V. N. Karazin
Kharkiv National
University
(since 1804)

ROBUST AND UNBIASED ESTIMATIONS IN CHEMICAL DATA TREATMENT

Vladimir Ivanov, Yuriy Kholin
Materials Chemistry Department

http://www.bestnet.kharkov.ua/kholin

Svante Wold, Umea University:

Chemometrics: "How do we get chemically relevant information out of measured chemical data, how do we represent and display this information, and how do we get such information into data?"

A chemical model (**M**) relating experimentally determined variables, **X**, to each other, necessarily is associated with a model of noise (**E**). This model, **E**, describes the variability (noise)
X = M + E; Data = Chemical Model + Model of Noise

The problems and drawbacks not excluded even at using modern statistical packages:

– the lack of ability to detect the real outliers and, at the same time, the tendency to exclude suspicious measurements without sufficient justification,

– the inherent danger to over-fit data and to build models with good fitting but no predictive ability and to obtain the meaningless values of fitting parameters (the inevitable risk at solving ill-posed problems),

– the application of common statistical methods for data handling at situations when the statistical premises are certainly disturbed (for instance, for data arrays for QSAR or QSPR analysis).

Nobody can guarantee that the standard statistical procedures give the trustworthy results

Data analysis includes every procedures of data handling that can not be completely formalized (the complete algorithmization is impossible):

- consideration of data at variation of hypotheses about statistical and another properties of measurements,

- the set of models rather than the only one model are used to interpret data.

Y. Adler, 1982:

Data analysis invents procedures to use completely intrinsic (endogene) information about measurements but also it is aimed at the highest possible involvement of the external information.

The aim of this report is to demonstrate a few procedures of data treatment used in our practice that follow the methodology of data analysis and allow to overcome some drawbacks of conventional statistical procedures.

Topics to be discussed

1. Robust regression on the base of Huber's

M-estimates *vs.* ordinary least squares.

2. Some aspects of data handling in QSAR.

3. Regularization of ill posed problems: where does its power end? What tools can be used else? (analysis of the energetic heterogeneity of sorbents).

Robust regression on the base of Huber's M-estimates *vs.* ordinary least squares

$$A_k = \tilde{f}(\boldsymbol{X}_k, \theta) + \varepsilon_k = \hat{A}_k + \varepsilon_k, \;\; k = 1, 2, ..., \; N$$

The conventional ordinary least-squares (OLS) for obtaining θ estimates:

$$\theta = \arg\min \sum_{k=1}^{N} \left[\hat{A}_k - A_k \right]^2$$

Robust estimators

$$\boldsymbol{\theta} = \arg\min \sum_{k=1}^{N} \rho(\xi_k)$$

$$\xi_k = \widehat{A}_k - A_k$$

Examples of loss functions

L_q – estimates: $\rho(\xi) = |\xi|^q, 1 \le q < 2$

Welsch estimates: $\rho(\xi) = d^2 / 2 \cdot [1 - \exp(-\xi/d)^2], d > 0,$

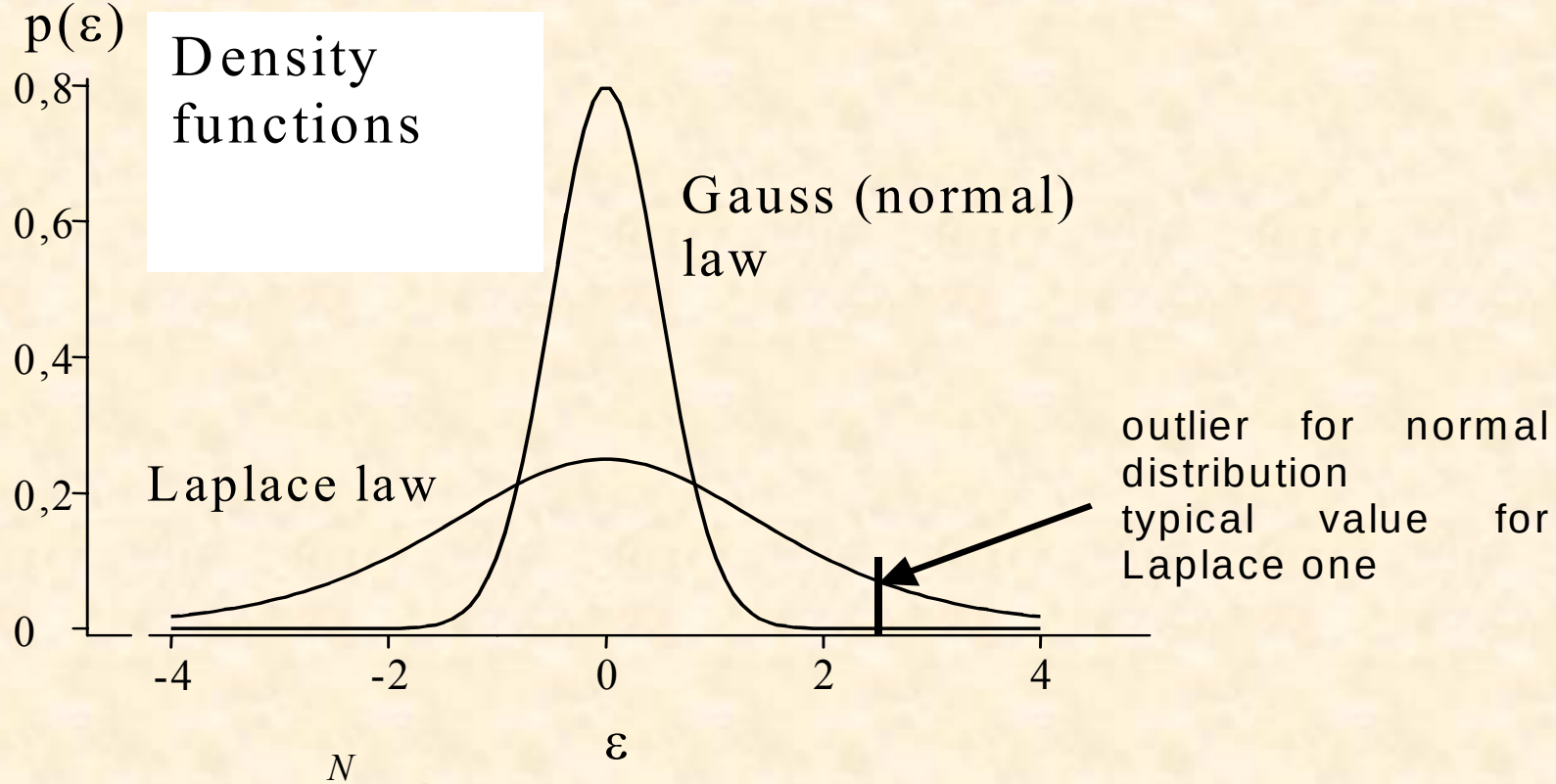
Demidenko estimates: $\rho(\xi) = \xi^2 / (\xi + d), d > 0,$

