Методы многомерной калибровки: PCR/PLS

Андрей Богомолов

Российское хемометрическое общество

Тема лекции

Многомерная калибровка
Multivariate Calibration

Анализ многомерных данных (Хемометрика)

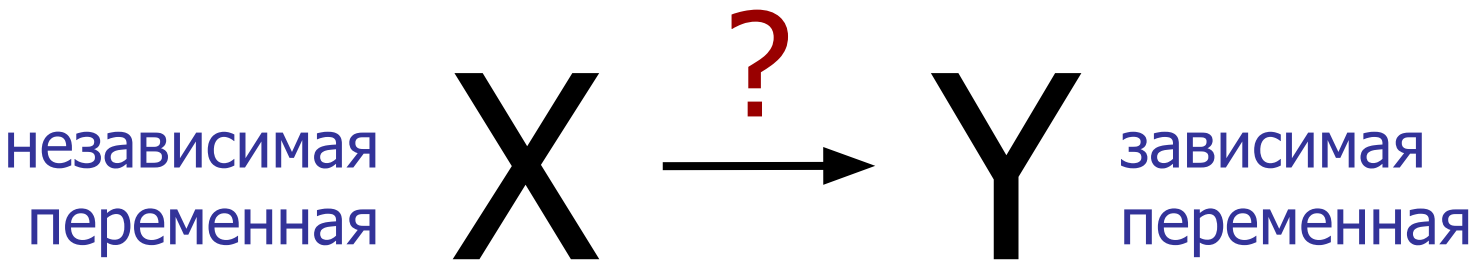
Multivariate Data Analysis (Chemometrics)

11:28

К вопросу о русской терминологии

- родной язык хемометрики - английский
- терминология за 30 лет устоялась: статьи, учебники, книги, конференции
- устоявшиеся аббревиатуры: PCA, PCR, PLS, SIMCA, RMSEP, etc. - не нуждаются в расшифровке
- русская терминология создается сейчас
- почему нужен перевод?
- в настоящей лекции - параллельная терминология

Регрессионный анализ



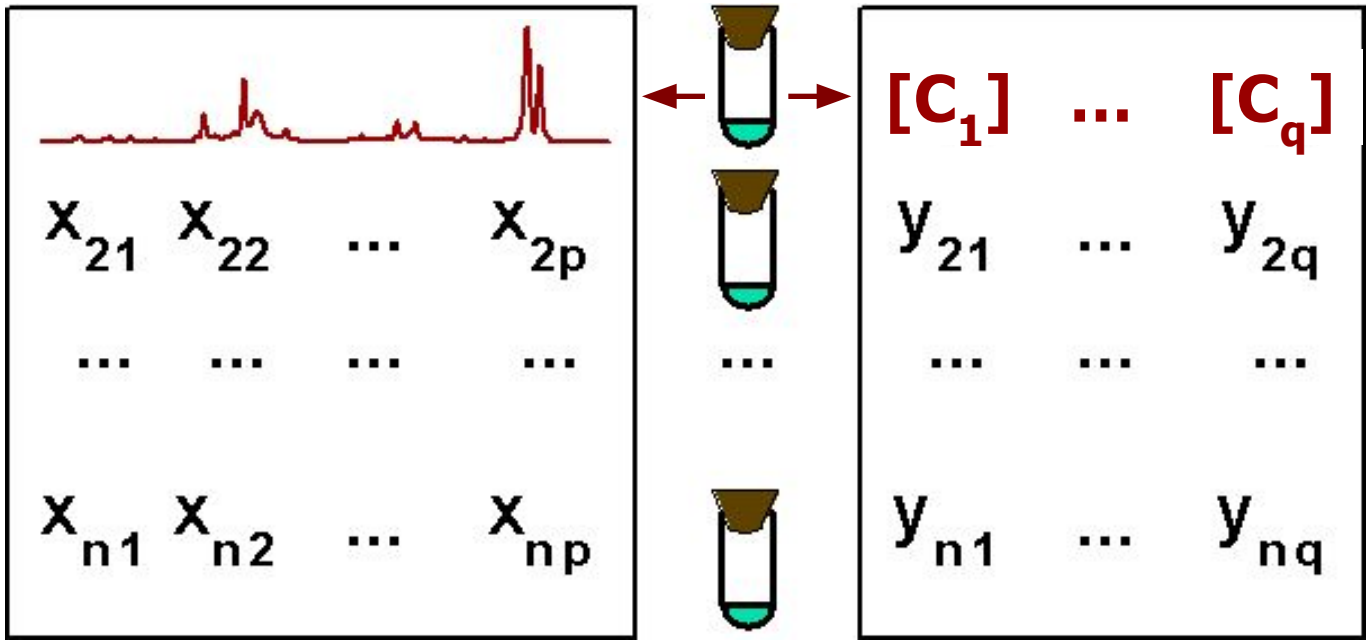
- линейная регрессия

$$\mathbf{Y = XB + E}$$

- МГК - моделирование (**X**)
- Регрессия - моделирование (**X,Y**)

Спектроскопические данные

Спектры (X) **Концентрации (Y)**



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

Для чего нужна калибровка?

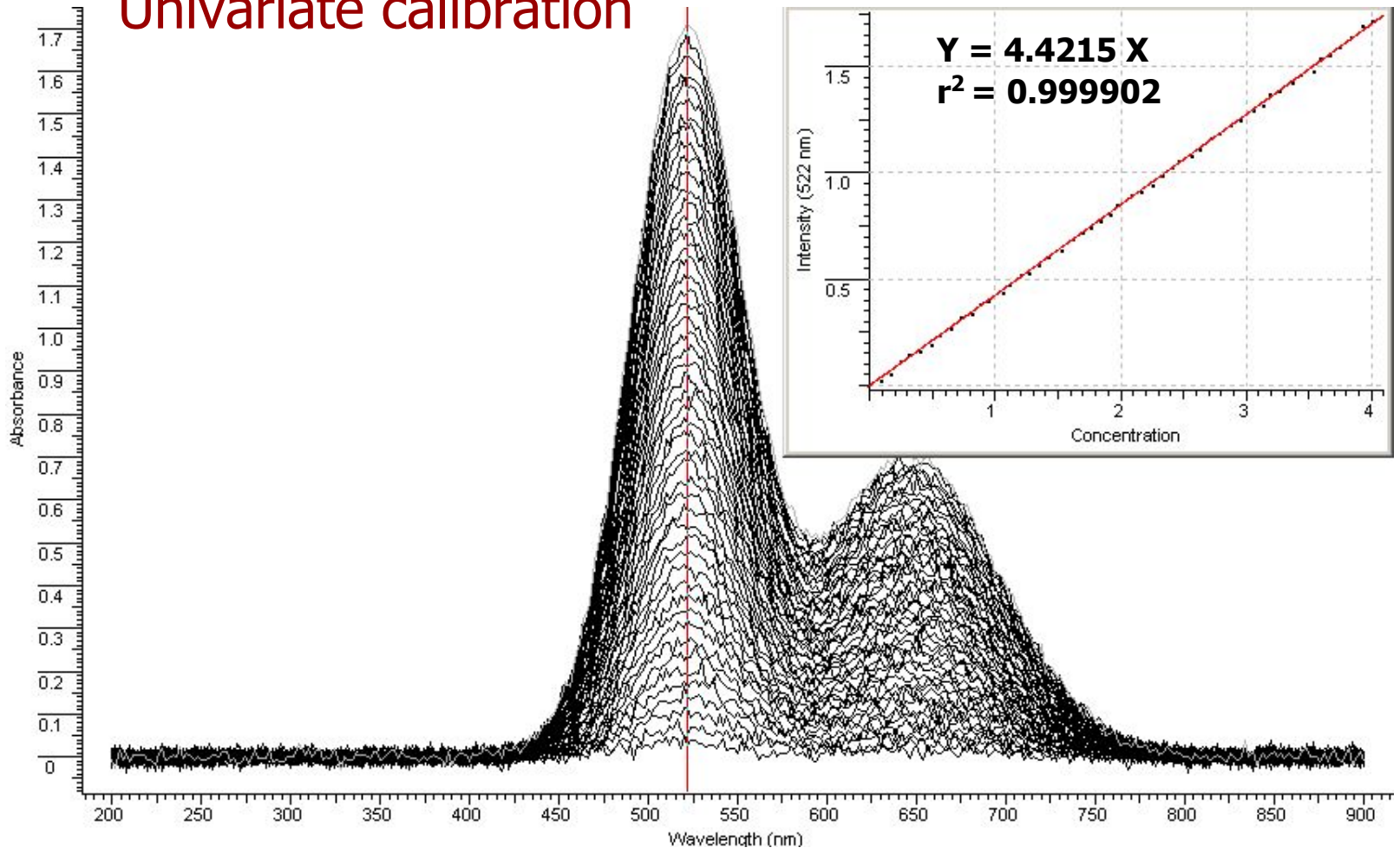
- замена прямого измерения интересующего свойства, измерением другого, коррелирующего с первым
- такая потребность возникает если прямое измерение интересующего свойства нежелательно:
  - дорого
  - трудоемко
  - занимает много времени
  - этически нежелательно
  - эксперимент невозможен, и т. п.
- в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

