Iterative methods (e.g., stochastic gradient descent)
More math solutions: Support Vector Machines (SVMs), Lasso, ...

QSAR Model for Hansch's HIV Data: Linear Regression

$$\vec{\boldsymbol{w}} = (\boldsymbol{X}_{NM}^T \boldsymbol{X}_{NM})^{-1}\, \boldsymbol{X}_{NM}^T \vec{\boldsymbol{y}} = (\boldsymbol{X}_{NM}^T \boldsymbol{X}_{NM})^{-1} \boldsymbol{X}_{NM}^T \vec{\boldsymbol{y}}$$

- 64 data records, 184 MOE descriptors (features)
- Random split into 50 training data and 14 validation (test) data
- Build model: Apply linear regression on training data with small ridge parameter
- Test model performance on validation (test) data
- Results





Conclusion: Works perfect on training data.

Does not work well on validation data. Can we do better?

What is PLS? Projection to Latent Structures

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*

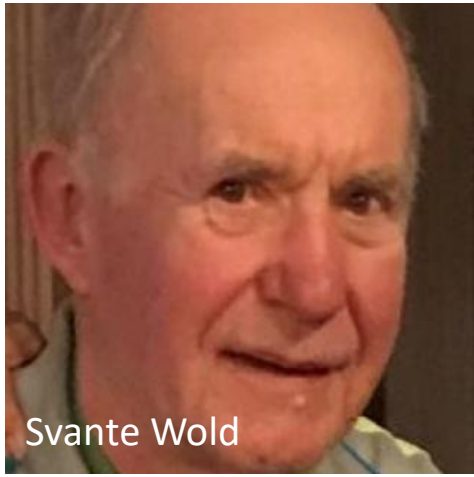
Abstract

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.

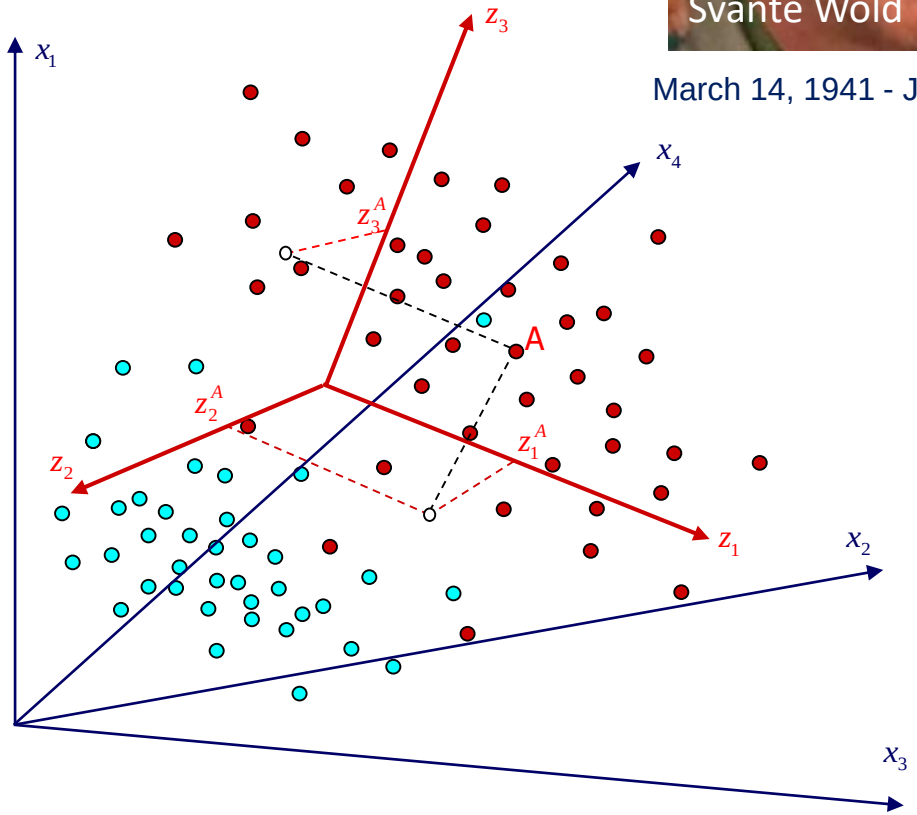
This article reviews PLSR as it has developed to become a standard tool in chemometrics and used in chemistry and engineering. The underlying model and its assumptions are discussed, and commonly used diagnostics are reviewed together with the interpretation of resulting parameters.

