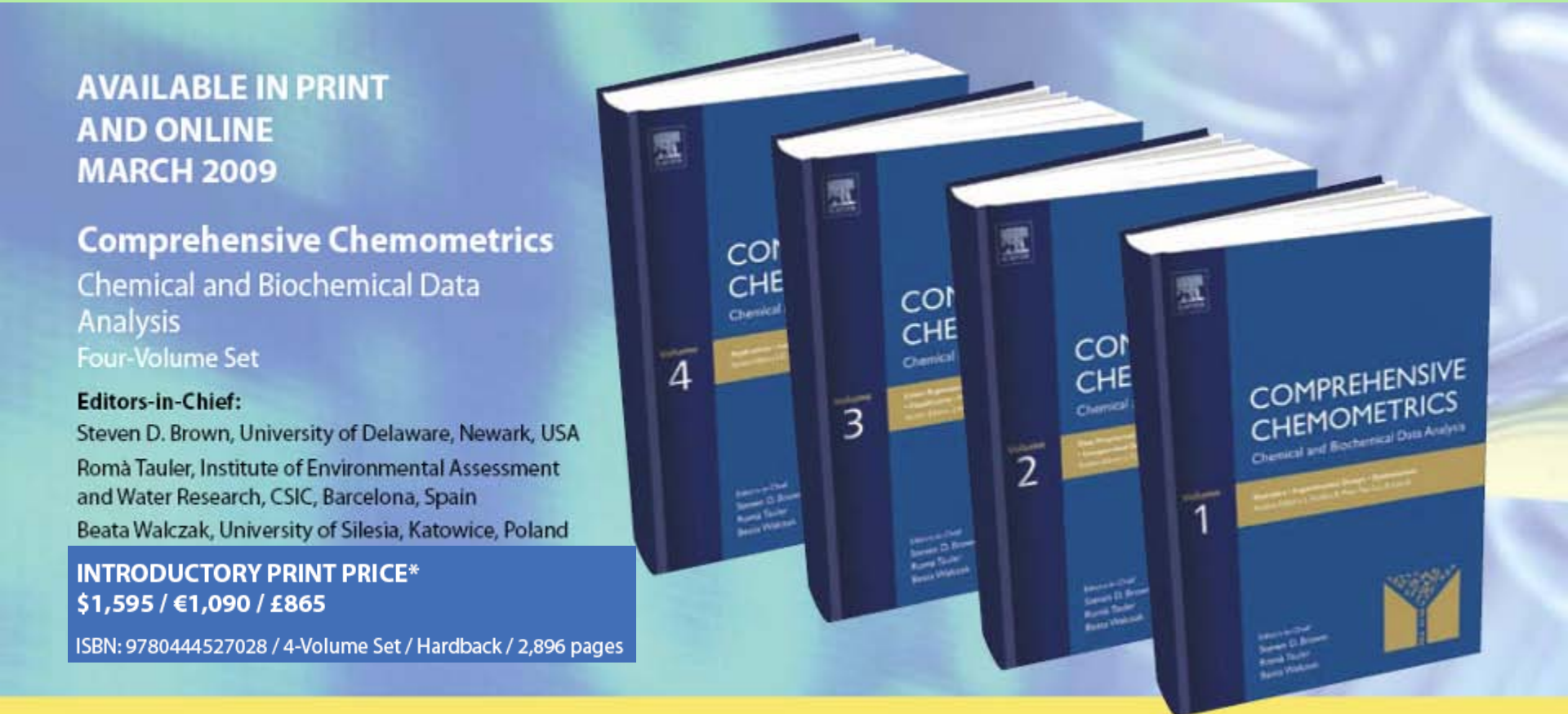
Хемометрика

Алексей Померанцев,
Оксана Родионова

Институт химической физики РАН

Российское хемометрическое общество



Содержание

Хемометрика: два определения

Дедуктивное

Хемометрика - это научная дисциплина, находящаяся на стыке химии и математики, предметом которой являются математические методы исследования химических данных

сайт Российского хемометрического общества

Индуктивное

Хемометрика – это то, что делают хемометрики.

сайт Международного хемометрического общества

Хемометрики – это такие люди, которые все время пьют пиво и воруют идеи у математиков

Svante Wold

А если серьезно ....

- Хемометрика имеет дело с **данными** (зачастую с очень большими), поэтому хемометрика - это подраздел информатики (Data mining)
- Данные, которые исследует хемометрика по большей части происходят из **химии**, поэтому хемометрика - это подраздел химии (Analytical chemistry)
- Методы, которые использует хемометрика ориентированы на **формальное** моделирование (Soft modeling)

Два «не» и три «да»

1. Хемометрика ≠ химическая метрология
2. Хемометрика ≠ статистика в химии

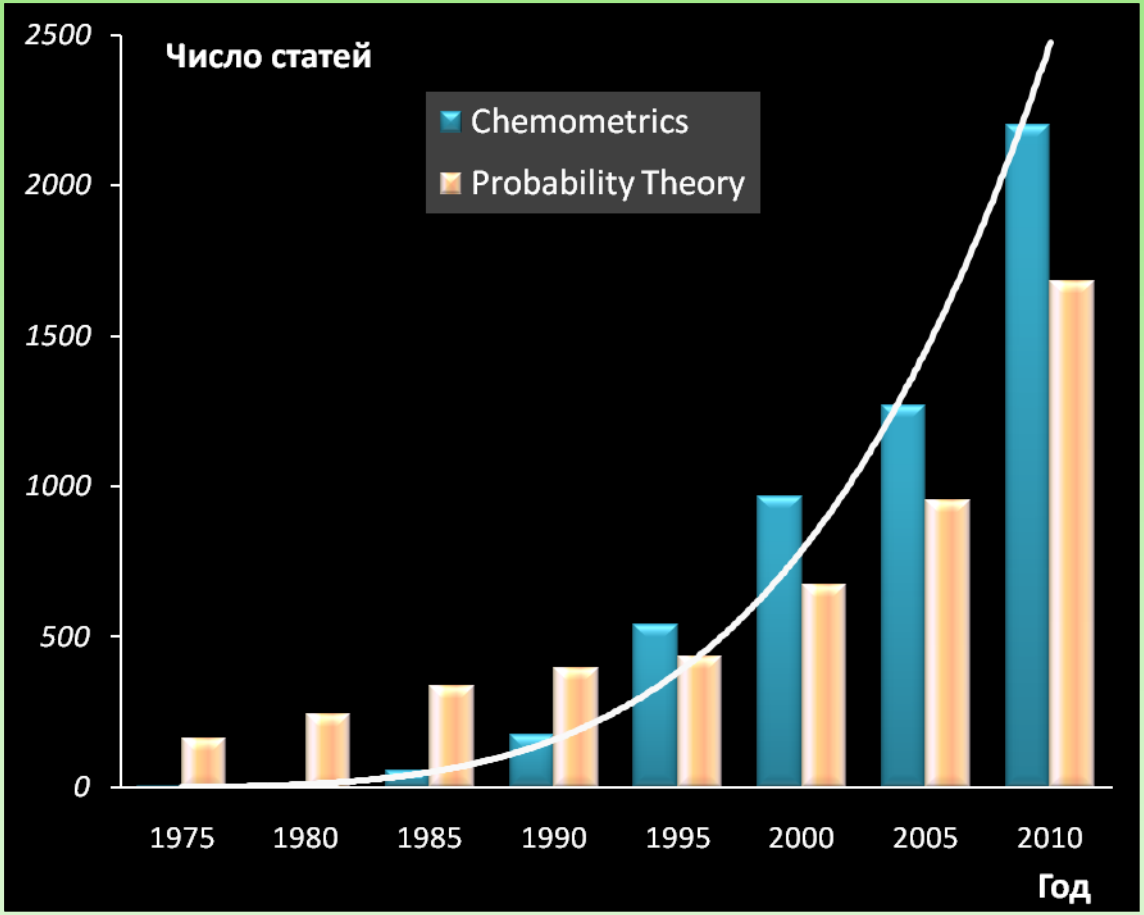
Хемометрика решает следующие задачи в области химии:

(1) как получить химически важную информацию из химических данных,

(2) как организовать и представить эту информацию,

(3) как получить данные, содержащие такую информацию.

Прогресс хемометрики



Число статей с ключевыми словами *chemometrics* и *Probability Theory* по базе Scopus

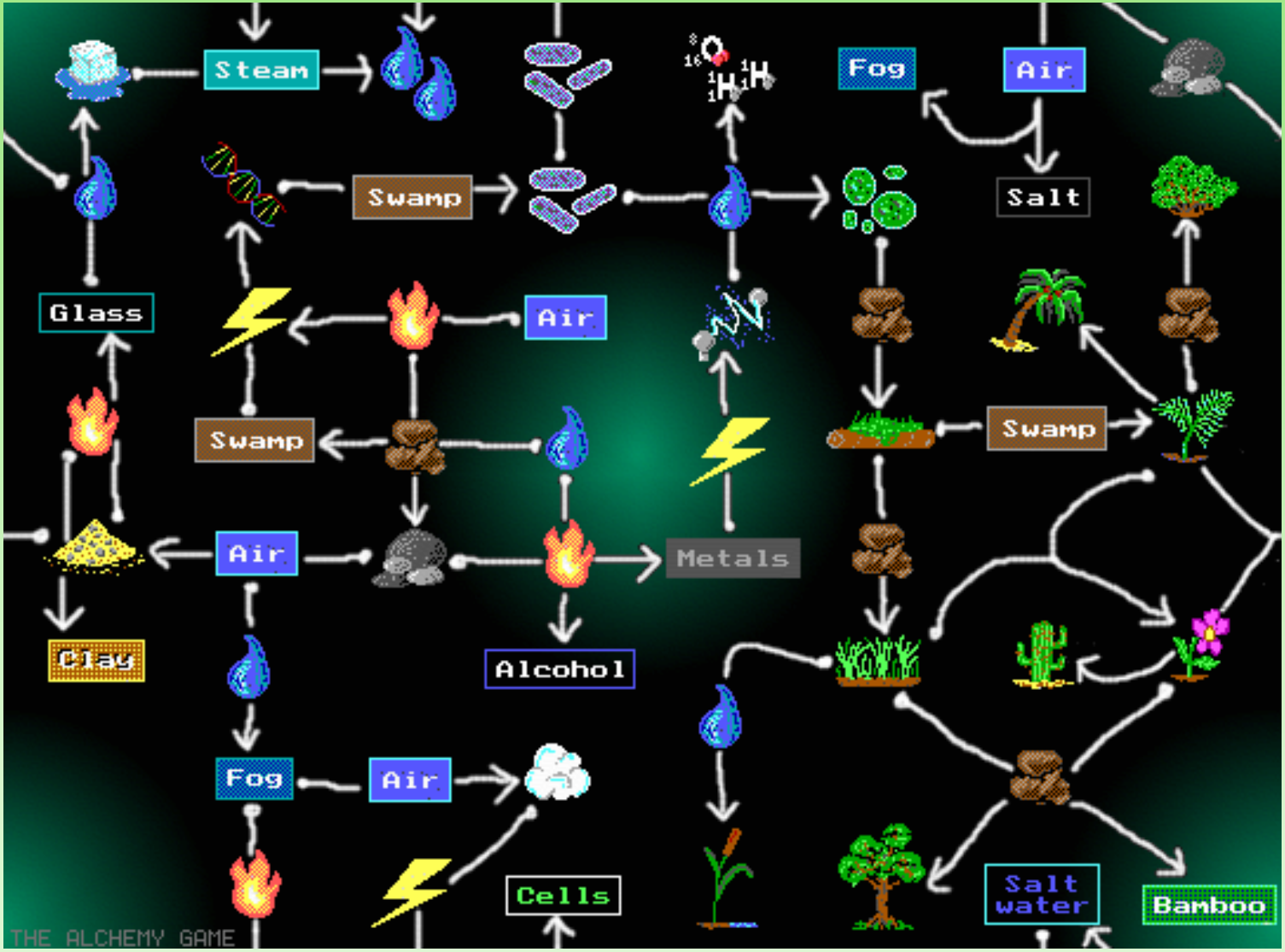
Почему «хемо-» ?

- Хемометрика родилась из задачи анализа химических **спектров**
- Спектроскопия – наилучший метод получения информации по ходу процесса (**on-line**) в режиме реального времени: быстро и без влияния на процесс
- «Хемо» подчеркивает **практическую**, а не статистическую значимость применяемых методов

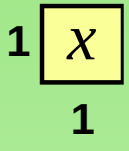
Почему «-метрика» ?

- Хемометрические методы легко и плодотворно **переносятся** в другие области, например, в психологию, биологию, геологию, и т. д.
- Хемометрика активно эксплуатирует **математику** статистику, линейную алгебру
- ‘It is easier to teach a chemist statistics that to teach chemistry to a statistician.’ (Svante Wold)

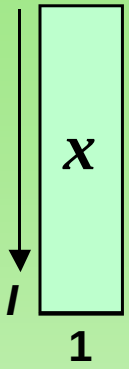
Данные


Steam
Fog
Air
Swamp
Salt
Glass
Air
Swamp
Swamp
Air
Metals
Clay
Alcohol
Fog
Air
Cells
Salt water
Bamboo
THE ALCHEMY GAME

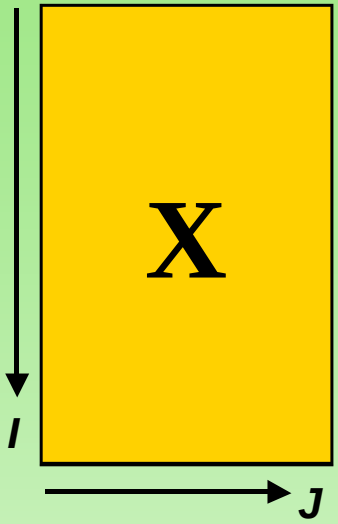
Данные



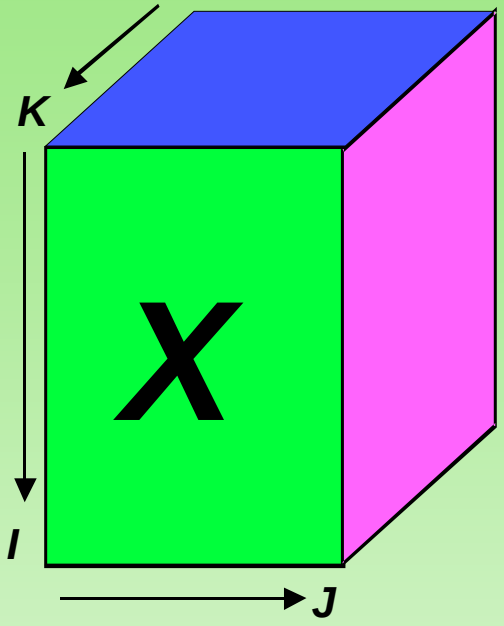
0D



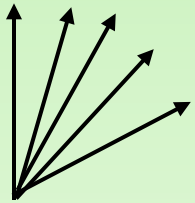
1D



2D



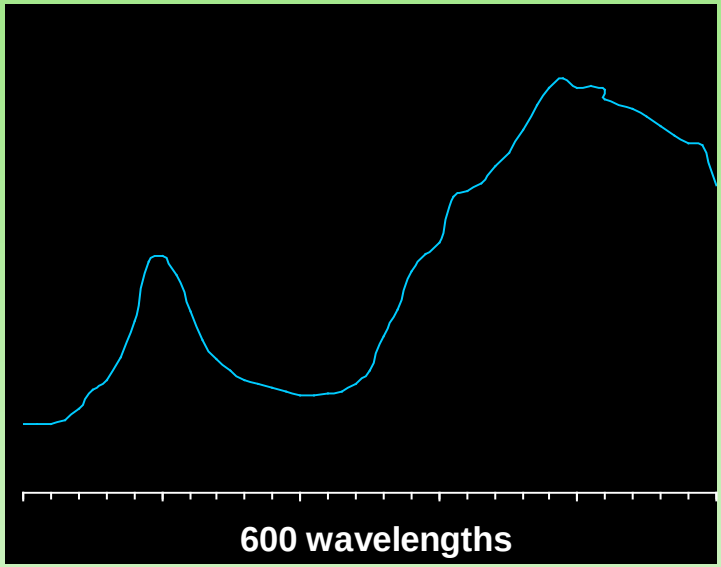
3D



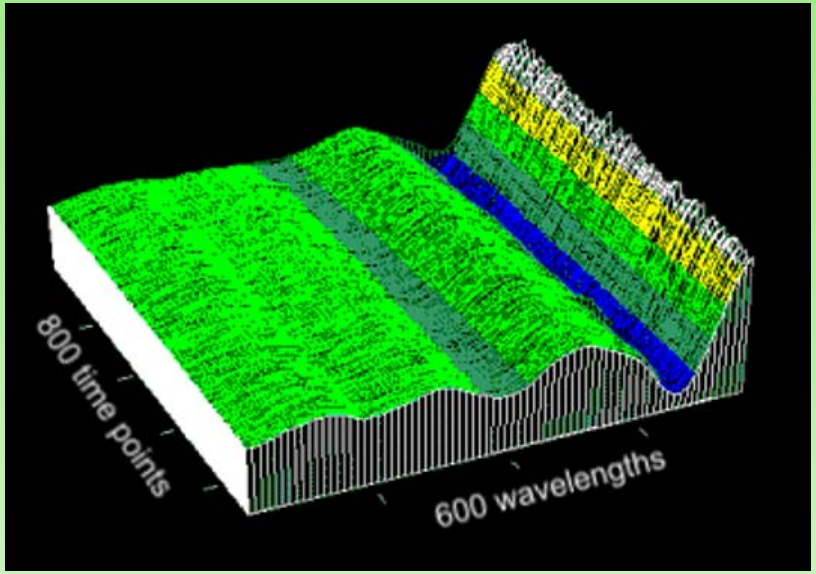
ND

Много переменных и много измерений

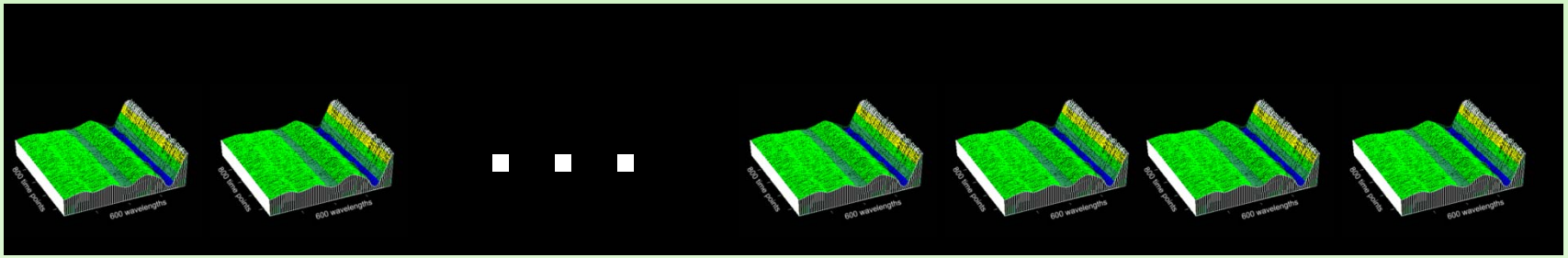
Одно измерение – спектр (600 точек)