Примеры из различных областей

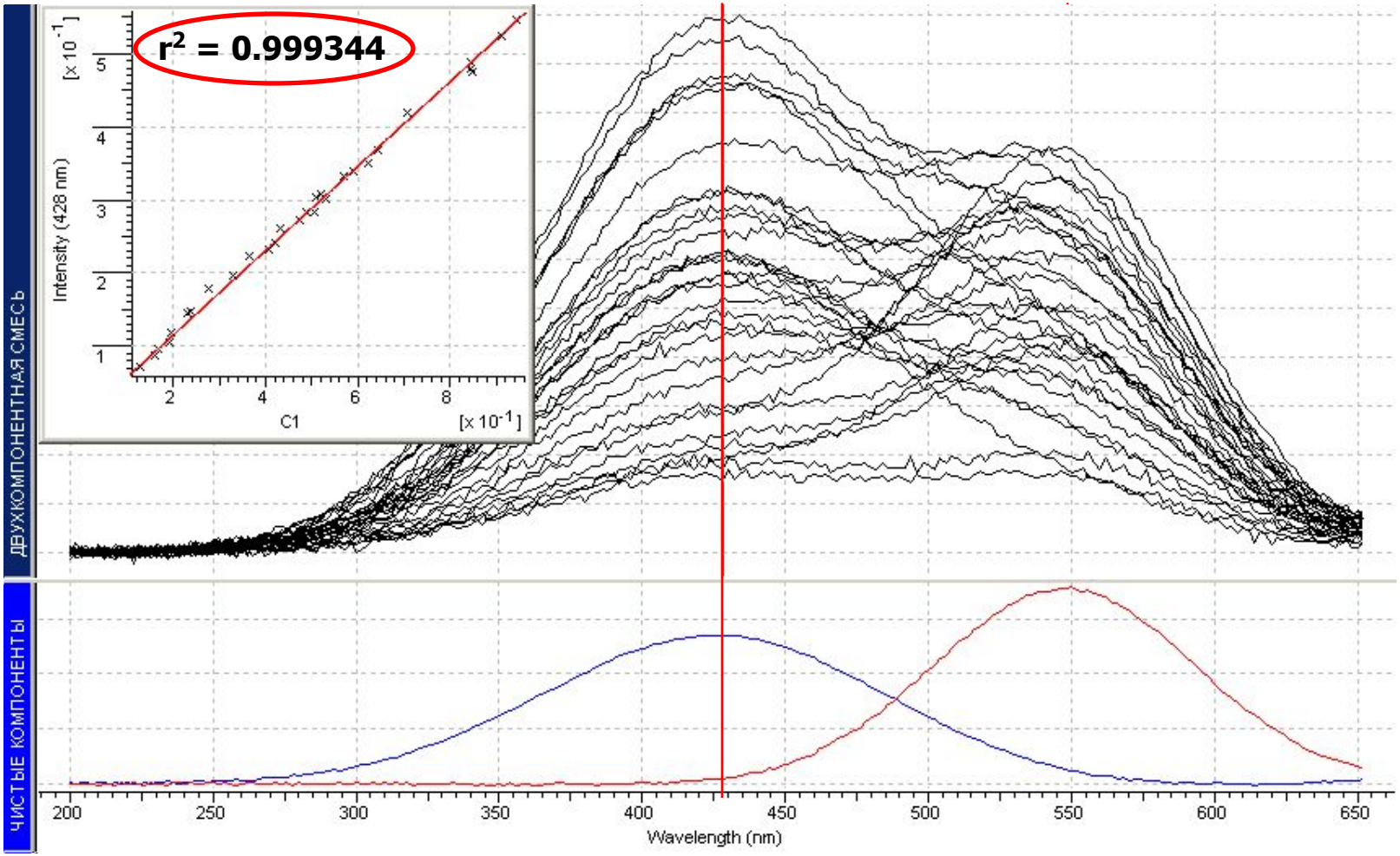
- ХИМИЯ: калибровка – инструмент №1 количественного анализа
- БИОЛОГИЯ: непосредственный анализ может быть губителен для живых существ
- МЕДИЦИНА: неинвазивный анализ, например, определение сахара в крови спектроскопически (ближний ИК)
- ПСИХОЛОГИЯ: анализ личности может потребовать длительных наблюдений,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

Одномерная калибровка: один компонент

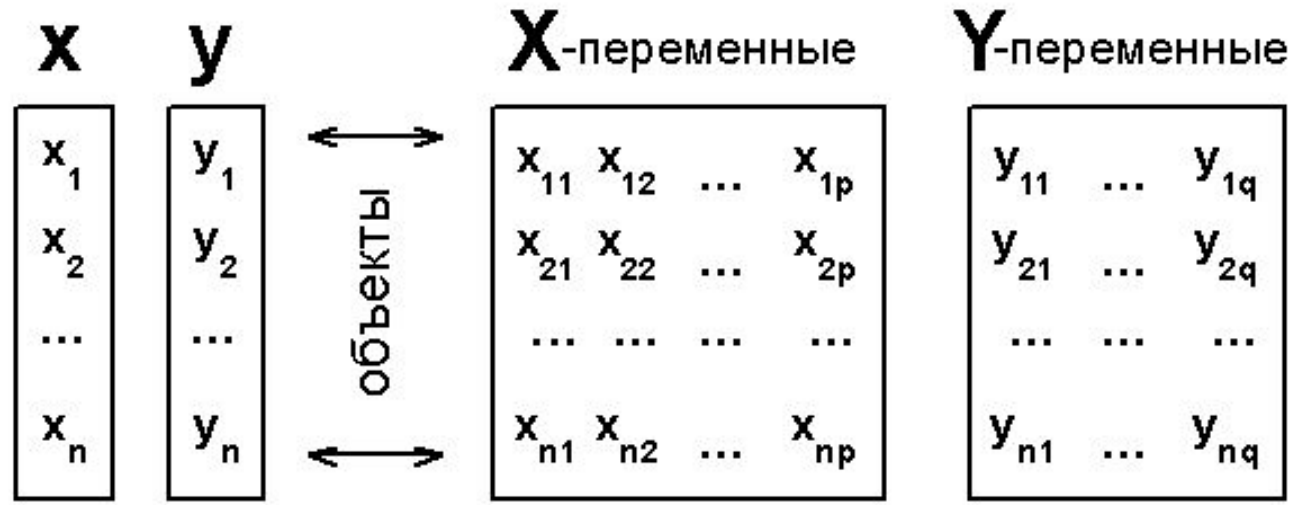
Univariate calibration



Одномерная калибровка: многокомпонентная смесь



Многомерная калибровка



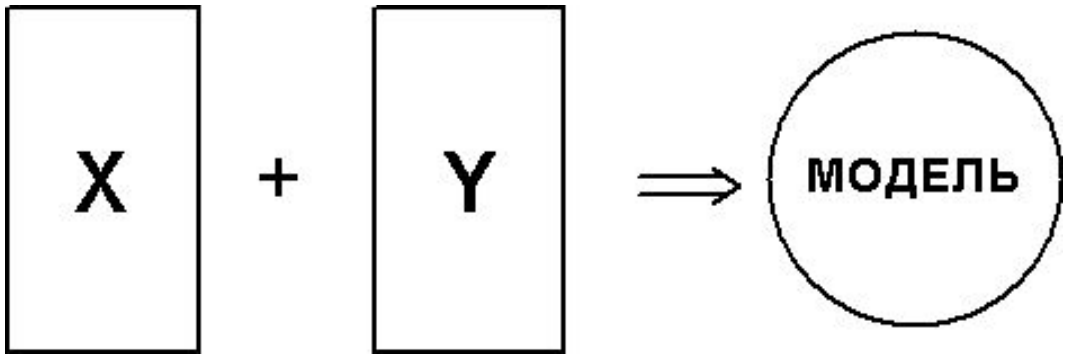
$\mathbf{y}=\mathbf{x}b+\mathbf{e}$ $\mathbf{Y}=\mathbf{XB}+\mathbf{E}$

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 (например, NIR)

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)