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

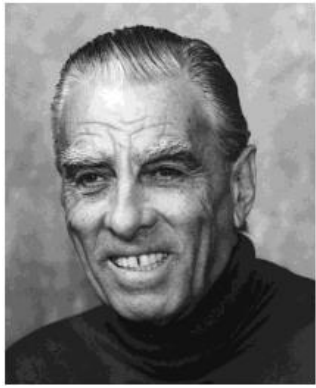


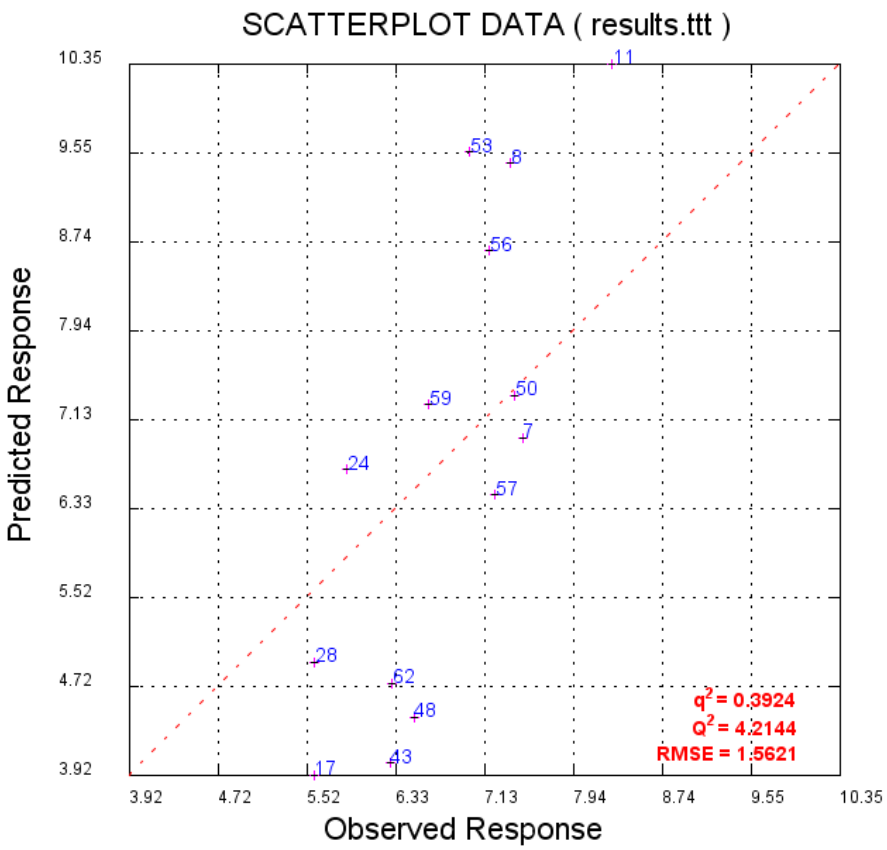
March 14, 1941 - January 4, 2022



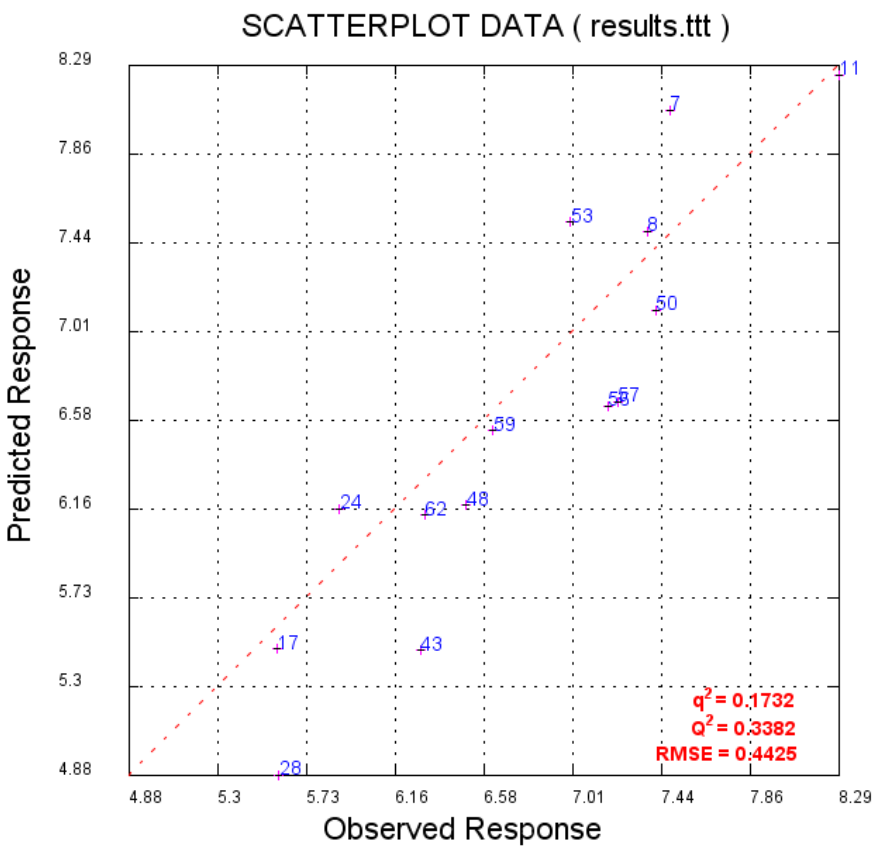
PLS: Partial Least Squares
PLS: Projection to Latent Structures
PLS: Please Listen to Svante

PLS: Partial Least Squares (Hansch HIV data)





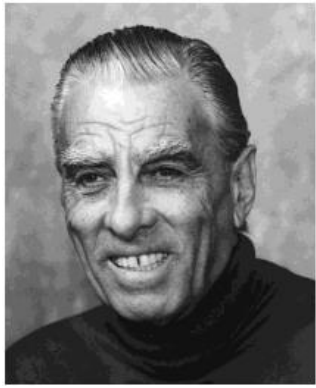
Linear Model with matrix inverse and $\lambda = 0.001$

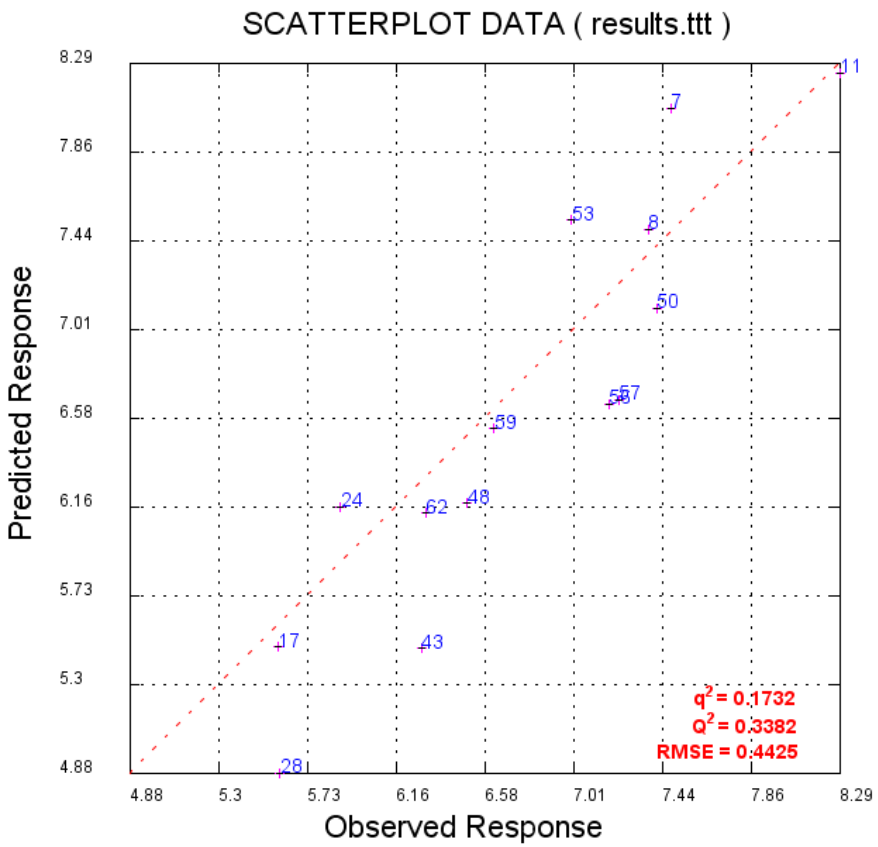


Linear PLS Model with 2 latent variables

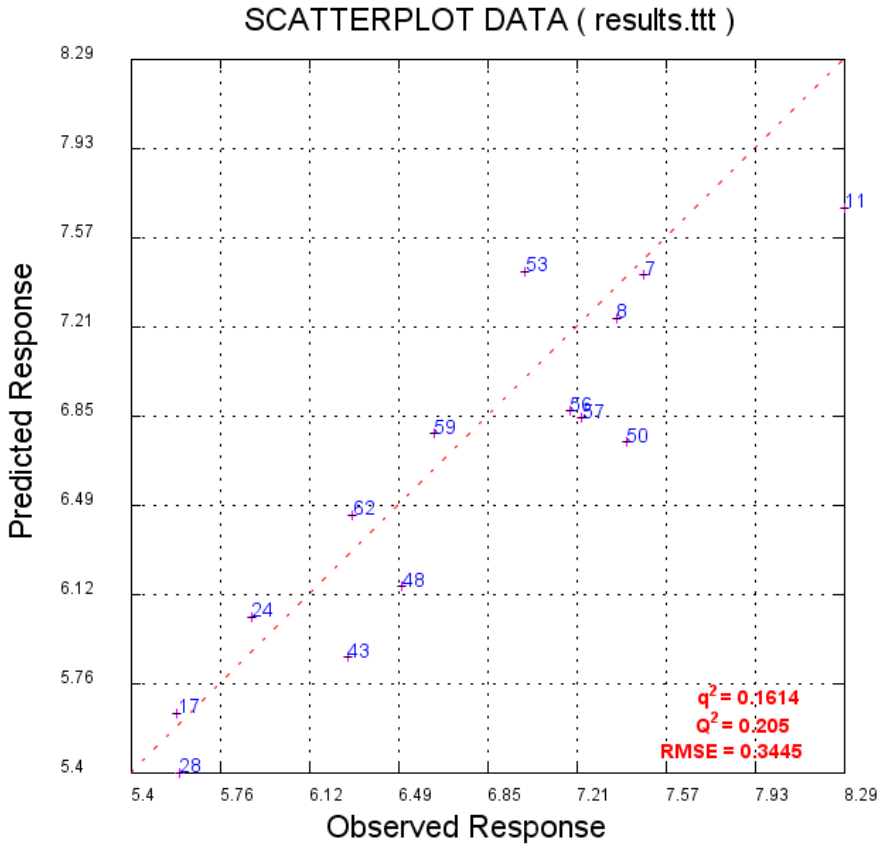
q2	Q2	MSE	MAE	Method
0.3924	4.2144	1.562	1.369	Standard linear method
0.1836	0.3139	0.426	0.363	Linear with regularization (lambda = 500)
0.3278	0.5642	0.572	0.454	PLS 3 latent variables
0.1732	0.3382	0.443	0.371	PLS 2 latent variables
0.2805	0.3673	0.461	0.347	PLS 1 latent variable