Andrews estimates:

$$\rho(\xi) = \begin{cases} (2\lambda)^{-1} \{1 - \cos[(2\lambda)^{1/2} \xi]\}, & |\xi| < \pi(2\lambda)^{-1/2} \\ \lambda^{-1}, & |\xi| \ge \pi(2\lambda)^{-1/2} \end{cases}, \lambda > 0.$$

Laplace density function

$$p(\varepsilon) = \frac{1}{2\lambda} \exp\left(-|\varepsilon|/\lambda\right)$$



$$\boldsymbol{\theta} = \arg\min \sum_{k=1}^{N} |\hat{A}_k - A_k|$$ (robust L_1-estimate)

The model of gross errors

$$p(\varepsilon) = [(100 - \delta)\cdot\varphi(\varepsilon) + \delta\cdot h(\varepsilon)] / 100,$$

$\varphi(\varepsilon)$ – density function of the normal distribution,

$h(\varepsilon)$ – long-tails density function ("gross errors" density function),

δ – intensity of "gross errors" density function, %.

$$\rho(\xi) = \begin{cases} (1/2)\xi^2 & \text{at } |\xi| \le c_{out} \\ c_{out}\cdot|\xi| - (1/2)c_{out}^2 & \text{at } |\xi| > c_{out} \end{cases}$$

constant c_{out} depends on the intensity of the "gross errors" density function

at $\delta = 0$ M-estimates are identical to the ordinary least squares,

at $\delta \to 100\%$ M-estimates coincide with the estimates of the least absolute values method.

Merits and demerits of M-estimates

An experimenter can not distinguish the conditions of measurements leading to the normal or the Laplace distributions of errors.

Merits

•– both normal and Laplace distributions are limiting laws (statistics theory),

•– conditions of their formation are close (experiment),

•– **Robust** Huber's M-estimates have the MLM substation (theory),

•– data analysis approach may be realized through changing hypothesis on δ, intensity of "gross errors" (practice of data handling).

The main **demerit** consists of rather complicated calculations.

Non-linear regression model of protolytic equilibria of m-Aminobenzoic acid (HQ) studied by the spectrophotometric method