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p + e$$

$$y = X * b + e$$

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

n - число объектов (спектров)

p - число переменных (длин волн)

$n \geq p$

Решение: $b = (X^T X)^{-1} X^T y$

Недостатки МЛР

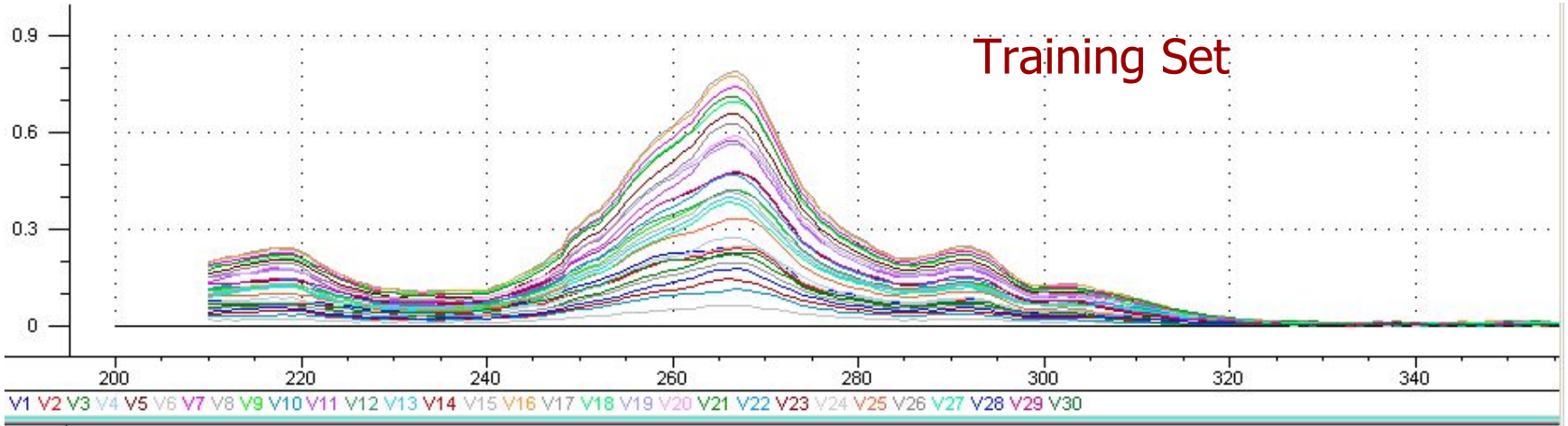
МЛР может не сработать, если:

- 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

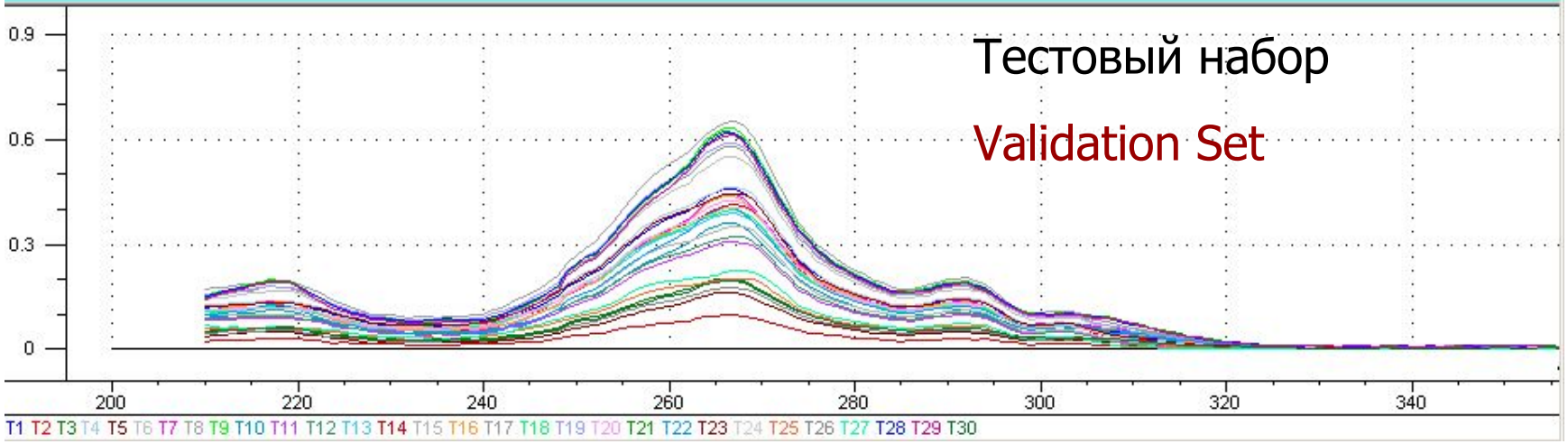
Обучающий набор

Training Set

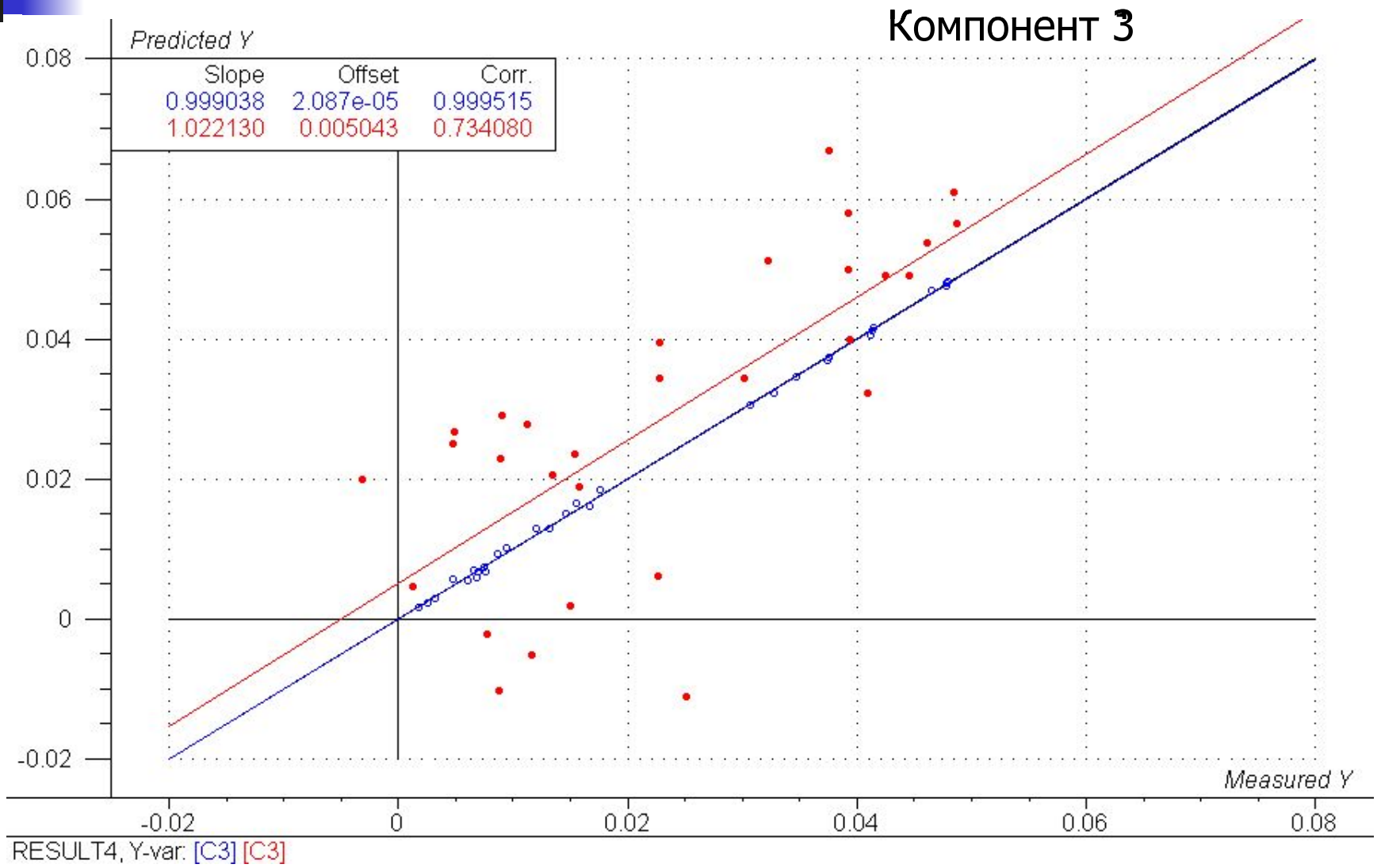


Тестовый набор

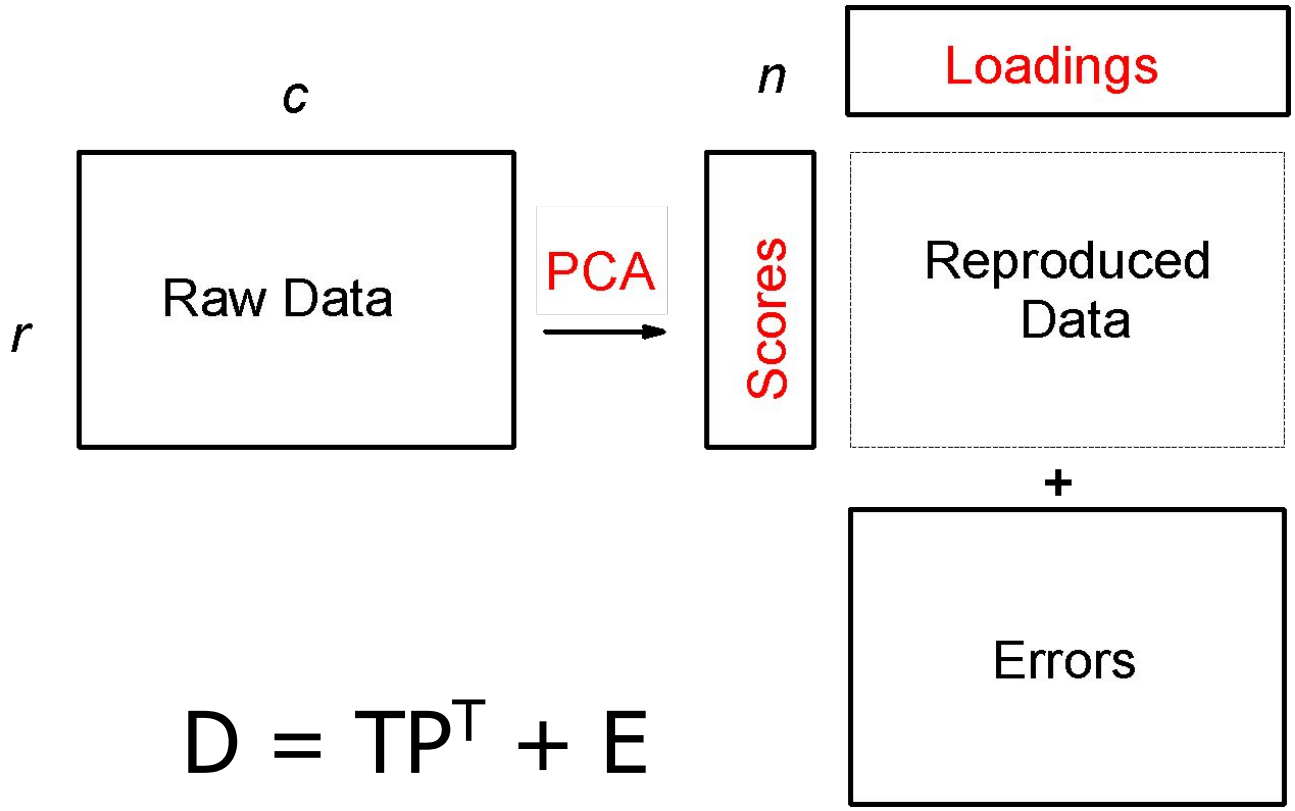
Validation Set



МЛР-калибровка

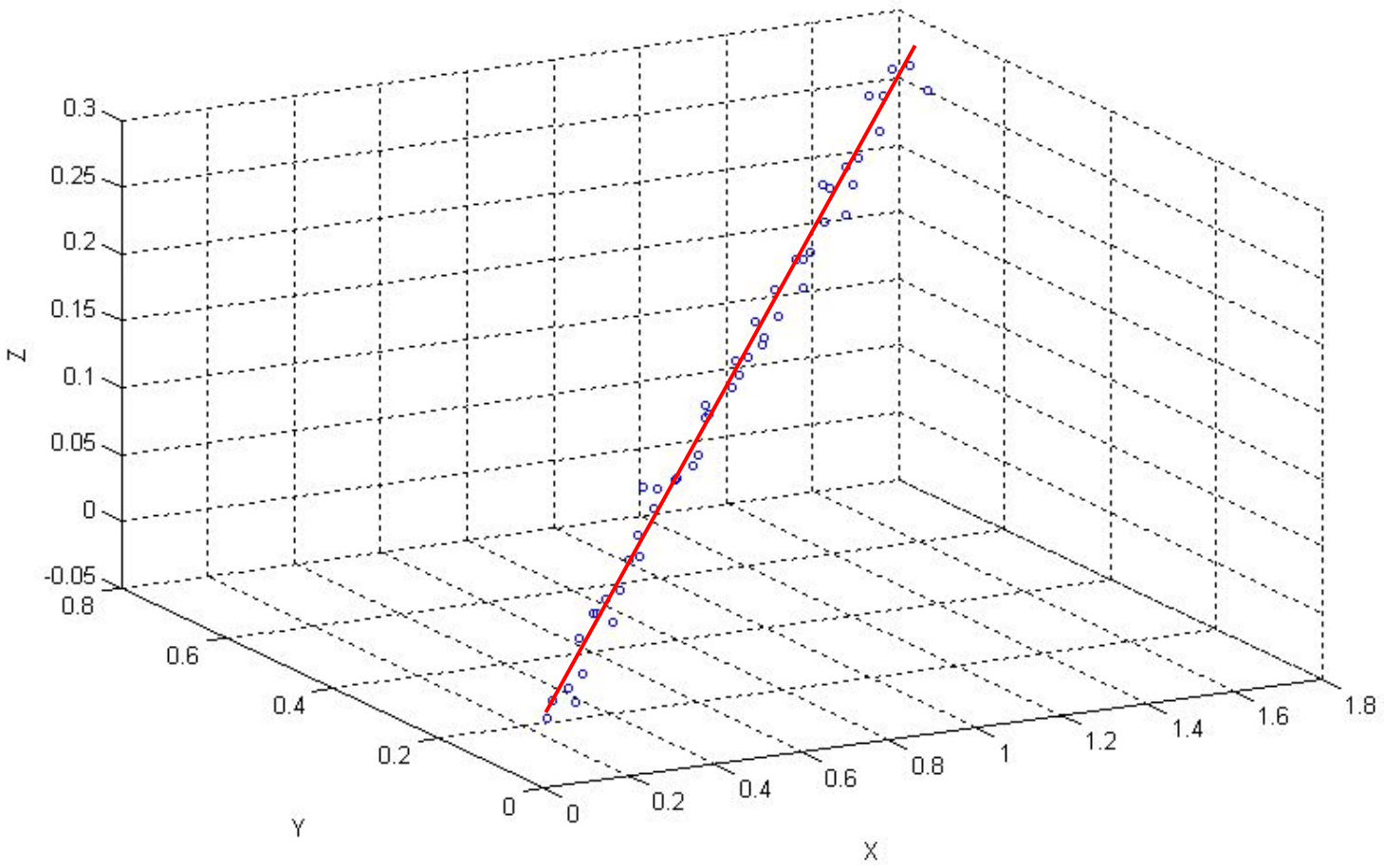


МГК (PCA) - оружие против коллинеарности



$$D = TP^T + E$$

Концепция PCA «на пальцах»

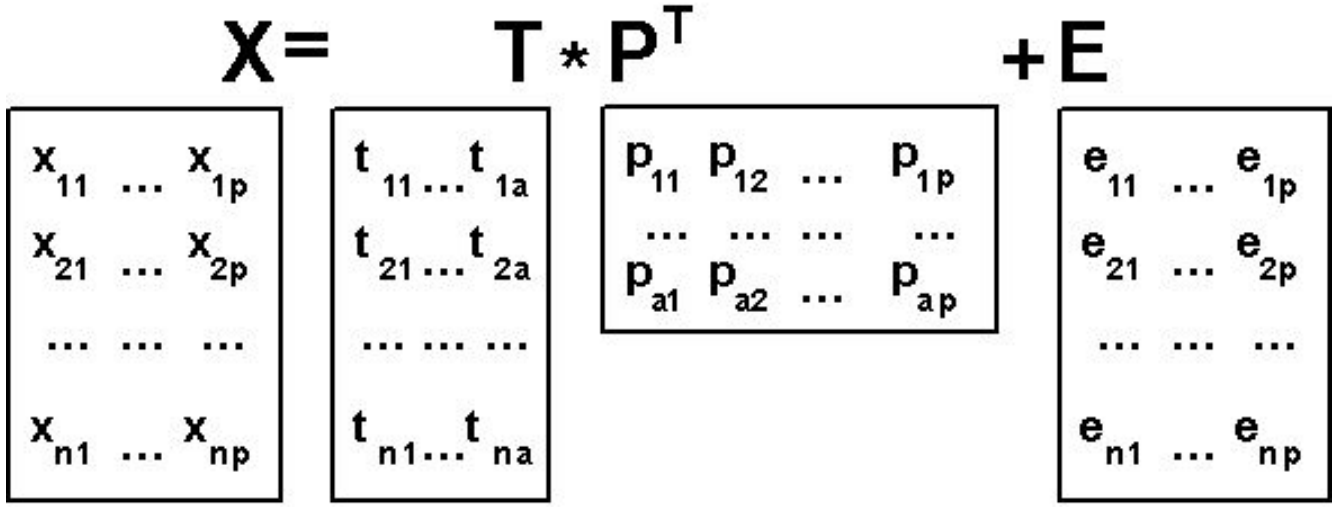


PCA + MLR = PCR !

