Проекционные методы в линейном регрессионном анализе: РГК/ПЛС

Андрей Юрьевич Богомолов

Российское хемометрическое общество

European Molecular Biology Laboratory (EMBL)

Методы многомерной калибровки

Андрей Юрьевич Богомолов

Российское хемометрическое общество

European Molecular Biology Laboratory (EMBL)

Тема лекции

Многомерная калибровка
Multivariate Calibration

Анализ многомерных данных (Хемометрика)

Multivariate Data Analysis (Chemometrics)

К вопросу о русской терминологии

- родной язык хемометрики - английский
- терминология за 30 лет устоялась: статьи, учебники, книги, конференции
- привычные аббревиатуры: PCA, PCR, PLS, SIMCA, RMSEP, etc. - не нуждаются в расшифровке
- русская терминология создается сейчас
- нужен ли перевод? – да!
  - например: “scores and loadings” (!?)
  - нужно время, чтобы русские термины вошли в обиход
- в настоящей лекции - параллельная терминология

Калибровка или градуировка?

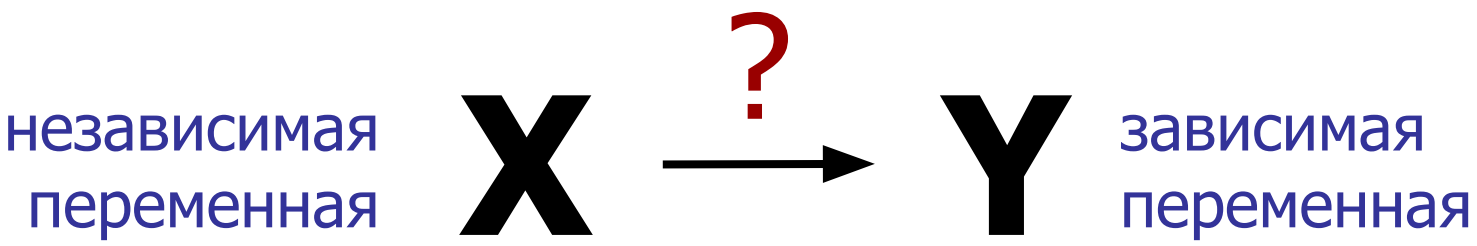
- в русском языке – два сходных термина:
  - «КАЛИБРОВКА (средств измерений) – совокупность операций, выполняемых с целью определения и подтверждения действительных значений метрологических характеристик и (или) пригодности к применению средств измерений…»
  - «ГРАДУИРОВКА – метрологическая операция, при помощи которой устанавливается значение меры или делениям шкалы измерительного прибора придаются значения...»
- на английский оба переводятся как calibration
- «градуировка» – официальный термин
- в лекции будет использоваться **некорректный** термин «калибровка»

Регрессия & Калибровка

- "Regression is an approach for relating two sets of variables to each other"
  Kim Esbensen
- "Calibration is a process of constructing a mathematical model to relate the output of an instrument to properties of samples"
  Kenneth Beebe

- Калибровка ~ Регрессия

Регрессионный анализ

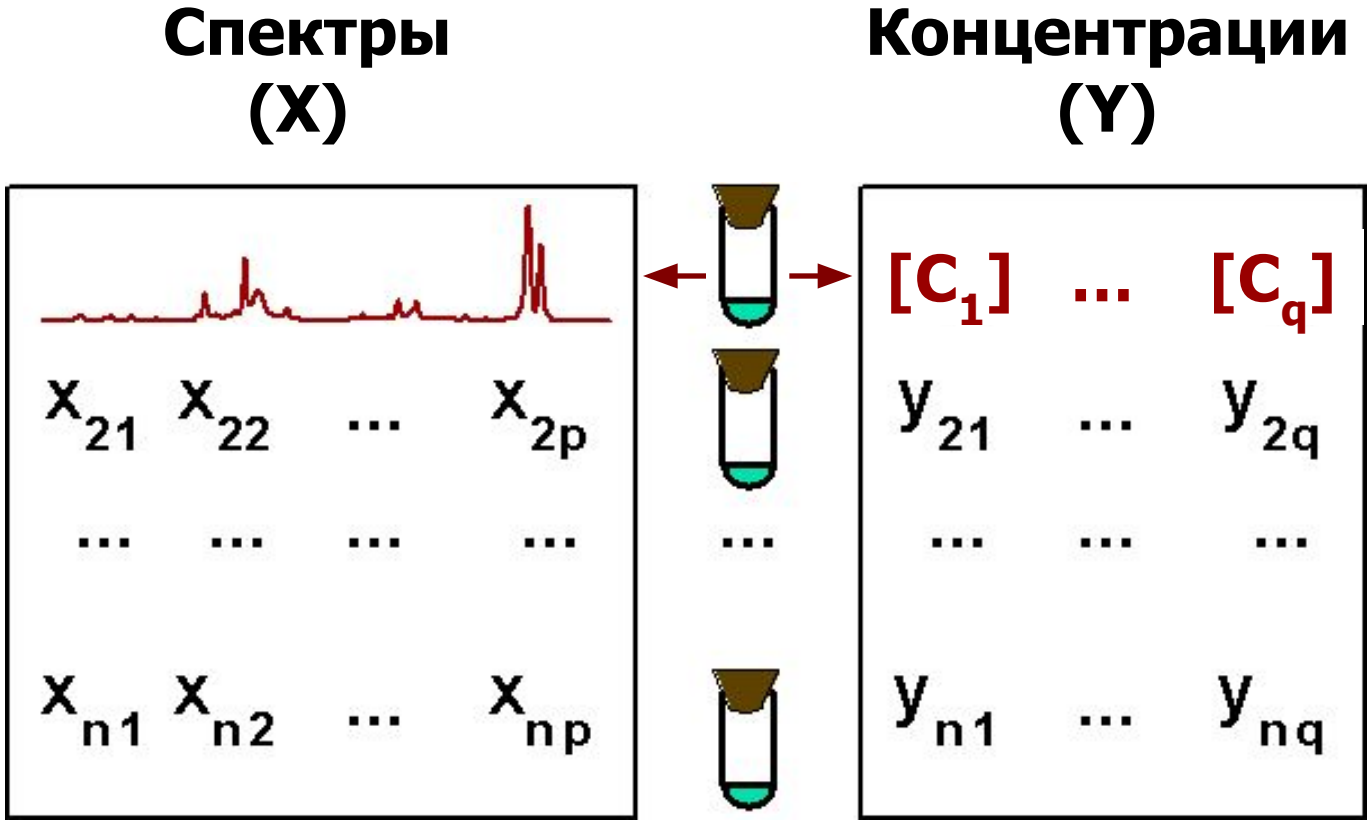


- линейная регрессия

$$Y = XB + E$$

- МГК (PCA) – моделирование (**X**)
- регрессия – моделирование (**X,Y**)

Спектральные данные



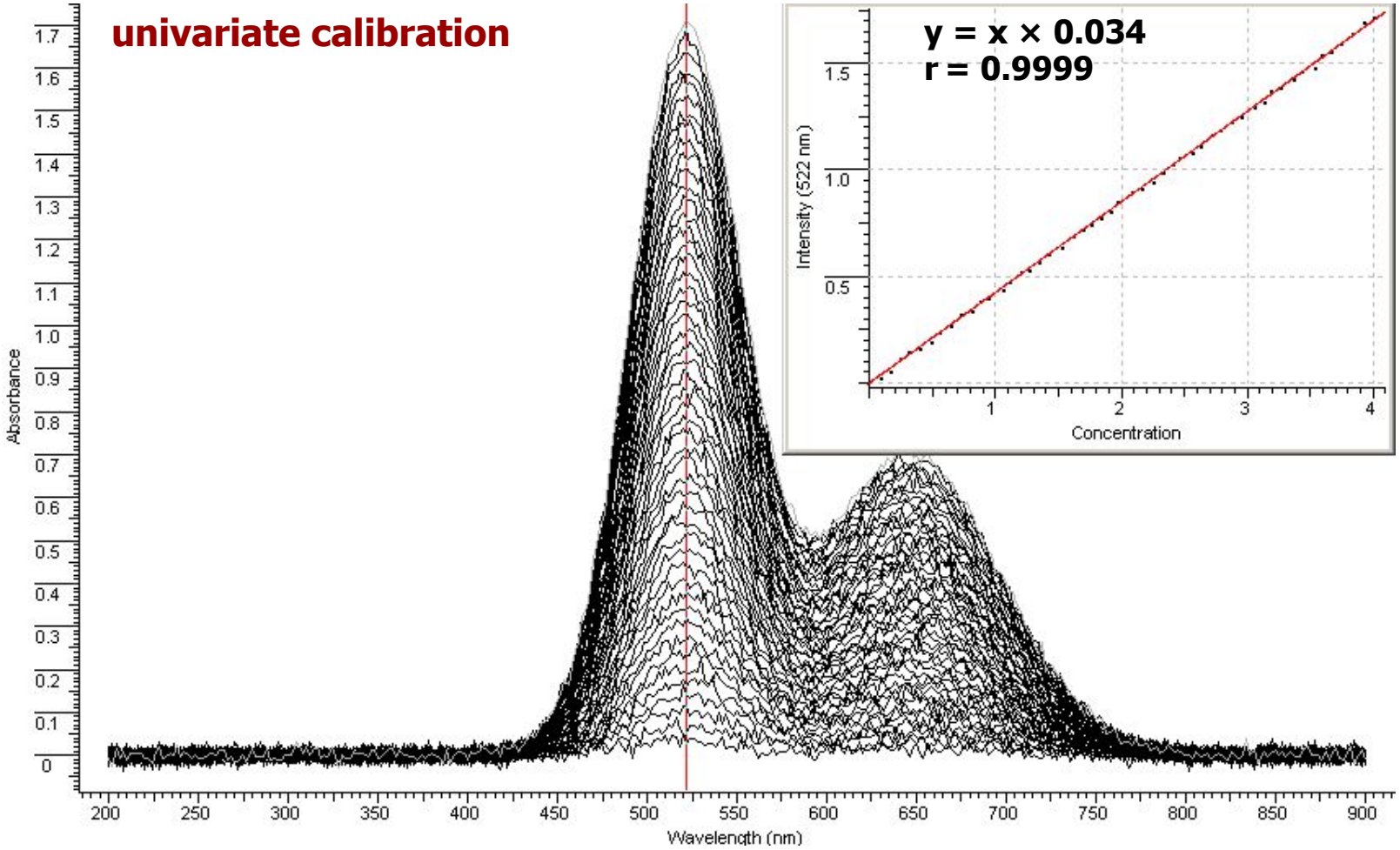
Для чего нужна калибровка?

- замена прямого измерения интересующего свойства, измерением другого, коррелирующего с первым
- такая потребность возникает если прямое измерение интересующего свойства нежелательно:
  - дорого
  - трудоемко
  - занимает много времени
  - этически нежелательно
  - эксперимент невозможен, и т. п.
- в подавляющем числе практических ситуаций такая замена оправдана!

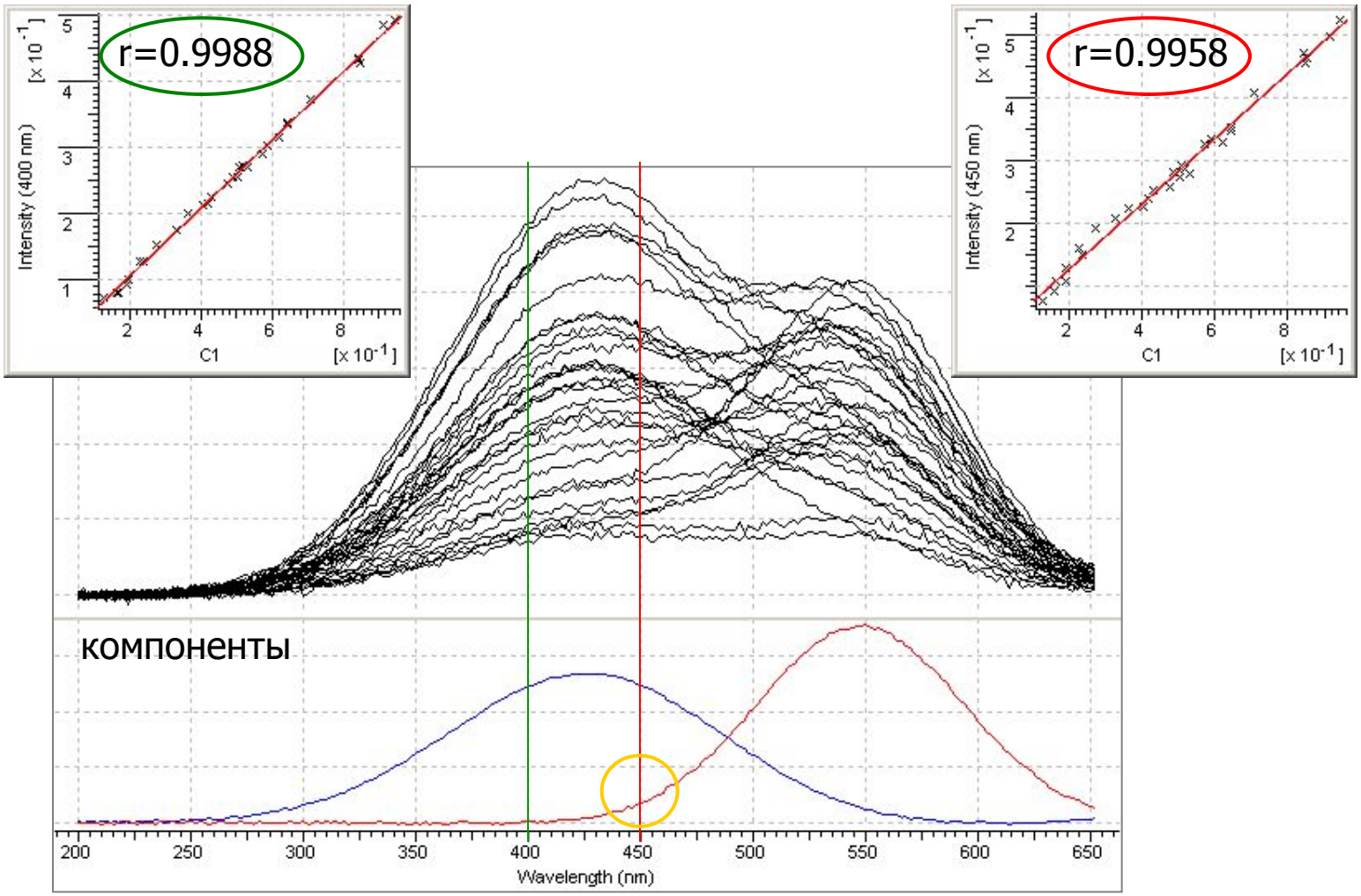
Примеры из различных областей

- ХИМИЯ: калибровка – инструмент №1 количественного анализа
- БИОЛОГИЯ: непосредственный анализ может быть губителен для живых существ
- МЕДИЦИНА: неинвазивный анализ, например, определение сахара в крови спектроскопически (ближний ИК)
- ПСИХОЛОГИЯ: анализ личности может потребовать длительных наблюдений, желательно использовать косвенные данные
- СОЦИОЛОГИЯ и ФИНАНСЫ: предсказание может быть основано только на исторических данных