Chemical model

$$\mathbf{HQ + H^+ \overset{K_H}{=} H_2Q^+},$$

$$\mathrm{HQ \overset{K_a}{=} Q^- + H^+}$$

Data

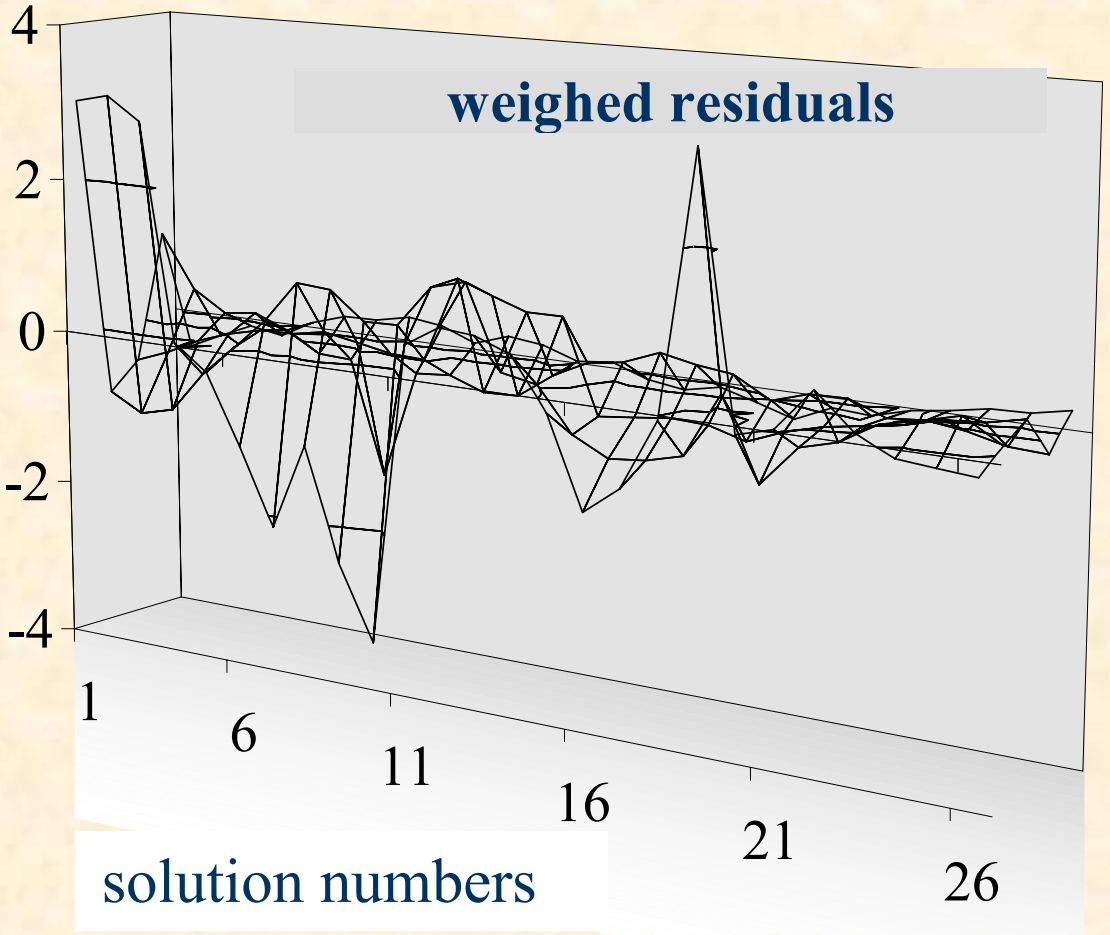
Absorbances of 26 solutions ($1.7 < \mathrm{pH} < 6.2$) at 5 wavelengths (200-300 nm).

The ordinary weighed non-linear least squares method:

$\mathrm{p}K_a = 4.376 \pm 0.004$

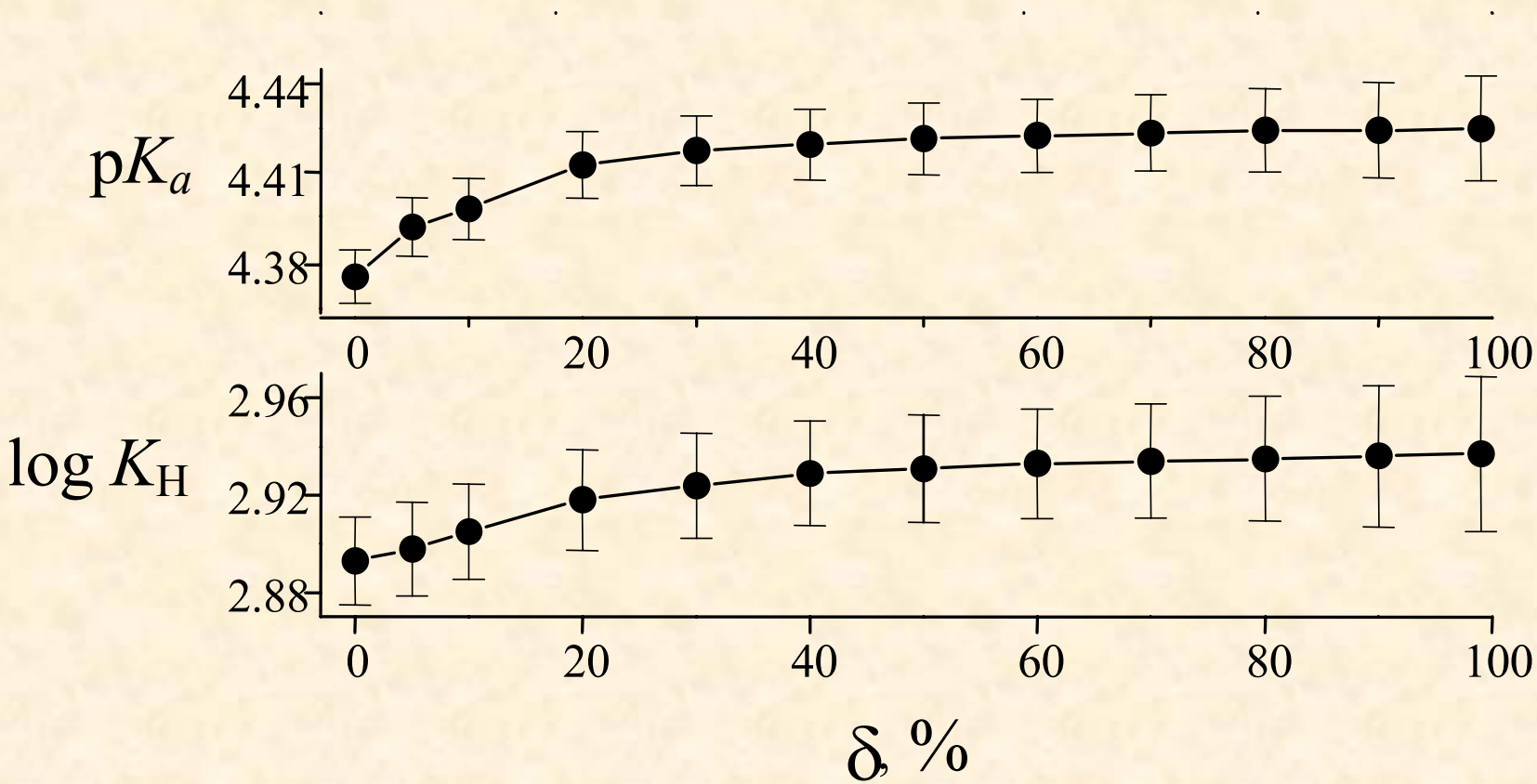
$\log K_H = 2.893 \pm 0.007$

The weighed residuals for $\delta = 0$ (OLS) point to the possible presence of outliers



M-estimates of fitting parameters

δ, %	Parameters	
	log K_H	pK_a
0	**2.893**	**4.376**
5	**2.898**	**4.393**
10	**2.905**	**4.399**
20	2.919	4.417
30	2.924	4.419
40	2.929	4.421
50	2.931	4.423
60	2.933	4.424
70	2.934	4.425
80	2.935	4.426
90	2.936	4.426
100	2.937	4.426



At $\delta \geq 20$ %, estimations of parameters are not changed significantly (but errors grow with increasing δ), the following M-estimates were accepted

$\log K_H = 2.93 \pm 0.01$; $pK_a = 4.425 \pm 0.005$.

QSAR

Svante Wold:

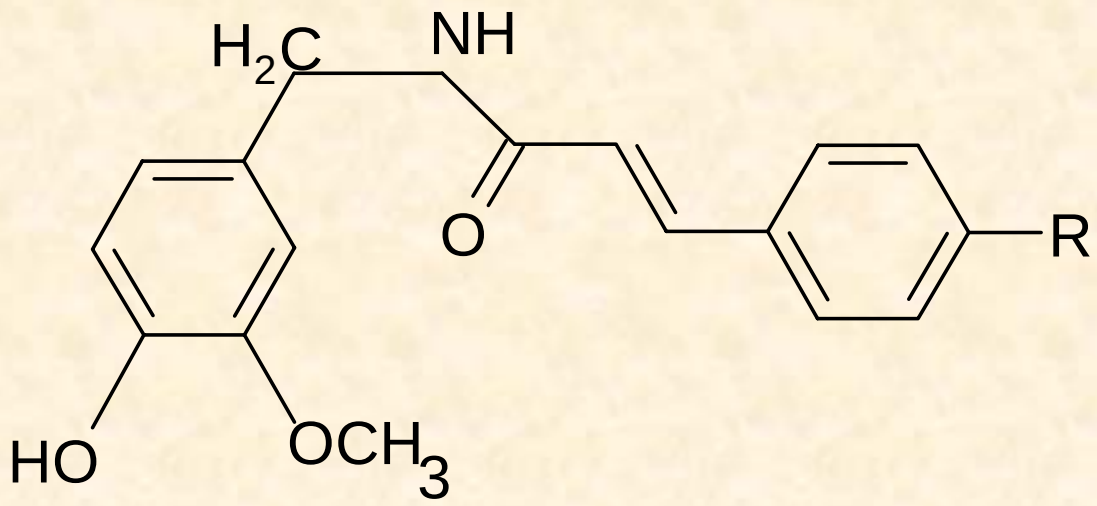
"Populations are relevant to genetics, but usually not to chemistry; we are not interested in the average boiling point of the "population" of organic esters, but rather in the boiling point of each individual ester as a function of its structure.

We are not willing to think of chemical properties, such as melting points, viscosities, or enzymatic reaction rates, as drawn from multivariate distributions".

Suppositions

- handling data from small designed sets of chemical objects (QSAR modeling) has specific features (objects do not belong to the same population and are not exchangeable),
- in such tasks the robust estimation can be simplified: it is possible to compare only the OLS and L_1 estimations,
- cross-validation (CV) can be applied as heuristic but useful tool.