Logistic Regression (Hansch HIV data)





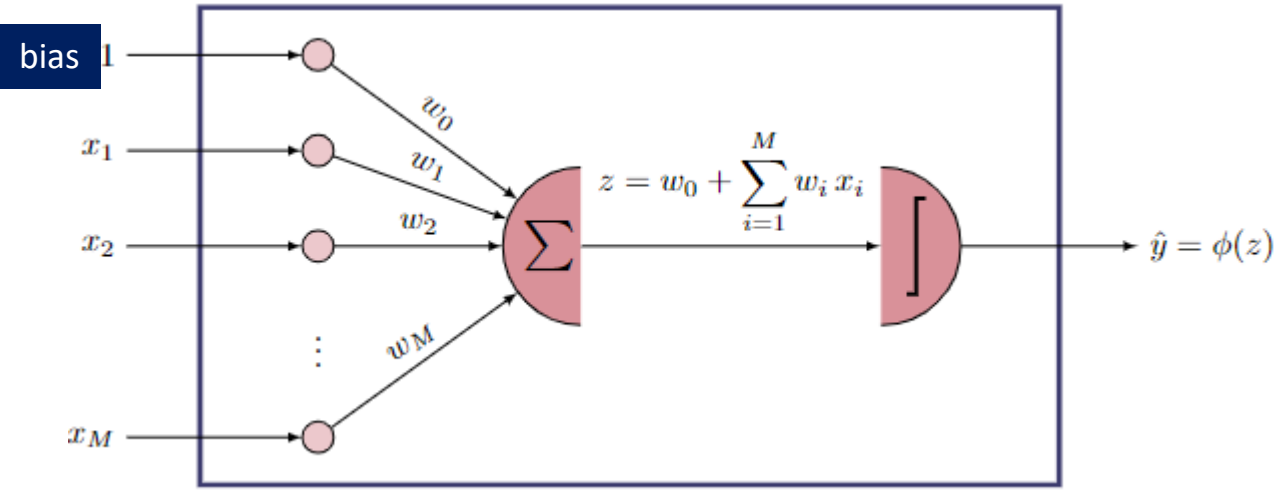
Linear PLS Model with 2 latent variables



Logistic regression model

q2	Q2	MSE	MAE	Method
0.1836	0.3139	0.426	0.363	Linear with regularization (lambda = 500)
0.1732	0.3382	0.443	0.371	PLS 2 latent variables
0.1614	0.2050	0.345	0.288	Logistic regression (3 its, early stopping)
0.7291	4.4725	1.609	1.436	Classic second-order logistic regression
0.4834	1.3731	0.892	0.730	Deep Learning

Logistic Regression: Basic Idea



Why a cross-entropy loss function?
Answer: Predictions are now class probabilities

Logistic Regression Model

$$z = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_M x_M$$

$$z = \sum_{j=0}^{M} w_j x_j \qquad \hat{y} = \frac{1}{1 + e^{-z}}$$

How to make a logistic regression model?

Gradient Descent

$$\vec{\boldsymbol{w}}^{(I+1)} = \vec{\boldsymbol{w}}^{(I)} - \eta \frac{dE}{d\vec{\boldsymbol{w}}}$$

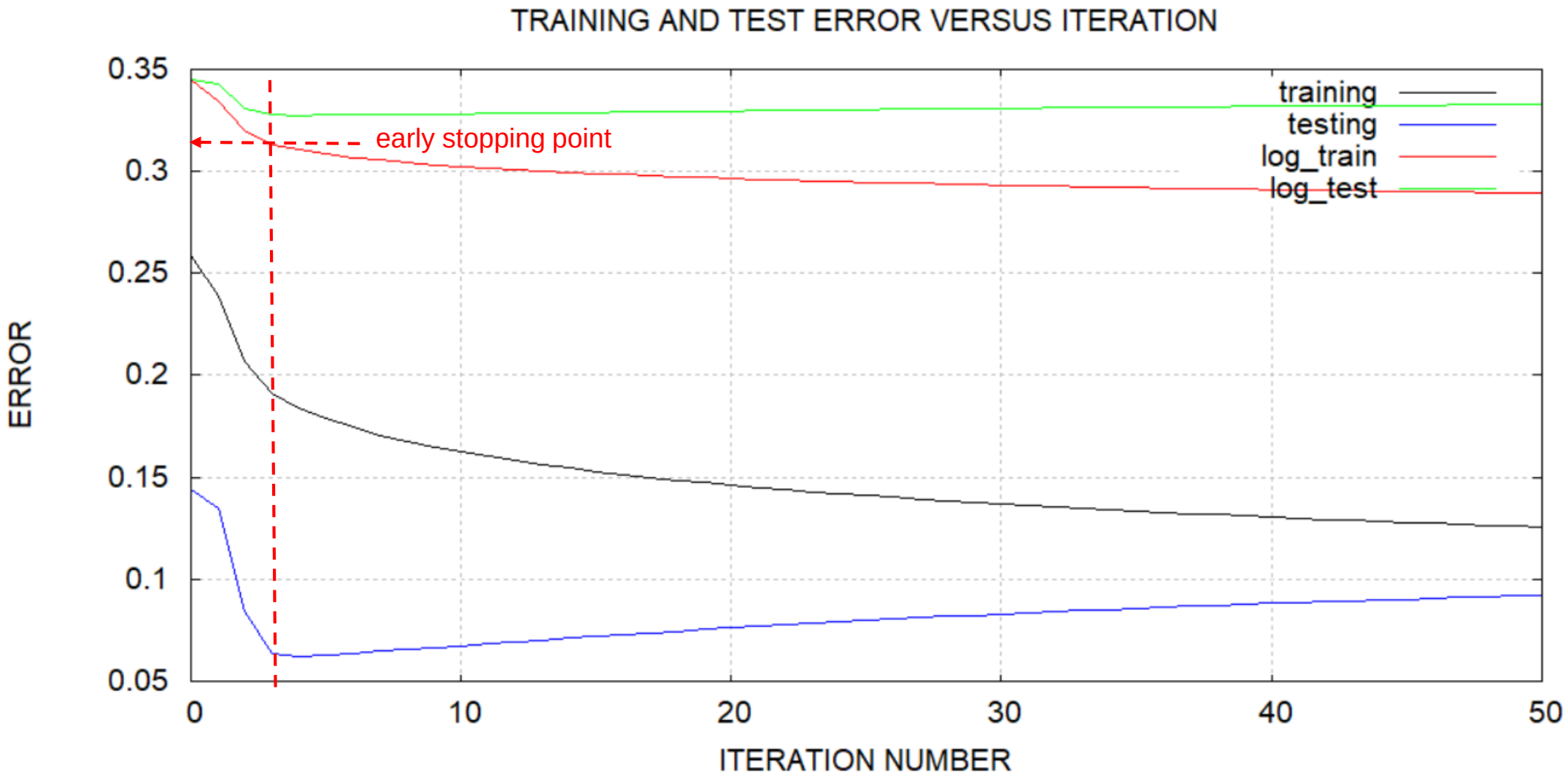
$$\Delta w_j = \eta\left(y_j - \hat{y}_j\right) x_j$$