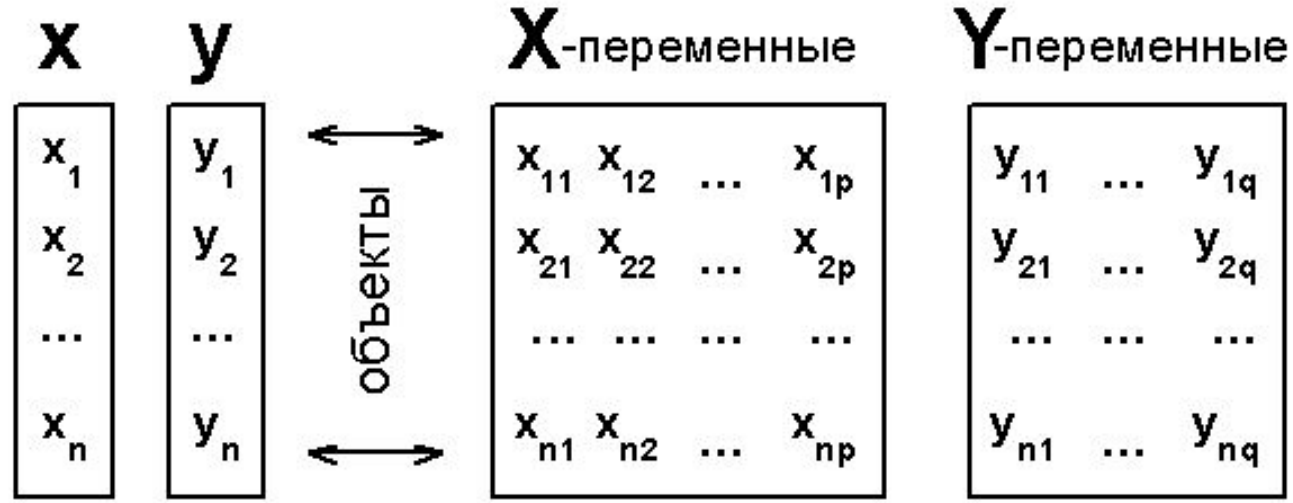
Одномерная калибровка: один компонент



Одномерная калибровка: многокомпонентная смесь



Многомерная калибровка



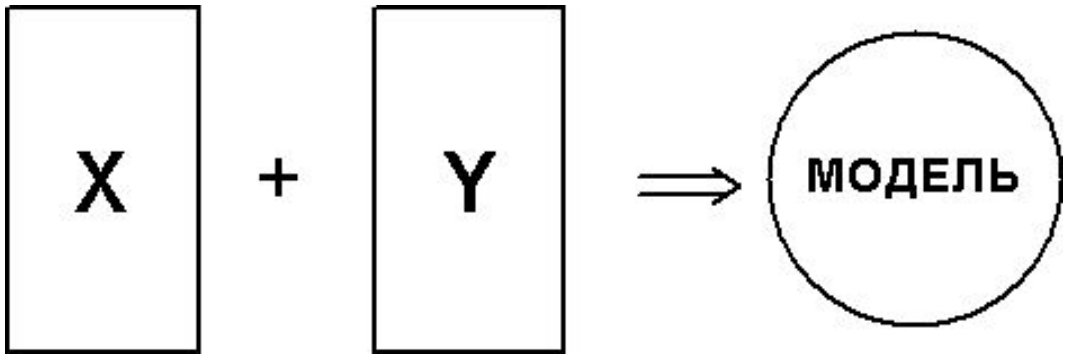
$\mathbf{y}=\mathbf{x}b+\mathbf{e}$ $\mathbf{Y}=\mathbf{XB}+\mathbf{E}$

Преимущества многомерной калибровки

- возможность анализировать несколько компонентов одновременно
- выигрыш в точности от усреднения при использования «избыточных», в т.ч. сильно коррелирующих измерений (спектры)
- возможность диагностики «плохих» образцов в процессе предсказания
- «парадигматический сдвиг» в подходах к решению проблем
  - с появлением ПЛС регрессии (PLS-R) спектроскопия ближнего ИК стала одним из наиболее популярных методов анализа

Калибровка и предсказание

Калибровка (Calibration)



Предсказание (Prediction)

Классические и инверсные методы

- Два основных подхода в многомерной калибровке:
- Классический МНК (Classical Least Squares, CLS) основан на прямом решении уравнения Бугера-Ламберта-Бера

$$\mathbf{A = C\varepsilon \mid X = Y\varepsilon}$$

- Инверсный МНК (Inverse Least Squares, ILS) решают уравнение вида

$$\mathbf{C = Ab \mid Y = Xb}$$

- В настоящей лекции – только ILS

Множественная линейная регрессия (МЛР)

Multiple Linear Regression (MLR)

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + e$$

$$y = X * b + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} * \begin{bmatrix} b_1 \\ \cdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

n - число объектов (спектров)

p - число переменных (длин волн)

$n \geq p$

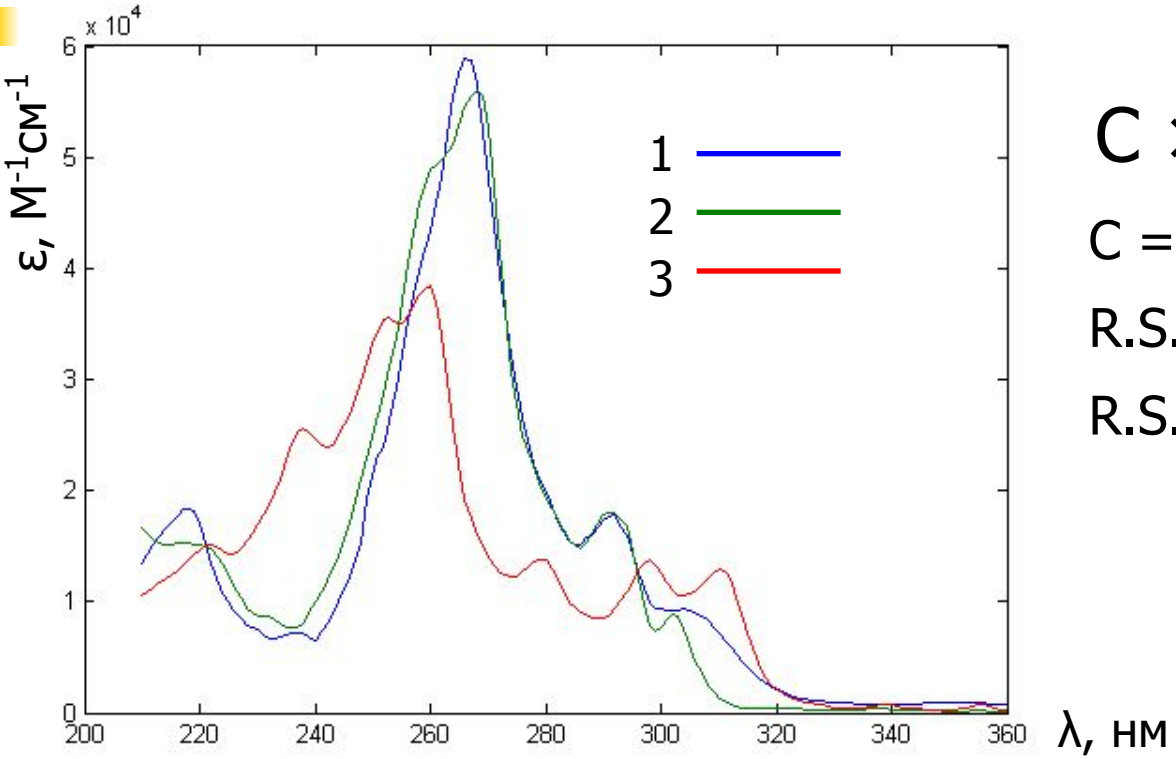
Решение: $b = (X^T X)^{-1} X^T y$

Недостатки МЛР

МЛР может не сработать, если:

- высока коллинеарность в **X** (спектры)
  - неустойчивое решение для коллинеарных даных обусловлено преобразованием $\mathbf{(X^T X)^{-1} X^T}$
- высокий уровень шума, ошибки в **X**
- переменных больше, чем образцов (типично для спектральных данных)
- есть линейная зависимость между переменными внутри **X**
- визуальная интерпретация МЛР-моделей затруднительна

Пример спектральных данных: полиароматические углеводороды



$$C \times \varepsilon + E = D$$

$C = C_0 + E_c$

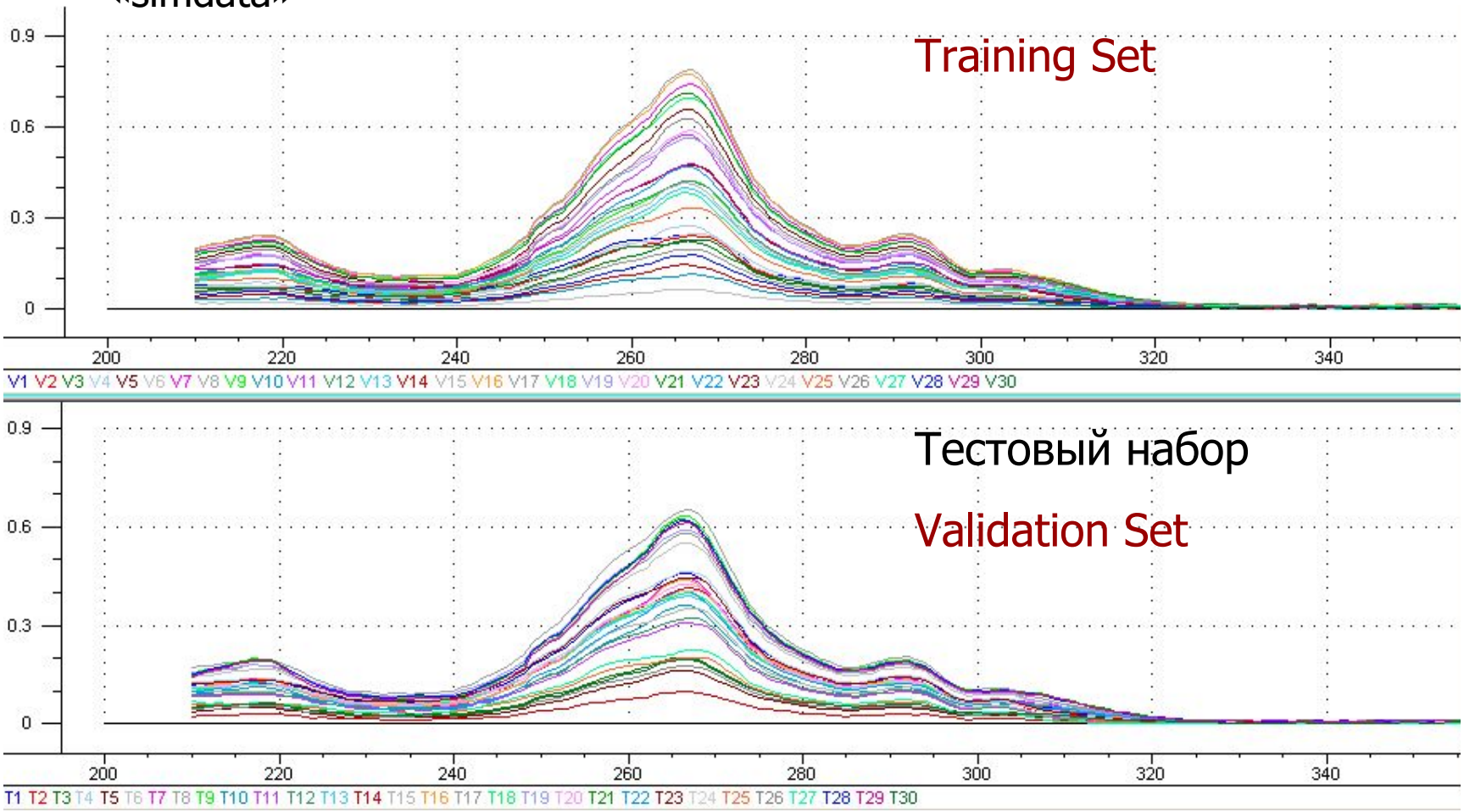
R.S.D. (E_c) = 5% (C_{max})

R.S.D. (E) = 0.001

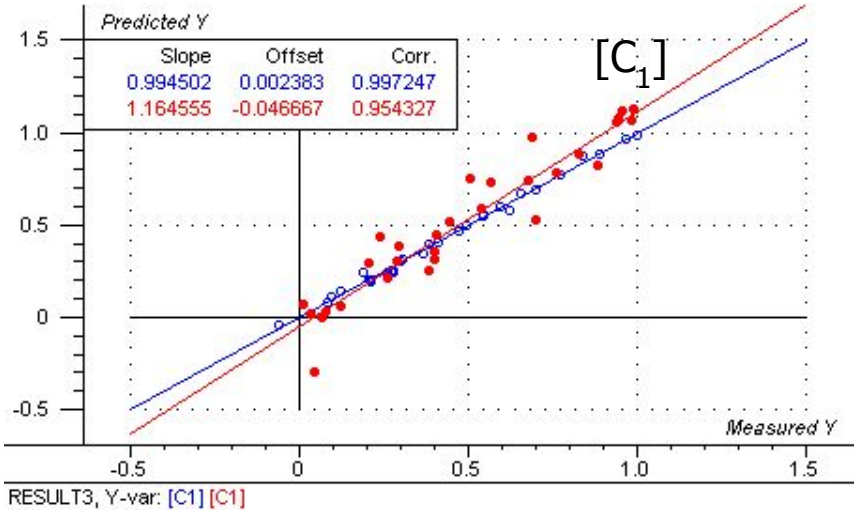
n	Компонент	$[C_n]$, М
1	2-ацетофенантрен	0 - 1
2	2-ацетиламинофенантрен	0 - 0.5
3	3-ацетиламинофенантрен	0 - 0.05

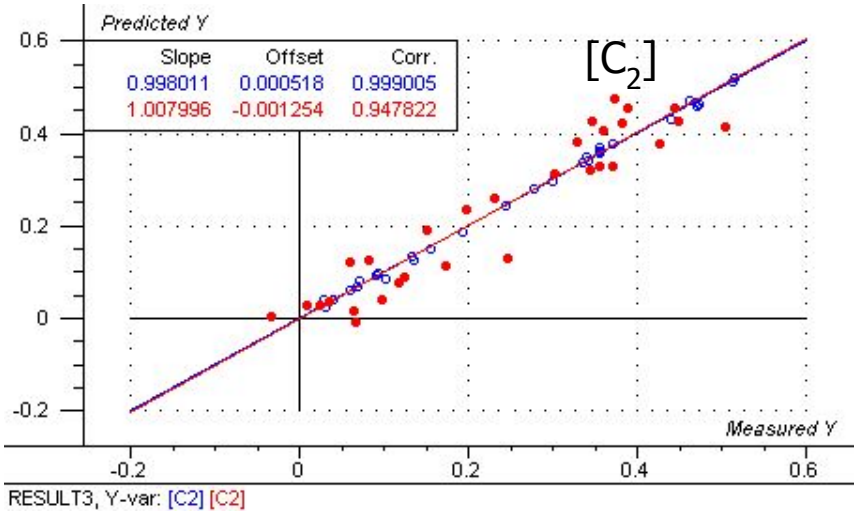
Полиароматические углеводороды: обучающий и тестовый наборы

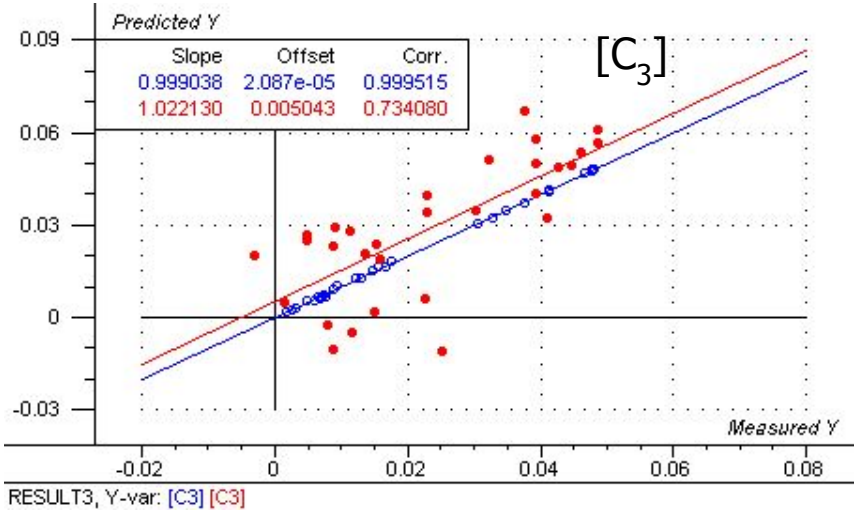
«simdata»



МЛР-калибровка (Simdata)