Один цикл – 800 спектров (времен)



Один массив данных – 200 образцов (циклов)

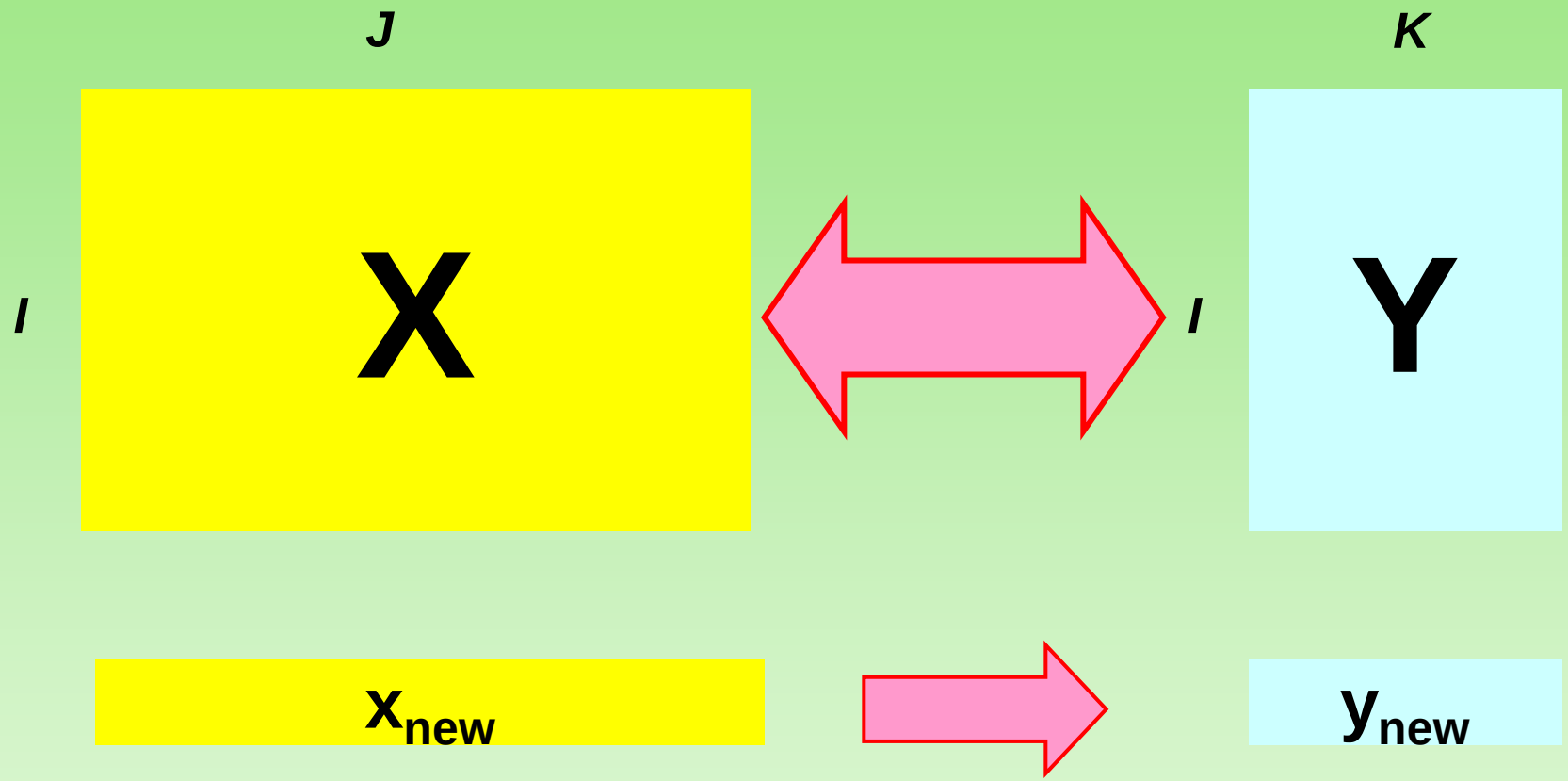


Формальные и содержательные модели

	Содержательные “Hard” models	Формальные “Soft” models
Откуда	физика, химия, ….	из данных
Модель	нелинейная	(квази)линейная
Параметры	имеют смысл	физически бессмысленны
Проблемы	построить модель	интерпретировать данные
Назначение	экстраполяция	интерполяция
Пример	хим. кинетика	ANOVA

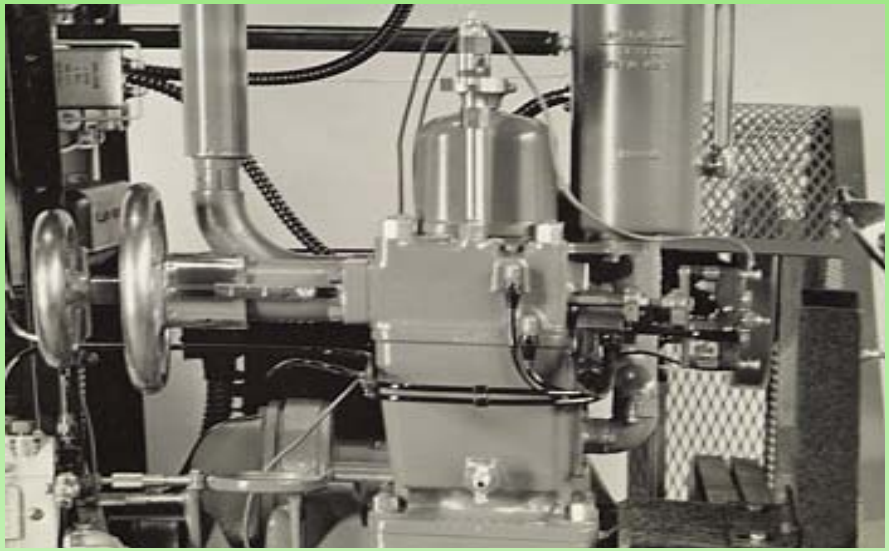
Задачи

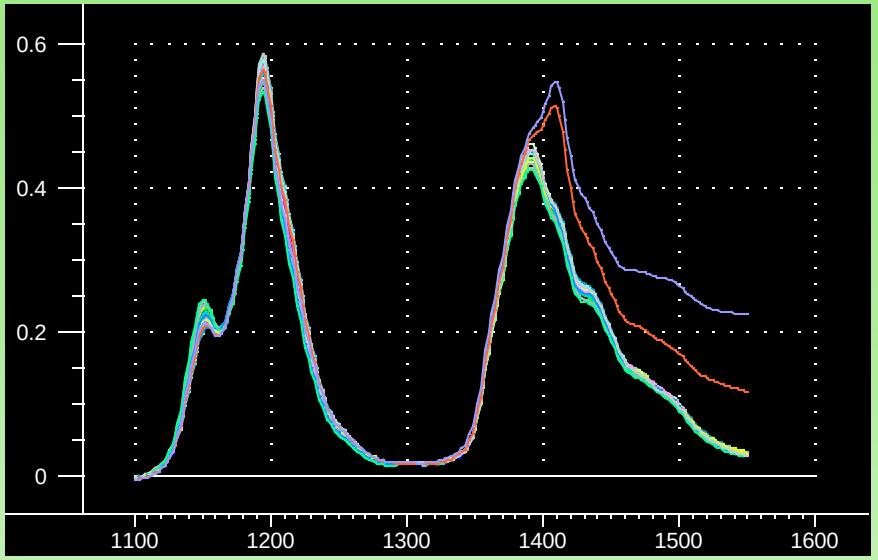
Калибровка

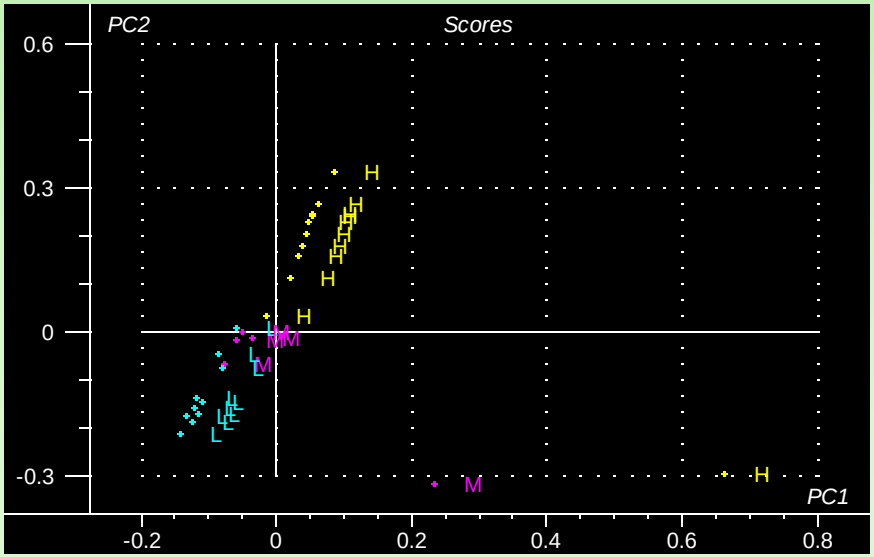


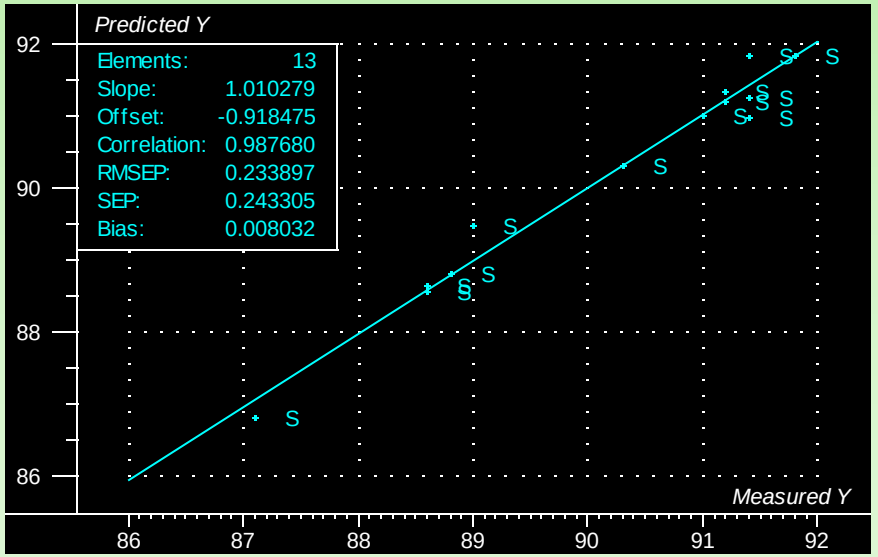
$5<I<10^{+9}$ $1<J<10^{+7}$ $0<K<10^{+3}$

Определение ОЧ бензина

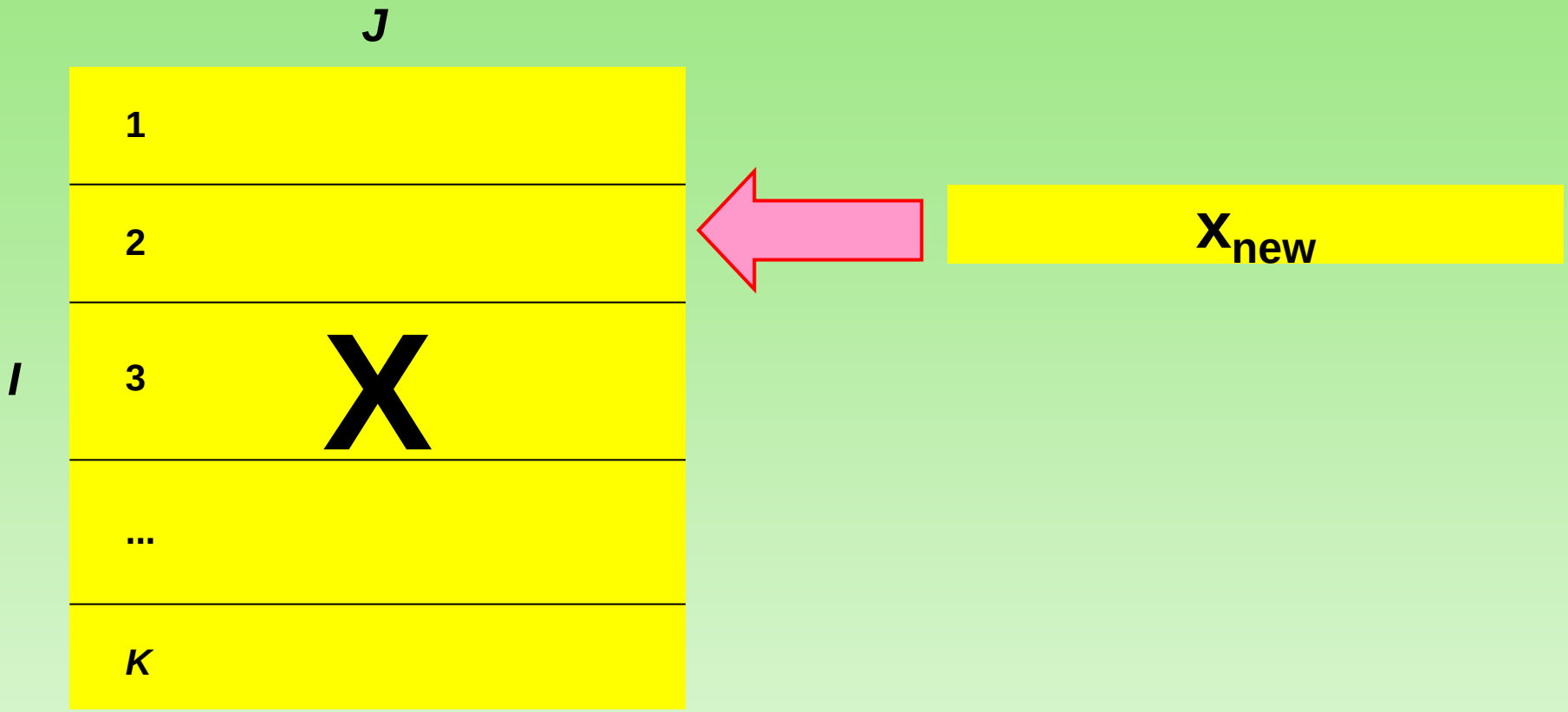








Классификация



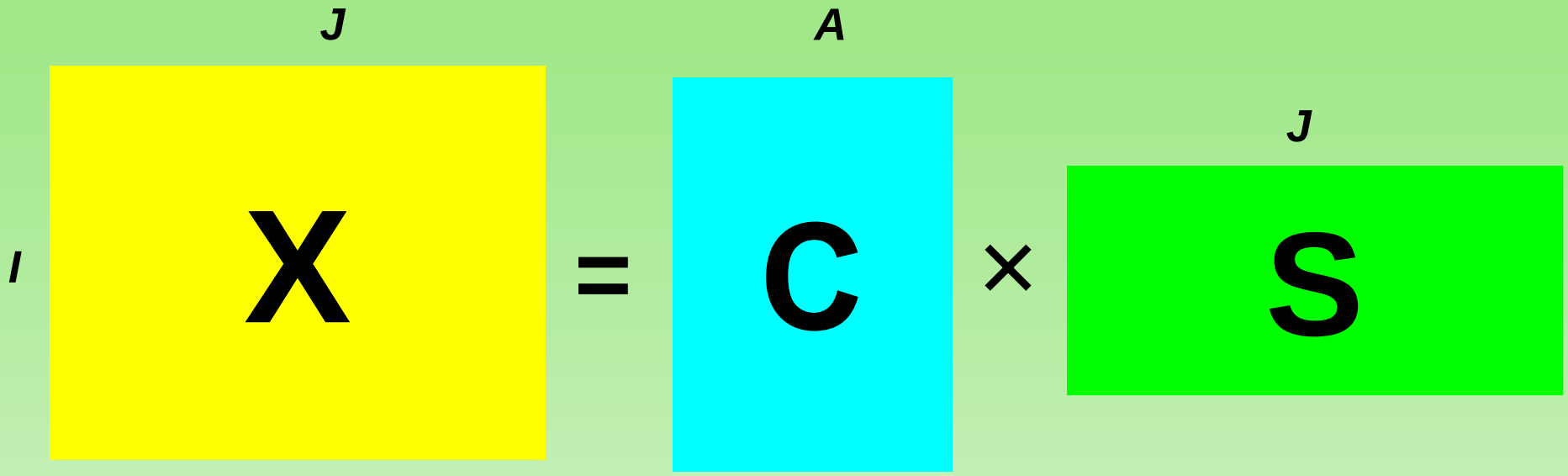
$5<I<10^{+9}$ $1<J<10^{+7}$ $0<K<10^{+3}$

Апрель 2009

Mildronate

Listenon

Разрешение кривых (MCR)



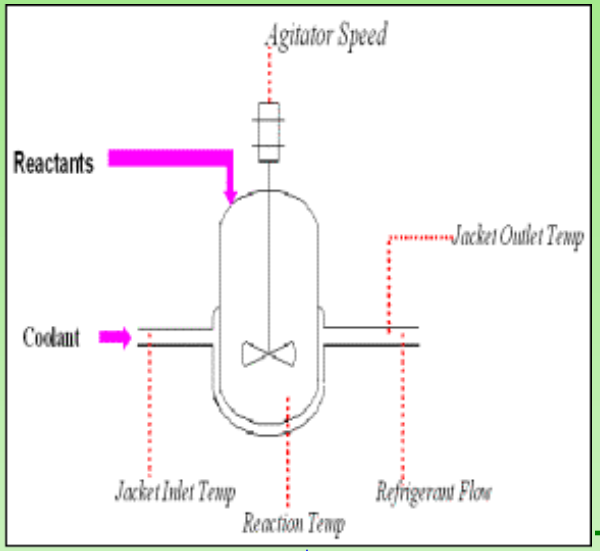
A=?