How good is the model? ➔ Loss (error) function

Cross-entropy loss function

$$L_{Entropy} = -\sum_{\mu=1}^{N} y^{\mu} log(\hat{y}^{\mu}) + (1 - y^{\mu}) log(1 - \hat{y}^{\mu})$$

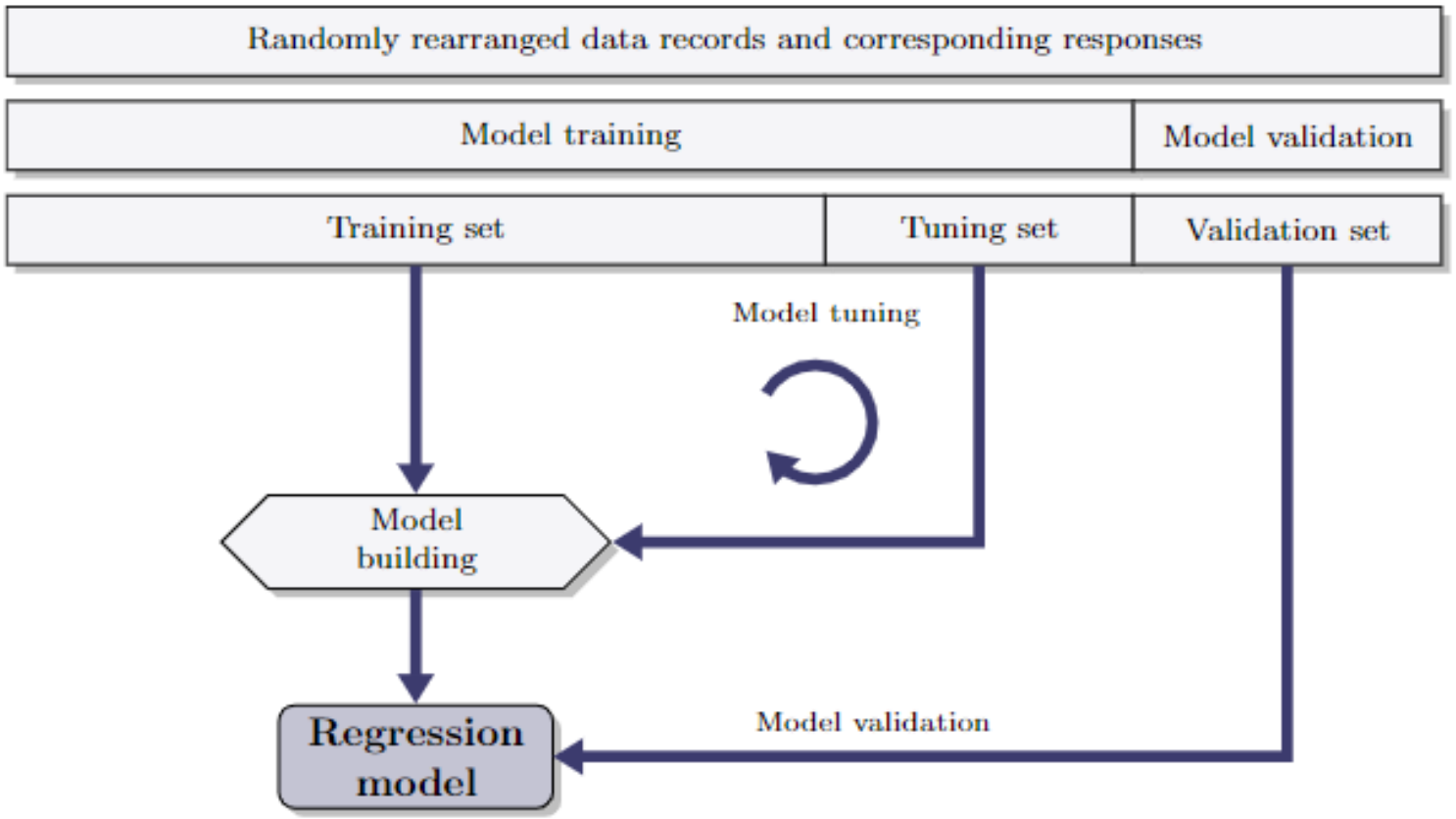
Logistic regression with batch regression and early stopping



Working with scripts (logistic regression example)

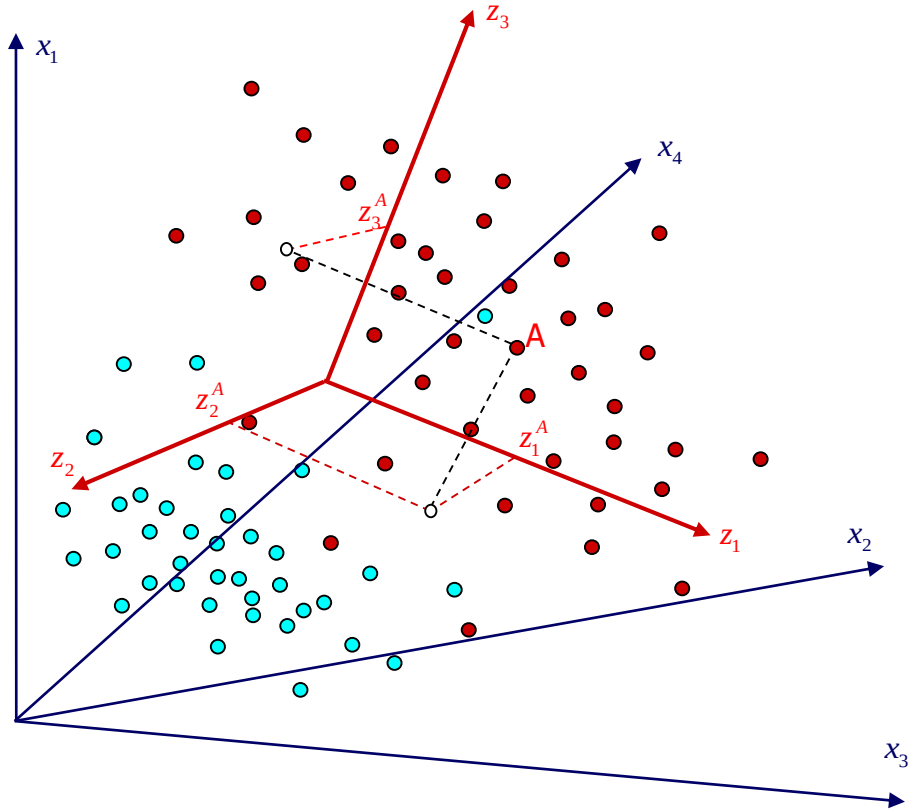
```
REM MAKE LOGISTIC REGRESSION MODEL
REM SCALE DATA
mje han --SCALE_LOGISTIC
REM SPLIT DATA
mje han.txt --SPLITT_BIG
REM MAKE LOGISTIC REGRESSION MODEL
mje han.pat --IRS2
pause
REM ERROR PROGRESSION
mje han.pat --ERR
REM DO METRICS
mje resultss.ttt --DESCALE_LOGISTIC
mje results.ttt --METRICS
mje results.ttt --RESIDUAL
mje results.ttt --SCATTER_PLOT
REM FACTOR ANALYSIS
mje han.pat --FAC_LR2
```

Regularization in a Linear Regression Model: Split data into: training set, tuning set, validation set



- How many principal components to include in a PCR model?
- How many latent variables to consider in a PLS model?
- What is a good value for the regularization parameter or λ?
- Early stopping point?