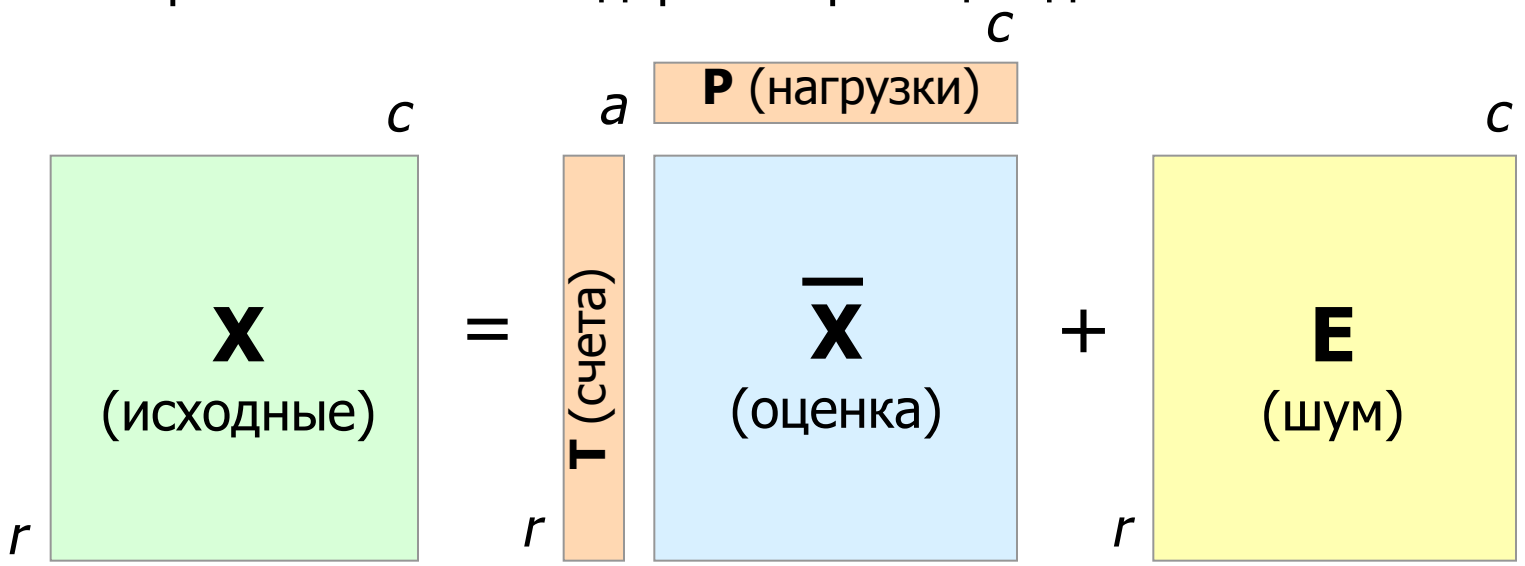




точность МЛР-модели для [C_3] (3-го компонента смеси ПАУ) неудовлетворительна

Метод Главных Компонент (МГК) - оружие против коллинеарности

- МГК (Principle Component Analysis) - преобразование: $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}$
  - счета **T** (scores) и нагрузки **P** (loadings) определяют пространство клавных компонент
  - **T** ортогональны и содержит проекции данных на ГК

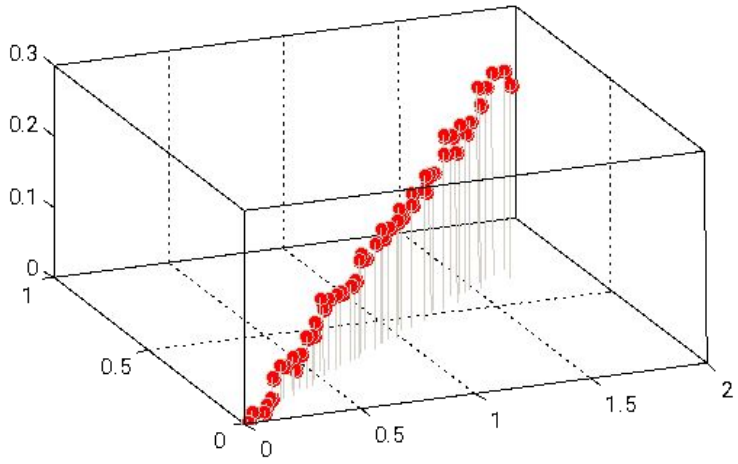


- **T** можно использовать вместо **X** для анализа (!)

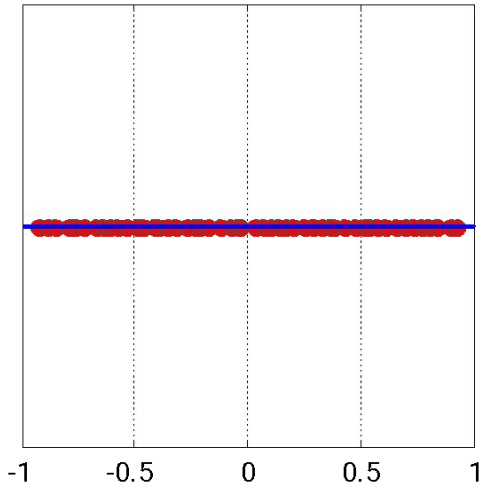
Концепция PCA «на пальцах»

X=A(522 nm)
Y=A(644 nm)
Z=A(714 nm)

X=A(430 nm)
Y=A(550 nm)
Z=A(750 nm)



МГК + МЛР = РГК! (PCA + MLR = PCR)



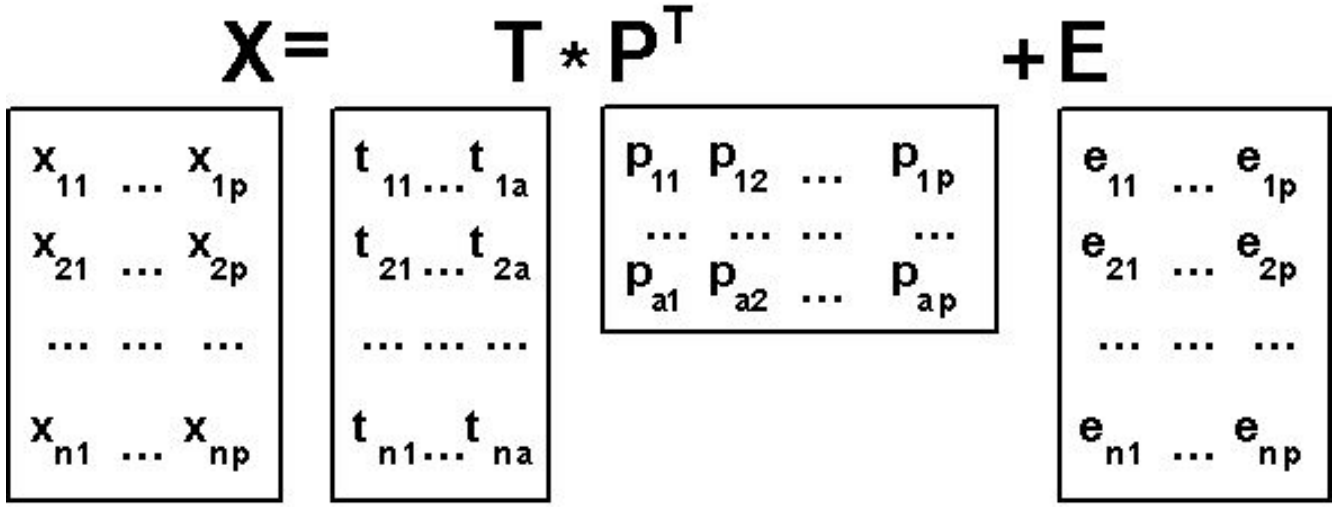
- МГК-счета (PCA scores) **T** можно использовать вместо **X** для построения МЛР-модели (MLR):

MLR: $y=\mathbf{X}b+e$ (I)	$b=[\mathbf{X}\mathbf{X}^T]^{-1}\mathbf{X}^T y$	$y_{new}=\mathbf{X}_{new}b$
PCR: $y=\mathbf{T}b+e$ (II)	$b=[\mathbf{T}\mathbf{T}^T]^{-1}\mathbf{T}^T y$	$y_{new}=\mathbf{T}_{new}b$

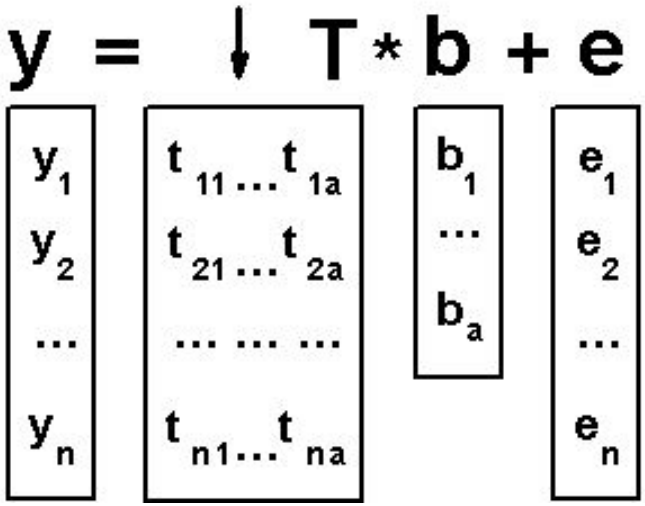
- Метод называется: регрессия на главные компоненты,

Схема РГК (PCR) – подробнее

PCA:



MLR:



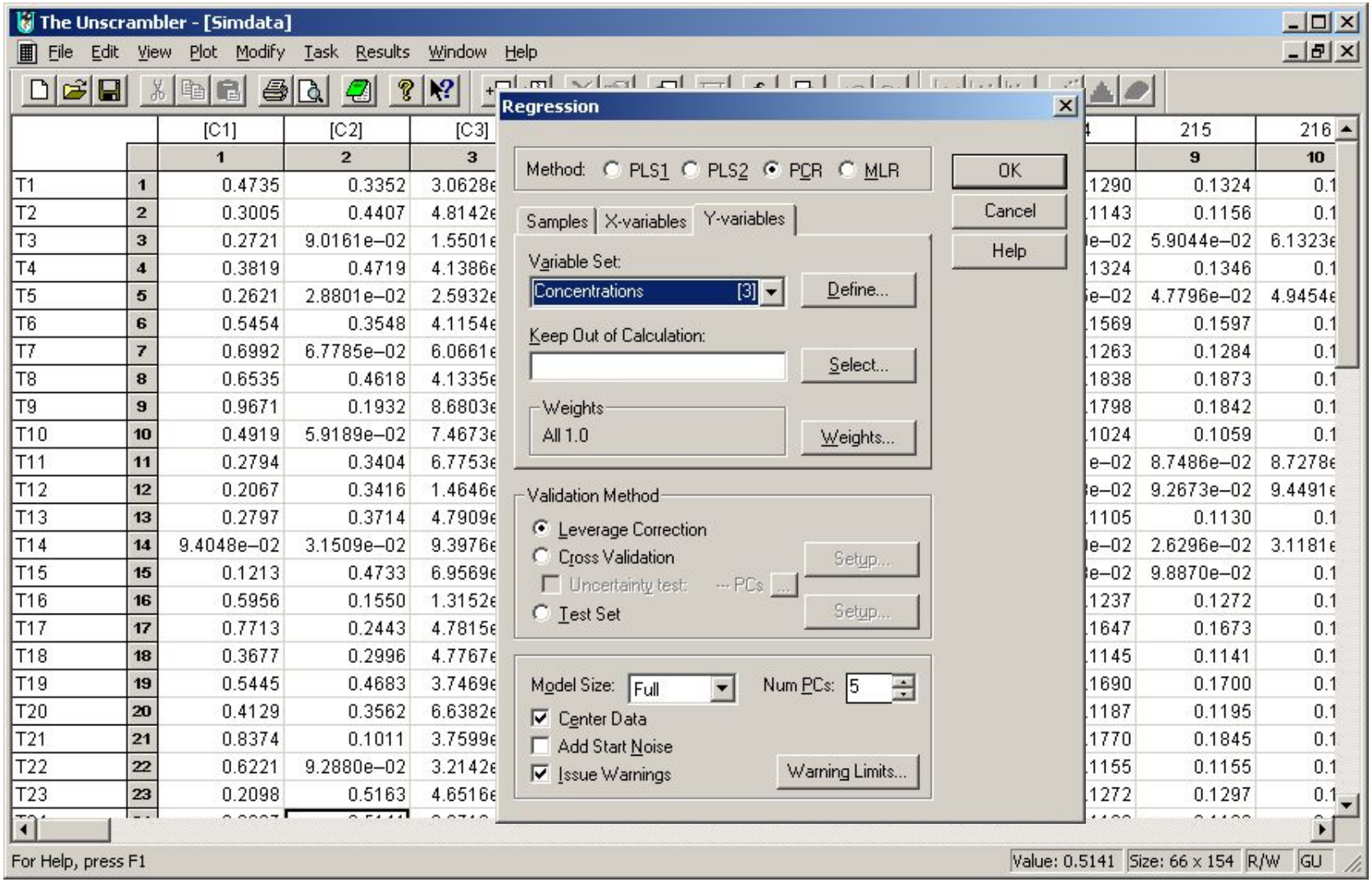
n - объектов
p - переменных
a - главных компонент

$a \leq \min(n,p)$

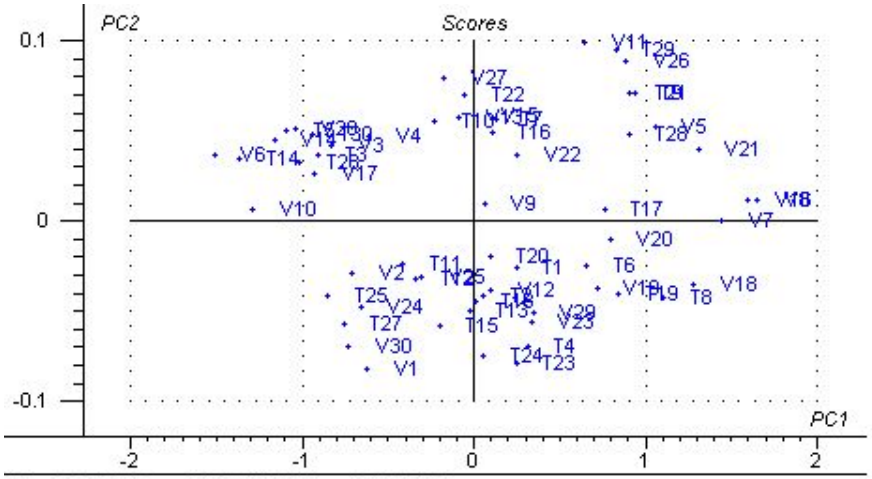
Интерпретация РГК-модели

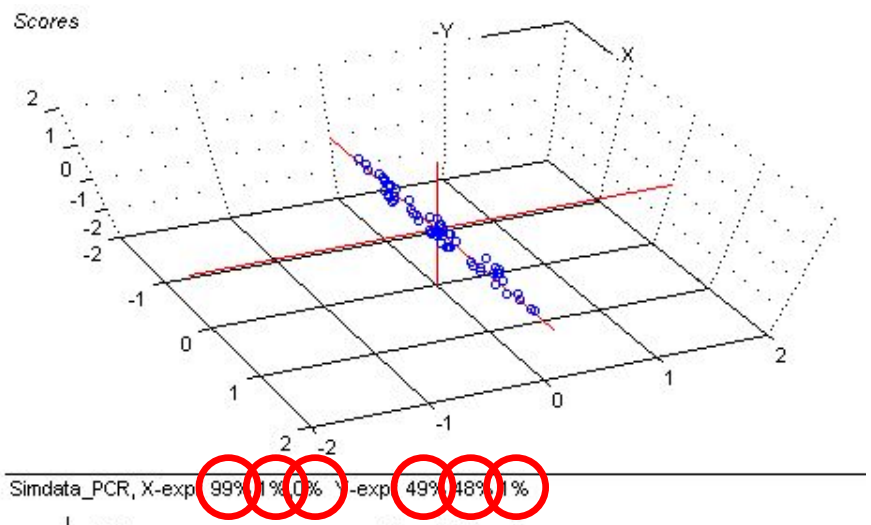
- интерпретация модели служит для изучения внутренней структуры данных:
  - группы
  - выбросы
  - связь между **X** и **Y**
- инструменты диагностики МГК (PCA) работают в РГК (PCR):
  - график счетов (scores)
  - график нагрузок (loadings)
  - график счетов и нагрузок вместе (bi-plot)
  - график остатков (residuals)
- инструменты диагностики РГК:
  - совместный график нагрузок **X** и **Y**