Analgetic Activity of the Capsaicin ANALOGUES



R = H, Cl, NO_2, CN, C_6H_5, $N(CH_3)_2$, I, NHCHO

EC_{50} **(μM) is the effective concentration of the capsaicin leading to decreasing Ca^{2+} influx on 50%**

	r	q
OLS	0.726	0.554
L_1	0.725	0.715

OLS: $$-\log EC_{50} = 1.137 - 0.068 \cdot MR$$

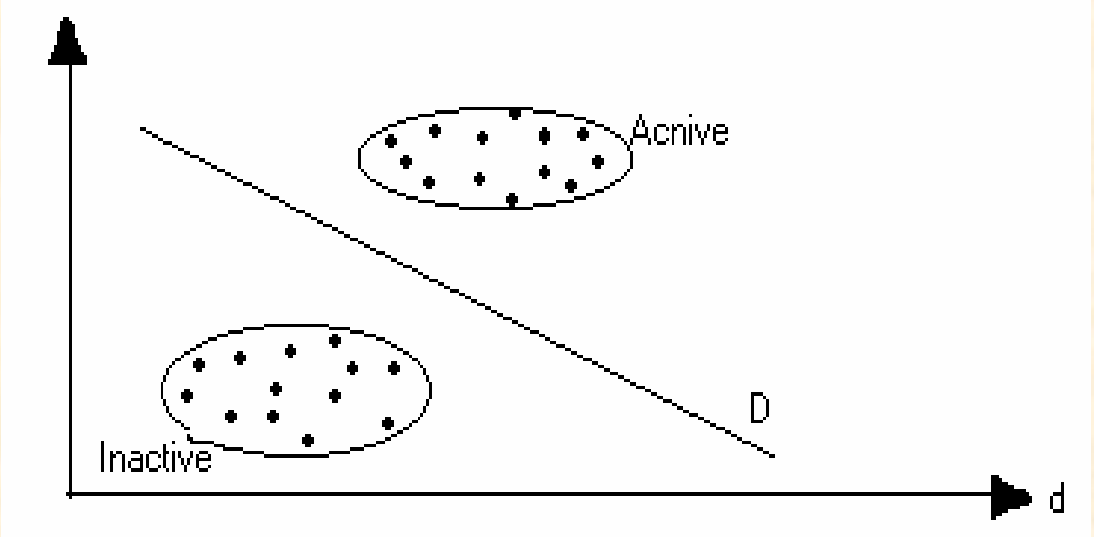
L_1: $$-\log EC_{50} = 1.142 - 0.070 \cdot MR$$

Using the hydrofobic substituent constants (π)

OLS: $-\log EC_{50} = 0.770 - 0.815 \cdot \pi$, **r = 0.943, q = 0.877**

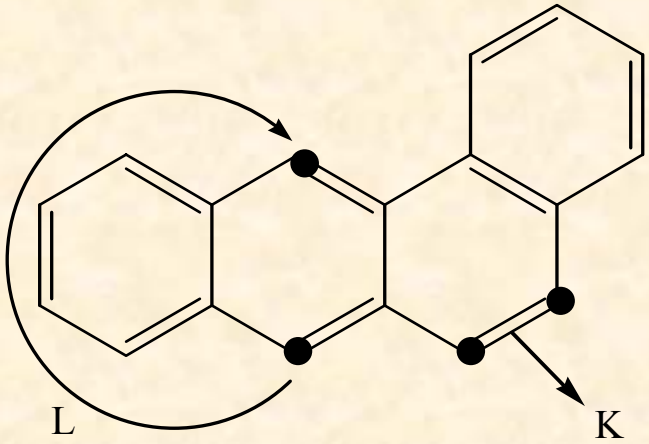
L_1: $-\log EC_{50} = 0.767 - 0.708 \cdot \pi$, r = 0.940, q = 0.846

TWO EXAMPLES OF CROSS-VALIDATION IN THE QSAR DISCRIMINANT ANALYSIS



$$D(d_1, d_2, \ldots) = x_0 + \sum_i x_i d_i + \sum_{i,j} x_{ij} d_i d_j + \ldots$$

Carcinogenic activity of the condenced hydrocarbons



Interaction of cells with K-region leads to **cancer** while the interaction with L-region **deactivates** hydrocarbons

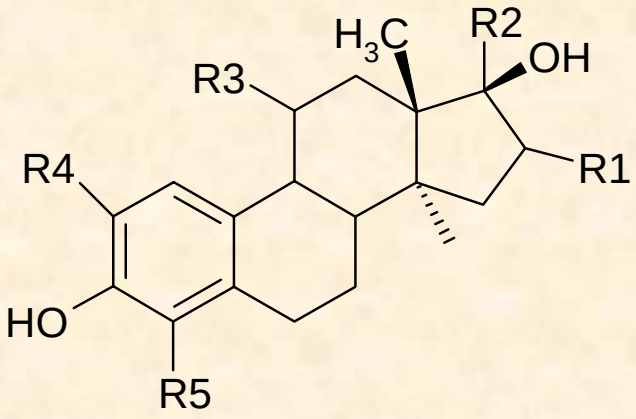
The activity of K– and L– regions can be estimated with using quantum chemistry indices (electrophylic superdelocalizability, S_e).