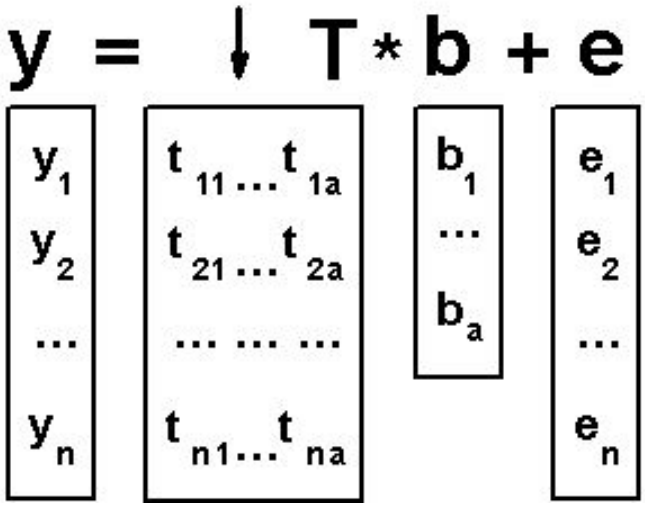
- В результате РГК (PCA):
  - Происходит компрессия данных
  - уменьшается размерность данных
  - коллинеарность обращается во благо;
  - уменьшается ошибка;
  - РГК-нагрузки (PCA-scores) T ортогональны
  - содержат информацию о концентрациях компонентов
- T можно использовать для построения MLR-модели, вместо **X**; этот метод называется…
- регрессия на главные компоненты, РГК (Principal Component Regression, PCR)

Схема РГК (PCR) – подробнее

PCA:



MLR:



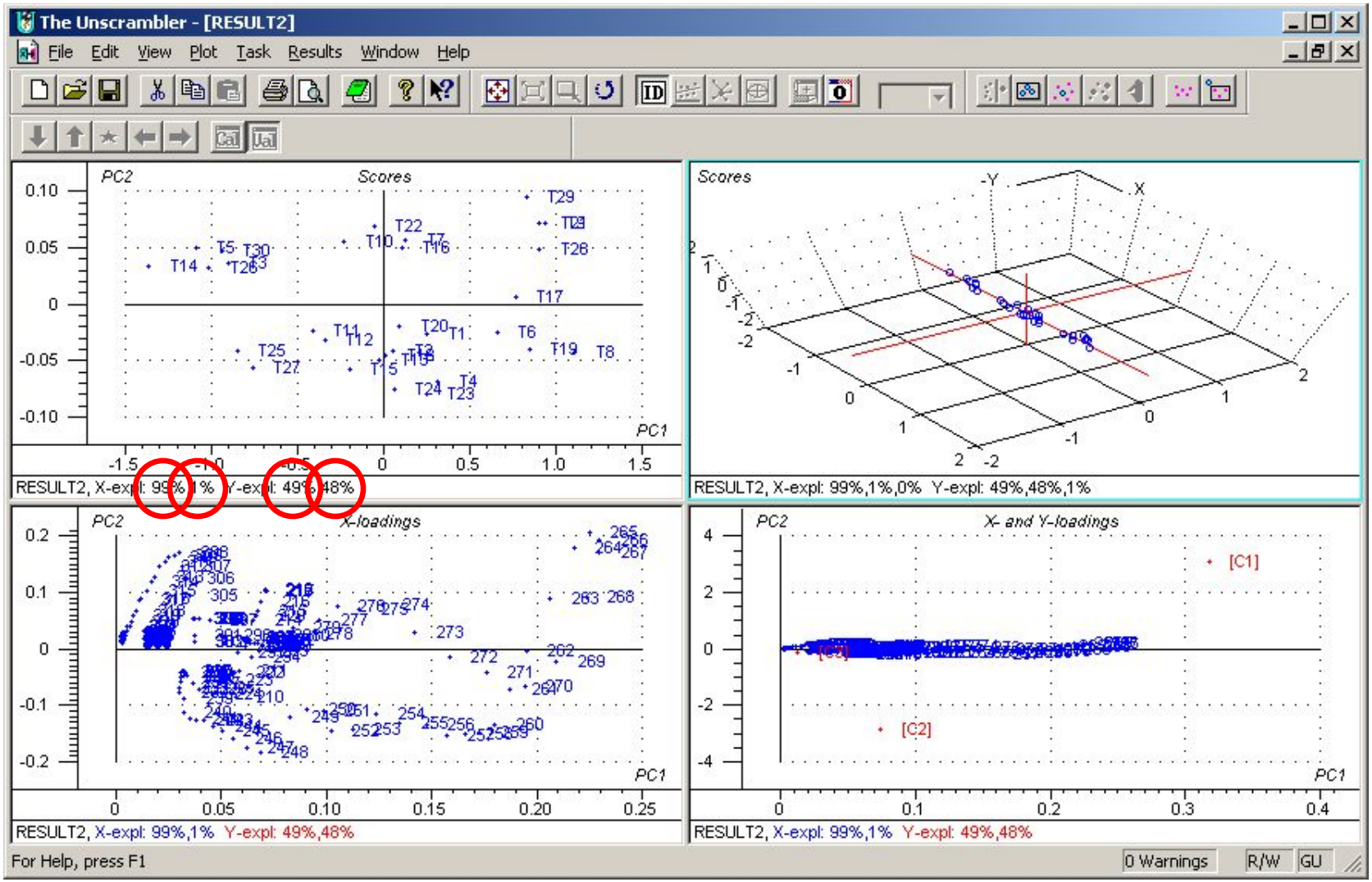
n - объектов
p - переменных
a - главных компонент

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

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)