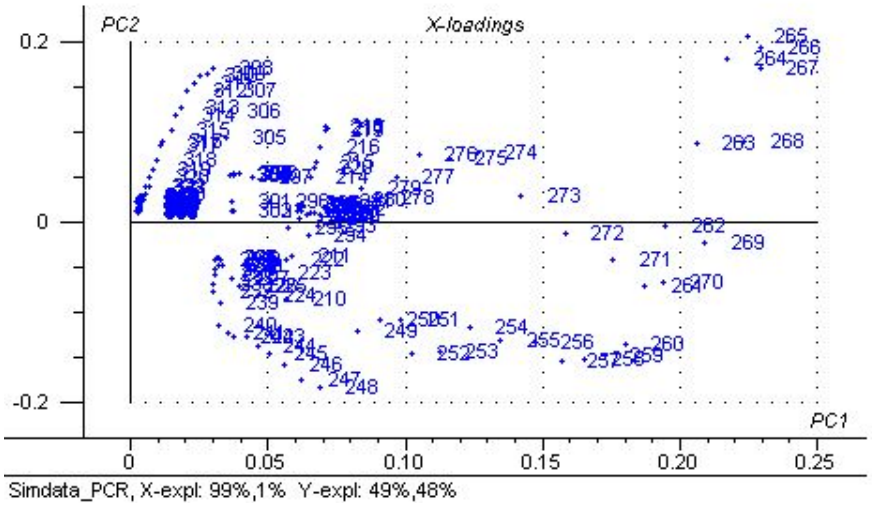
Строим РГК-модель (Simdata)

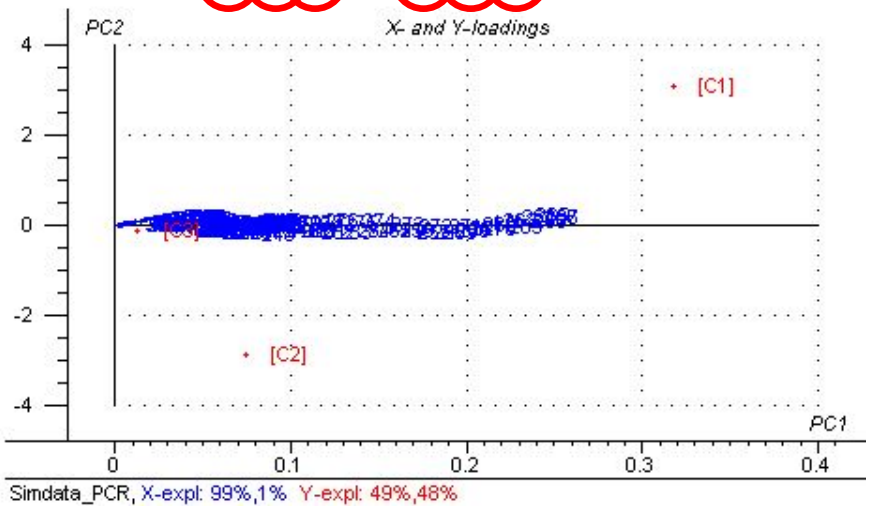


Строим РГК-модель (simdata)









Проверка (валидация) модели

- проверка (validation) модели служит для:
  - определения размерности модели (числа ГК)
  - оценки предсказательной способности модели
- проверка модели производится с помощью тестовых данных:
  - того же диапазона и того же качества что обущающие данные (та же генеральная выборка)
  - достаточно представительные
- или кросс-валидации (cross-validation)
  - полная (leave-one-out, LOO)
  - сегментная (например, Venetian blind)

Среднеквадратичная ошибка предсказания (RMSEP)

- RMSEC = Root Mean Square Error of Calibration

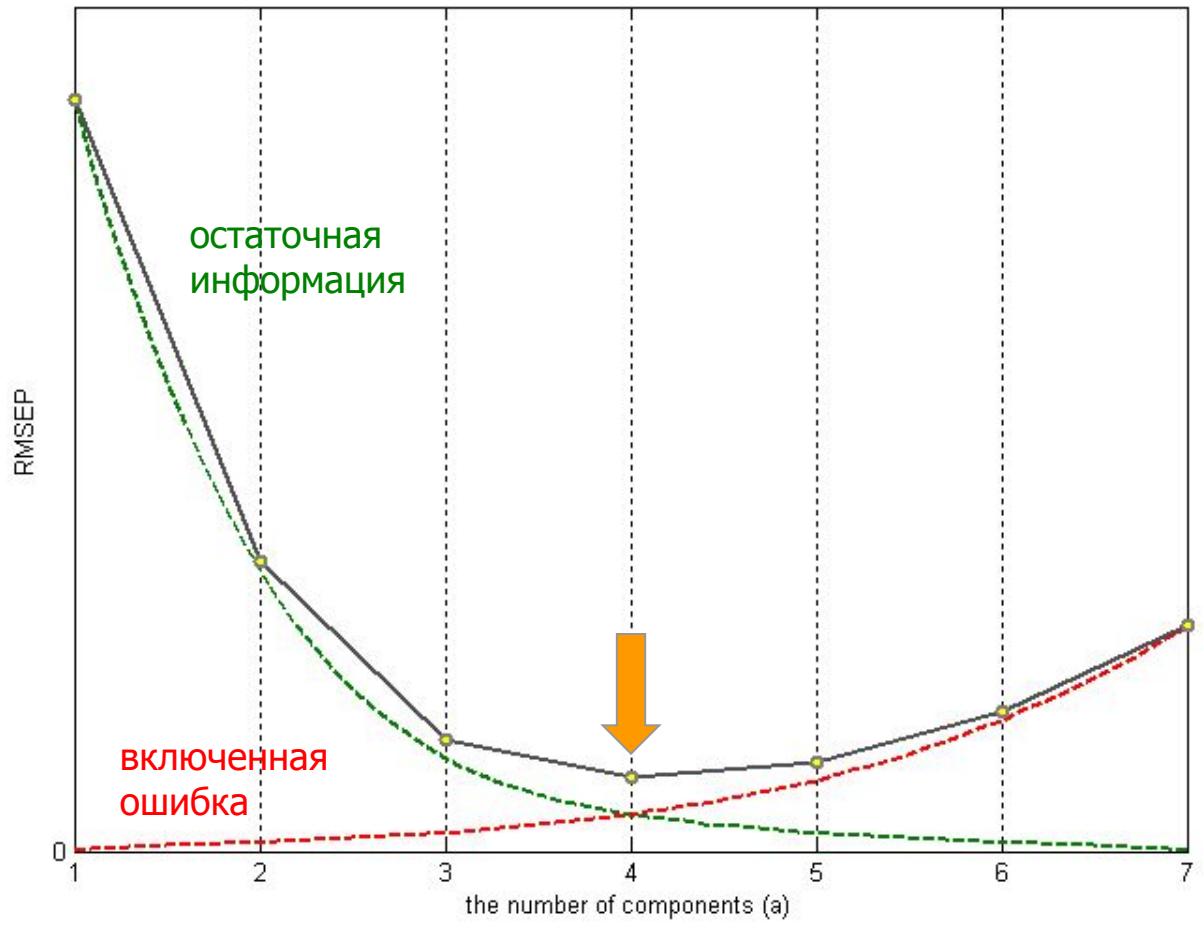
$$RMSEC = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i^{cal} - y_i^{cal}\right)^2}{n}}$$

- RMSEP = Root Mean Square Error of Prediction

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i^{val} - y_i^{val}\right)^2}{n}}$$

- минимум на кривой RMSEP – основной индикатор числа ГК
- RMSEP – оценка точности в единицах измерения (!)
- RMSEP используется для сравнения моделей

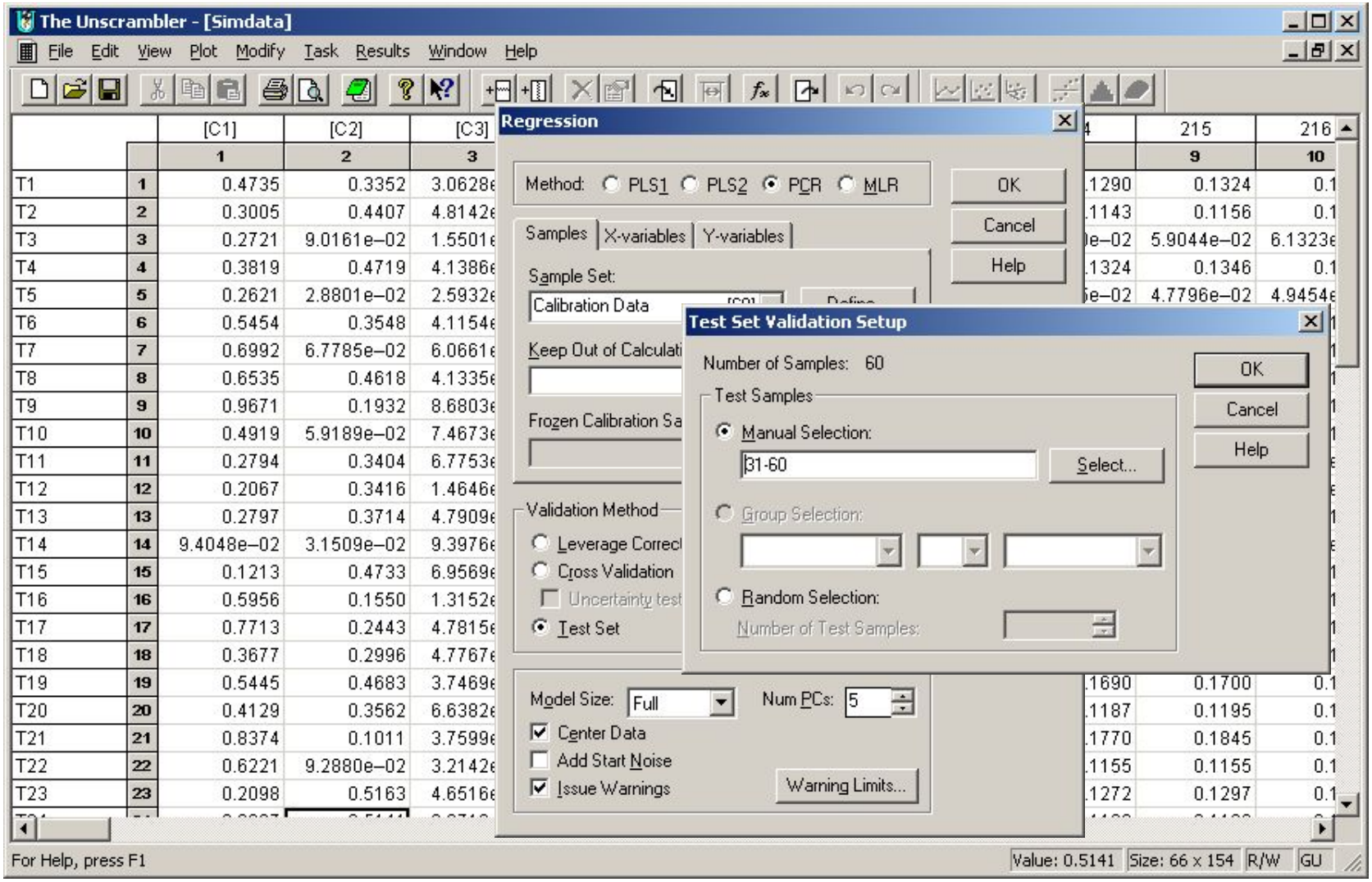
Число компонент: почему минимум на кривой RMSEP?



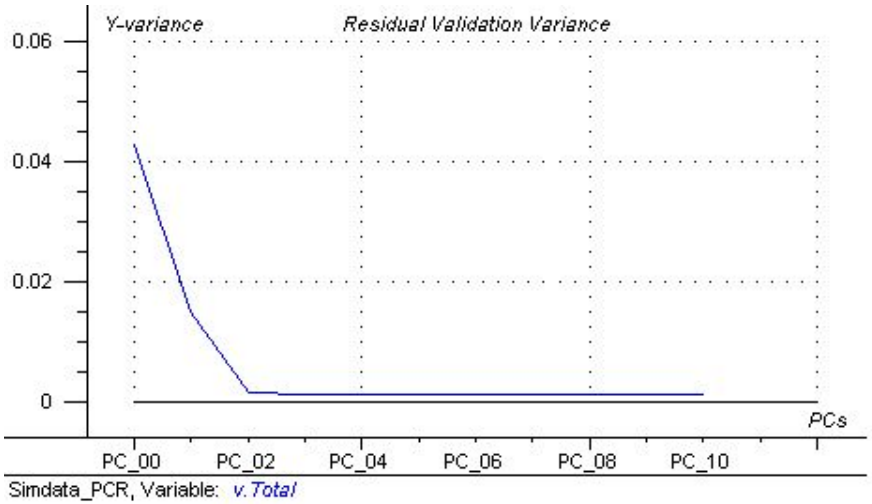
Оценка числа компонент в РГК

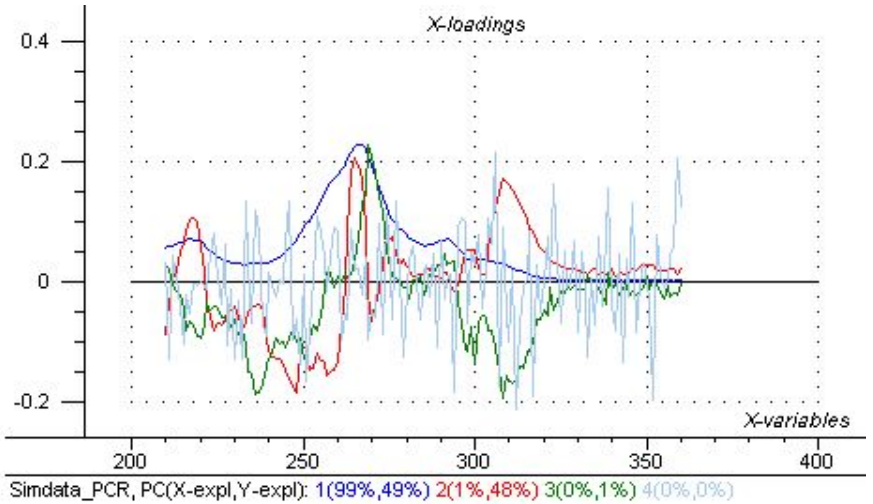
- правильный выбор числа главных компонент (principle components, PC) - ключевая проблема многомерной калибровки
  - модель с недостаточным числом ГК (underfitting) не использует всей полезной информации из данных
  - модель с избыточным числом ГК (overfitting) начинает моделировать шум (ошибку)
- найти оптимальную размерность помогают тестовые данные (validation set)

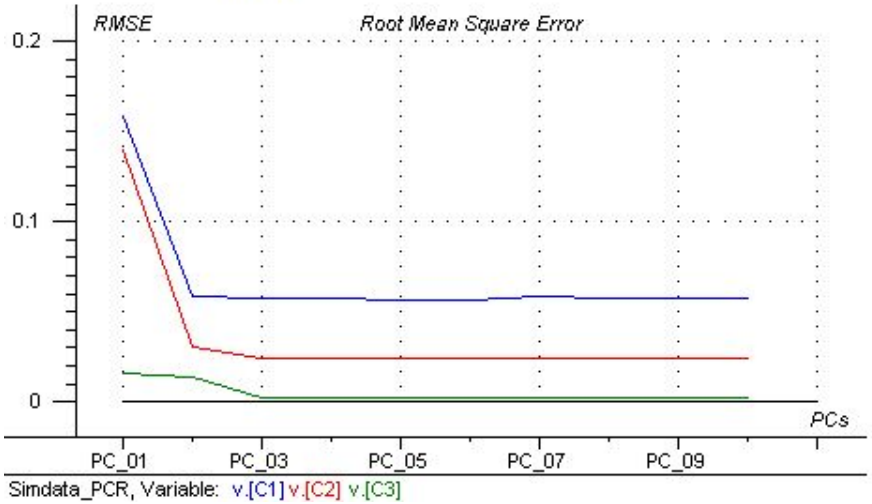
Число компонент: РГК - simdata

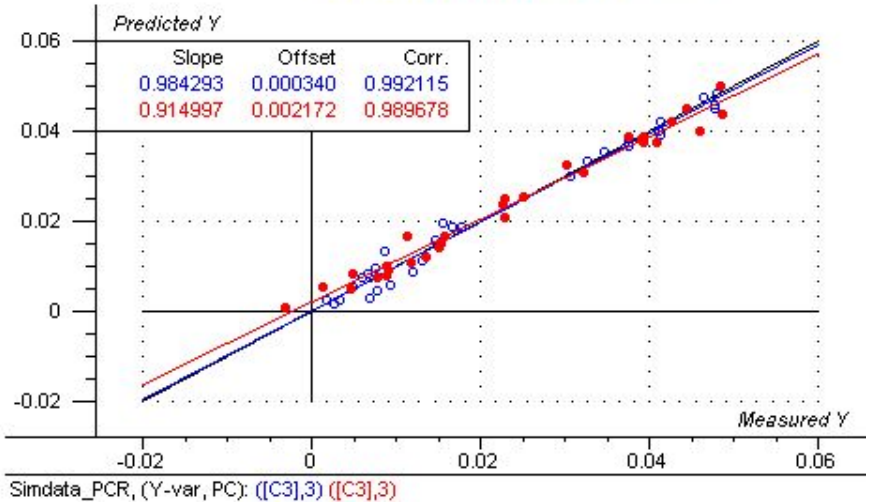


Число компонент: РГК - simdata









Оценка числа ГК в РГК: особенности

- число главных компонент (размерность модели) определяется в РГК (PCR) нуждами калибровки, и не обязательно совпадает с результатом МГК (PCA)
- Особенности:
  - в РГК есть RMSEP
  - активно используется тестовые данные (test set)
- минимум на кривой RMSEP - основной индикатор числа ГК
- для спектральных данных показательной может быть форма X-нагрузок (X-loadings)
- решение всегда за экспертом!