The discriminant function:

$$-24.1+17.7\,S_e(\mathrm{K})-4.2\,S_e(\mathrm{L})>0$$

All training set: efficiency 100%
LOO-CV: 80%

Relative binding affinity of the estradioles



estradiol

$$1.032 - 1.49\varepsilon_{HOMO} - 0.089\alpha + \\ +1.41\log P - 0.17(\log P)^2 > 0$$

All training set (150 substances): efficiency 88%
LOO-CV: 75%

Regularization of ill posed problems: where does its power end? What tools can be used else?

$$M + \overline{Q} \overset{K}{=} \overline{MQ}$$

$$f([M]) = \int_0^\infty \theta^{\text{local}}([M], K) \cdot p(K)\, dK$$

$$\theta^{\text{local}}([M], K) = \frac{K \cdot [M]}{1 + K \cdot [M]}$$

Regularization technique

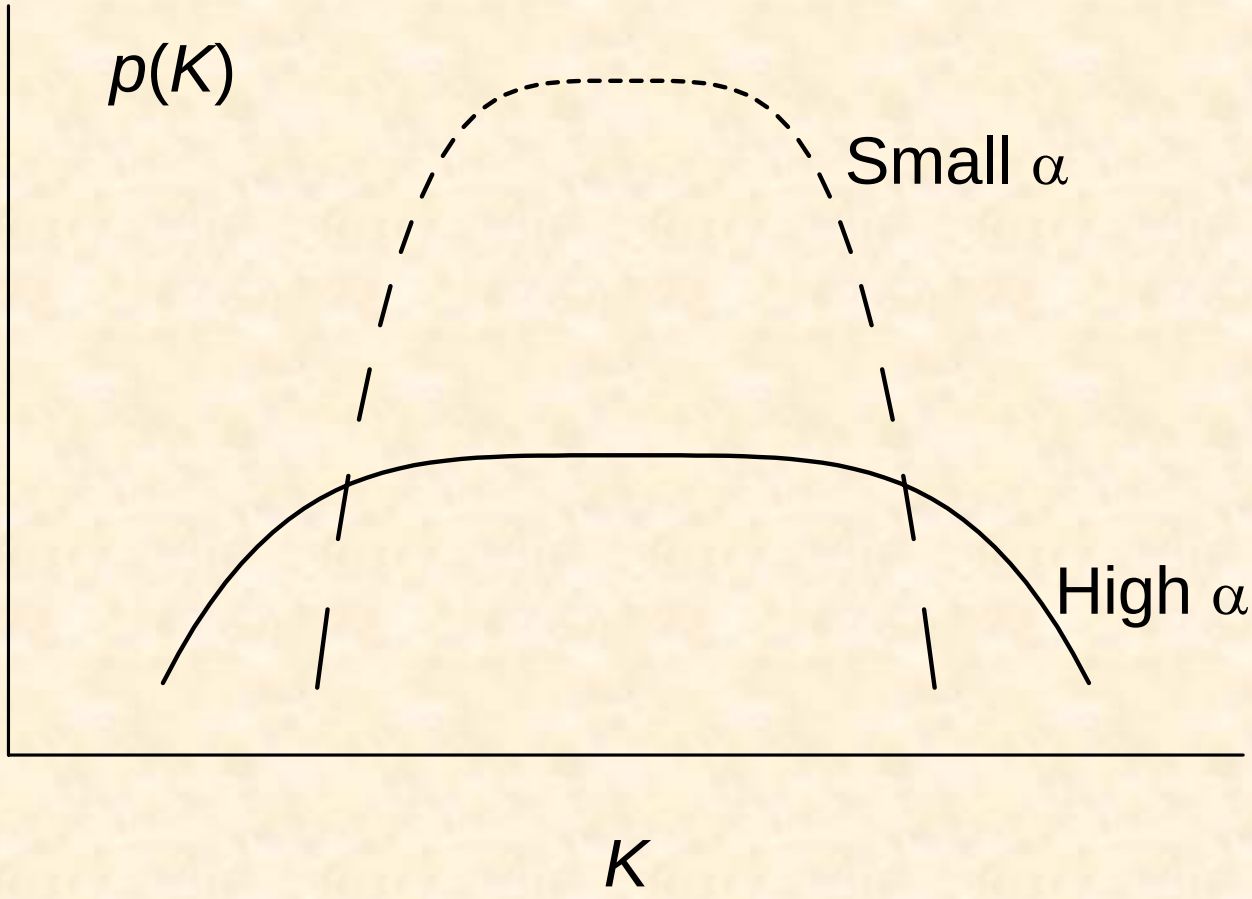
$$U_{\alpha} = U + \alpha \cdot \Omega(p)$$