C≥0 S≥0 + доп. ограничения

$5<I<10^{+9}$ $1<J<10^{+7}$ $0<A<10$

Аналитический контроль процессов

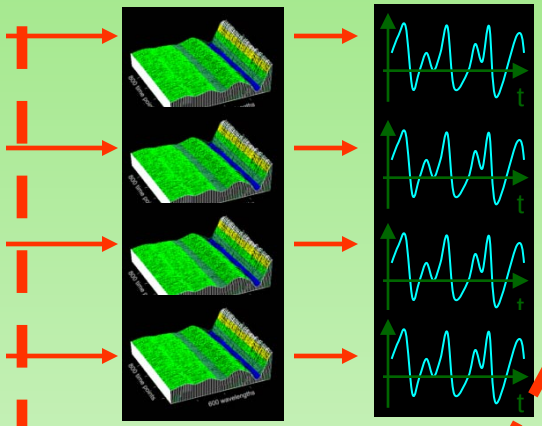
Процесс



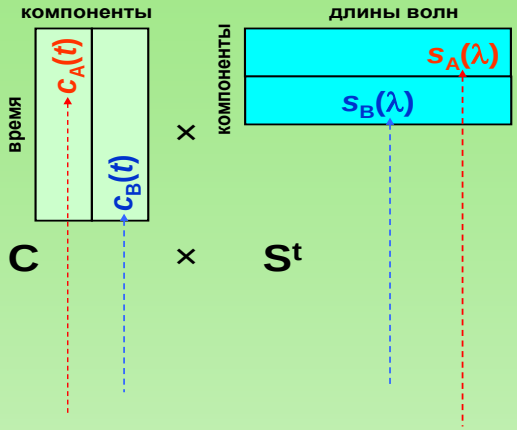
История процесса

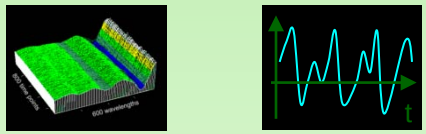
Спектры

Макро



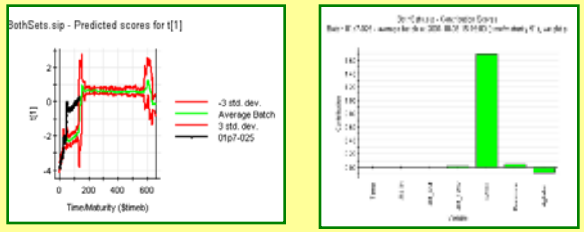
MCR модель





Текущий цикл

MSPC Контроль



Обратная связь

Прочие задачи

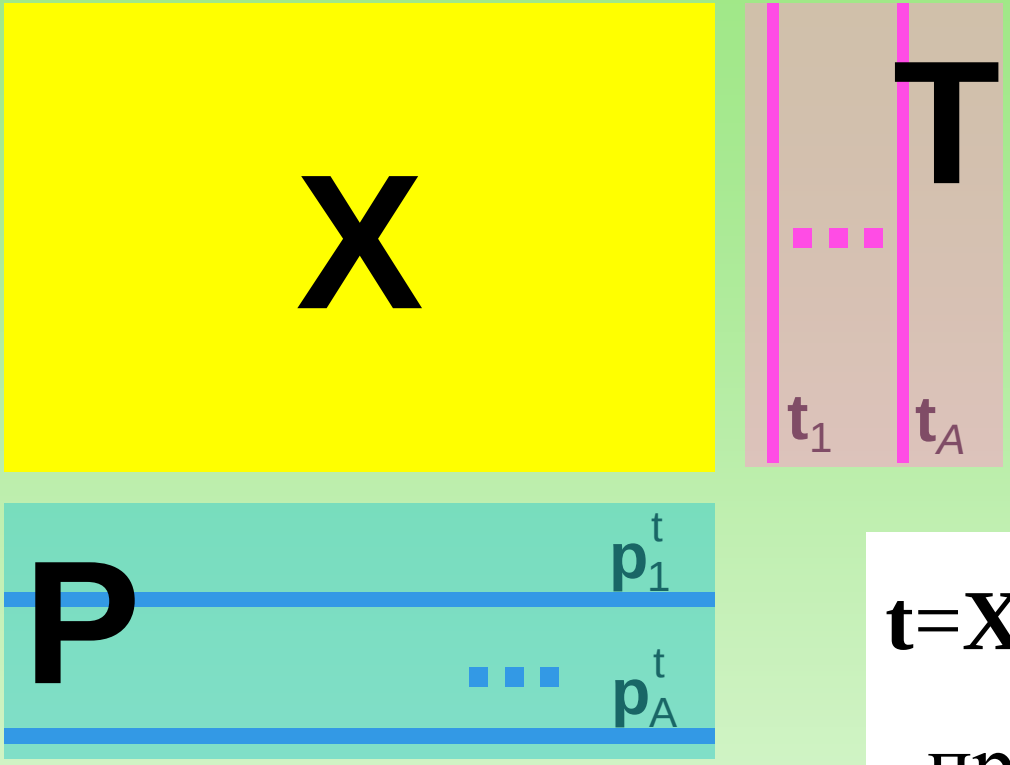
- робастные методы
- подготовка данных (фильтрация, выбросы, ...)
- представительный отбор образцов и переменных
- планирование эксперимента
- пропущенные измерения
- визуализация и картирование
- и т.д.

Методы

Методы калибровки

- Principal Component Regression (PCR)
- **Projection on Latent Structures (PLS)**
- **SIC Regression**
- Ridge Regression (RR)
- Support Vector Machine Regression (SVM-R)
- Projection Pursuit (PP)
- Artificial Neural Network (ANN)

Регрессия на главные компоненты (PCR)

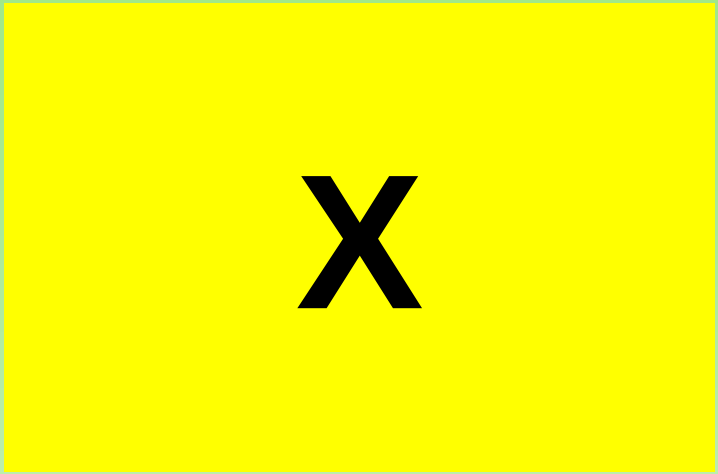


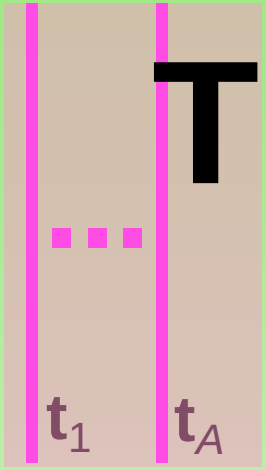
$$\mathbf{t}=\mathbf{X}\mathbf{w} \Leftarrow \max|\mathbf{X}\mathbf{w}|^2$$

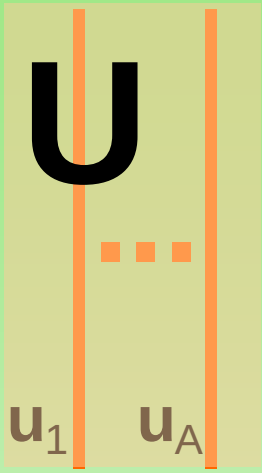
при условии $|\mathbf{w}|=1$

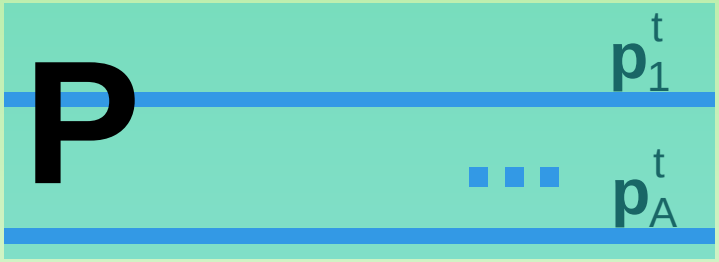
$$\Leftrightarrow \quad \mathbf{X}^t\mathbf{X}\mathbf{w} = \lambda\mathbf{w}$$

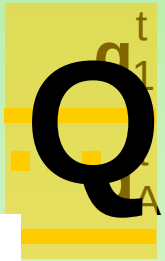
Проекция на латентные структуры (PLS)

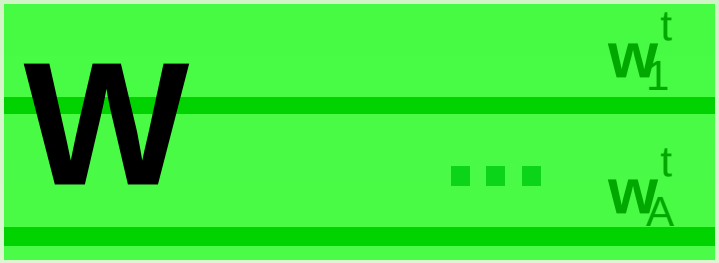










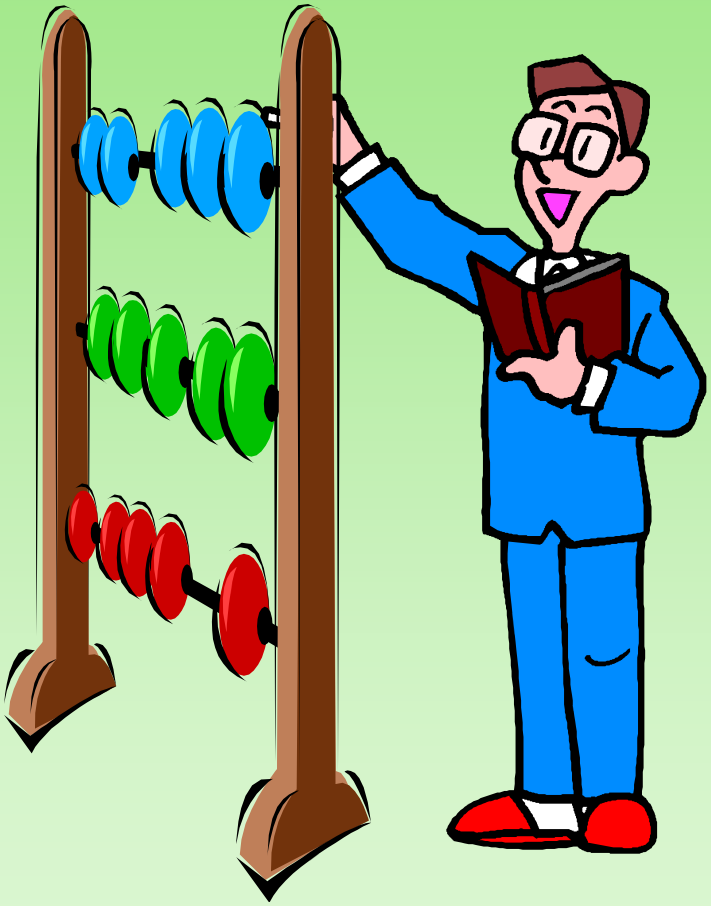


$$\mathbf{t}=\mathbf{Xw} \Leftarrow \max|\mathbf{Y}^t\mathbf{Xw}|^2$$

при условии $|\mathbf{w}|=1$

$$\Leftrightarrow \quad \mathbf{X}^t\mathbf{Y}\mathbf{Y}^t\mathbf{X}\mathbf{w} = \lambda\mathbf{w}$$

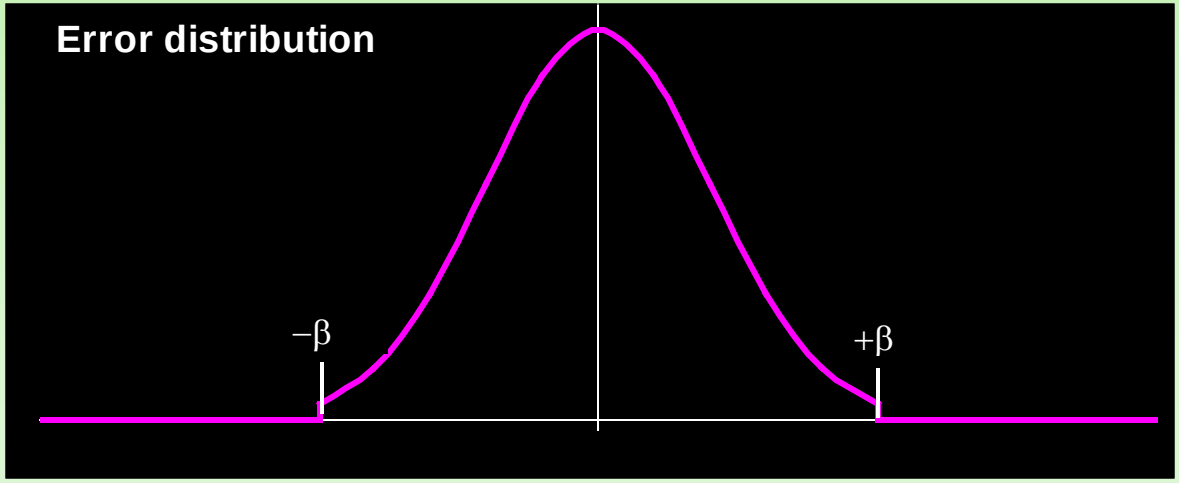
Пример: SIC метод



Простое интервальное оценивание

Все ошибки ограничены!

$$\exists \beta > 0: \mathrm{Prob}\{ |\varepsilon| > \beta \} = 0$$



Оценка параметров - область

Линейная модель $\mathbf{y} = \mathbf{X}\mathbf{a}+\boldsymbol{\varepsilon}$, I образцов, J переменных

Ошибка ограничена, поэтому:

$$y_i - \beta \le \mathbf{x}_i \mathrm{a} \le y_i + \beta,\ i=1,...,I$$

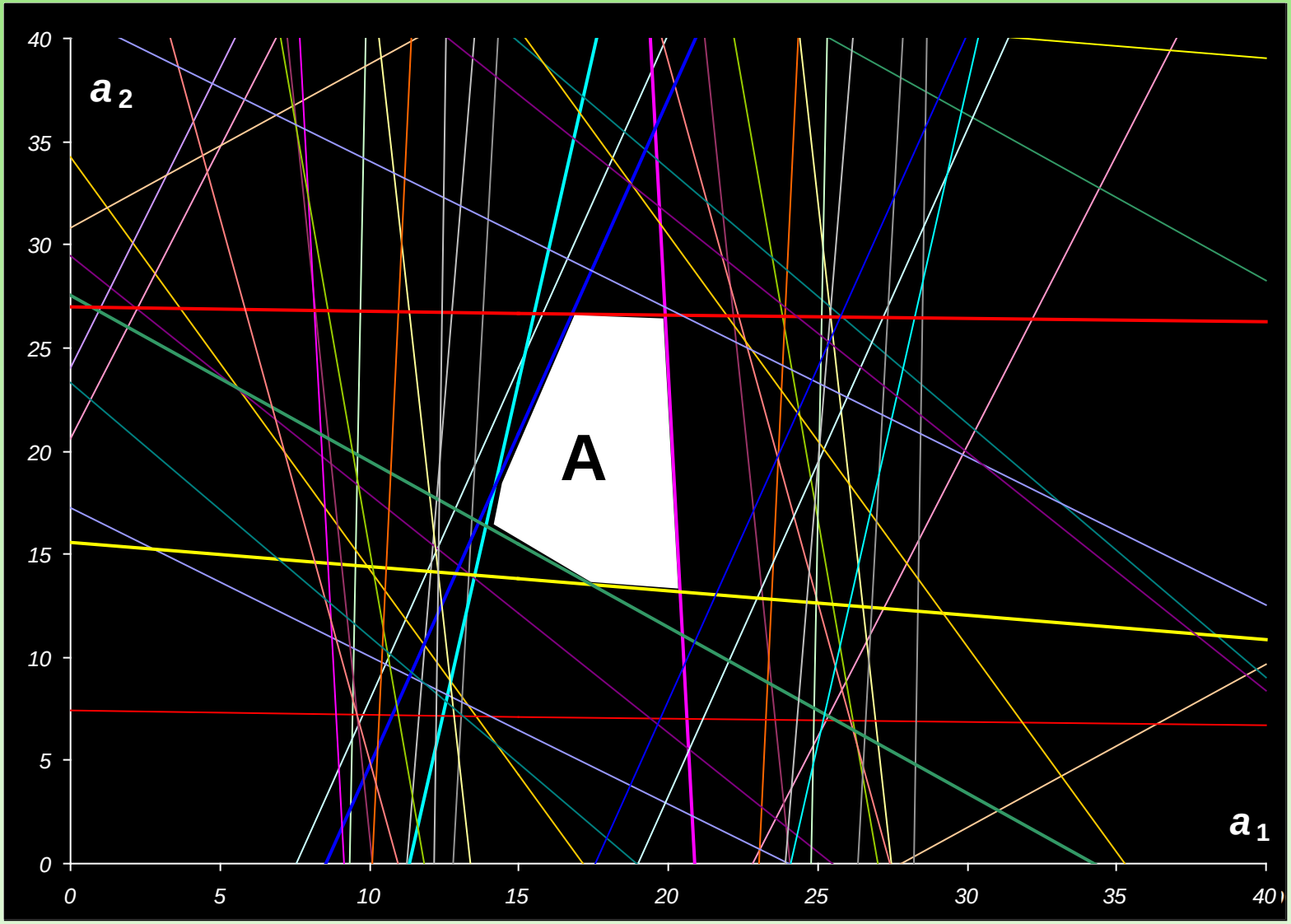
Все $\mathbf{a}$ удовлетворяющие этим неравенствам образуют полосу