Несовершенства РГК

- РГК (PCR) – мощный метод многомерной калибровки
- имеет безусловные преимущества перед MLR
- однако, не вполне оптимизирован для калибровки
- пространство ГК не учитывает структуры **Y** и связи между **X** и **Y**
- можно ли учесть эту связь при построении проекционной модели?
- да, это делает PLS!

Факторные пространства

- уравнение PCA имеет универсальный смысл:

  $\mathbf{X = TP^T + E}$

- преобразование называется факторной компрессией, проекцией данных на факторное пространство (factor space)
  - парные вектора в **T** и **P** называются факторами (factors)
- главные компоненты – важный пример факторного пространства, но не единственный
- факторное пространство можно оптимизировать для решения конкретной задачи
  - ГК (PC) оптимальны для исследования структуры **X**
  - как оптимизировать пространство для калибровки?

PLS – мощная альтернатива PCR

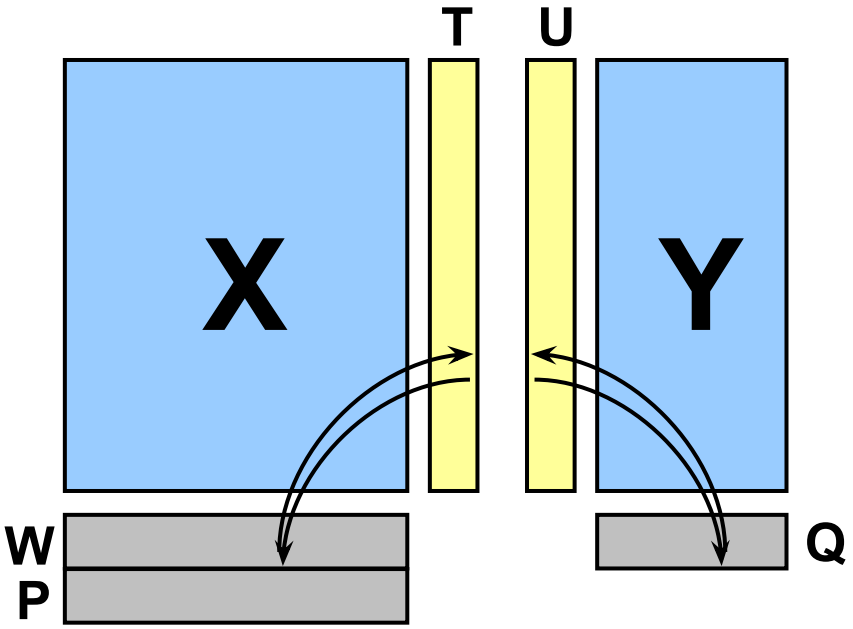
- Метод проекции на латентные структуры (ПЛС) и ПЛС-регрессия (ПЛС-Р)
  - PLS = Partial Least Squares ->
  - = Projection on Latent Structures
- ПЛС-пространство создается при участии двух переменных **X** и **Y** одновременно
  - критерий – моделирование той структуры (информации) в **X**, которая коррелирует с **Y**
  - например, спектральные полосы (**X**), которые отвечают за концентрацию компонента(ов), заданные в **Y**, получат в подели больший вес
- метод ПЛС оптимизирован для регрессионного анализа

ПЛС-регрессия: схематическое представление

- участвуют обе матрицы **X** и **Y**
- факторы рассчитываются по очереди – алгоритм NIPALS
- => 2 набора счетов (scores) **T**, **U** и нагрузок (loadings) **P**, **Q** плюс матрица **W** взвешенных нагрузок (loading-weights)
- итерационное улучшение модели, чтобы максимизировать *cov(T,U)*
- Предсказание: $\hat{Y} = T_{new} B_t$

  $\hat{Y} = X_{new} B$

  $B = W(P^T W)^{-1} Q^T$



$$X = TP^T + E_x$$

$$Y = UQ^T + E_y$$

[1] S. Wold, H. Martens, H. Wold, Lecture Notes Math. **973** (1983) 286–293

Две разновидности ПЛС: ПЛС1 и ПЛС2

- существуют две популярных разновидности ПЛС: ПЛС1 (PLS1) и ПЛС2 (PLS2)
- ПЛС1 модель строится для единственной переменной **y** (свойства), например, для концентрации одного компонента смеси
- если нужна калибровка по нескольким свойствам, строится несколько независимых моделей
- ПЛС2 рассчитывается для нескольких свойств одновременно
- расчетные алгоритмы методов отличаются соответственно

Основы алгоритма ПЛС

- ПЛС-декомпозиция производится алгоримом NIPALS
  - NIPALS = Non-linear Iterative Partial Least Squares
- факторы находятся по очереди, один за другим, расчет всех факторов (как в SVD) не обязателен
- итерационная замена векторов $\mathbf{u}_f$ -> $\mathbf{t}_f$ и $\mathbf{u}_f$ -> $\mathbf{t}_f$ для нахождения текущего фактора *f* - алгоритмическая основа ПЛС2
- алгоритм работает до выполнения критерия сходимости
- ознакомимся с принципиальной схемой, начиная с более общего ПЛС2

NIPALS алгоритм для ПЛС2

0.	$\mathbf{u_f}$	выбор начального приближения **u**
1.	$\mathbf{w_f} = \mathbf{X_f^T u_f} / \lvert \mathbf{X_f^T u_f} \rvert$	расчет нормализованного вектора взвешенных нагрузок **w**
2.	$\mathbf{t_f} = \mathbf{X_f^T w_f}$	расчет вектора весов **t**
3.	$\mathbf{q_f} = \mathbf{Y_f^T t_f} / \lvert \mathbf{Y_f^T t_f} \rvert$	расчет нормализованного вектора нагрузок **q**
4.	$\mathbf{u_f} = \mathbf{Y_f^T q_f}$	расчет вектора счетов **u**
5.	$\lvert \mathbf{t_{f.new}} - \mathbf{t_{f.old}} \rvert < \mathbf{lim?}$	проверка сходимости: да -> go to 1.
6.	$\mathbf{p_f} = \mathbf{X_f^T t_f} / \mathbf{t_f^T t_f}$	расчет вектора весов **p**
7.	$b_f = \mathbf{u_f^T t_f} / \mathbf{t_f^T t_f}$	расчет внутненнего коэффициента регрессии b
8.	$\mathbf{X_{f+1}} = \mathbf{X_{f+1}} - \mathbf{t_f^T p_f}$ $\mathbf{Y_{f+1}} = \mathbf{Y_{f+1}} - b_f \mathbf{t_f q_f^T}$	расчет остатка **X** и **Y**
9.	f = f + 1	Переход к следующему фактору

NIPALS алгоритм для ПЛС1

1.	$\mathbf{w_f} = \mathbf{X_f^T y_f} / \lvert \mathbf{X_f^T y_f} \rvert$	расчет нормализованного вектора взвешенных нагрузок **w**
2.	$\mathbf{t_f} = \mathbf{X_f^T w_f}$	расчет вектора весов **t**
3.	$q_f = \mathbf{y_f^T t_f} / \lvert \mathbf{t_f^T t_f} \rvert$	расчет нагрузки q (скаляр) фактора f
4.	$\mathbf{p_f} = \mathbf{X_f^T t_f} / \mathbf{t_f^T t_f}$	расчет вектора весов **p**
5.	$\mathbf{X_{f+1}} = \mathbf{X_{f+1}} - \mathbf{t_f}^T \mathbf{p_f}$ $\mathbf{y_{f+1}} = \mathbf{y_{f+1}} - q_f \mathbf{t_f}$	расчет остатка **X** и **y**
6.	f = f + 1	переход к следующему фактору

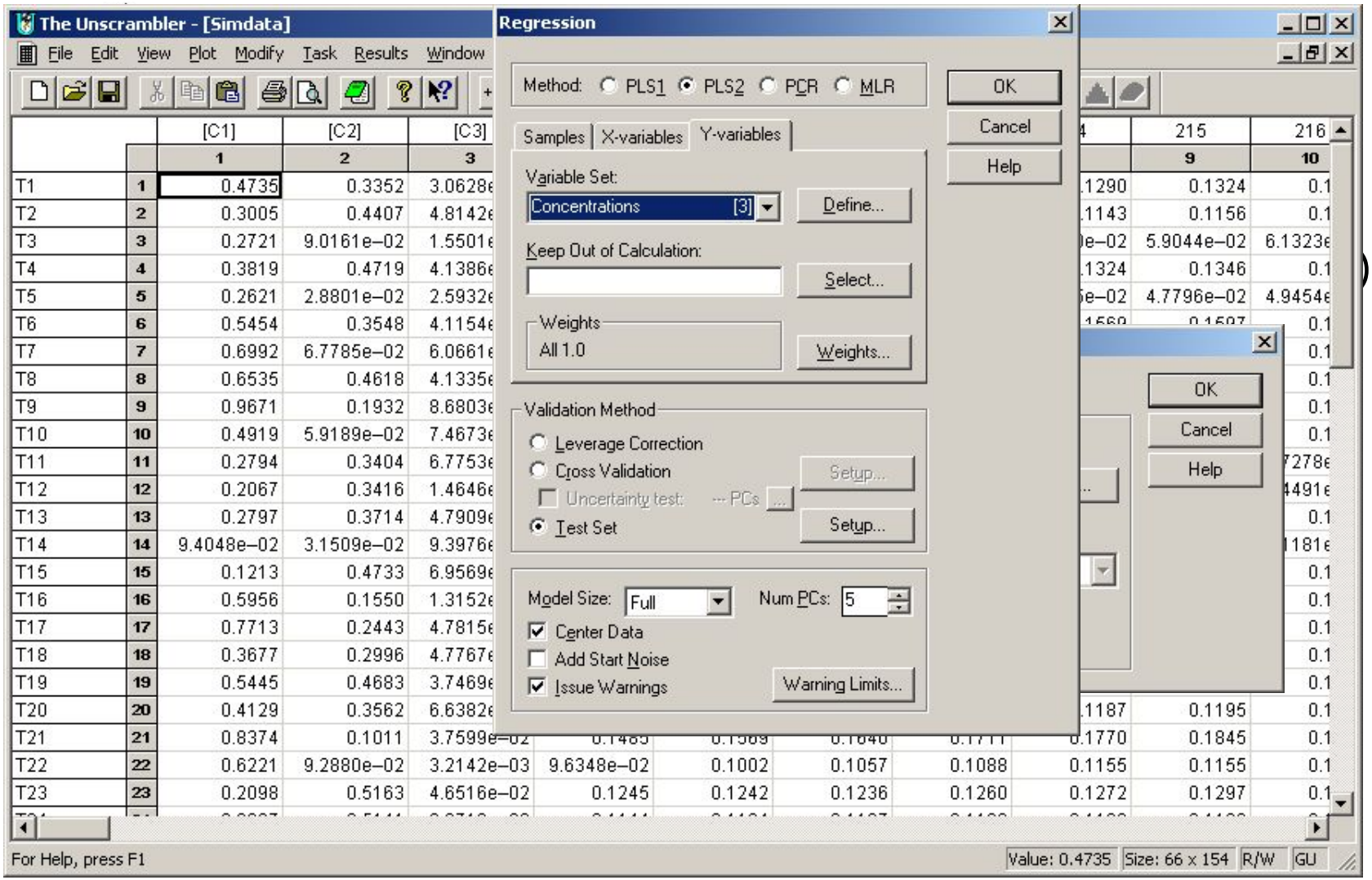
NIPALS алгоритм для ПЛС1

1.	$\mathbf{w_f} = \mathbf{X_f^T y_f} / \lvert \mathbf{X_f^T y_f} \rvert$	расчет нормализованного вектора взвешенных нагрузок **w**
2.	$\mathbf{t_f} = \mathbf{X_f^T w_f}$	расчет вектора весов **t**
3.	$q_f = \mathbf{y_f^T t_f} / \lvert \mathbf{t_f^T t_f} \rvert$	расчет нагрузки q (скаляр) фактора f
4.	$\mathbf{p_f} = \mathbf{X_f^T t_f} / \mathbf{t_f^T t_f}$	расчет вектора весов **p**
5.	$\mathbf{X_{f+1}} = \mathbf{X_{f+1}} - \mathbf{t_f^T p_f}$ $\mathbf{y_{f+1}} = \mathbf{y_{f+1}} - q_f \mathbf{t_f}$	расчет остатка **X** и **y**
6.	f = f + 1	переход к следующему фактору

ПЛС1 и ПЛС2

- ПЛС1 моделирует только одну переменную **y** «за раз»
- ПЛС2 позволяет моделировать любую комбинацию переменных **Y** без их разделения – совместно
  - он кажется более подходящим при калибровке нескольких свойств…
  - однако, ПЛС1 дает по отдельной модели на каждое из интересующих свойств, возможно, с различным числом факторов
- не будет ли набор независимых моделей всегда лучшим решением?
  - однозначного ответа нет…
  - сравним методы на практике!