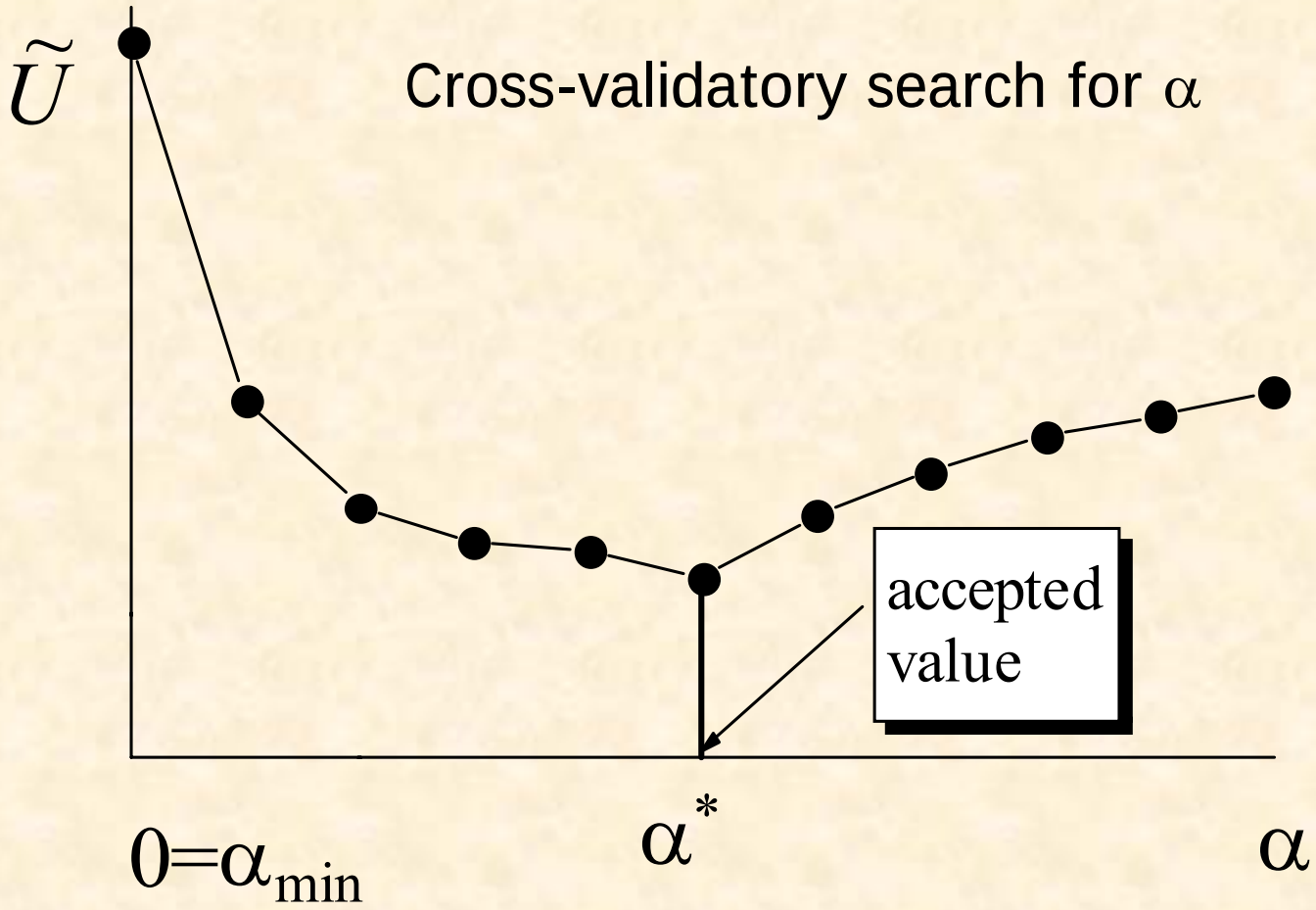
$$U = \sum_{k=1}^{N} \left(\tilde{f}_k - f_k \right)^2$$

$$\Omega(p) = \int_0^{\infty} \left\{ p^2(K) + n \cdot \left[\frac{d\, p(K)}{d\, K} \right]^2 \right\} d\, K$$

$$\Omega(p) = \| p(K) \| = \int_0^{\infty} p^2(K)\, d\, K$$

Influence of α on the shape of density functions




$\tilde{U}$
Cross-validatory search for α
accepted value
$0=\alpha_{min}$
α^*
α

Maxent approach

Experimental data are considered as restrictions imposed on the searched density function $p(K)$

$$\tilde{f}_k = \sum_{j=1}^{J} p(K_j) \frac{K_j \cdot [M]_k}{1 + K_j \cdot [M]_k} \qquad k = 1, 2, \ldots, N$$

Restrictions in the form of inequalities

$$\sum_{j=1}^{J} p(K_j) \times \frac{K_j \cdot [M]_k}{1 + K_j \cdot [M]_k} \geq \tilde{f}_k - \Delta, \; \boldsymbol{k = 1, 2, \ldots, N,}$$

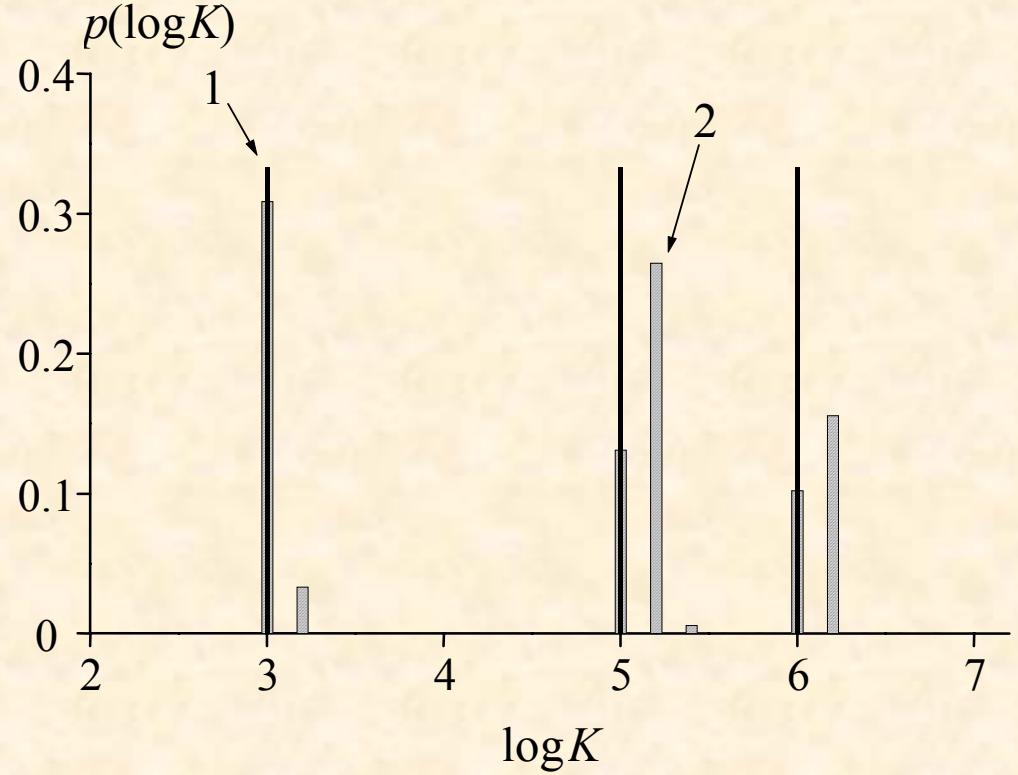
$$\sum_{j=1}^{J} p(K_j) \times \frac{K_j \cdot [M]_k}{1 + K_j \cdot [M]_k} \leq \tilde{f}_k - \Delta, \; \boldsymbol{k = 1, 2, \ldots, N,}$$