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
  - Сегментная

RMSEP

- RMSEC = Root Mean Square Error of Calibration

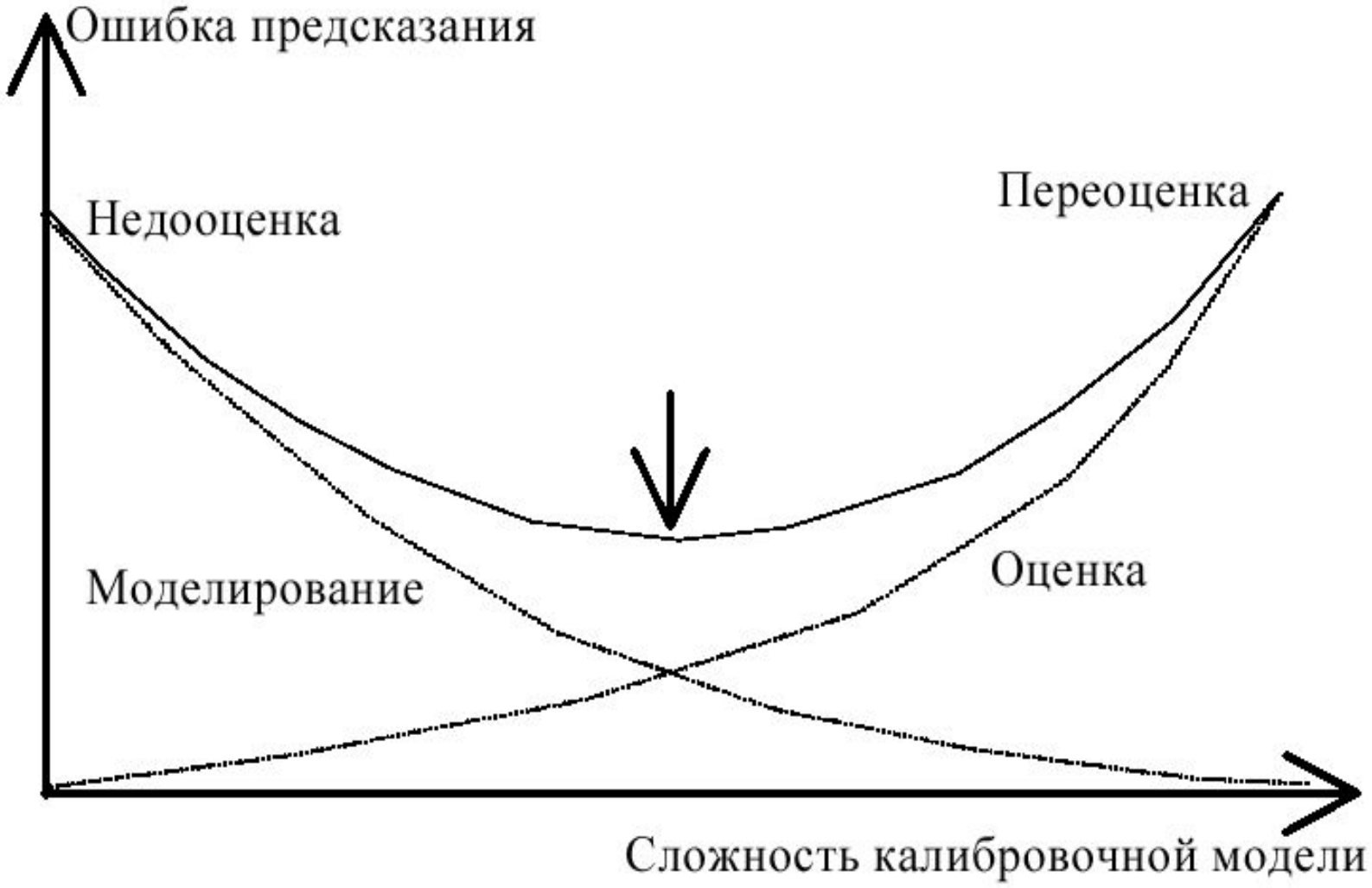
$$RMSEC = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_{i,cal} - y_{i,cal})^2}{n}}$$

- RMSEP = Root Mean Square Error of Prediction

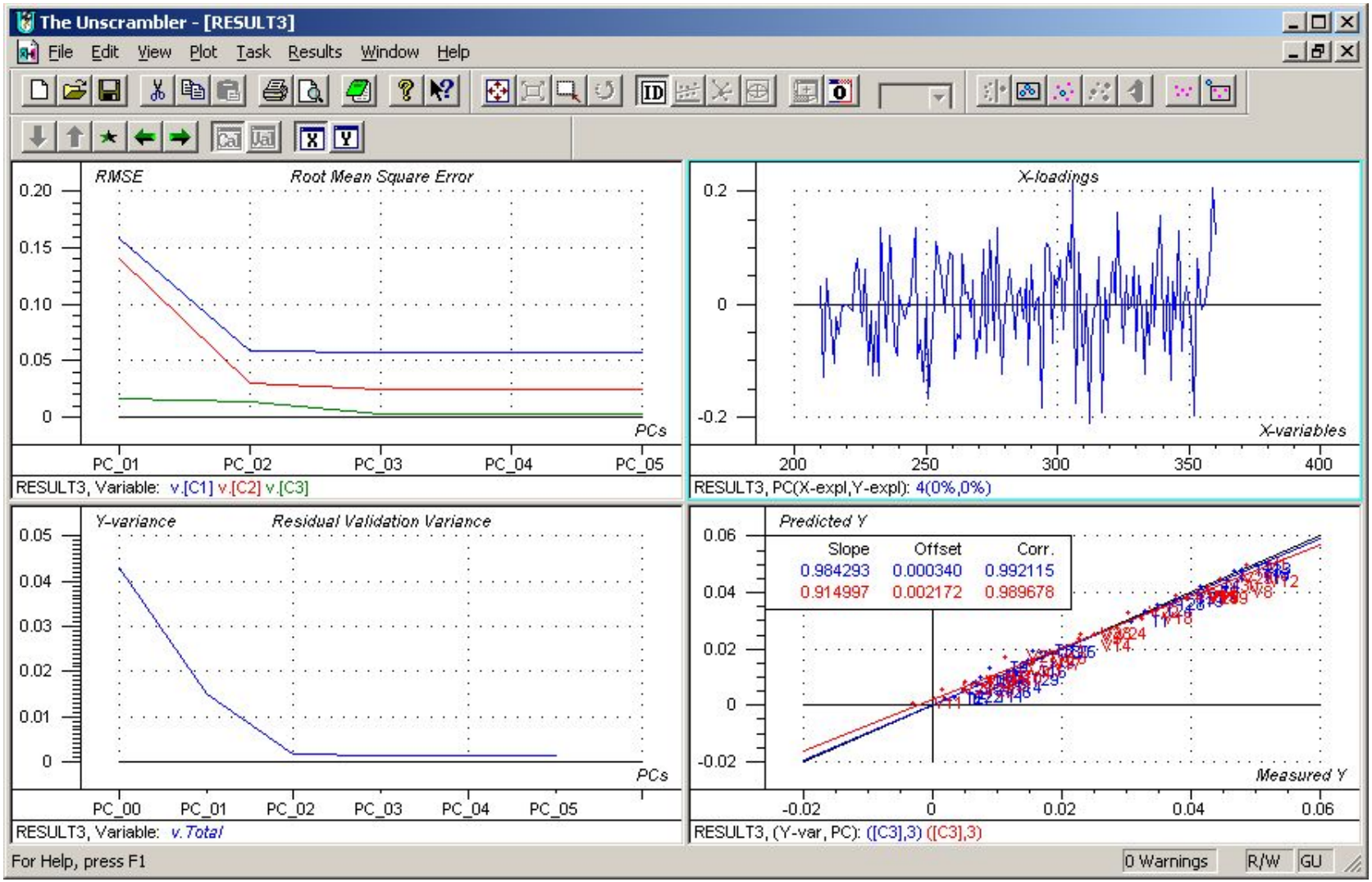
$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_{i,val} - y_{i,val})^2}{n}}$$

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

Оценка числа компонент в РГК



Число компонент (Simdata)



Оценка числа ГК в РГК: особенности

- число ГК (размерность модели) определяется нуждами калибровки, и не обязательно совпадает с результатом МГК
- активно используется тестовые данные (Test Set)
- RMSEP = Root Mean Square Error of Prediction
- 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 - мощный метод многомерной калибровки
- имеет безусловные преимущества перед MLR
- однако, не вполне оптимизирован для калибровки
- пространство ГК оптимально для моделирования внутренней структуры данных матрицы **X**, но не учитывает структуры **Y** и связи между **X** и **Y**
- можно ли учесть эту связь при построении проекционной модели?
- да, использовать PLS!

Факторные пространства

- существует бесконечное множество способов декомпозиции данных вида

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

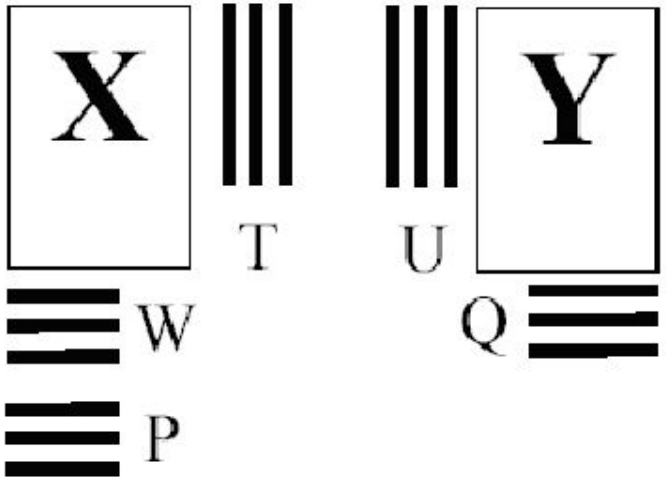
- парные вектора в **T** и **P** называются факторами (factors), а преобразование - проекцией данных на факторное пространство (factor space) или факторной компрессией
- пространство главных компонент один из наиболее важных вариантов факторного пространства
- для задания факторного пространства нужен критерий, например, МГК (PCA) использует критерий максимальной остаточной дисперсии

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; критерием является моделирование той структуры (информации) в **X**, которая имеет корреляцию с **Y**
- например, спектральные полосы (**X**), которые отвечают за концентрацию компонента(ов), заданные в **Y**
- ПЛС-модель специально оптимизирована для регрессионного анализа

ПЛС-регрессия: схематическое представление

- ПЛС-декомпозиция затрагивает обе матриц **X** и **Y**
- в результате - 2 набора счетов (scores) и нагрузок (loadings)
- плюс дополнительная матрица взвешенных нагрузок **W** (loading-weights)
- критерий: максимальная ковариация между **T** и **U**