Строим ПЛС2-модель (Simdata)

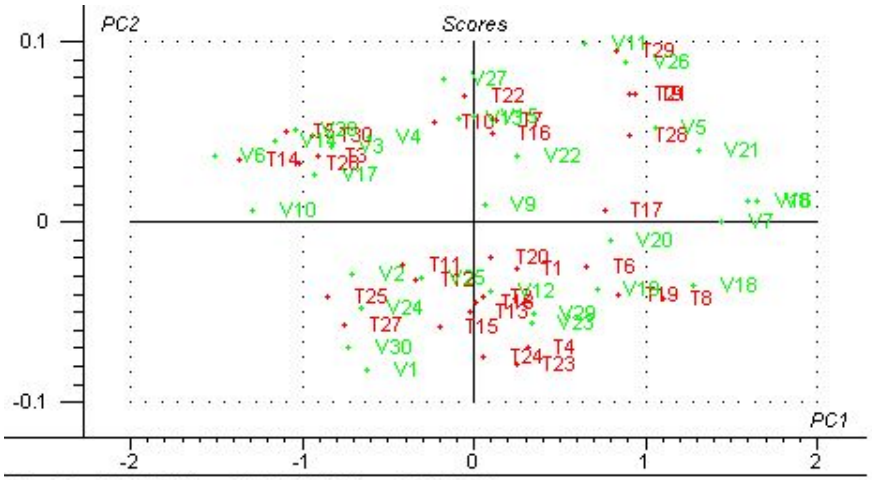


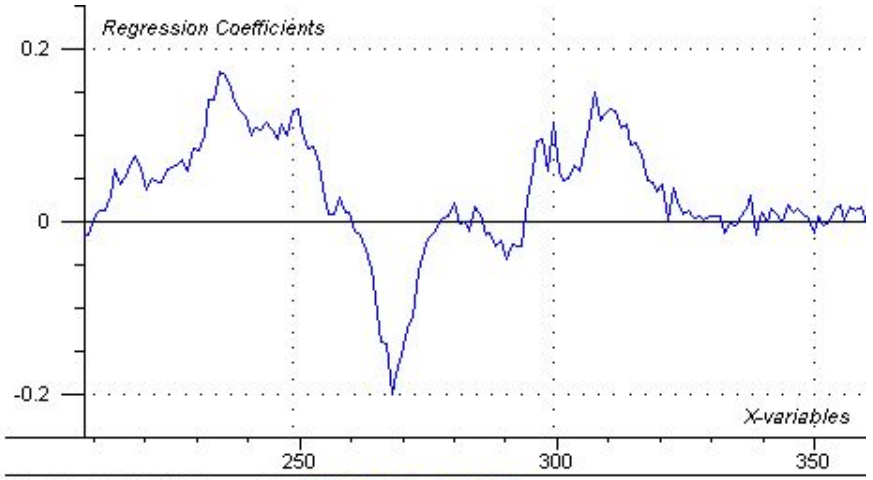
Интерпретация ПЛС-моделей

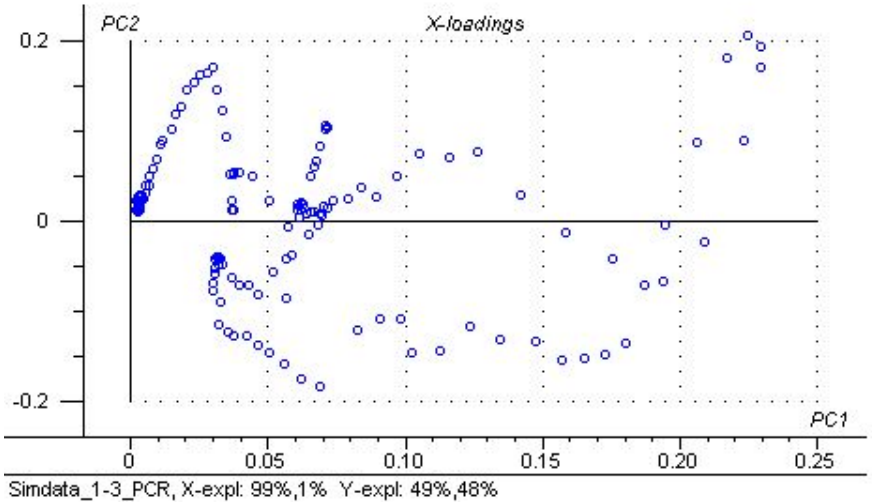
- Интерпретация модели служит для изучения внутренней структуры данных
  - группы
  - выбросы
  - взаимовсвязи
- Сходство с РГК (PCR):
  - X-счета и нагрузки (scores & loadings)
- Особенности:
  - график **t** – **u** : метод обнаружения выбросов (outliers)
  - графики нагрузок **w** – **w** : карта переменных
  - сравнение двух X-нагрузок **p** – **w** : насколько Y повлияла на декомпозицию X
  - график **w** – **q**

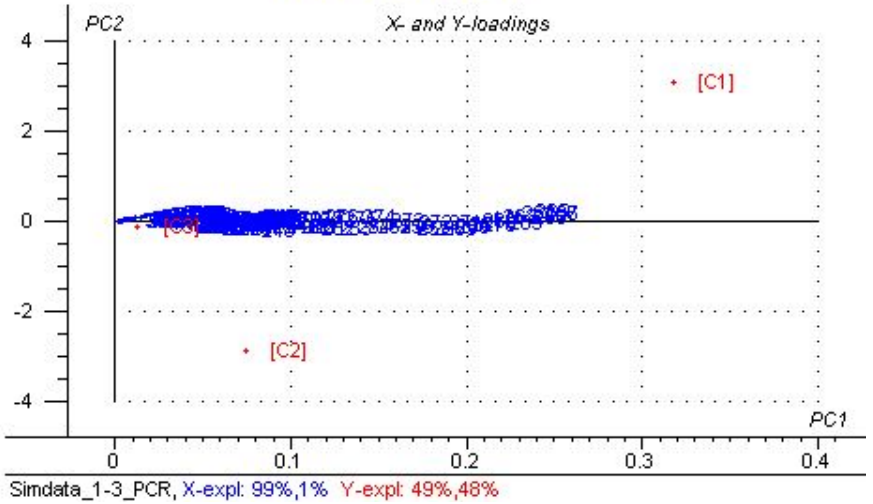
Интерпретация моделей: ПЛС2 против РГК

PCR

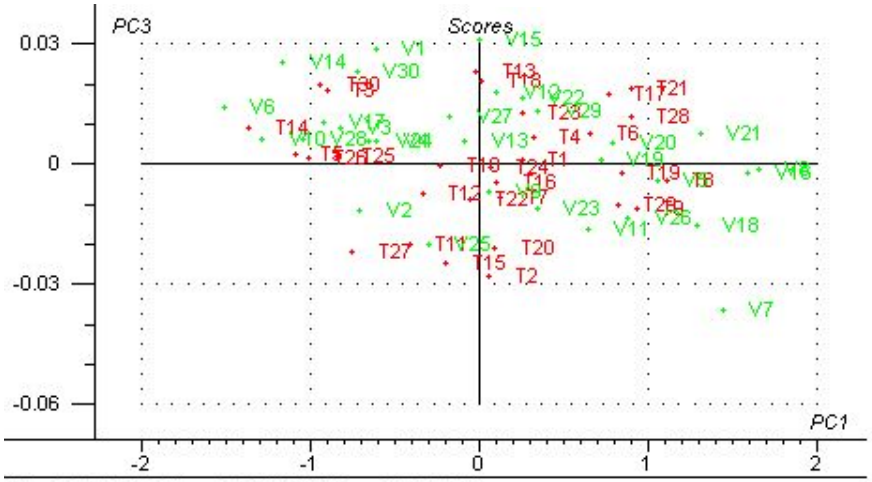


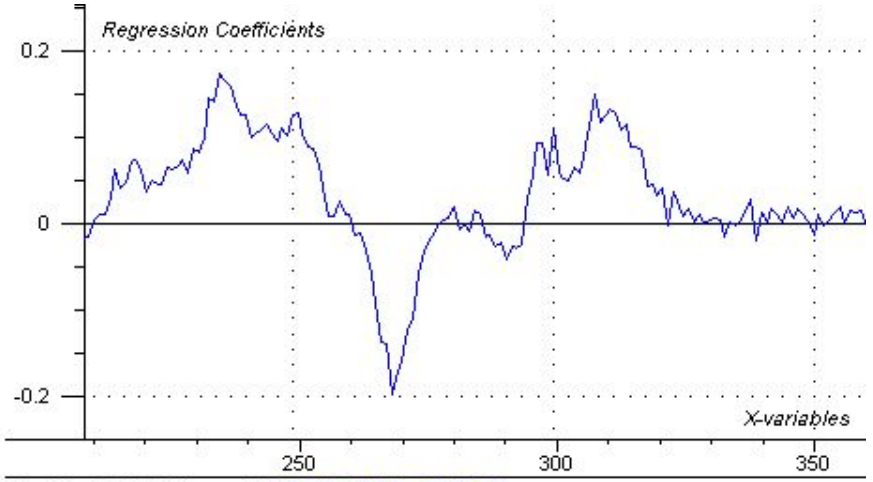


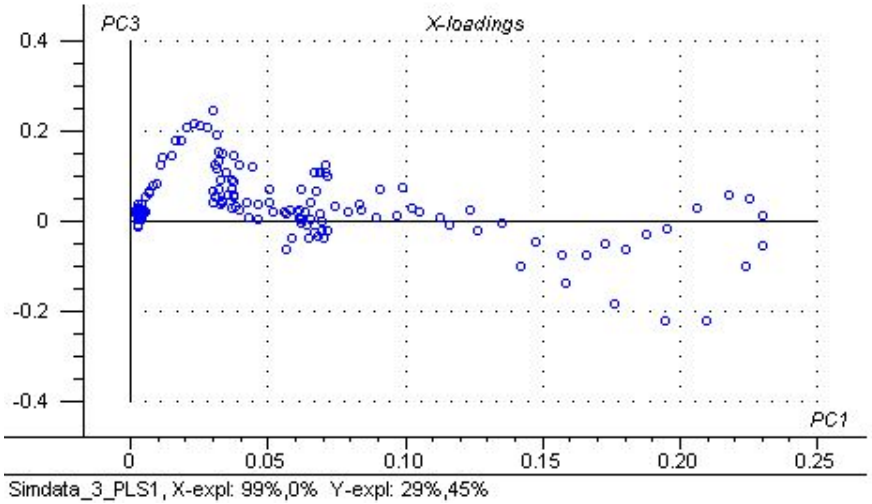


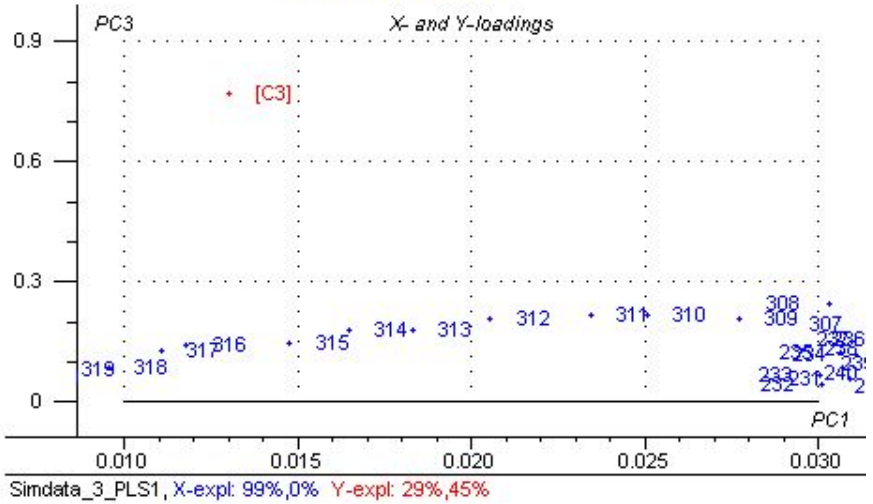


Интерпретация моделей: ПЛС1 против ПЛС2

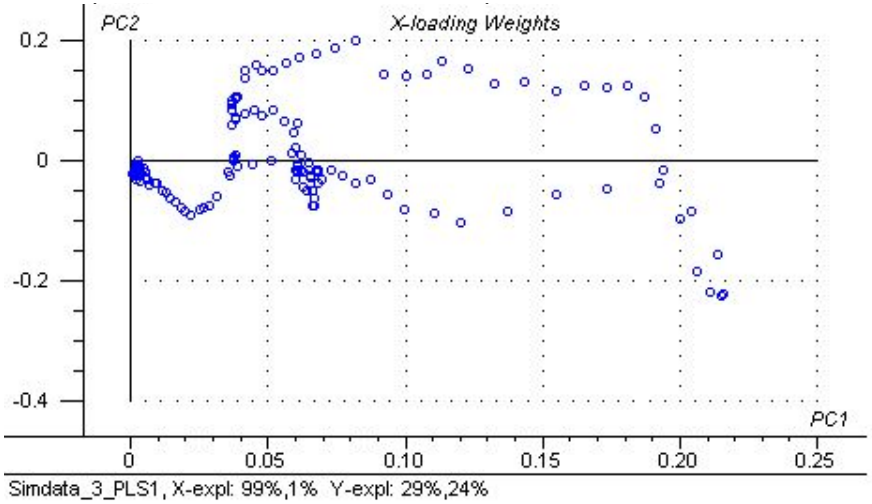


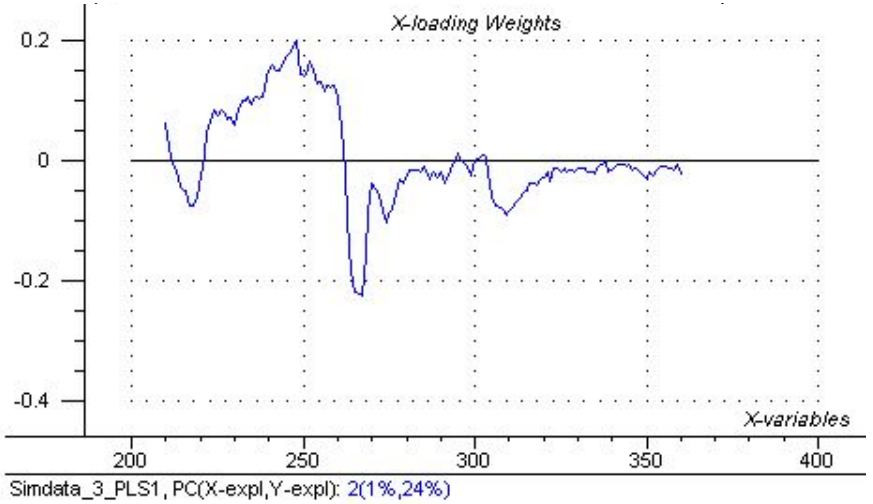


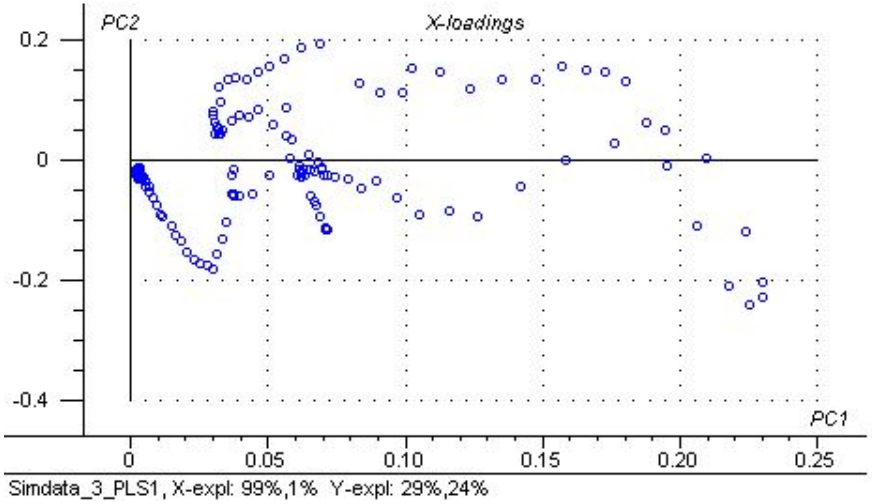


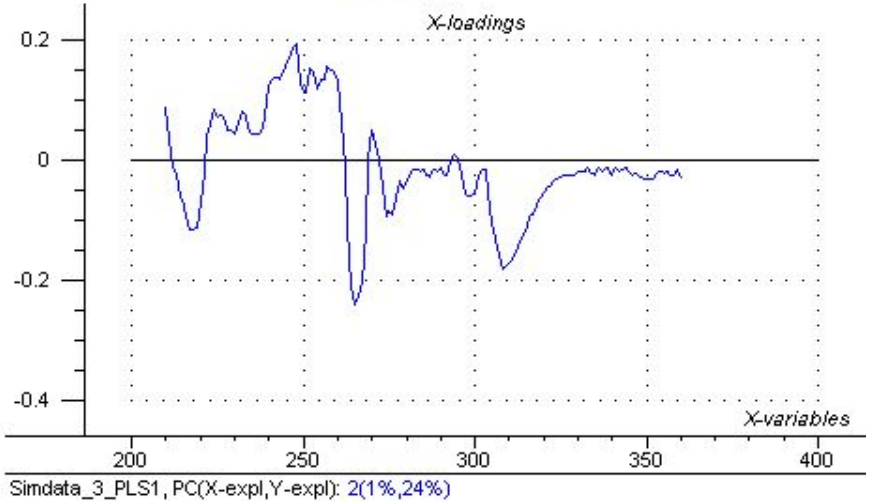


Интерпретация ПЛС-моделей: связь **X** и **Y** (Simdata)