$$S(\mathbf{x}_i, y_i) \subset \mathrm{R}^{\mathrm{J}}$$

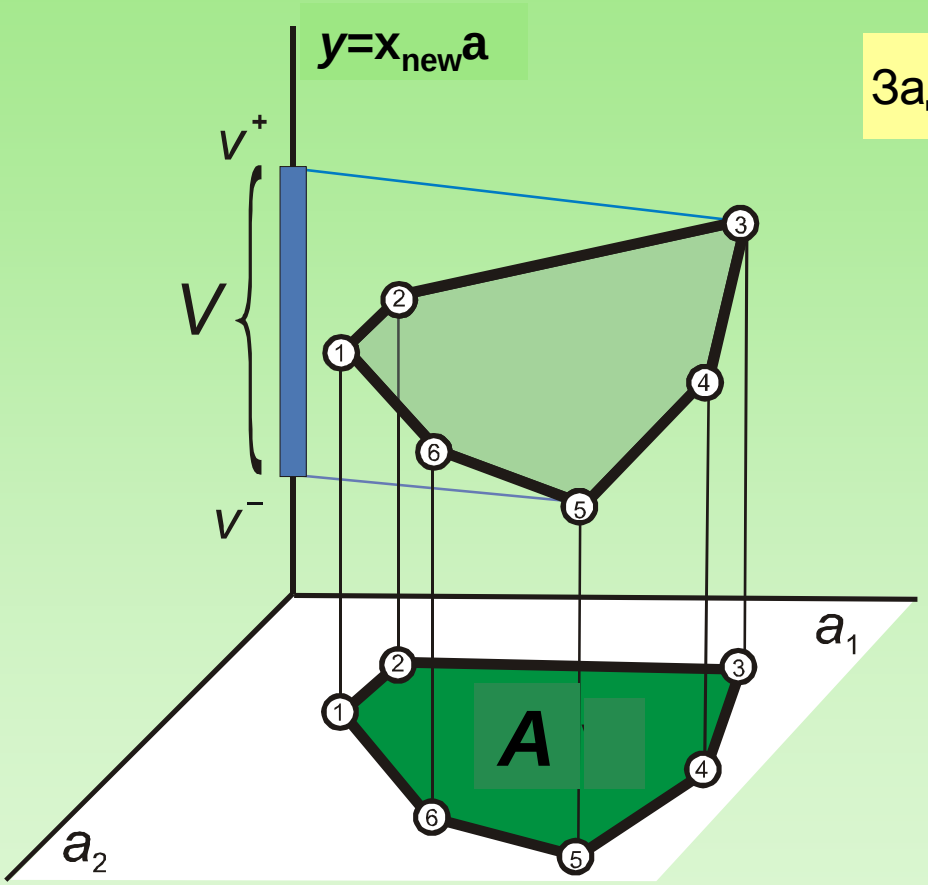
Решение A – это пересечение всех полос

$$A = \bigcap_{i=1}^{I} S(\mathbf{x}_i, y_i)$$

Область допустимых значений *A*



SIC прогноз



Задача линейного программирования

$$v^- = \min_{\mathbf{a}\in A} \mathbf{x}_{new}\mathbf{a}$$

$$v^+ = \max_{\mathbf{a}\in A} \mathbf{x}_{new}\mathbf{a}$$

SIC-остаток и SIC-размах

Эти величины характеризуют взаимоотношения измерения и прогноза

SIC-остаток –

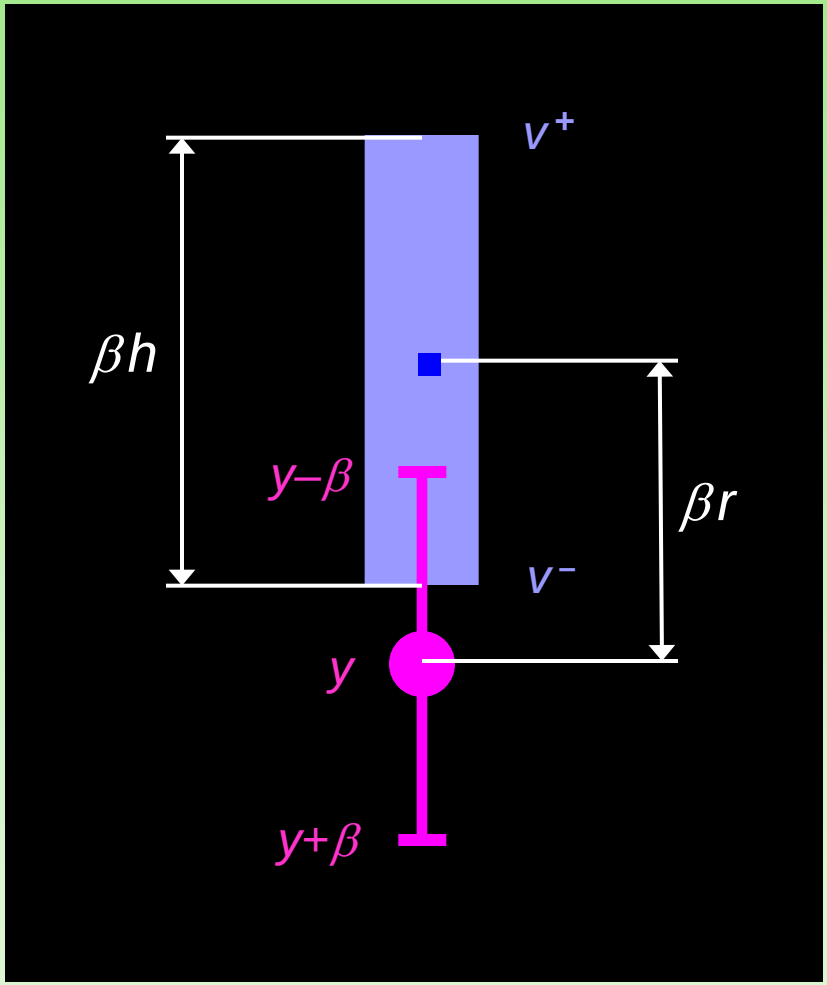
$$r(x,y)=\frac{1}{\beta}\left(y-\frac{v^{+}(x)+v^{-}(x)}{2}\right)$$

характеризует смещение

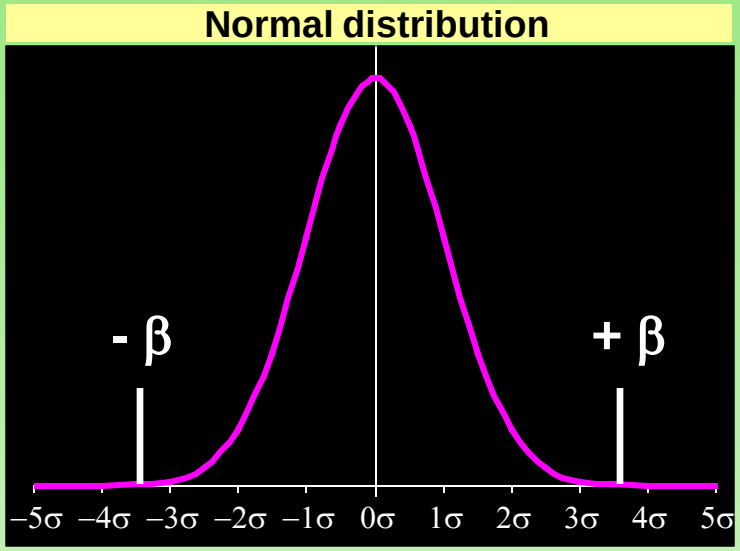
SIC-размах–

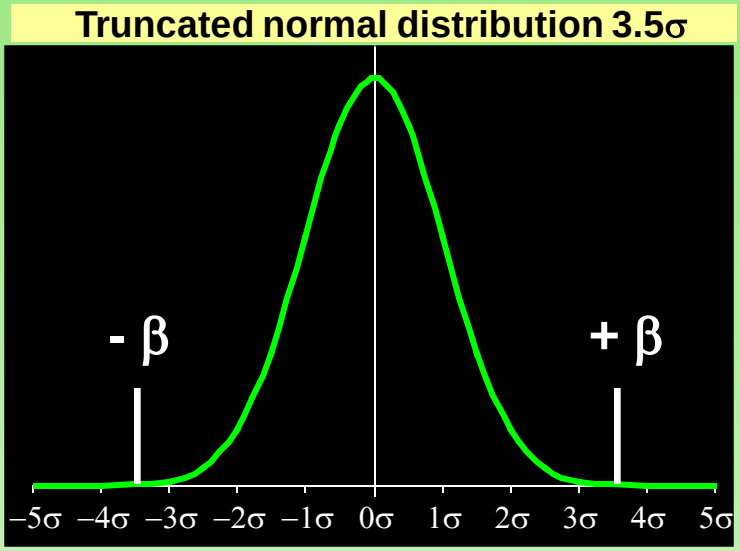
$$h(x)=\frac{1}{\beta}\left(\frac{v^{+}(x)-v^{-}(x)}{2}\right)$$

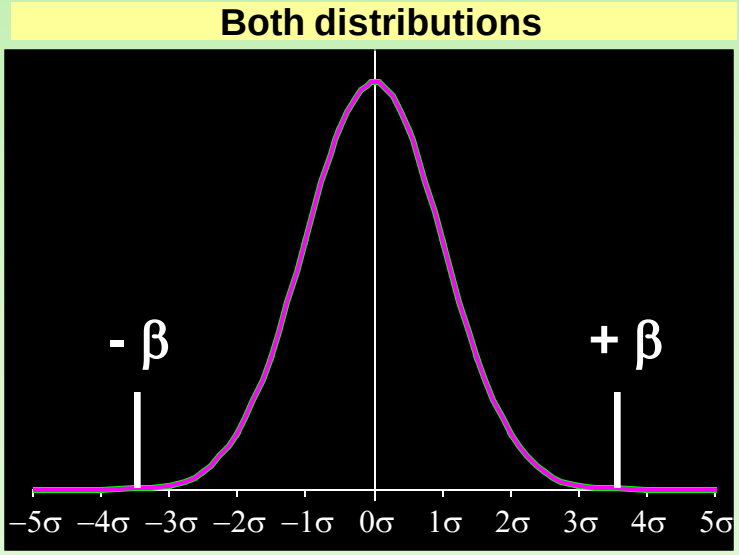
характеризует точность



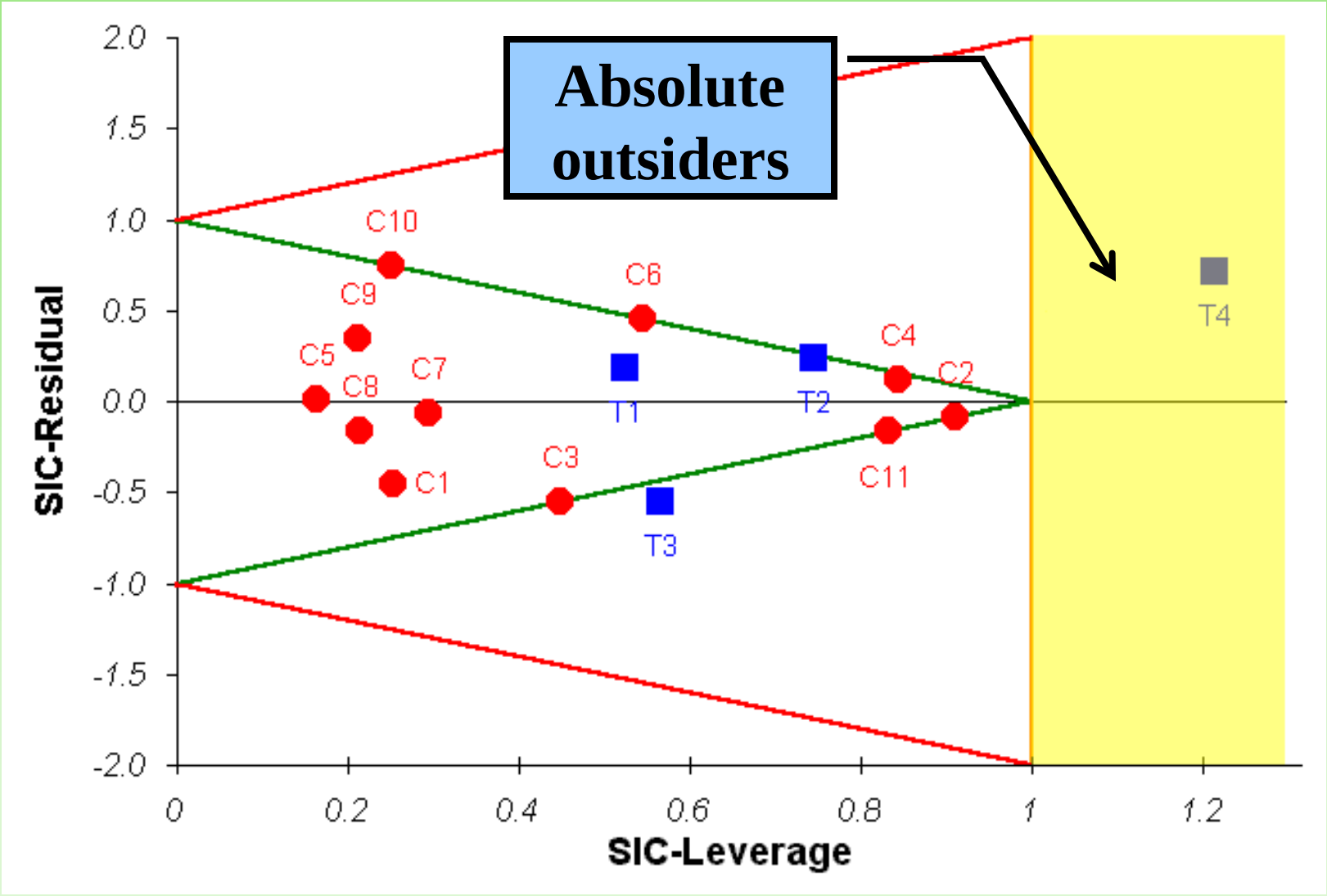
Проблема: как найти β?







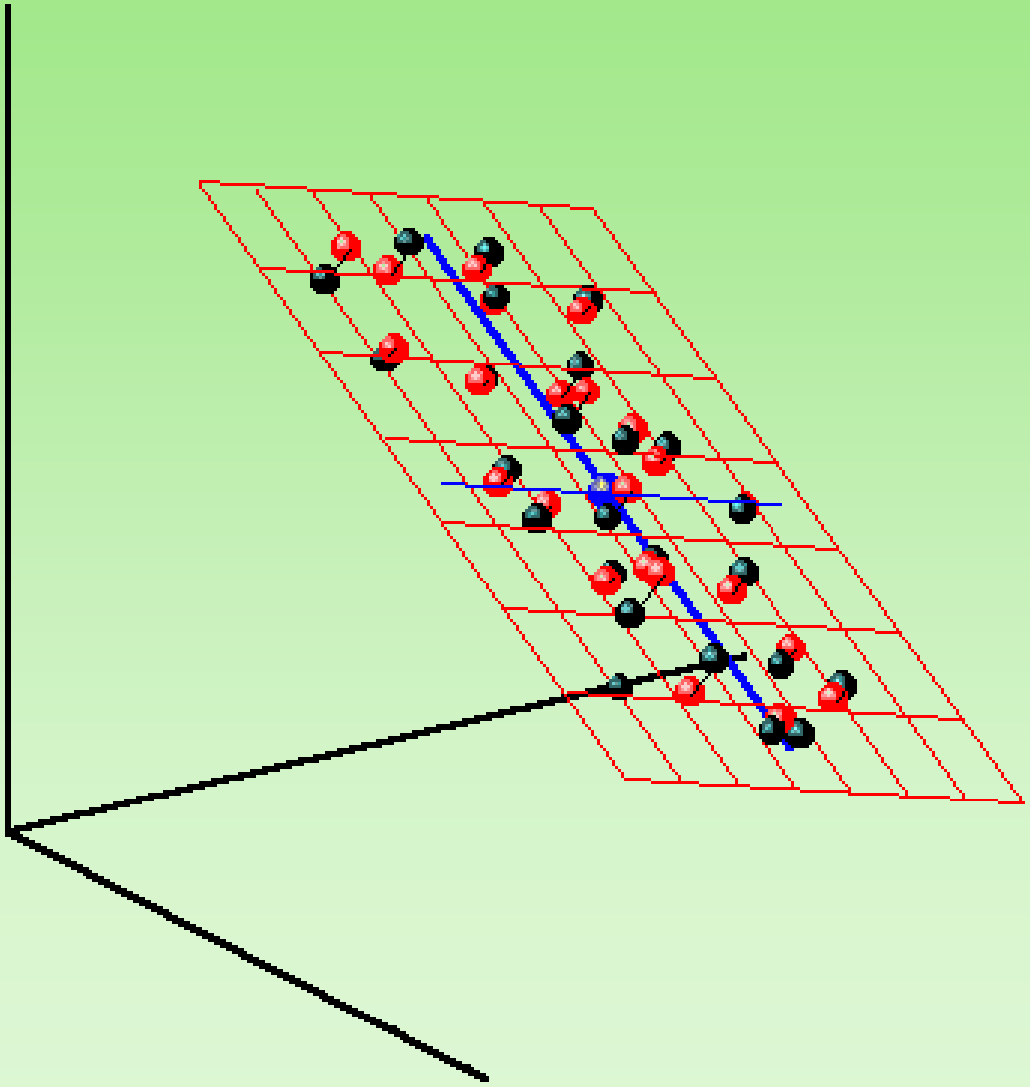
Классификация статуса образцов



Методы классификации

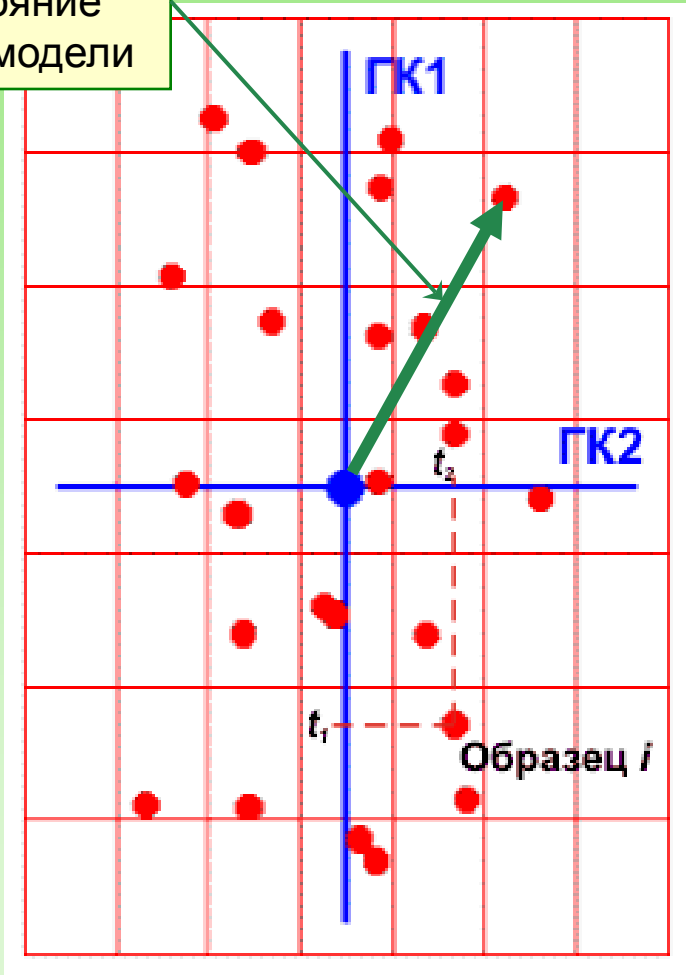
- **Soft Independent Modeling of Class Analogy (SIMCA)**
- PLS Discriminant Analysis (PLS-DA)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Support Vector Machine (SVM)
- Artificial Neural Network (ANN)
- k-Nearest Neighbor (KNN)