Principal Component Regression (PCR)



Correlation Matrix

$$R_{MM} = \frac{1}{N-1} \boldsymbol{X}_{NM}^T \boldsymbol{X}_{NM}$$

Eigenvector Decomposition of data matrix

$$X_{NM} = T_{NH} B_{HM}$$
$$T_{NH} = X_{NM} B_{NH}^T$$

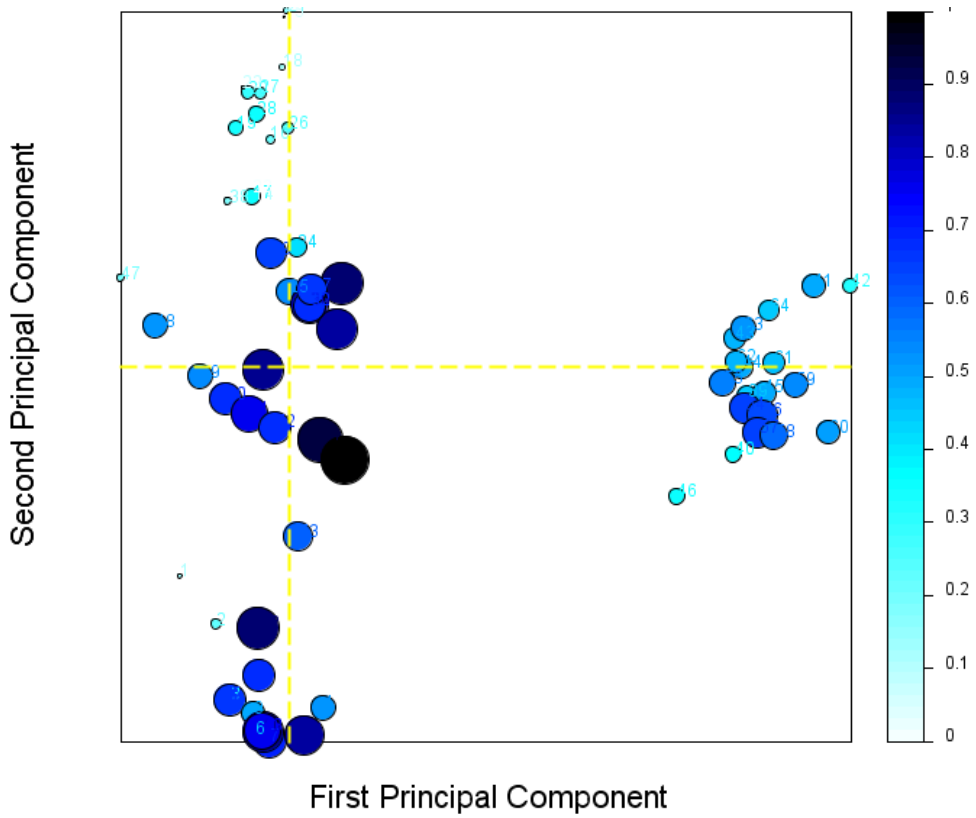
Problem Reformulation

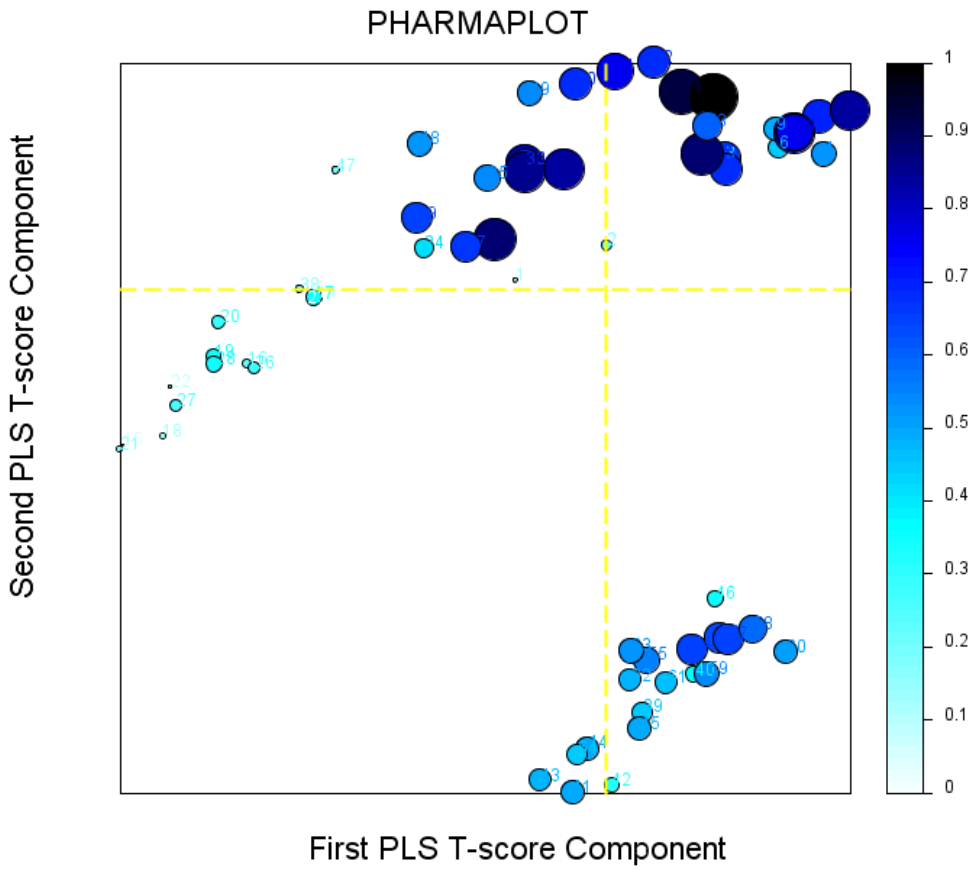
$$\vec{\boldsymbol{w}} = (\boldsymbol{T}_{HN}^T \boldsymbol{T}_{NH})^{-1} T_{NH}^T \vec{\boldsymbol{y}}$$

Key Question: How many principal components?

Basic Idea: Orthogonal coordinate transformation
Project data on eigenvectors of the correlation matrix
Question: How many eigenvectors to consider?

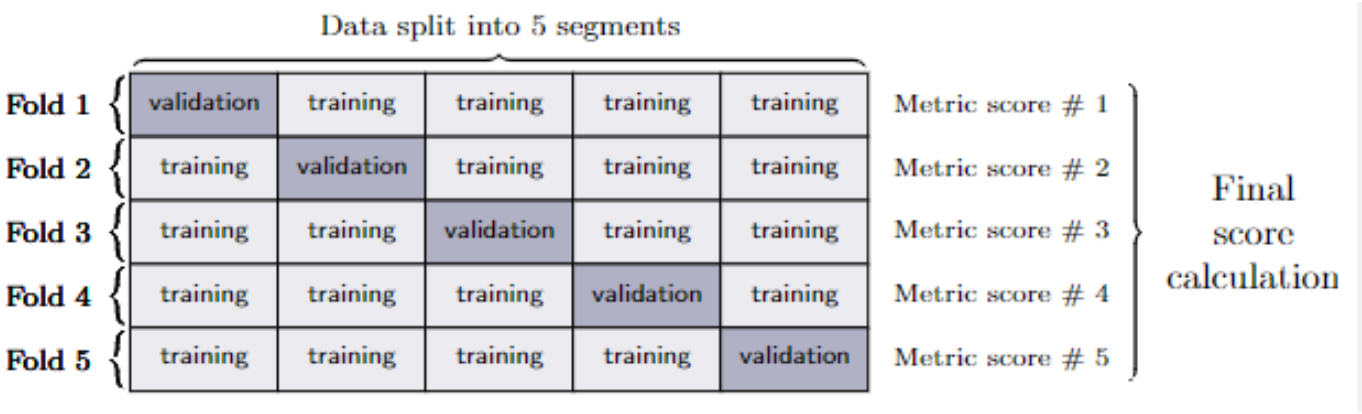
Principal Component and PLS-Score Plots for Hansch Dataset