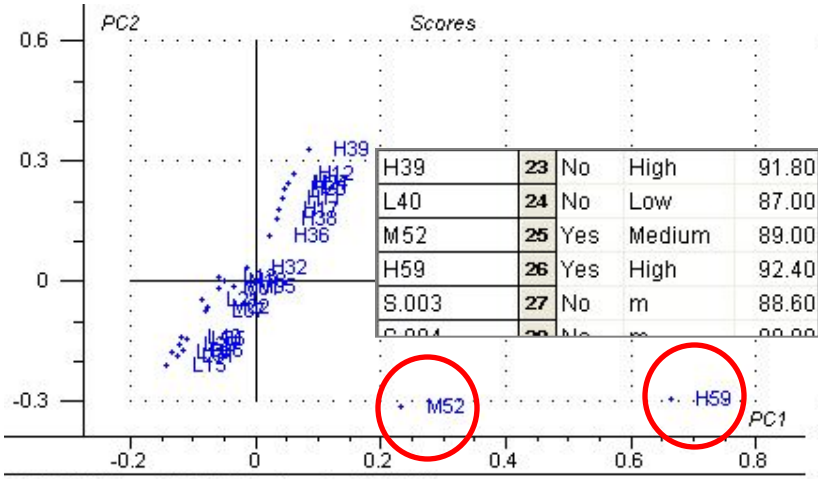




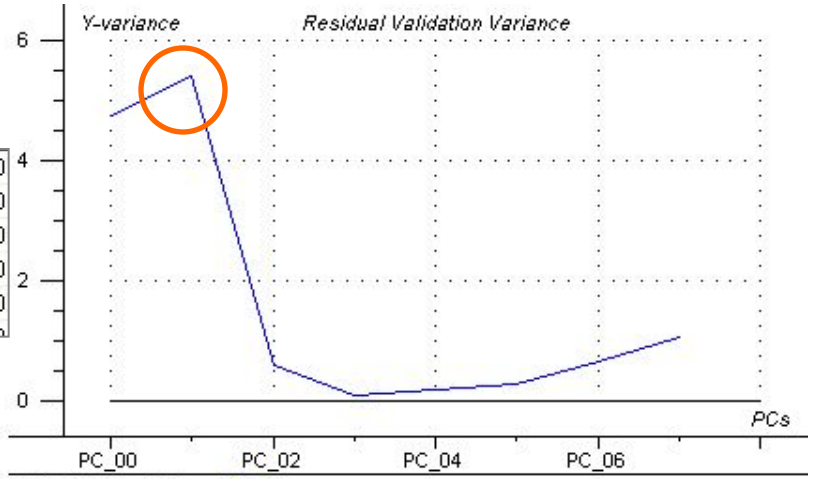


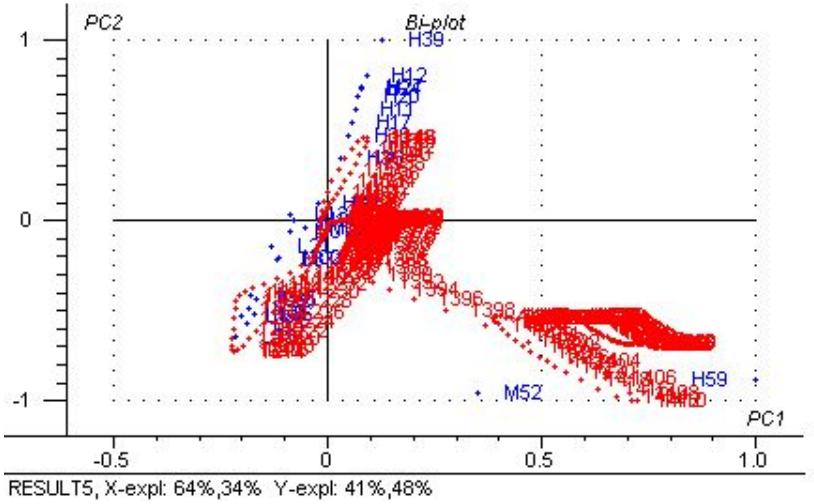


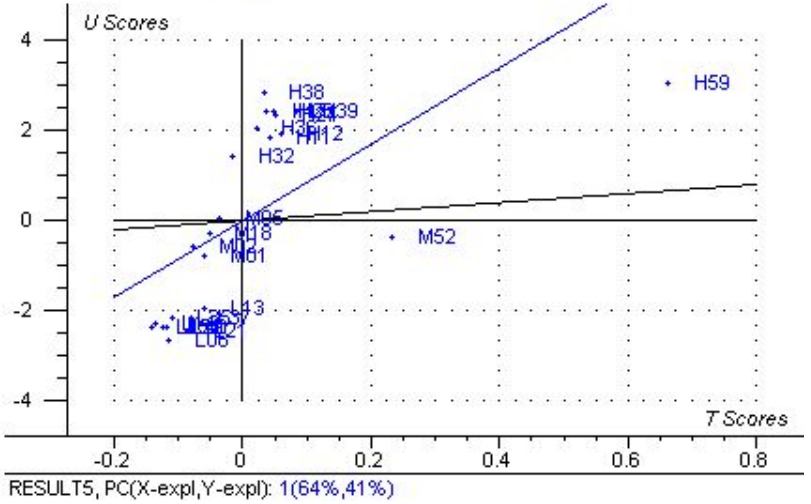
Интерпретация ПЛС-модели: выбросы (Octane)



H39	23	No	High	91.80
L40	24	No	Low	87.00
M52	25	Yes	Medium	89.00
H59	26	Yes	High	92.40
S.003	27	No	m	88.60



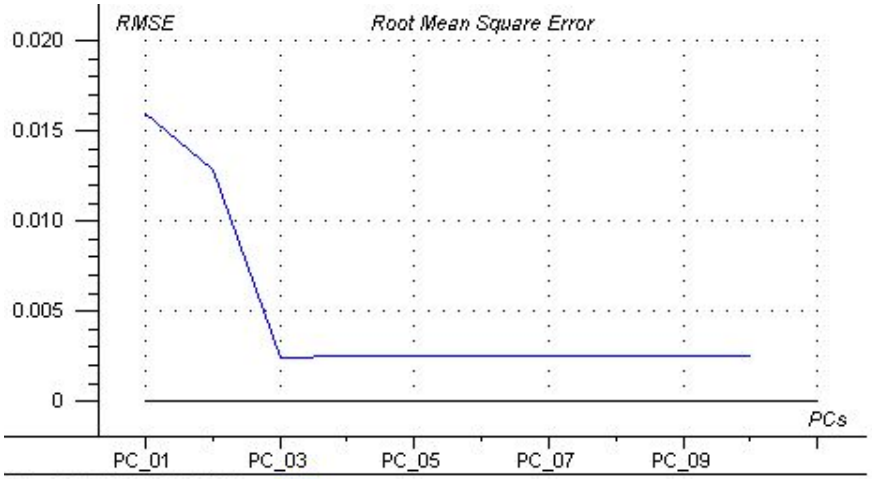


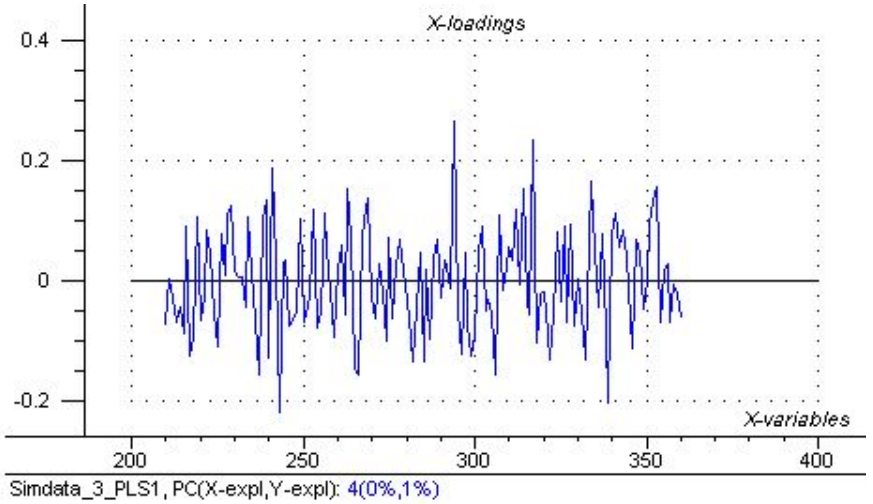


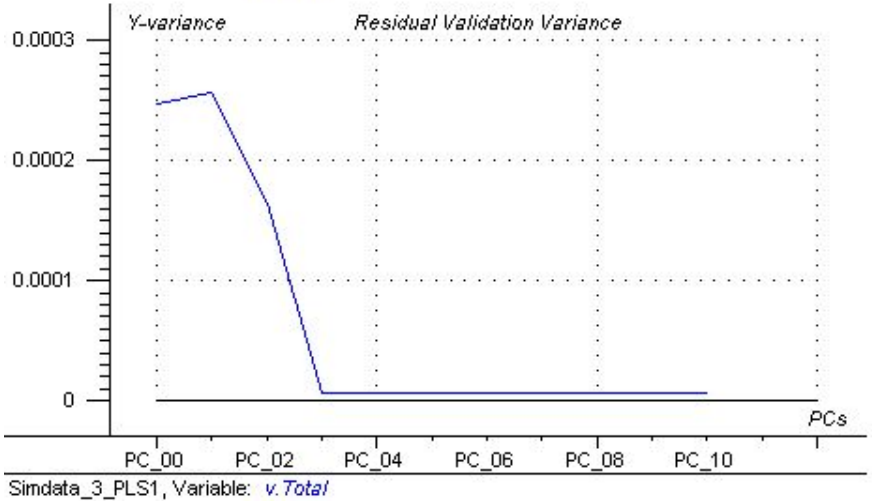
Проверка регрессионных моделей

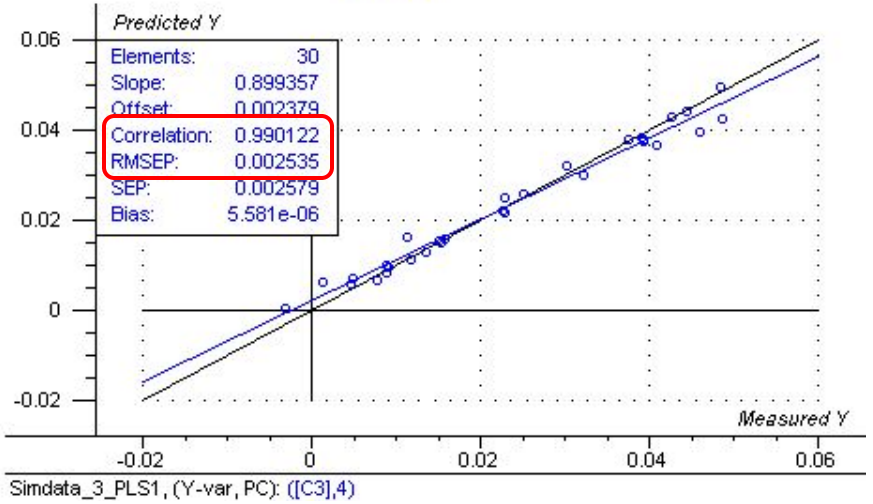
- Проверка (validation) модели преследует две основные цели:
- Определение оптимального числа компонент
  - Меньше факторов чем в РГК
  - Минимум RMSEP
- Оценка предсказательной способности модели:
  - График “предсказанние относительно измерения” (predicted vs measured)
  - RMSEP

Проверка регрессионных моделей: simdata – ПЛС1









Сравнение моделей: Simdata

Сравнение моделей калибровки трехкомпонентной смеси ПАУ (simdata)

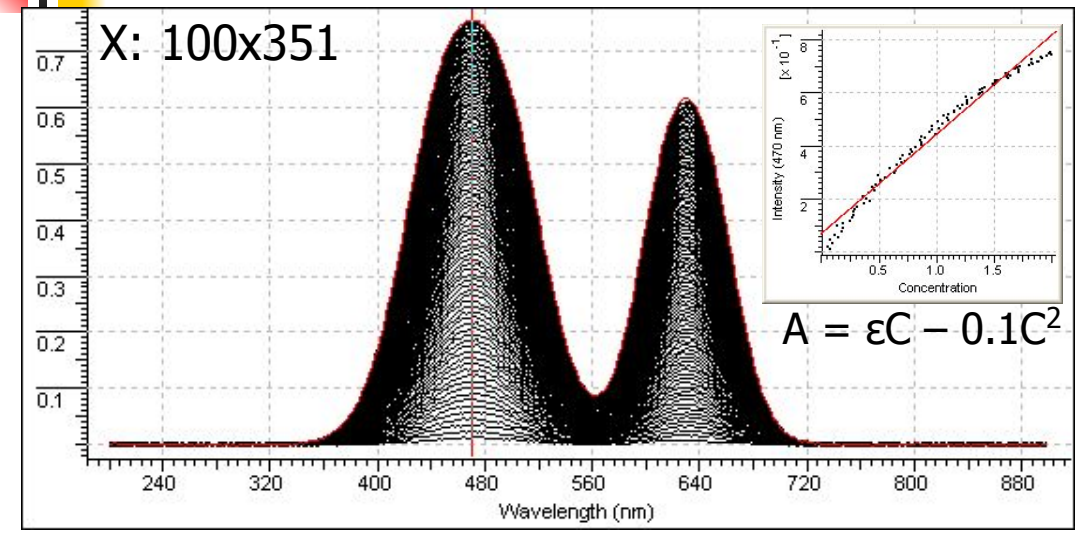
	МЛР MLR	РГК PCR	ПЛС1-Р PLS1-R	ПЛС2-Р PLS2-R
$[C_1]$	0.1312	0.0576	0.0575	0.0575
$[C_2]$	0.0527	0.0241	0.0245	0.0245
$[C_3]$	0.01579	0.00246	0.00246	0.00249

- вывод: модели РГК, ПЛС1-Р, ПЛС2-Р примерно одинково хороши для калибровки этих данных (без осложнений)
- результаты МЛР значительно хуже, для $[C_3]$ - неудовлетворительные

Сравнение методов калибровки

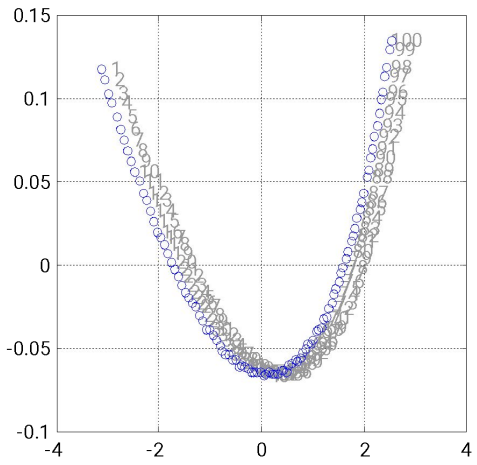
- МЛР (MLR) плохо пригоден для спектроскопических данных
- РГК (PCR) имеет недостатки, но хорошо работает при отсутствии осложнений
- ПЛС регрессия (PLS-R) является лучшим решением для большинства практических задач
- PLS1 или PLS2?
- Как выбрать метод? – пробовать!
- Как сравнивать разные модели? RMSEP