SIMCA

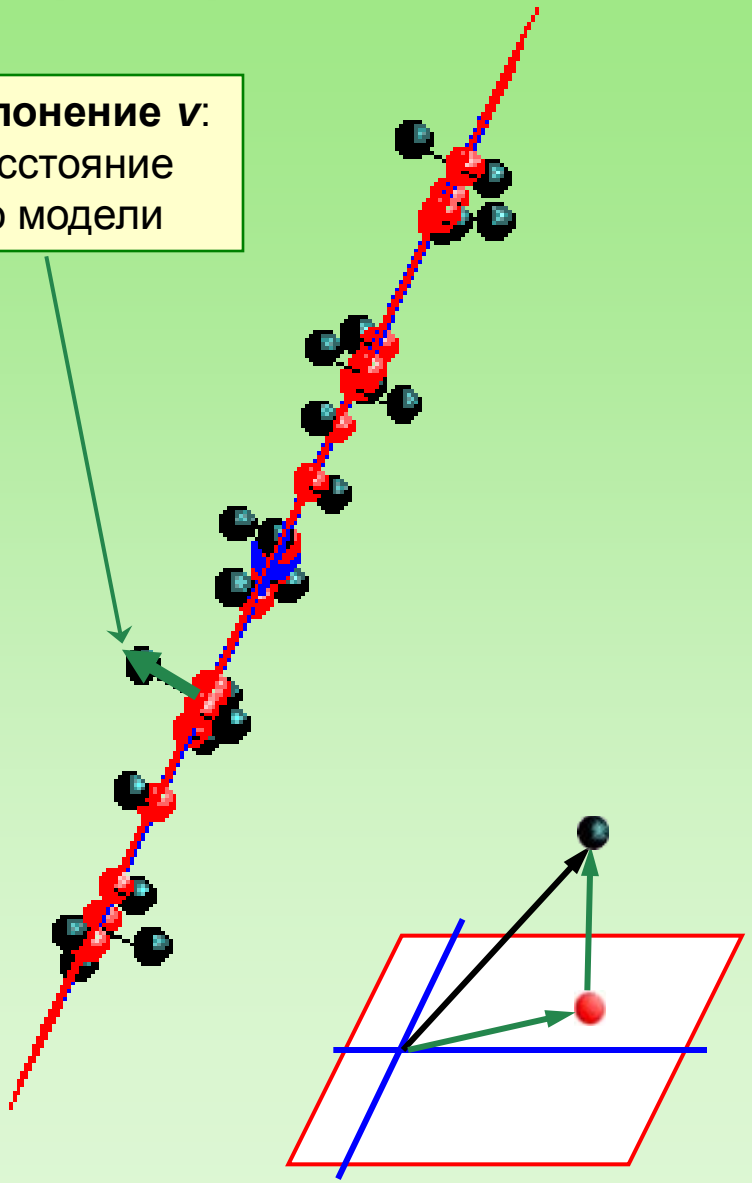


SIMCA: размах и отклонение

Размах *h*:
расстояние
внутри модели

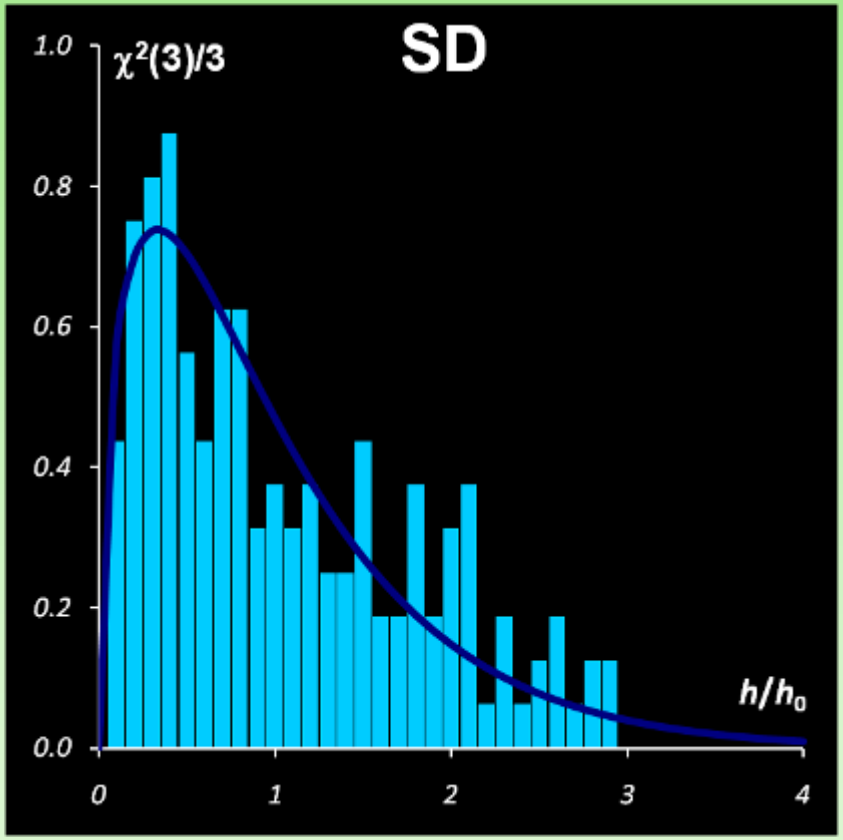


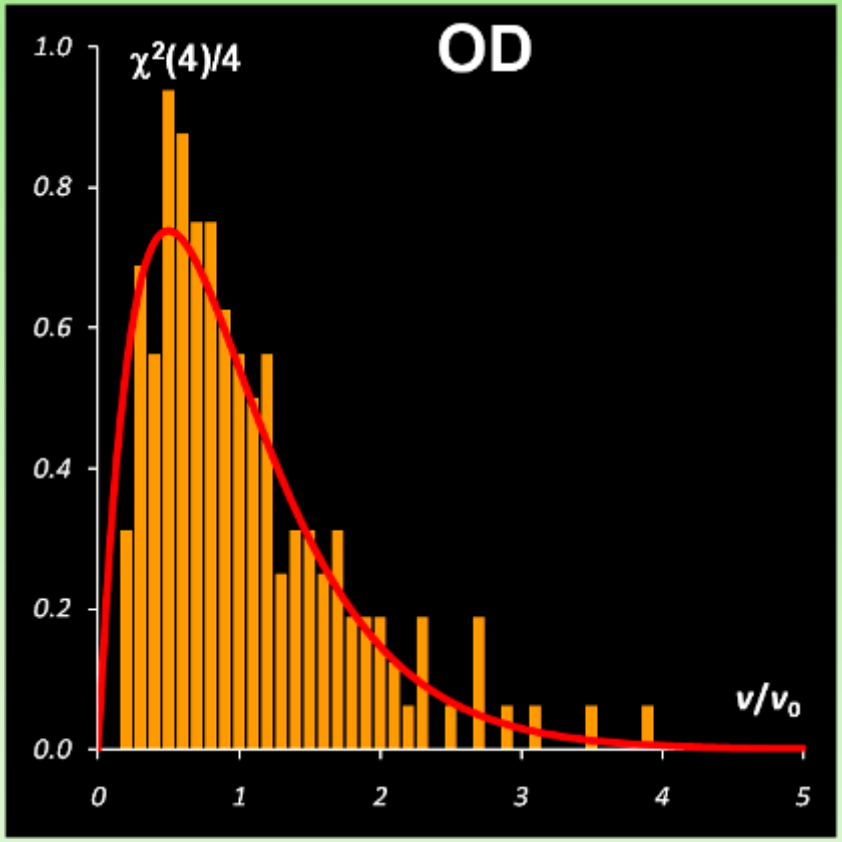
Отклонение *v*:
расстояние
до модели



Распределение расстояний: Пример

I=160 A=3 N_h=3 N_v=5





Проблема: как оценить DOF?

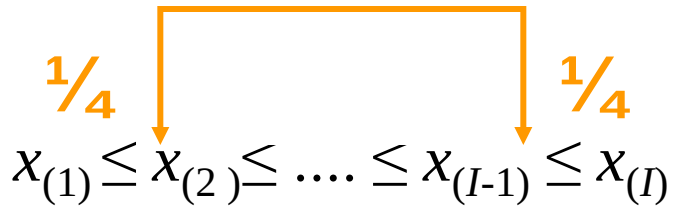
$$x = \begin{cases} = h/h_0 \\ = v/v_0 \end{cases} \qquad x_1, \ldots, x_I \sim \chi^2(N)/N \quad \Rightarrow \quad N = ?$$

Метод моментов

$$S^2 = \frac{1}{I}\sum_{i=1}^{I}(x_i - 1)^2$$

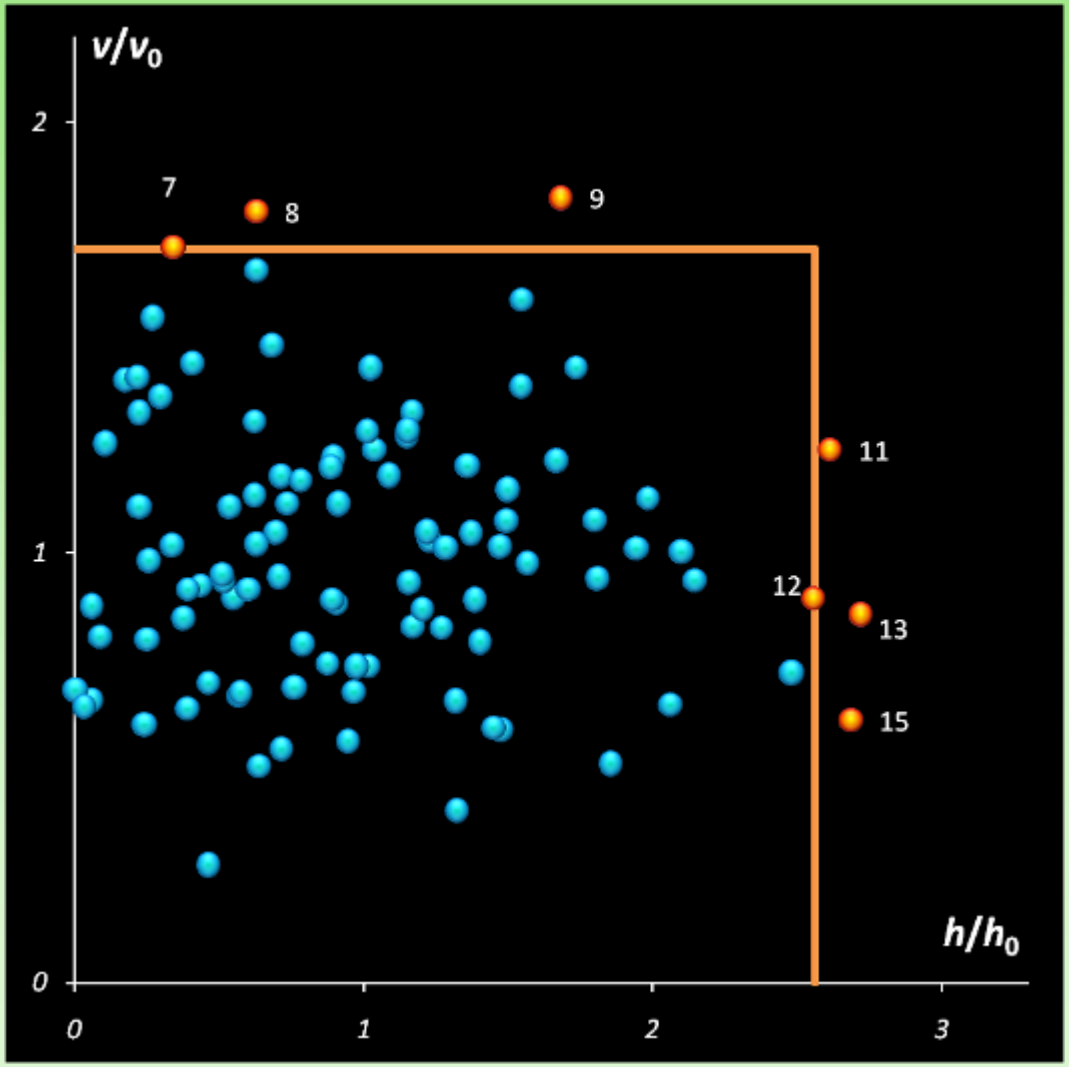
$$\hat{N} = \frac{2}{S^2}$$

Интерквартильный подход



$$\frac{\chi^{-2}(N, 3/4) - \chi^{-2}(N, 1/4)}{N} = IQR$$

Критическая область: обычная

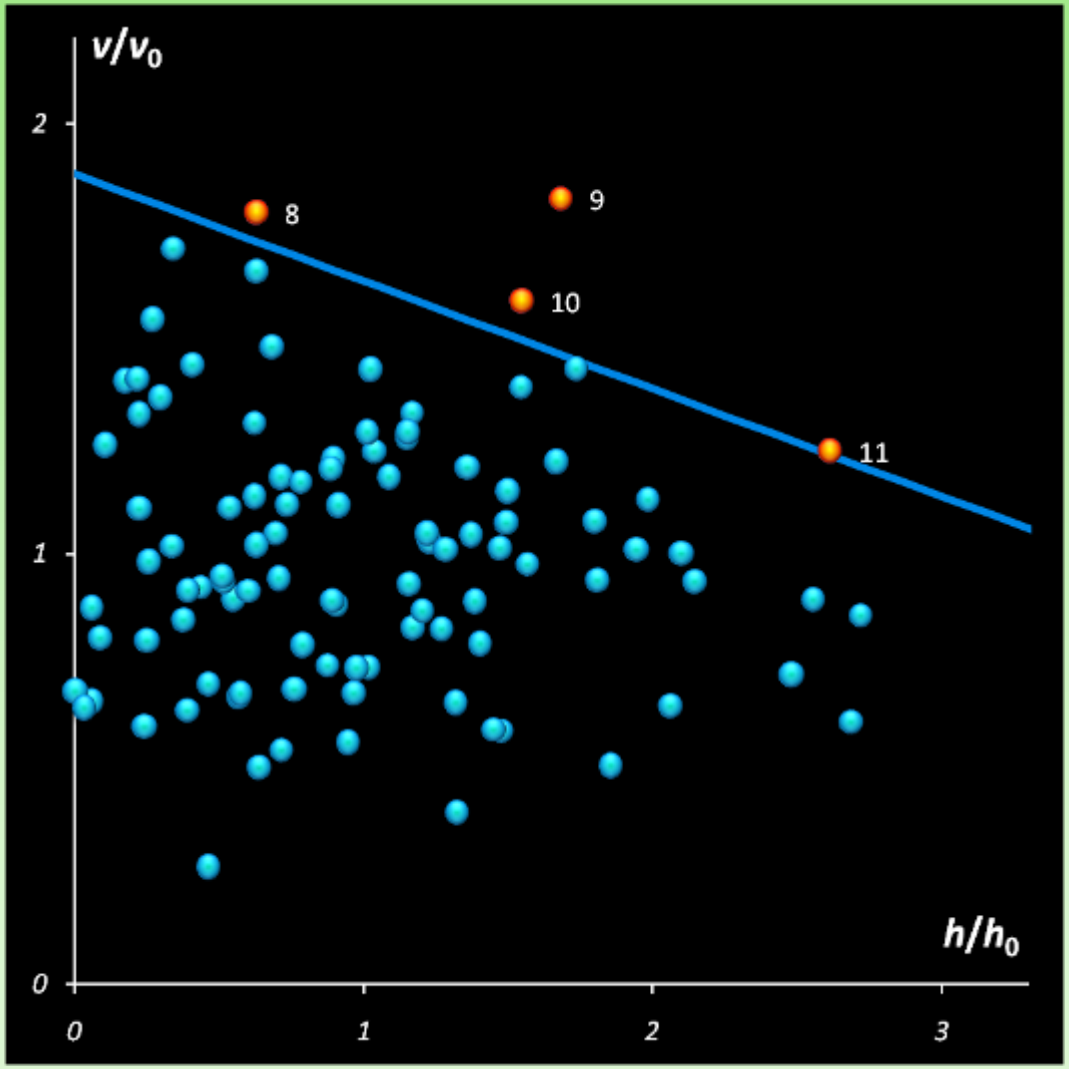


$$I=100 \quad \gamma=0.05$$

$$\alpha = 1-\sqrt{1-\gamma}$$

$$H_\gamma = \left[0, \frac{h_0}{N_h}\chi^{-2}(N_h,\alpha)\right] \otimes \left[0, \frac{v_0}{N_v}\chi^{-2}(N_v,\alpha)\right]$$