How to deal with small data sets? Cross-Validation

	A	B	C	D	E	F	G	H	I
1	0.23	0.31	-0.55	254.2	2.126	-0.02	82.2	8.5	1
2	-0.48	-0.6	0.51	303.6	2.994	-1.24	112.3	8.2	2
3	-0.61	-0.77	1.2	287.9	2.994	-1.08	103.7	8.5	3
4	0.45	1.54	-1.4	282.9	2.933	-0.11	99.1	11	4
5	-0.11	-0.22	0.29	335	3.458	-1.19	127.5	6.3	5
6	-0.51	-0.64	0.76	311.6	3.243	-1.43	120.5	8.8	6
7	0	0	0	224.9	1.662	0.03	65	7.1	7
8	0.15	0.13	-0.25	337.2	3.856	-1.06	140.6	10.1	8
9	1.2	1.8	-2.1	322.6	3.35	0.04	131.7	16.8	9
10	1.28	1.7	-2	324	3.518	0.12	131.5	15	10
11	-0.77	-0.99	0.78	336.6	2.933	-2.26	144.3	7.9	11
12	0.9	1.23	-1.6	336.3	3.86	-0.33	132.3	13.3	12
13	1.56	1.79	-2.6	366.1	4.638	-0.05	155.8	11.2	13
14	0.38	0.49	-1.5	288.5	2.876	-0.32	106.7	8.2	14
15	0	-0.04	0.09	266.7	2.279	-0.4	88.5	7.4	15
16	0.17	0.26	-0.58	283.9	2.743	-0.53	105.3	8.8	16
17	1.85	2.25	-2.7	401.8	5.755	-0.31	185.9	9.9	17
18	0.89	0.96	-1.7	377.8	4.791	-0.84	162.7	8.8	18
19	0.71	1.22	-1.6	295.1	3.054	-0.13	115.6	12	19



Basic idea of 5-fold cross-validation

Svante Wold's cartoon QSAR data:
N = 19 data records, M = 7 features

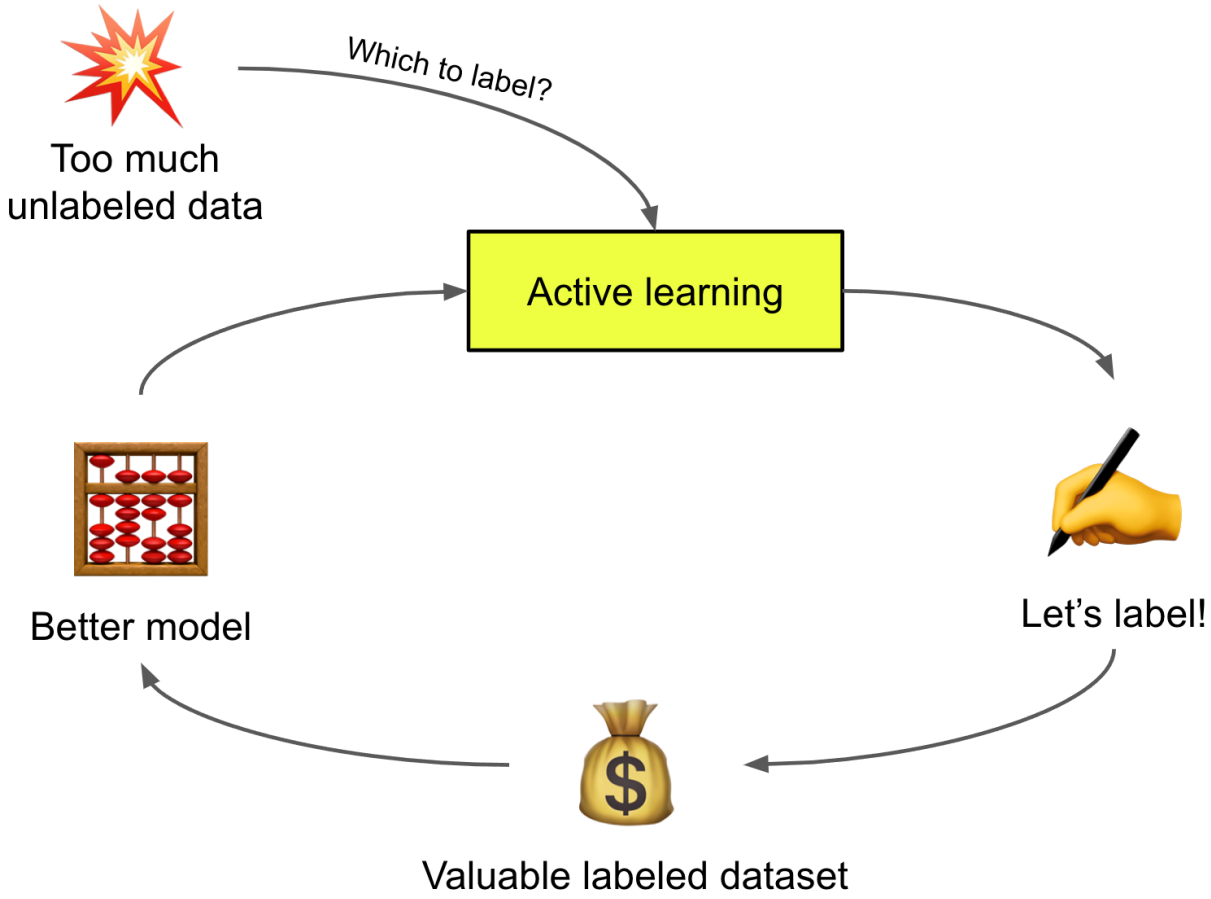
Alternatives:

- K-fold cross-validation
- Leave-one-out validation
- Bootstrapping

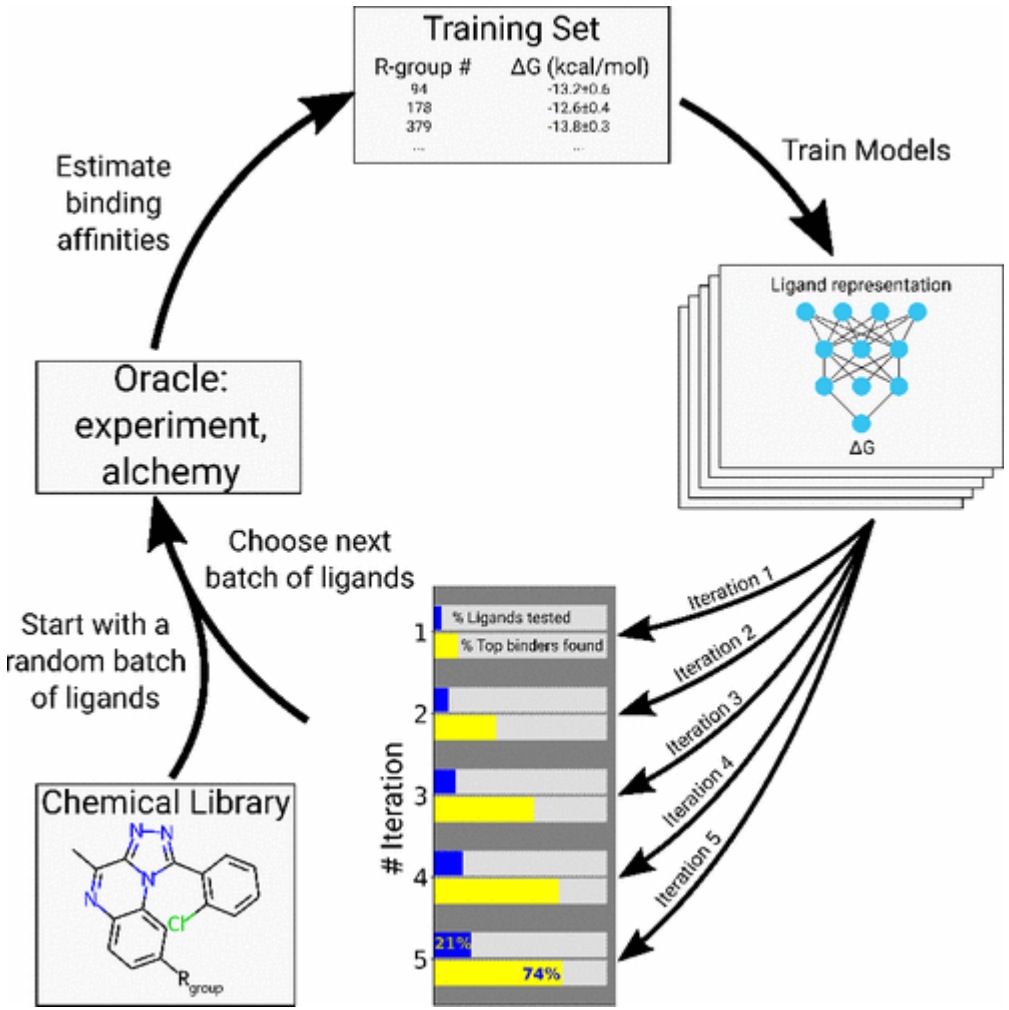
q2	Q2	MSE	MAE	# latent variables in PLS
0.6912	0.7011	2.248	1.758	1
0.6364	0.6443	2.155	1.787	2
0.5207	**0.5516**	**1.994**	**1.768**	**3**
0.6207	0.8943	2.539	1.882	4
0.6753	1.3177	3.081	2.106	5