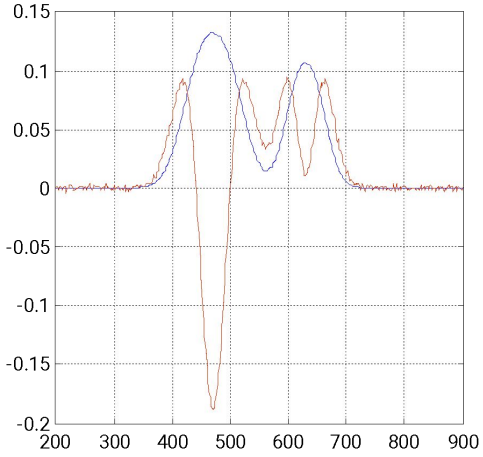
Линейная регрессия и нелинейность

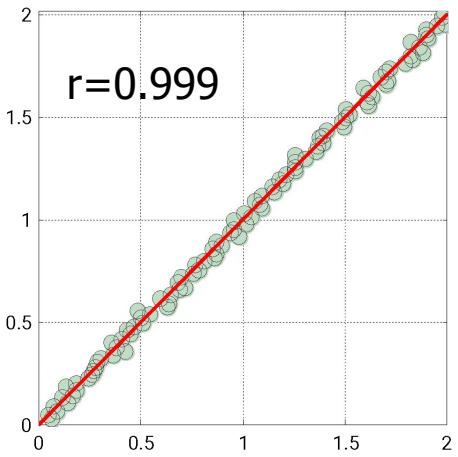


Y: 100x1

No.	Concentration
1	0.061618
2	0.044106
3	0.089839
4	0.067532
5	0.13498
6	0.10913
7	0.16404
8	0.17914
9	0.13095
10	0.17784
11	0.24147
...	
98	1.9919
99	1.9729
100	1.9823



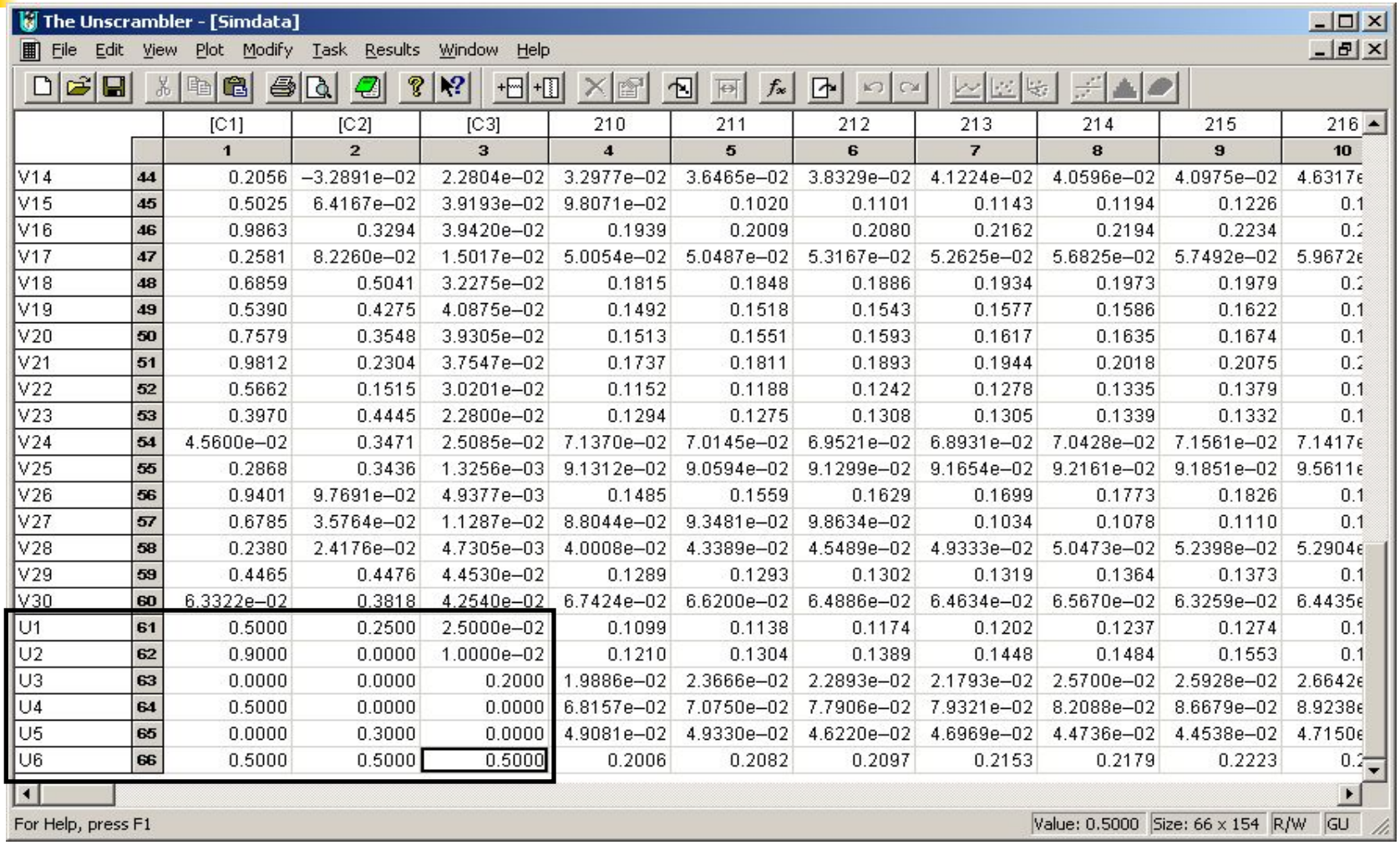




Предсказание: диагностика соответствия новых образцов

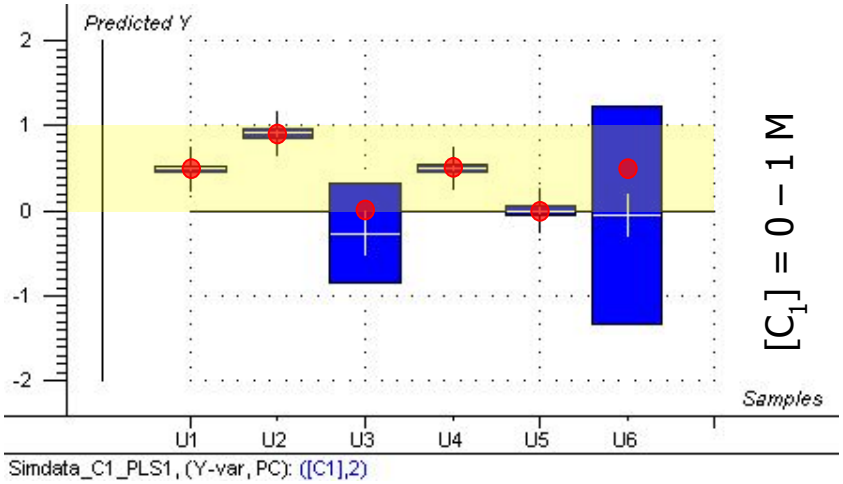
- не все проблемы заканчиваются с построением калибровочной модели!
- возможность выявления образцов, не соответствующих данной регрессионной модели является одним из преимуществ проекционного подхода
- Deviation - эмпирический параметр, характеризующий меру соответствия нового образца калибровочной модели
- рассмотрим наш пример…

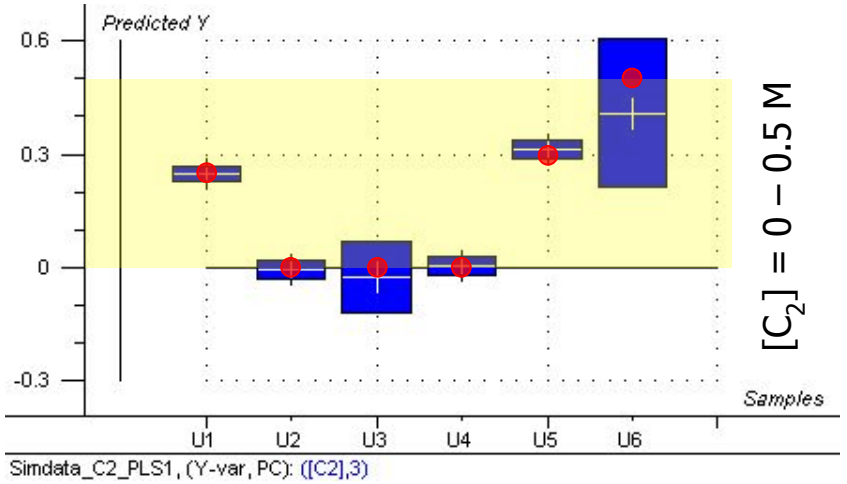
Диагностика предсказания (Simdata)

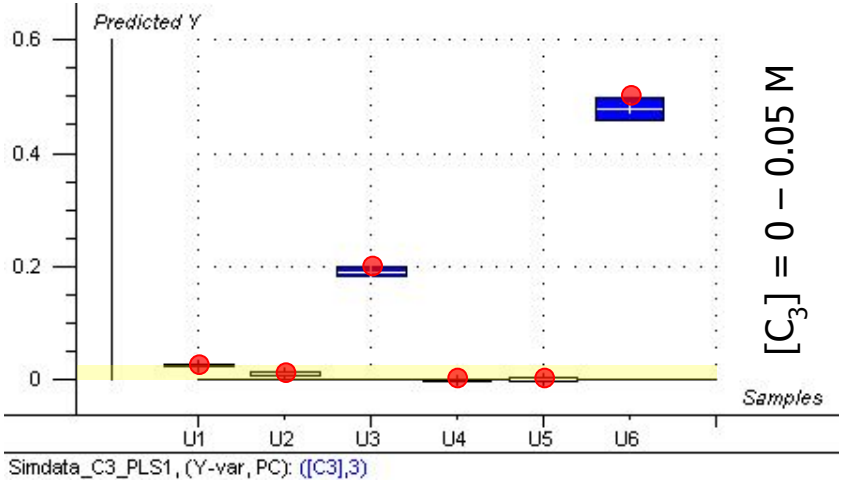


		[C1]	[C2]	[C3]	210	211	212	213	214	215	216
		1	2	3	4	5	6	7	8	9	10
V14	44	0.2056	−3.2891e−02	2.2804e−02	3.2977e−02	3.6465e−02	3.8329e−02	4.1224e−02	4.0596e−02	4.0975e−02	4.6317e
V15	45	0.5025	6.4167e−02	3.9193e−02	9.8071e−02	0.1020	0.1101	0.1143	0.1194	0.1226	0.1
V16	46	0.9863	0.3294	3.9420e−02	0.1939	0.2009	0.2080	0.2162	0.2194	0.2234	0.2
V17	47	0.2581	8.2260e−02	1.5017e−02	5.0054e−02	5.0487e−02	5.3167e−02	5.2625e−02	5.6825e−02	5.7492e−02	5.9672e
V18	48	0.6859	0.5041	3.2275e−02	0.1815	0.1848	0.1886	0.1934	0.1973	0.1979	0.2
V19	49	0.5390	0.4275	4.0875e−02	0.1492	0.1518	0.1543	0.1577	0.1586	0.1622	0.1
V20	50	0.7579	0.3548	3.9305e−02	0.1513	0.1551	0.1593	0.1617	0.1635	0.1674	0.1
V21	51	0.9812	0.2304	3.7547e−02	0.1737	0.1811	0.1893	0.1944	0.2018	0.2075	0.2
V22	52	0.5662	0.1515	3.0201e−02	0.1152	0.1188	0.1242	0.1278	0.1335	0.1379	0.1
V23	53	0.3970	0.4445	2.2800e−02	0.1294	0.1275	0.1308	0.1305	0.1339	0.1332	0.1
V24	54	4.5600e−02	0.3471	2.5085e−02	7.1370e−02	7.0145e−02	6.9521e−02	6.8931e−02	7.0428e−02	7.1561e−02	7.1417e
V25	55	0.2868	0.3436	1.3256e−03	9.1312e−02	9.0594e−02	9.1299e−02	9.1654e−02	9.2161e−02	9.1851e−02	9.5611e
V26	56	0.9401	9.7691e−02	4.9377e−03	0.1485	0.1559	0.1629	0.1699	0.1773	0.1826	0.1
V27	57	0.6785	3.5764e−02	1.1287e−02	8.8044e−02	9.3481e−02	9.8634e−02	0.1034	0.1078	0.1110	0.1
V28	58	0.2380	2.4176e−02	4.7305e−03	4.0008e−02	4.3389e−02	4.5489e−02	4.9333e−02	5.0473e−02	5.2398e−02	5.2904e
V29	59	0.4465	0.4476	4.4530e−02	0.1289	0.1293	0.1302	0.1319	0.1364	0.1373	0.1
V30	60	6.3322e−02	0.3818	4.2540e−02	6.7424e−02	6.6200e−02	6.4886e−02	6.4634e−02	6.5670e−02	6.3259e−02	6.4435e
U1	61	0.5000	0.2500	2.5000e−02	0.1099	0.1138	0.1174	0.1202	0.1237	0.1274	0.1
U2	62	0.9000	0.0000	1.0000e−02	0.1210	0.1304	0.1389	0.1448	0.1484	0.1553	0.1
U3	63	0.0000	0.0000	0.2000	1.9886e−02	2.3666e−02	2.2893e−02	2.1793e−02	2.5700e−02	2.5928e−02	2.6642e
U4	64	0.5000	0.0000	0.0000	6.8157e−02	7.0750e−02	7.7906e−02	7.9321e−02	8.2088e−02	8.6679e−02	8.9238e
U5	65	0.0000	0.3000	0.0000	4.9081e−02	4.9330e−02	4.6220e−02	4.6969e−02	4.4736e−02	4.4538e−02	4.7150e
U6	66	0.5000	0.5000	0.5000	0.2006	0.2082	0.2097	0.2153	0.2179	0.2223	0.2

Диагностика предсказания: ПЛС1 - Simdata







#	$[C_1]$	$[C_2]$	$[C_3]$
U1	0.5	0.25	0.025
U2	0.9	0	0.01
U3	0	0	0.2
U4	0.5	0	0
U5	0	0.3	0
U6	0.5	0.5	0.5

Правила построения «хорошей» калибровки

- правильно приготовить (собрать) образцы
- визуально изучить данные, если необходимо, применить предварительную обработку данных (pre-processing)
- если необходимо применить шкалирование/ взвешивание (scaling/weighting)
- интерпретировать модель, изучить структуру данных, выявить и удалить возможные выбросы
- тщательно оценить размерность модели, диагностировать модель
- диагностировать предсказание

План семинара

- Пример 1. Концентрационная калибровка трехкомпонентной смеси ПАУ по спектрам в УФ-видимой области (искусственные данные).
  - общие навыки калибровки, интерпретации и диагностики модели, предсказания на «идеальных» данных
- Пример 2. Определение октанового числа топлива по спектрам ближнего ИК.
  - калибровка на реальных данных, обнаружение и удаление выбросов
- Пример 3. Качество пшеницы (факультативно).
  - самостоятельное построение калибровки, MSC, выбор переменных

Рекомендуемая литература

- Richard Kramer
  Chemometric Tchniques for Quantitative Analysis *
- Kim H. Esbensen
  Multivariate Data Analysis - in Practice **
- Kenneth R. Beebee et al.
  Chemometrics: a Practical Guide **
- Harald Martens, Tormod Naes
  Multivariate Calibration **
- Richard G. Brereton
  Chemometrics: Data Analysis for the Laboratory and Chemical Plant ***
- Edmund R. Malinowski
  Factor Analysis in Chemistry ****

Пример 1:
Калибровка смеси ПАУ

Файл Simdata

Цель: выработка навыков калибровки с программой Unscrambler

- изучить наборы данные: обучающий, тестовый, «unknown» - в таблице, как серии спектров
- построить калибровки: РГК, ПЛС2 - сравнить модели
- построить ПЛС1 для каждого из 3-х компонентов, определить размерность моделей
- изучить графики scores, loadings, T-U, predicted vs measured, RMSEP, Variance для [C1] - [C3] с разным количеством факторов
- предсказать «неизвестные» образцы

Пример 2: Определение октанового числа бензина

стр. 139, файл Octane

Цель: работа с реальными данными, диагностика и устранение выбросов

преимущественно по книге:

- построить калибровку ПЛС1, диагностировать
- определить выбросы, удалить, обновить модель
- проверить модель различными способами, включая тестовый набор
- построить РГК, сравнить модели
- предсказать «неизвестные» образцы

Пример 3: Качество пшеницы

стр. 150, файл Wheat

Цель: самостоятельное построение калибровочной модели

- построение моделей ПЛС1/2, сравнение моделей
- определение и удаление выбросов
- применение MSC
- попробовать удаление переменных для улучшения модели