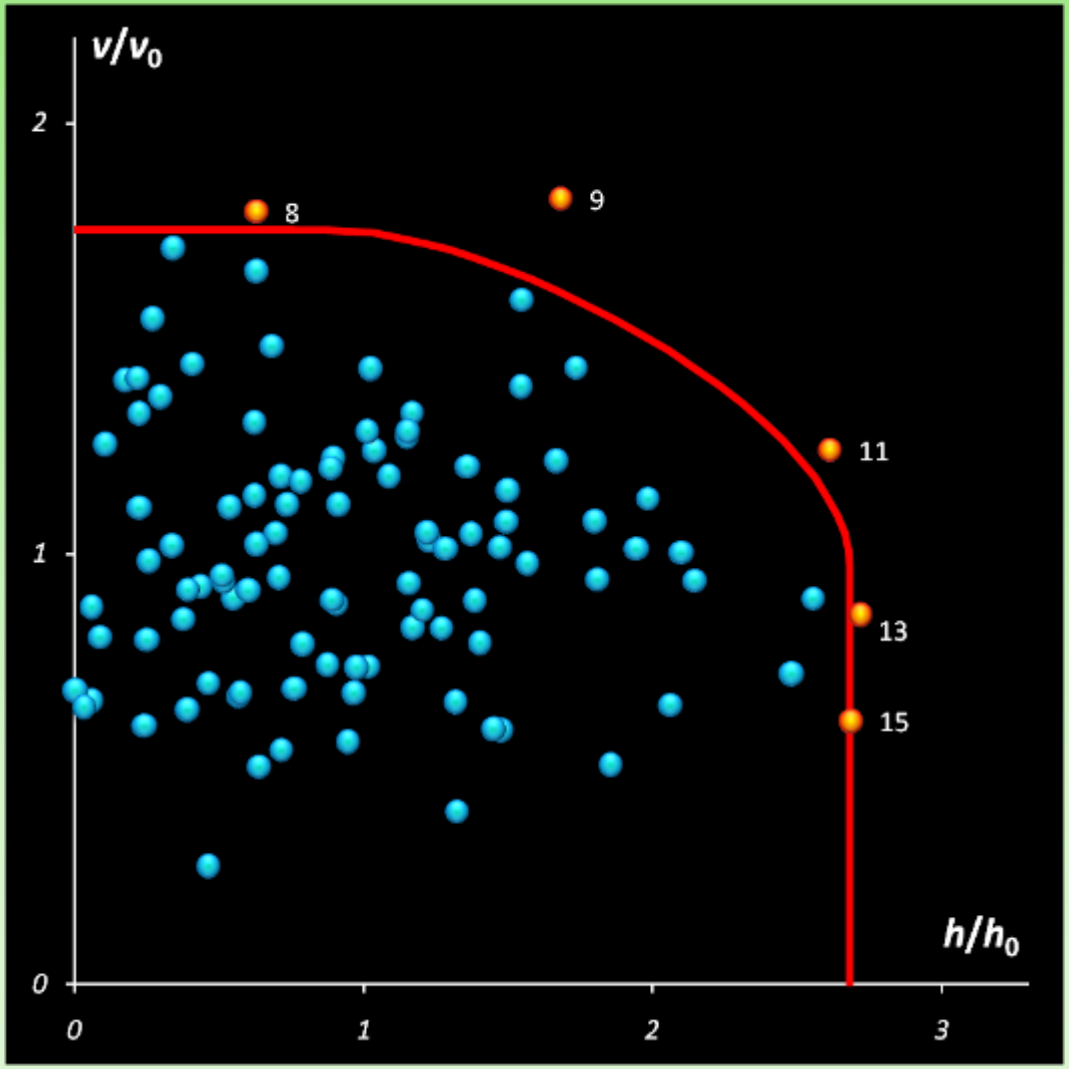
Критическая область: хи квадрат



I=100 γ=0.05

$$N_h \frac{h}{h_0} + N_v \frac{v}{v_0} \sim \chi^2(N_h + N_v)$$

Критическая область: Wilson-Hilferty



I=100 γ=0.05

$$\frac{h}{h_0} = \left(z\sqrt{2/9N_h} + 1 - 2/9N_h\right)^3$$

$$\frac{v}{v_0} = \left(w\sqrt{2/9N_v} + 1 - 2/9N_v\right)^3$$

Пример: входной контроль

Цель: проверка качества субстанций

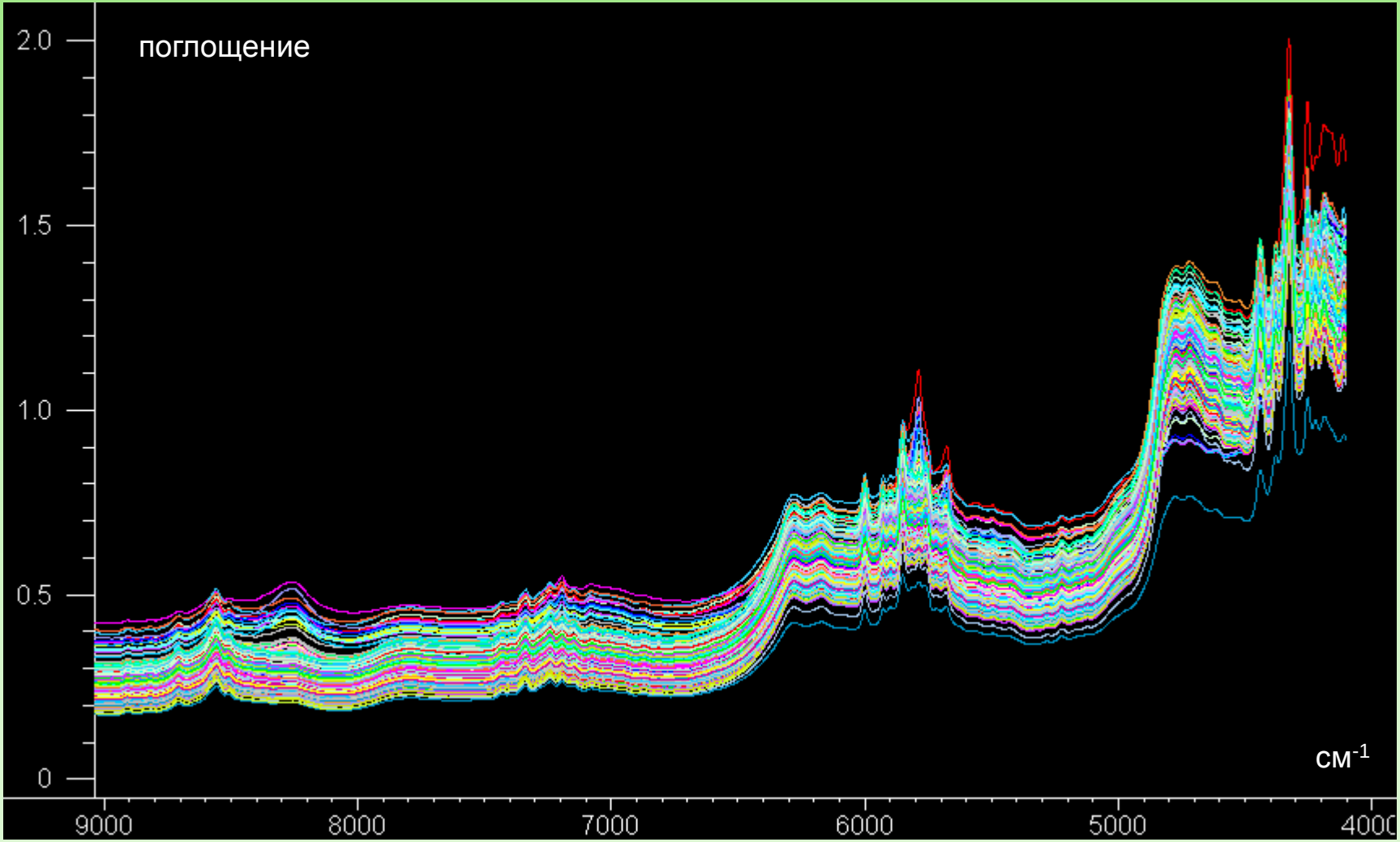
Инструмент: БИК щуп

Метод: SIMCA

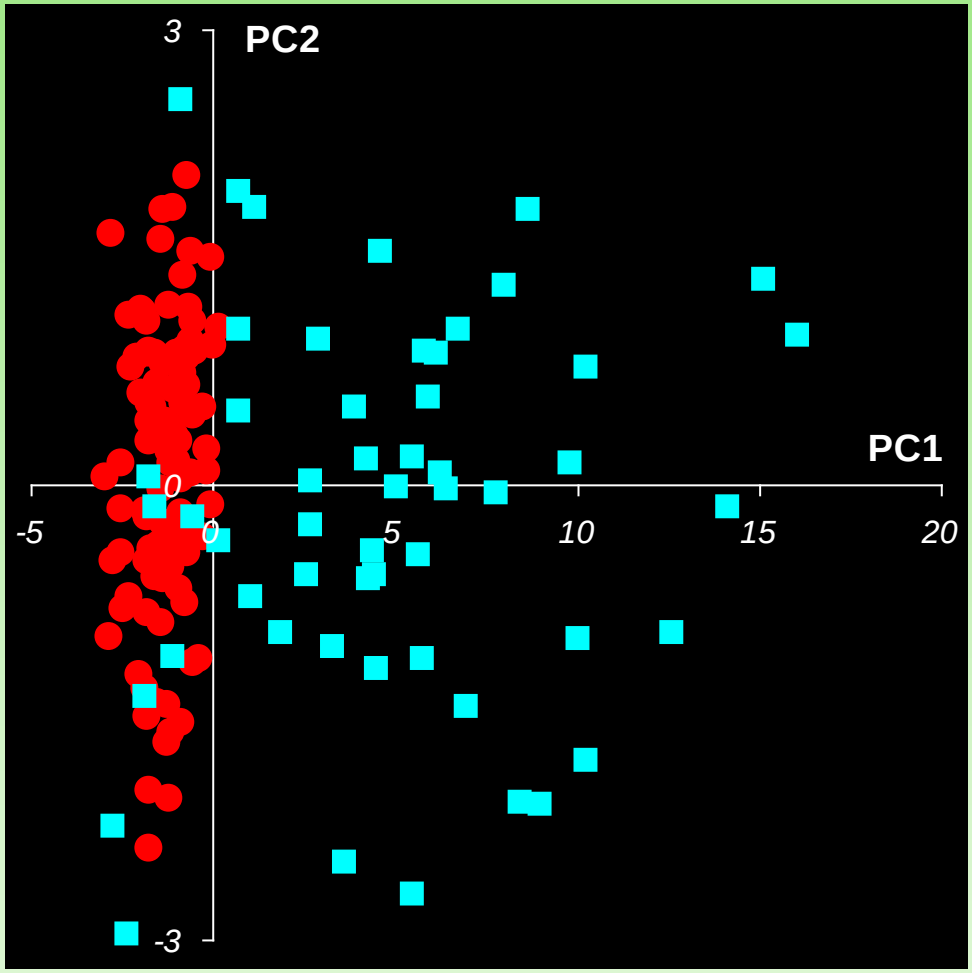
Объект: 82 бочки

Измерений: 82×3=246

Спектральные данные



Анализ данных (PCA)

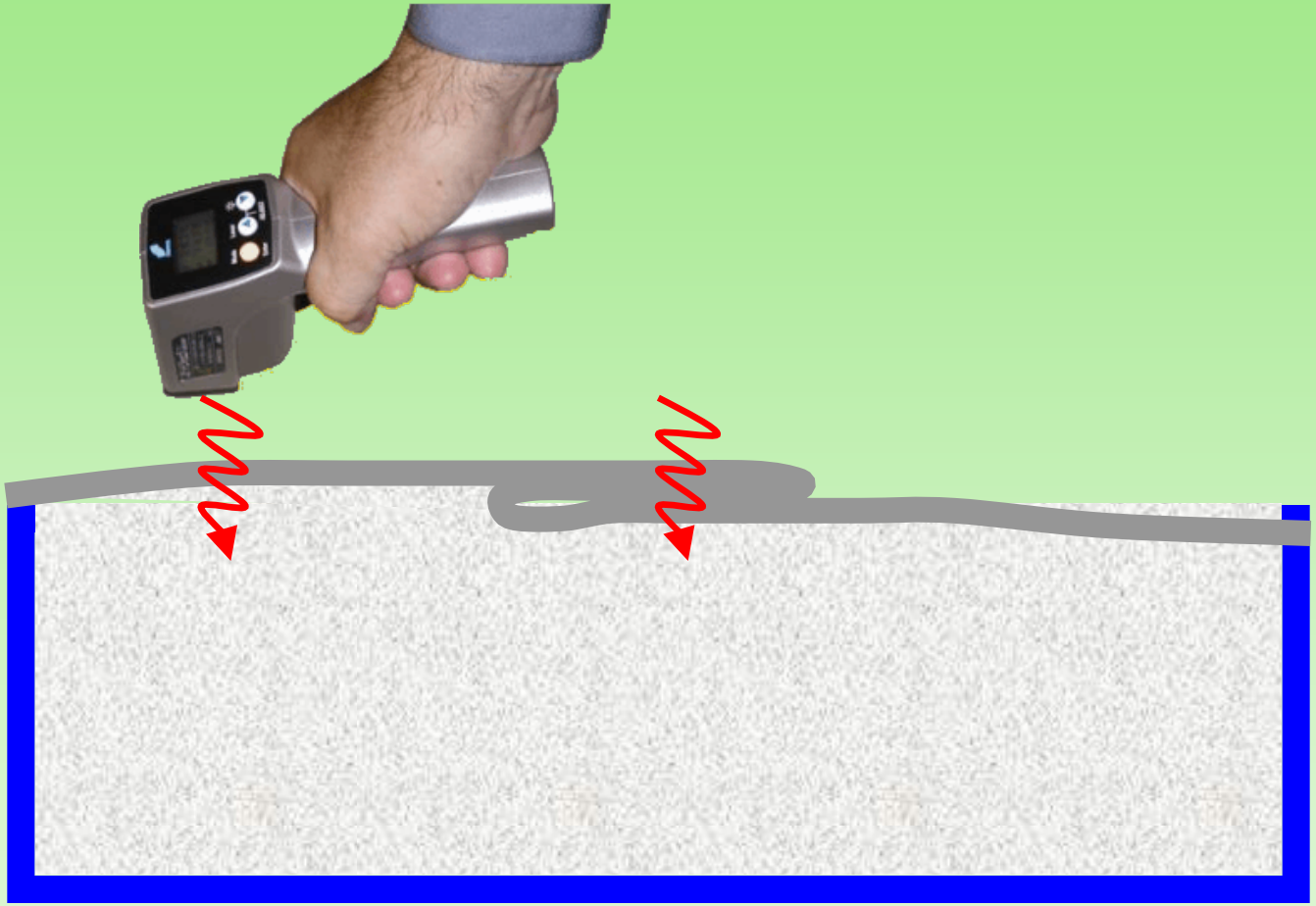


Проблема:

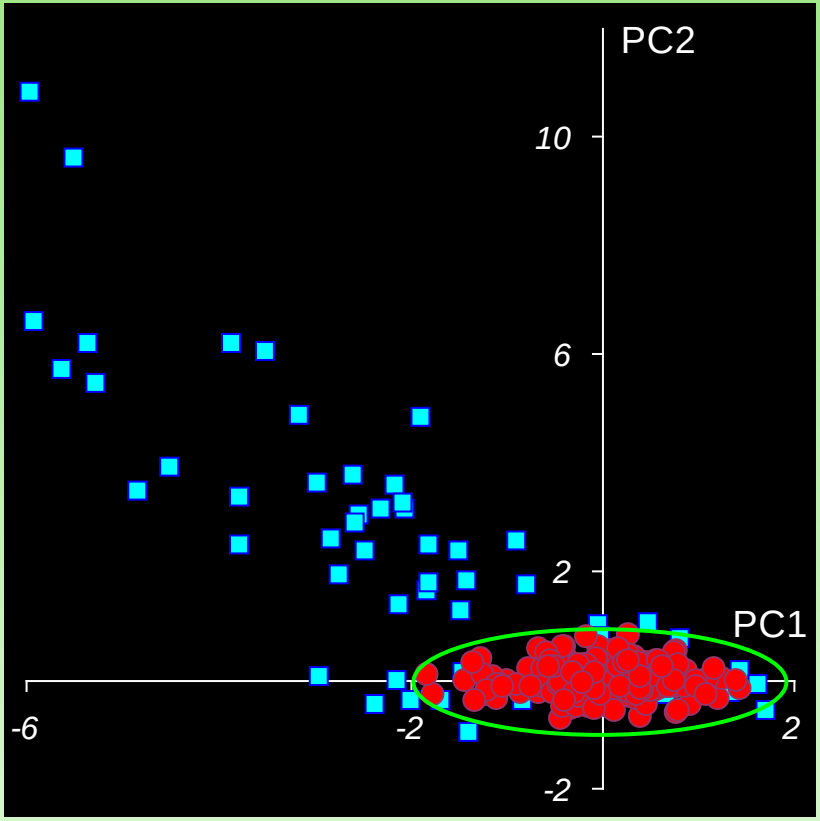
60 измерений из 246 никуда не годятся.

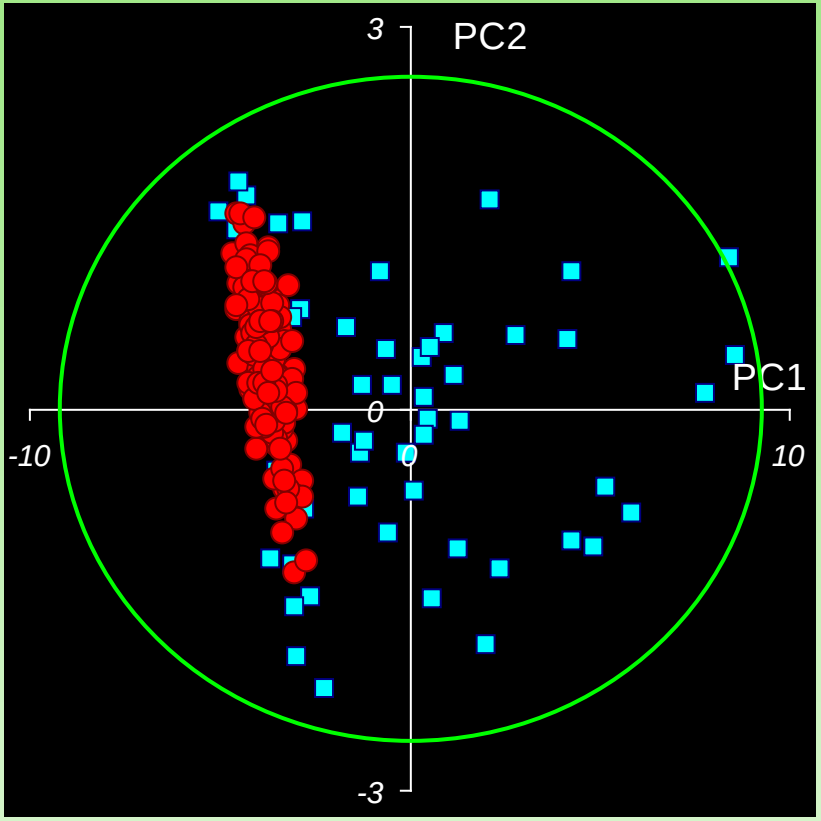
Неужели это выбросы?