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

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F}$$

Две разновидности ПЛС: ПЛС1 и ПЛС2

- существуют две популярных разновидности ПЛС: ПЛС1 (PLS1) и ПЛС2 (PLS2)
- ПЛС1 модель строится для единственной переменной **Y** (аналогия с МЛР), например, для концентрации одного компонента смеси
- если нужна калибровка по нескольким компонентам, строится несколько независимых моделей
- ПЛС2 рассчитывается для нескольких компонентов 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 ПЛС
- алгоритм работает до выполнения критерия сходимости
- детальное изучение алгоритмов не входит в задачу данной лекции, однако…
- ознакомимся с основными шагами на примере 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.	$\mathbf{t_{f.new} - t_{f.old}} \lvert < \mathbf{lim?}$	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.	$\mathbf{X_{f+1}} = \mathbf{X_{f+1}} - \mathbf{t_f^T p_f}$ $\mathbf{y_{f+1}} = \mathbf{y_{f+1}} - q_f \mathbf{t_f}$	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

Предсказание по ПЛС-модели

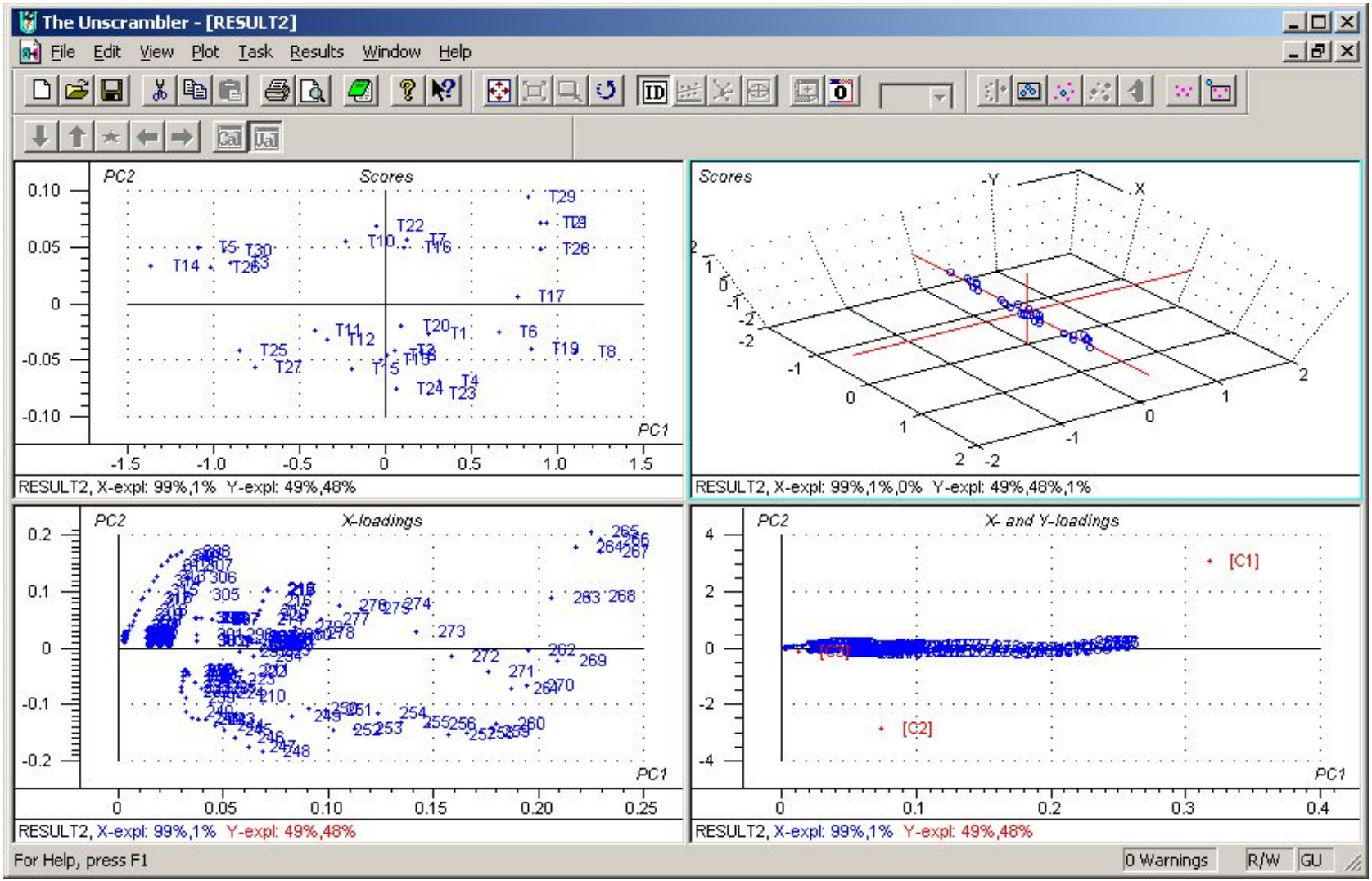
$$\hat{Y} = X_{new}B$$

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

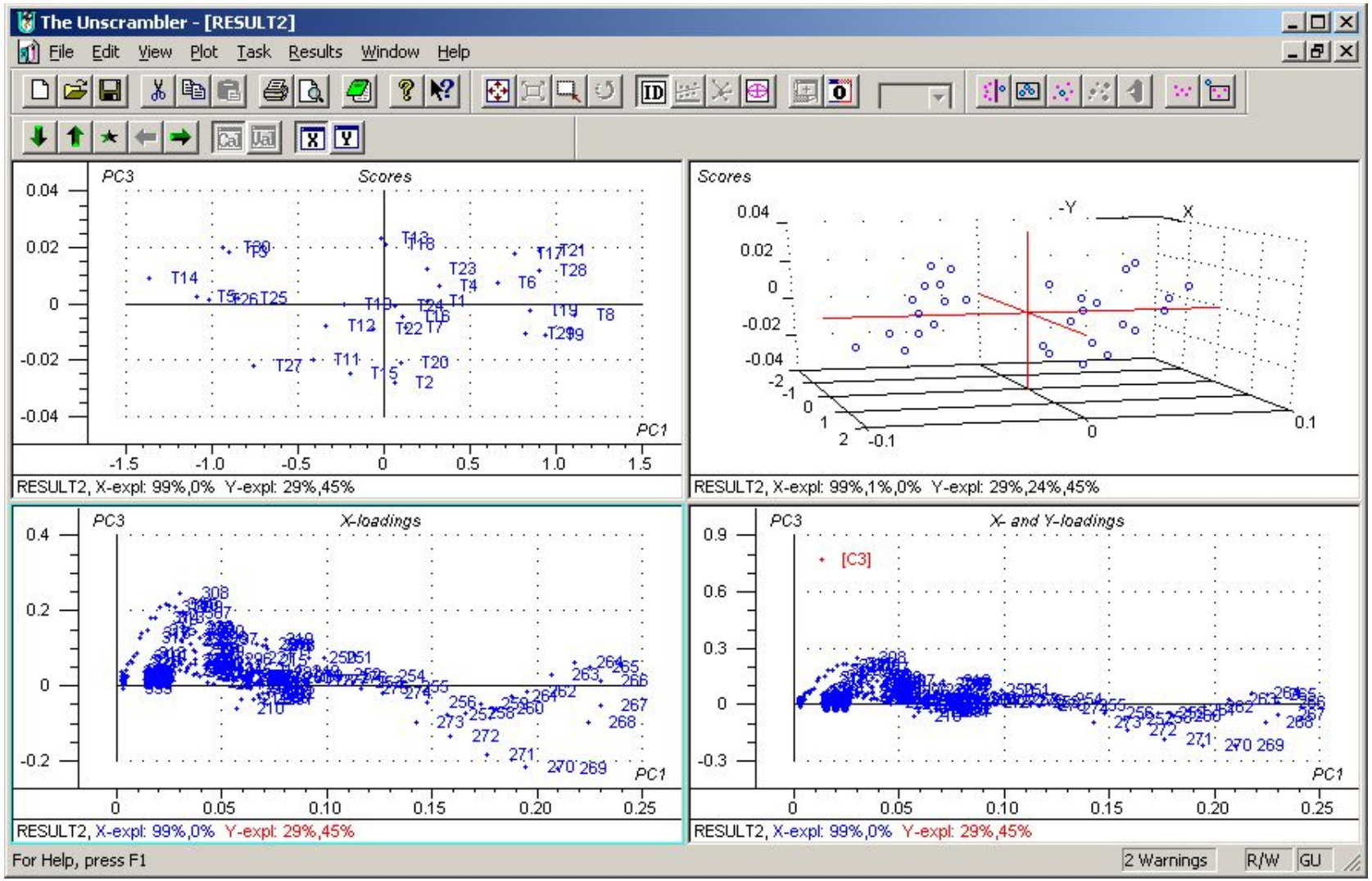
ПЛС1 и ПЛС2

- ПЛС1 моделирует только одну переменную **y** «за раз»
- в этом смысле ПЛС2 кажется гибче при калибровке нескольких свойств, позволяя моделировать любую комбинацию переменных без их разделения - совместно
- 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 пока нет…
- сравним методы на практике!