Svante Wold, Michael Sjoestroem, Lennart Eriksson [2001] PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, Vol. 58, pp. 109-130.

Active Learning

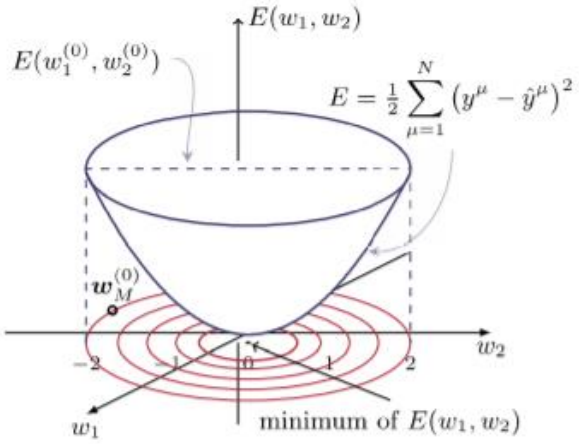


Active Learning

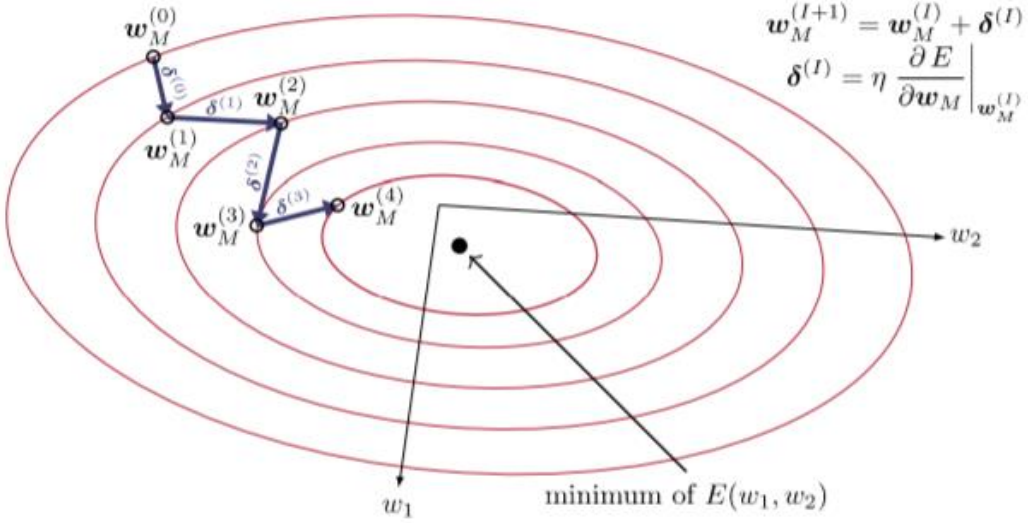


Yuri Khalak, Gary Tresadern, David F. Hahn, Bert L. de Groot, and Vytautas Gapsys [2022] Chemical space exploration with active learning and alchemical free energies. Journal of Chemical Theory and Computing, Vol 18(10), pp. 6259-6270.

Stochastic Gradient Descent: Delta Rule



(a) 3D plot of $E(w_1, w_2)$



(b) Path from $\boldsymbol{w}_M^{(0)}$ to $\boldsymbol{w}_M^{(4)}$ in the $w_1 w_2$-plane

$$L_{MSE} = \frac{1}{N}\sum_{\mu=1}^{N}(y^\mu - \hat{y}^\mu)^2 \rightarrow E = \frac{1}{2}\sum_{\mu=1}^{N}(y^\mu - \hat{y}^\mu)^2$$

$$w^{(I+1)} = w^{(I)} - \eta\frac{dE}{dw}$$

$$\Delta w_j = \eta(y - \hat{y})x_j$$

$$\Delta w_j = \eta\delta x_j$$

Initialize weights (random)
Apply one pattern
Update weights according to: $\Delta w_j = \eta\,(y_i - \hat{y}_i)x_i$

(I) is the iteration index
η is the learning parameter

a good choice is $\eta = \frac{1}{MN}$

Stochastic and Batch Gradient Descent for Linear Regression

Stochastic Gradient Descent

$$w_j^{(I+1)} = w_j^{(I)} - \eta \frac{dL_{ENT}}{dw}$$

$$w_j^{(I+1)} = w_j^{(I)} + \Delta w_j$$

$$\Delta w_j = \eta(y - \hat{y})x_j = \eta\delta x_j$$

- Update weights after each pattern
- Learning rate $\eta = \frac{1}{NM}$

Batch Gradient Descent

$$\boldsymbol{w}^{(I+1)} = \boldsymbol{w}^{(I)} - \eta\nabla L_{ENT}$$

$$\boldsymbol{w}^{(I+1)} = \boldsymbol{w}^{(I)} + \Delta\boldsymbol{w}$$

$$\Delta\boldsymbol{w} = \eta\boldsymbol{X}_{MN}^T\,\boldsymbol{\delta}$$

$$\boldsymbol{\delta} = (\boldsymbol{y} - \hat{\boldsymbol{y}})$$

- Update weights
  - after showing all patterns (epoch)
  - after showing mini-batch of patterns
- Learning rate $\eta = \frac{1}{N}$

Better models through Regularization: i.e., keep the weights as small as possible