Влияние позиции щупа



Две PCA модели





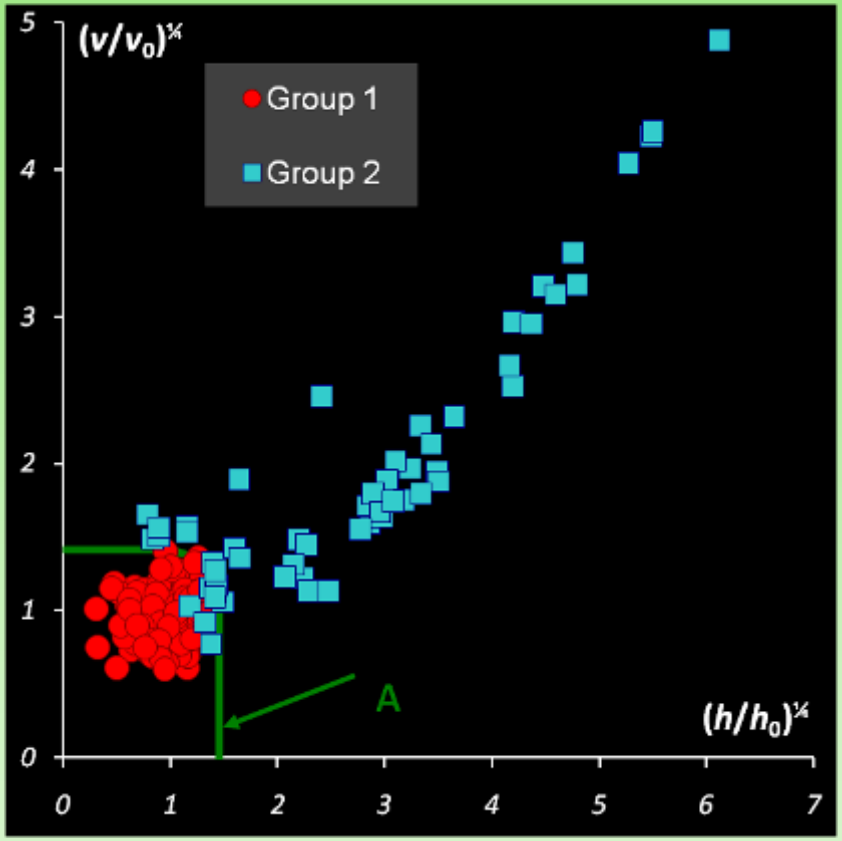
Модель 1

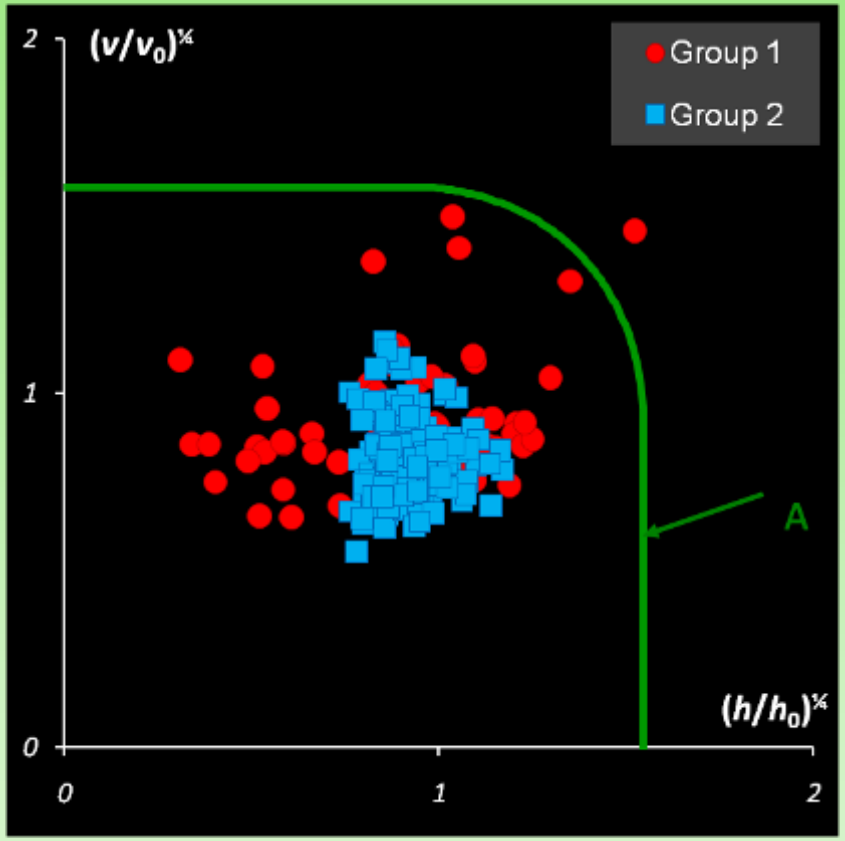
Круг – обучающий набор (группа 1),
Квадрат – тест (группа 2)

Модель 2

Квадрат – обучающий набор (группа 2),
Круг – тест (группа 1)

Две SIMCA модели





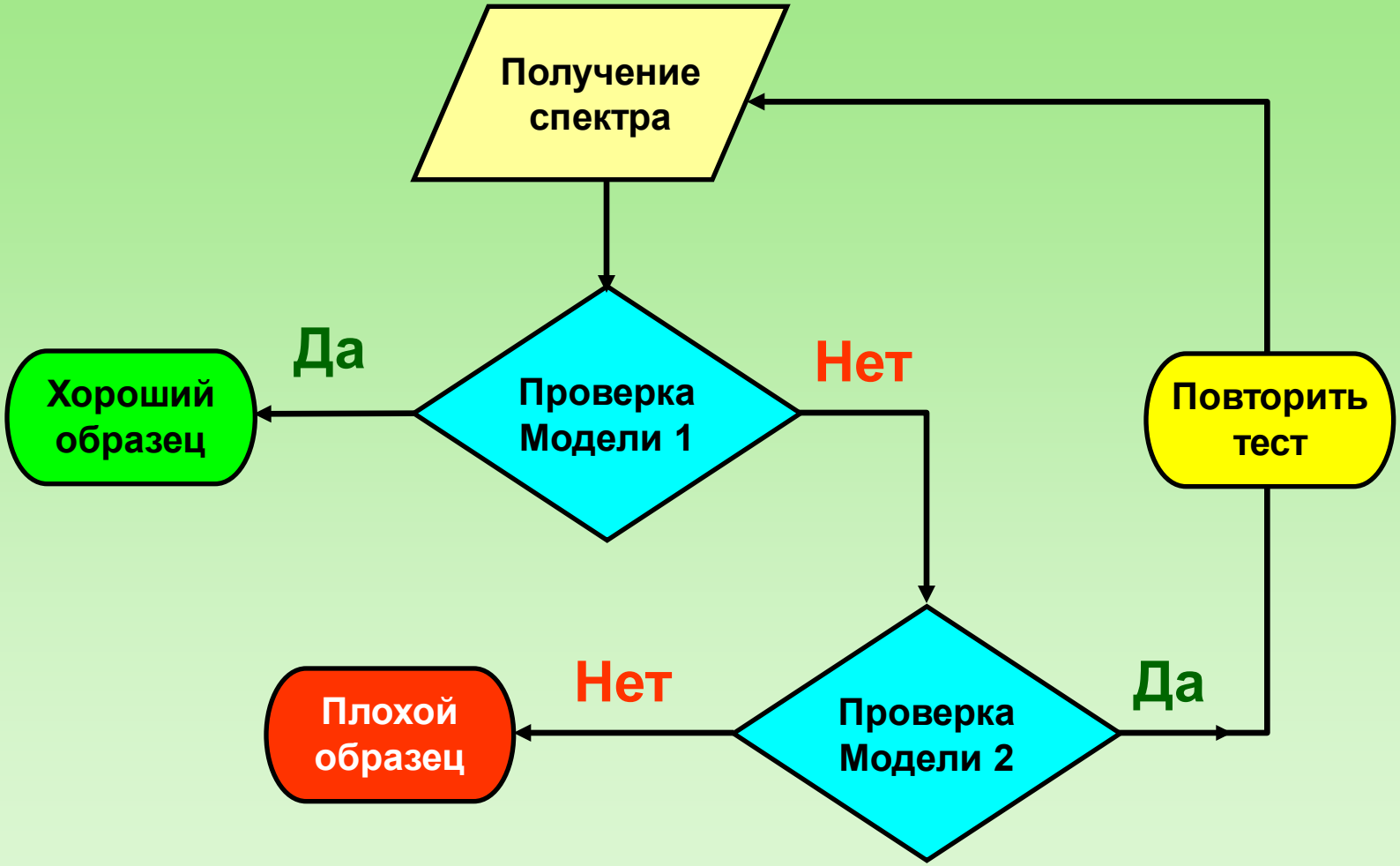
Модель 1

Круг – обучающий набор (группа 1),
Квадрат – тест (группа 2)

Модель 2

Квадрат – обучающий набор (группа 2),
Круг – тест (группа 1)

Блок схема входного контроля



Методы разрешения кривых

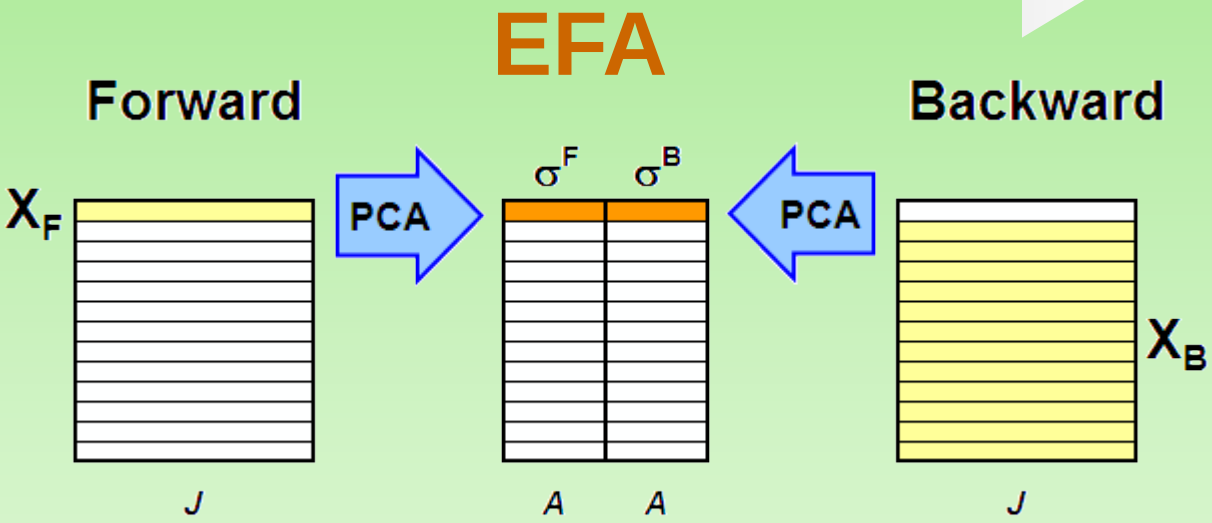
Факторные

- Прокрустово вращение
- Эволюционный факторный анализ (EFA)
- Оконный факторный анализ (WFA)

Итерационные

- Итерационный целевой факторный анализ (ITTFA)
- **Чередующиеся наименьшие квадраты (ALS)**

Многомерное разрешение кривых



TECHNOMETRICS VOL. 13, No. 3 AUGUST 1971

Self Modeling Curve Resolution

WILLIAM H. LAWTON AND EDWARD A. SYLVESTRE

Факторные методы

$$X \approx CS^t$$

$$X \approx TP^t$$

$$I = RR^t = \text{единичная}$$

$$X \approx T(RR^t)P^t = (TR)(PR)^t$$

$$\hat{C} \approx TR \qquad \hat{S} \approx PR$$

Итерационные методы (ALS)

Alternative Least Squares

0. Задать начальную **C**

1. Найти $\mathbf{S} = \mathbf{X}^t\mathbf{C}(\mathbf{C}^t\mathbf{C})^{-1}$

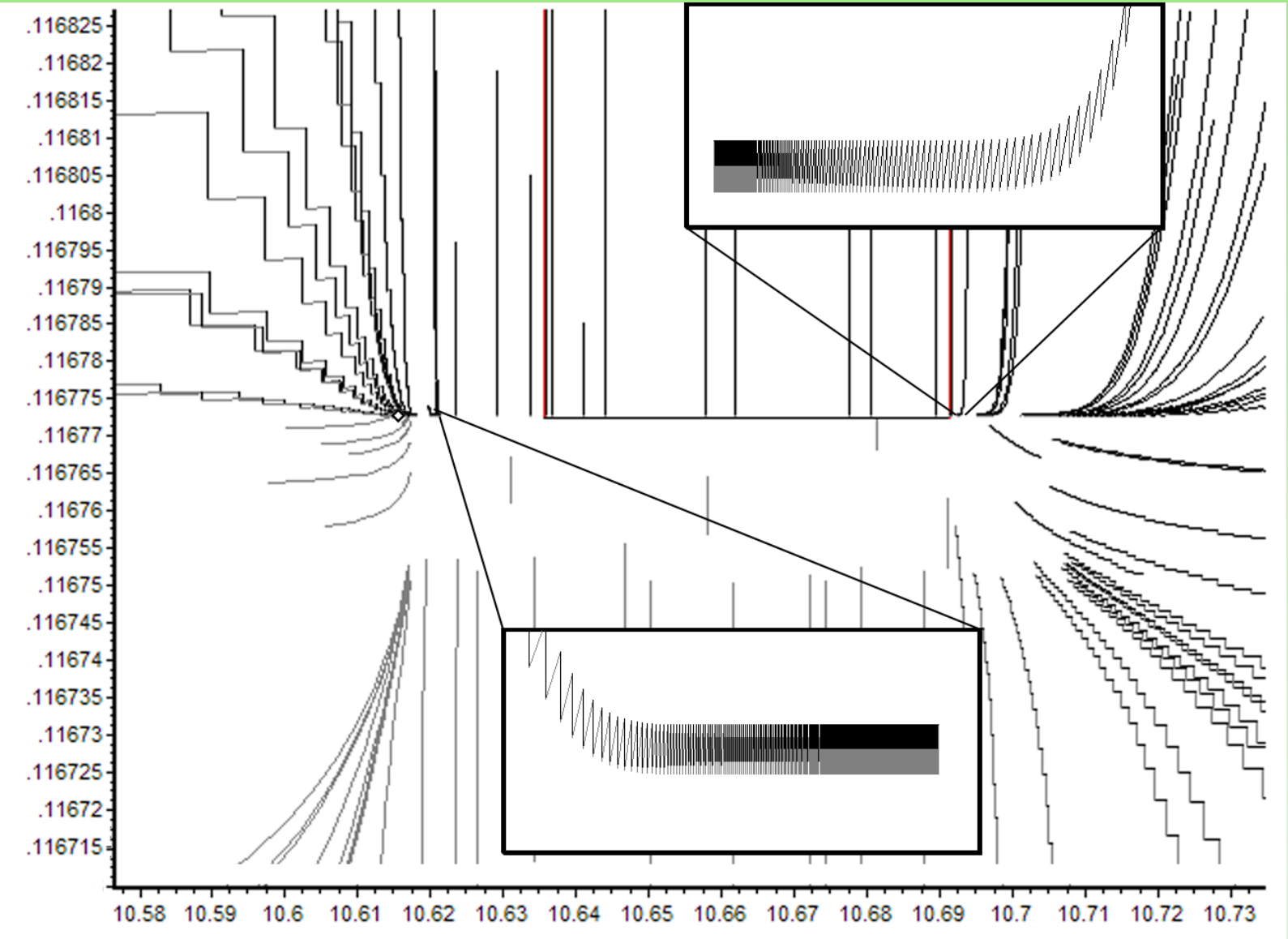
2. Подправить **S**, т.ч. **S** >0

3. Найти $\mathbf{C}=\mathbf{X}\mathbf{S}(\mathbf{S}^t\mathbf{S})^{-1}$

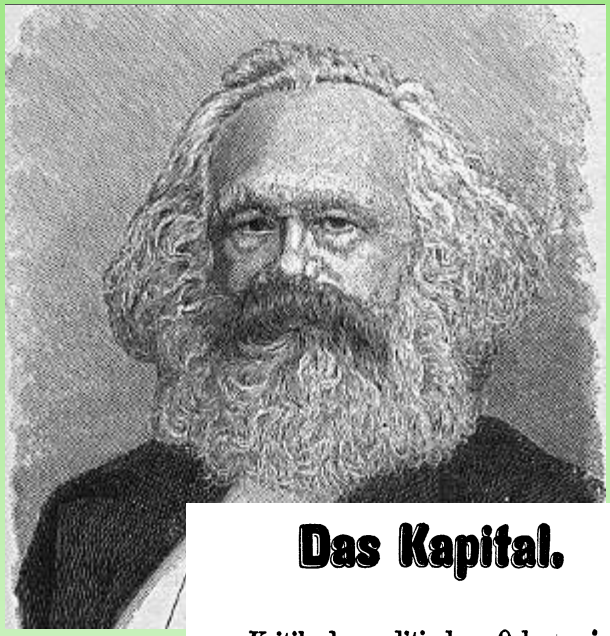
4. Подправить **C**, т.ч. **C** >0

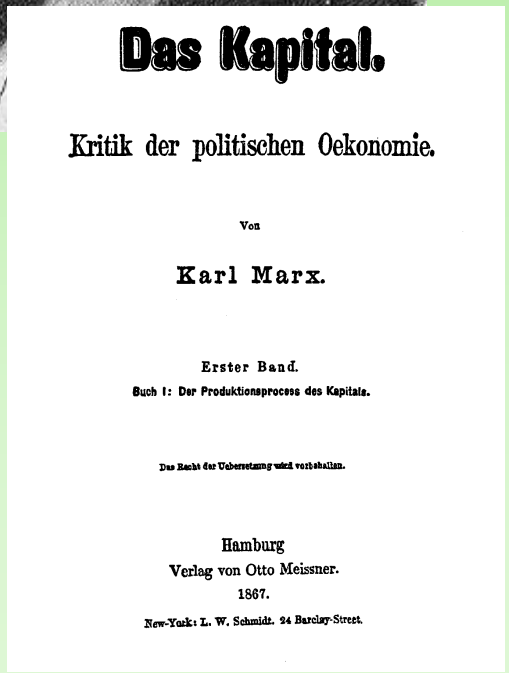
5. Повторить с 1 до сходимости

Проблема: сходимость



Заключение



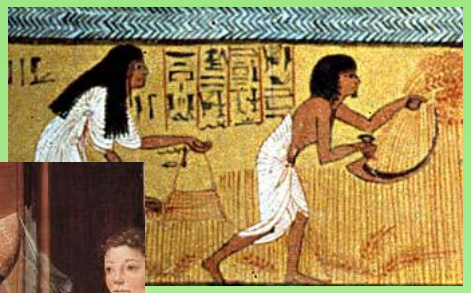


Нормальная наука — каждое новое открытие поддаётся объяснению с позиций господствующей теории.