$$\sum_{j=1}^{J} p(K_j) = 1, \;\; p(K_j) \geq 0,$$

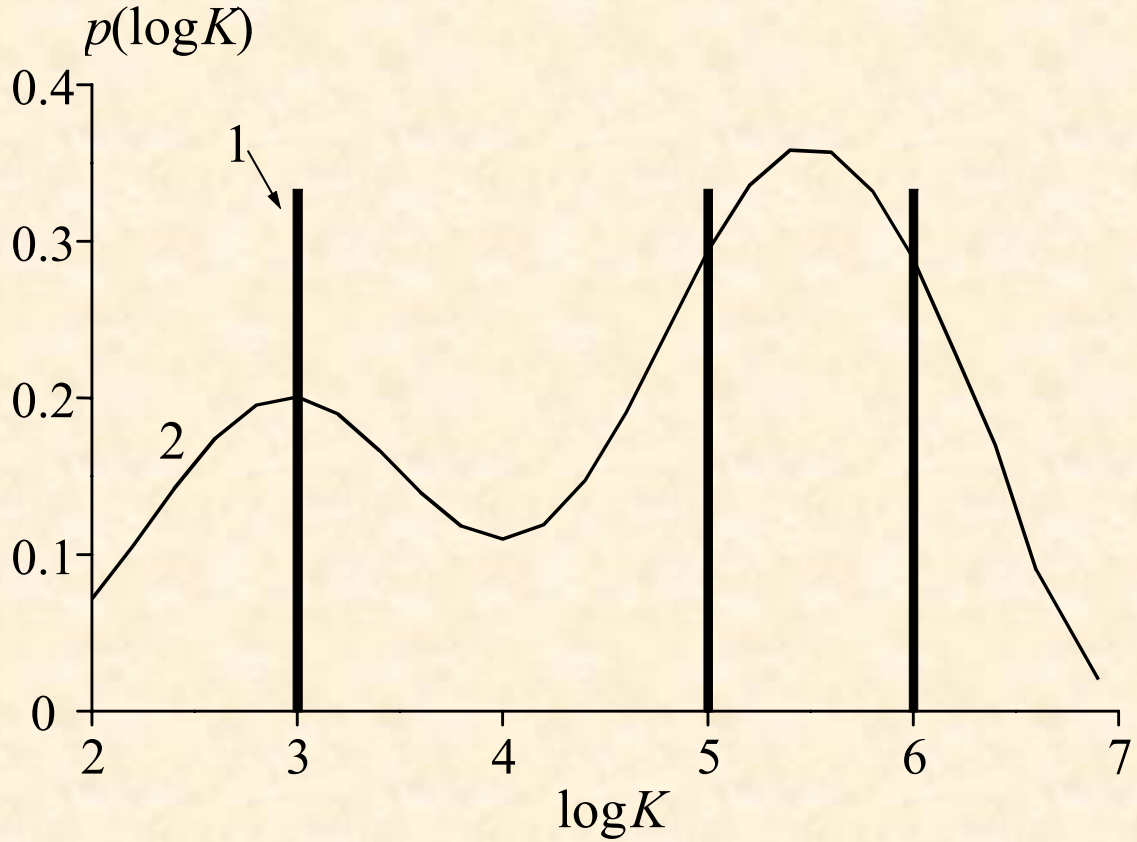
Numerical example

$$p(\log K = 2) = p(\log K = 5) = p(\log K = 6) = 1/3$$

The density function $p(\log K)$



The density function $p(\log K)$ calculated by the conventional α-regularization algorithm from exact data



Acknowledgements

- Dr. Lyudmila Sleta (QSAR)
- Dr. Sergey Myerniy (robust regression, M-estimates, Maxent approach)
- Dr Dmitriy Konyaev (robust regression, software)
- Prof. Alexander Korobov (discussions and advices)