Mean Squared Error Loss Function

$$L_{MSE} = \frac{1}{N}\sum_{\mu=0}^{N}(y^{\mu} - \hat{y}^{\mu})^2$$

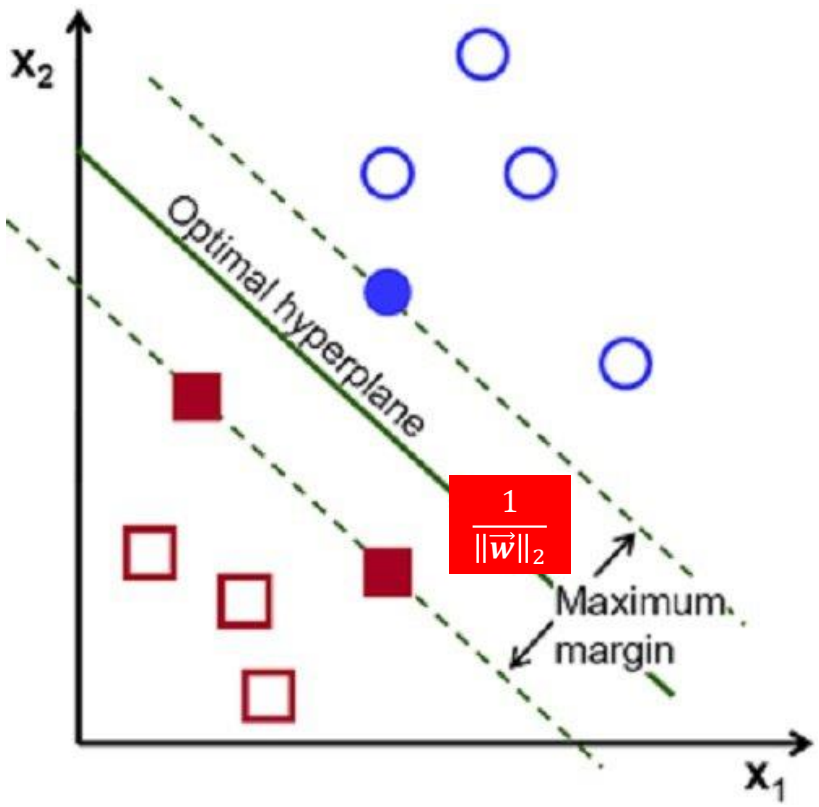
Mean Squared Error Loss Function with regularization

$$L_{SVM} = \frac{1}{N}\sum_{i=0}^{M}(y_i - \hat{y}_i)^2 + \lambda\|\vec{\boldsymbol{w}}\|_2$$

Second objective: Keep weights small

weight regularization objective: minimize $\|\vec{\boldsymbol{w}}\|_2$

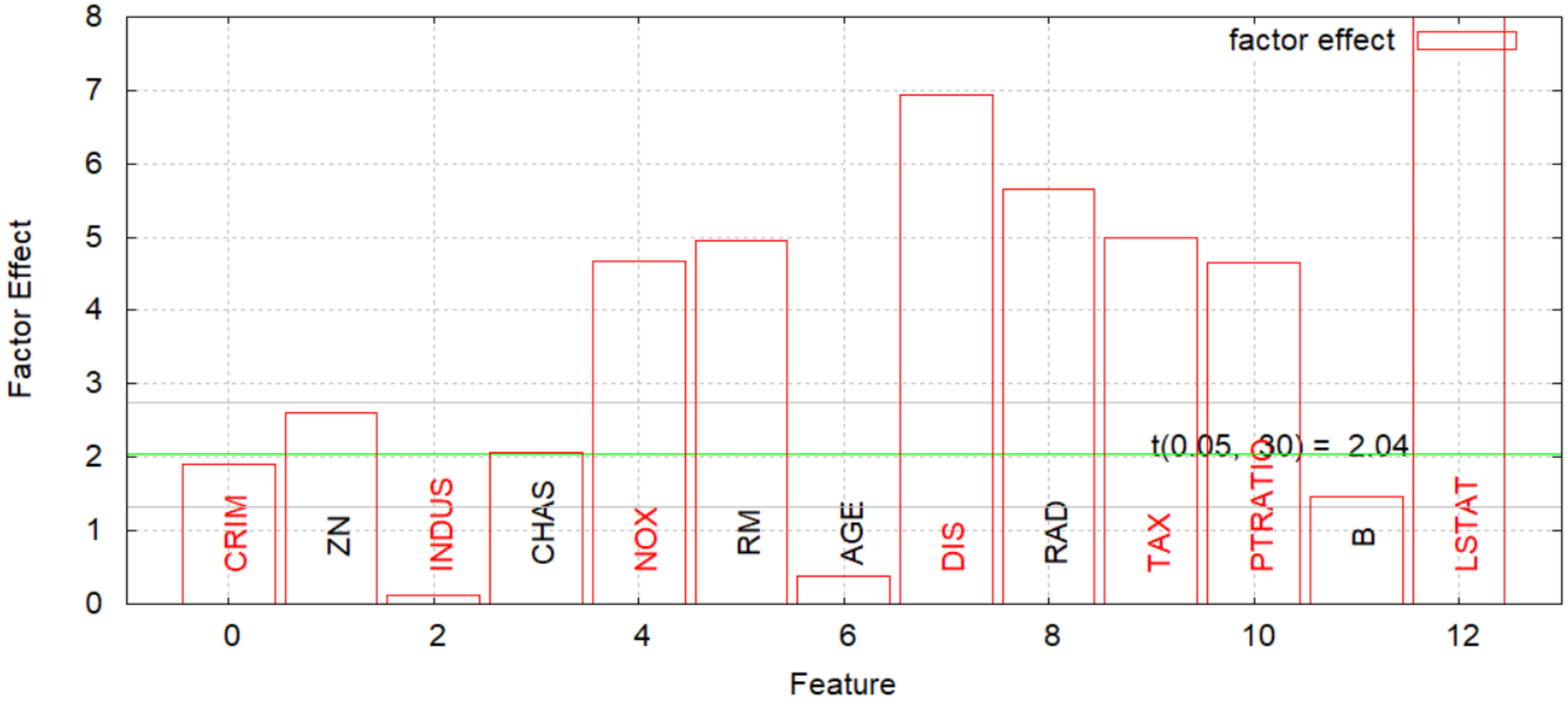
What is a good value for λ? Requires tuning for λ



This is similar to the maximum margin principle in support vector machines (SVM)

Learning from Data: Factor Analysis

Consider Boston housing data (506 data, 13 features or attributes)
The features associated with the weights in a linear model ➔ importance factors



How can we obtain better predictions (generalization)?

- Get more data: if possible?
- Get better data: active learning
- Use better descriptors
- K-fold cross-validation for small data sets
- Use better loss functions (e.g., regularization)
- Use better models by using advanced math
  - better linear models: regularization, PLS, SVM, lasso, ...
  - better nonlinear model: logistic regression, neural networks, deep learning

Take-aways

1. Linear regression with regularization is a powerful tool for small data sets
2. The weights in a linear regression model reflect the most important attributes
3. Keeping the weights small is a powerful regularization tool
4. PLS has inherent regularization and is a powerful linear method
5. PLS can also be used for data visualization
6. Neural networks might not work well on small data sets

Take-aways: Logistic Regression

1. Logistic regression can also be used for regression rather than classification
2. Traditionally logistic regression uses the IRLS algorithm (second-order method)
3. Gradient descent (a neural network method) can also be use for logistic regression
4. Gradient descent is not as aggressive as IRLS for logistic regression ➔ allows for early stopping

Linear Regression: Summary

- Data record (data pattern), attributes (features, descriptors), response
- Structuring QSAR data (with descriptors)
- Standardization (scaling), bias
- Training set, validation set (test set), tuning set
- Loss function (MSE)
- Error metrics: r2, R2, q2, Q2
- Improving generalization (on the validation data)
- Preventing overfitting with regularization and early stopping
- K-fold Cross-Validation (for small datasets)
- Factor Analysis (feature detection)
- Active Learning
- Methods
  - Matrix inverse method
  - Partial Least Squares (PLS)
  - Principal Component Regression (PCR)
  - Gradient Descent
  - Logistic Regression

I know what I know	I know what I don’t know ↑
I don’t know what I know	I don’t know what I don’t know

HELMHOLTZ
MUNICH