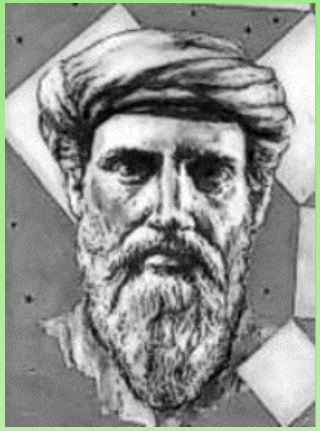
Экстраординарная наука — необъяснимых фактов. Появление альтернативных теорий и методов.

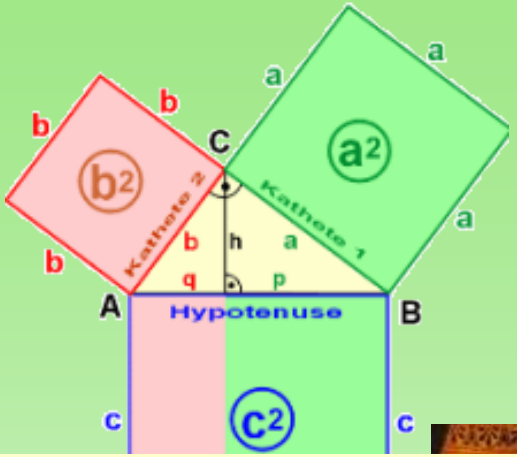
Научная революция — формирование новой парадигмы

Эволюция парадигм



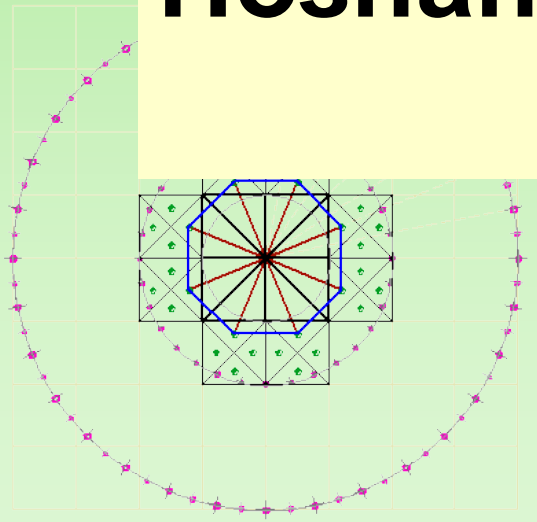
Геометрическая или аграрная

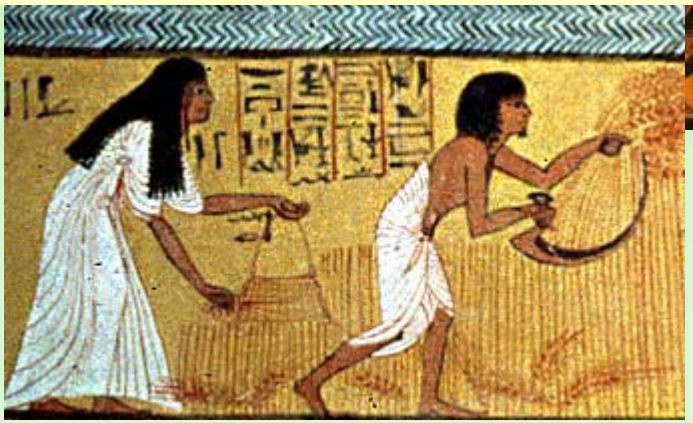




Познание –

это рисование

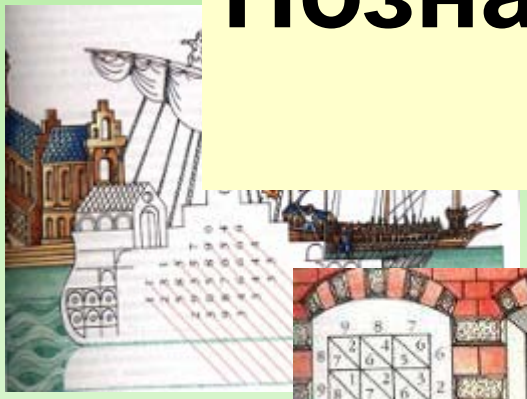




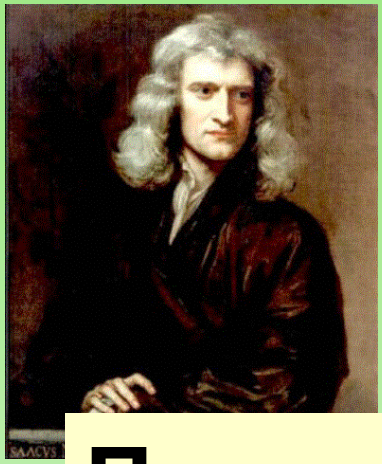
Алгебраическая или торговая

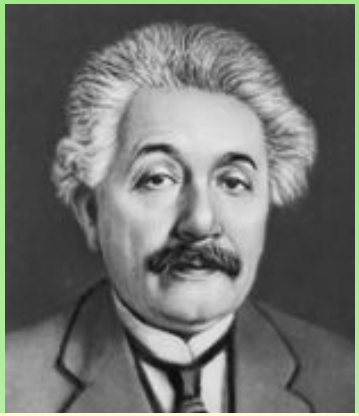
Познание –

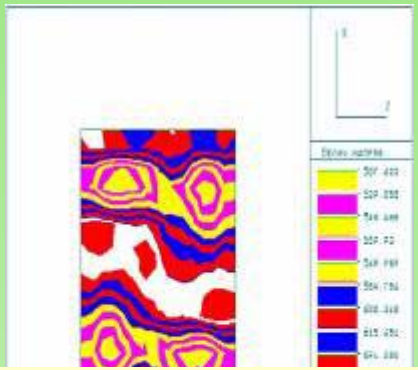
это вычисление



Дифференциальная или индустриальная

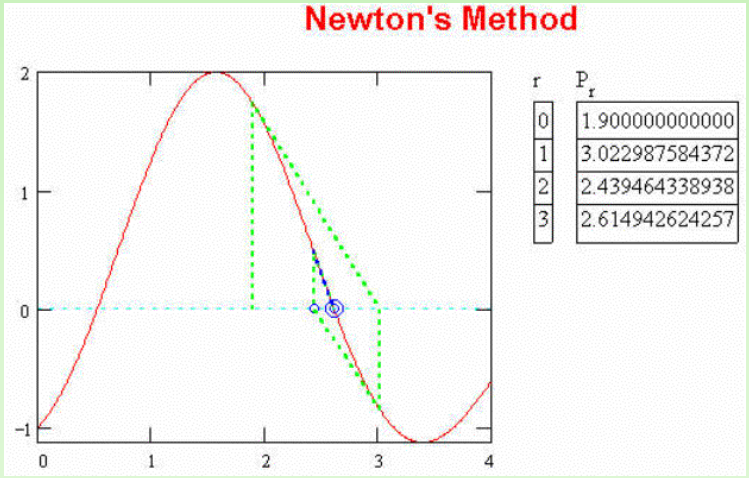




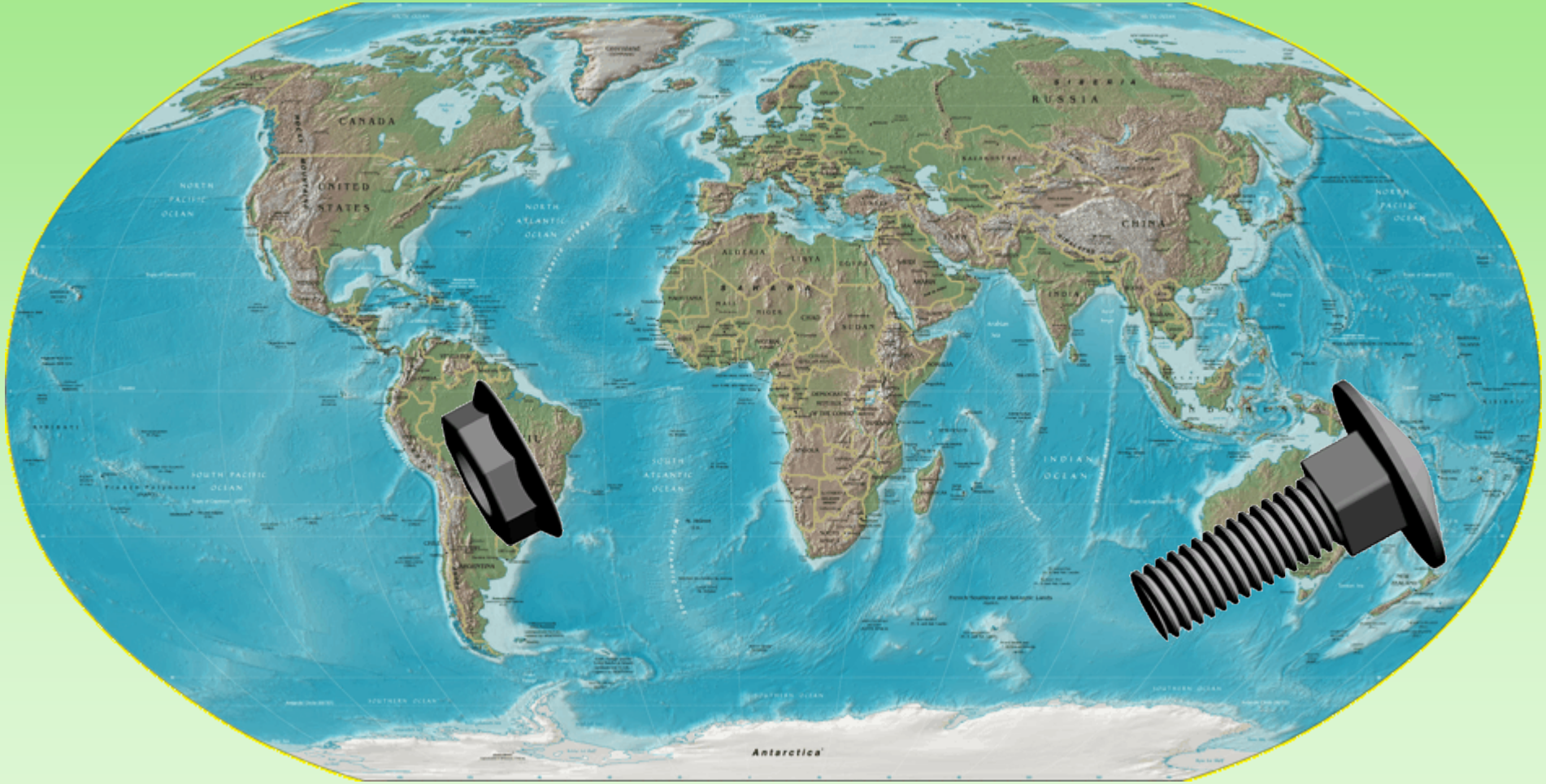


Познание –

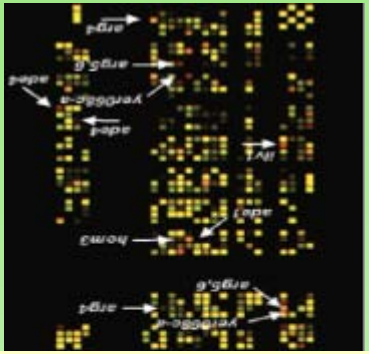
это дифференцирование

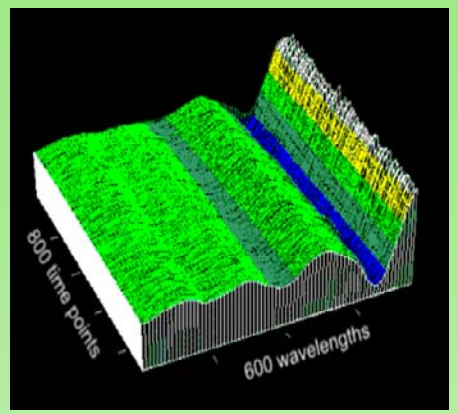


Дифференциальная парадигма: стандартизация



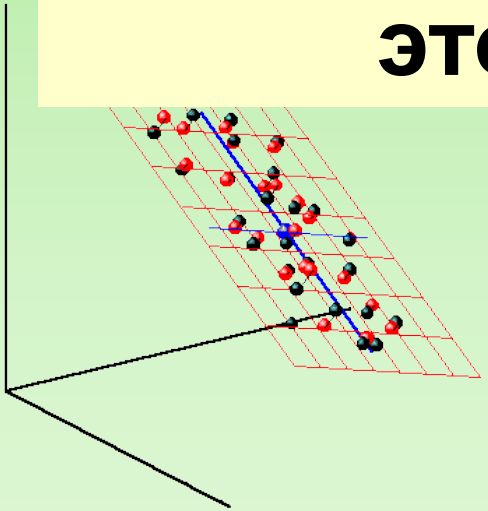
Новая парадигма, постиндустриальная

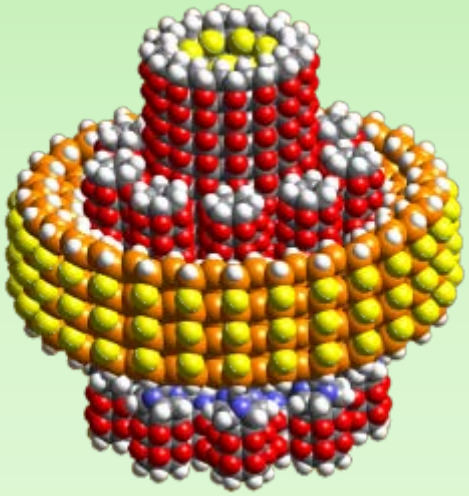




Познание –

это извлечение информации





Спасибо за внимание!

Информация

Российское Хемометрическое Общество

http://rcs.chph.ras.ru/index.html

Хемометрика в России

http://www.chemometrics.ru

Померанцев Алексей Леонидович

+7(903) 5618227

forecast@chph.ras.ru

http://rcs.chph.ras.ru/ICP/palrus.htm

Родионова Оксана Евгеньевна

+7(903) 1504855

rcs@chph.ras.ru

http://rcs.chph.ras.ru/ICP/royrus.htm

Литература

1. A.L. Pomerantsev "Confidence Intervals for Non-linear Regression Extrapolation", *Chemom. Intell. Lab. Syst*, **49**, 41-48, (1999), http://rcs.chph.ras.ru/papers/N49_041.pdf

2. E.V. Bystritskaya, A.L. Pomerantsev, O.Ye. Rodionova "Nonlinear Regression Analysis: New Approach to Traditional Implementations", *J. Chemometrics*, **14**, 667-692 (2000), http://rcs.chph.ras.ru/papers/J14_667.pdf

3. O. Ye. Rodionova, K. H. Esbensen, A.L. Pomerantsev, "Application of SIC (Simple Interval Calculation) for object status classification and outlier detection - comparison with PLS/PCR", *J. Chemometrics*, **18** , 402-413 (2004), http://rcs.chph.ras.ru/papers/J18_402.pdf

4. A.L. Pomerantsev, O.Ye. Rodionova, A. Höskuldsson, "Process control and optimization with simple interval calculation method", *Chemom. Intell. Lab.Syst*., **81** (2), 165-179 (2006), http://rcs.chph.ras.ru/papers/N81_165.PDF

5. Родионова О.Е., Померанцев А.Л. "Хемометрика: достижения и перспективы", *Успехи химии*, **75** (4) 302-317 (2006), http://rcs.chph.ras.ru/papers/UspKhim75_302.pdf

6. O.Ye. Rodionova, L.P. Houmøller, A.L. Pomerantsev, P. Geladi, J. Burger, V.L. Dorofeyev, A.P. Arzamastsev "NIR spectrometry for counterfeit drug detection", *Anal. Chim. Acta*, **549**, 151-158 (2005), http://rcs.chph.ras.ru/papers/ACA549_151.pdf

7. A. Pomerantsev "Acceptance areas for multivariate classification derived by projection methods", *J. Chemometrics*, **22**, 601-609 (2008), http://rcs.chph.ras.ru/papers/J22_601.pdf

8 O.Y. Rodionova, A.L. Pomerantsev "Subset selection strategy", *J. Chemometrics*, **22**, 674-685, (2008), http://rcs.chph.ras.ru/papers/J22_674.pdf

9. O.Ye. Rodionova, A.L. Pomerantsev "Simple view on Simple Interval Calculation (SIC) method" *Chemom. Intell. Lab.Syst*., **97** (1), 64-76 (2009), http://rcs.chph.ras.ru/papers/N97_064.pdf

10. O.Ye. Rodionova, Ya.V. Sokovikov, A.L. Pomerantsev " Quality control of packed raw materials in pharmaceutical industry" *Anal. Chim. Acta* , **642**(1-2), 222-227 (2009), http://rcs.chph.ras.ru/papers/ACA642_222.pdf

11. O.Ye. Rodionova, A.L. Pomerantsev,"NIR based approach to counterfeit-drug detection" *Trends Anal. Chem*., **29** (8), 781-938 (2010), http://rcs.chph.ras.ru/papers/TrAC29_795.pdf