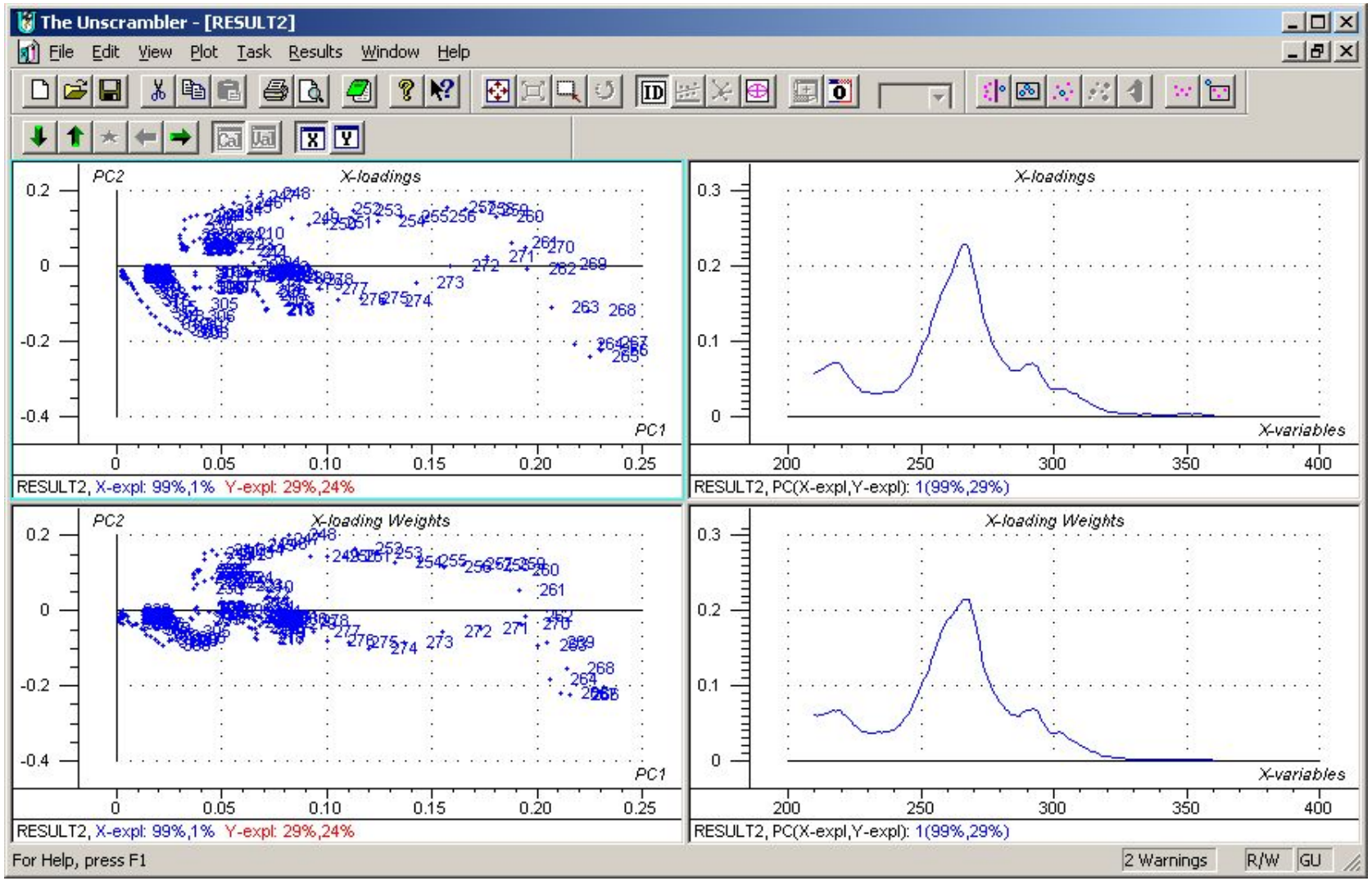
Строим ПЛС2-модель (Simdata)



Интерпретация ПЛС-моделей структура **X** (Simdata)

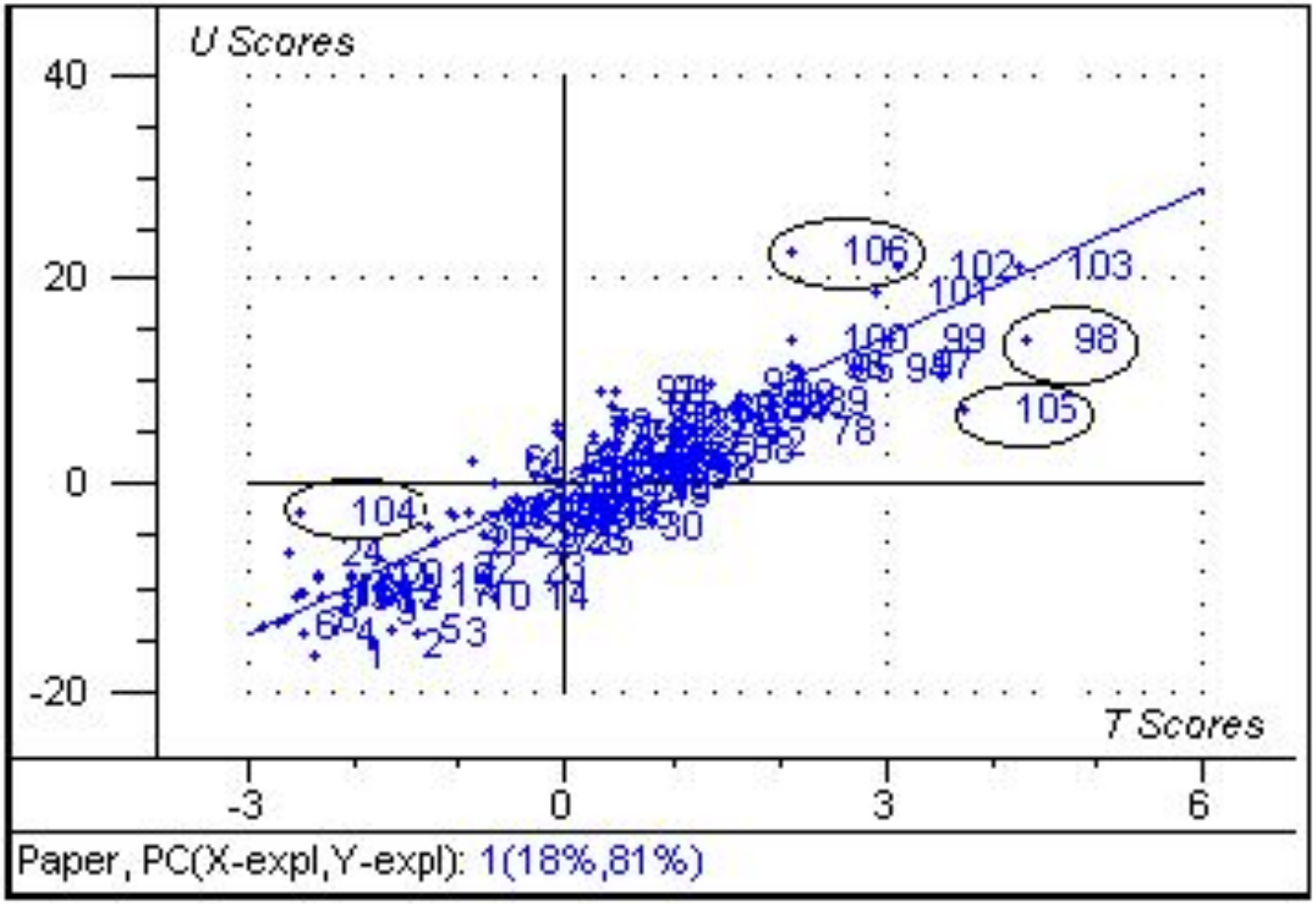


Интерпретация ПЛС-моделей: связь **X** и **Y** (Simdata)

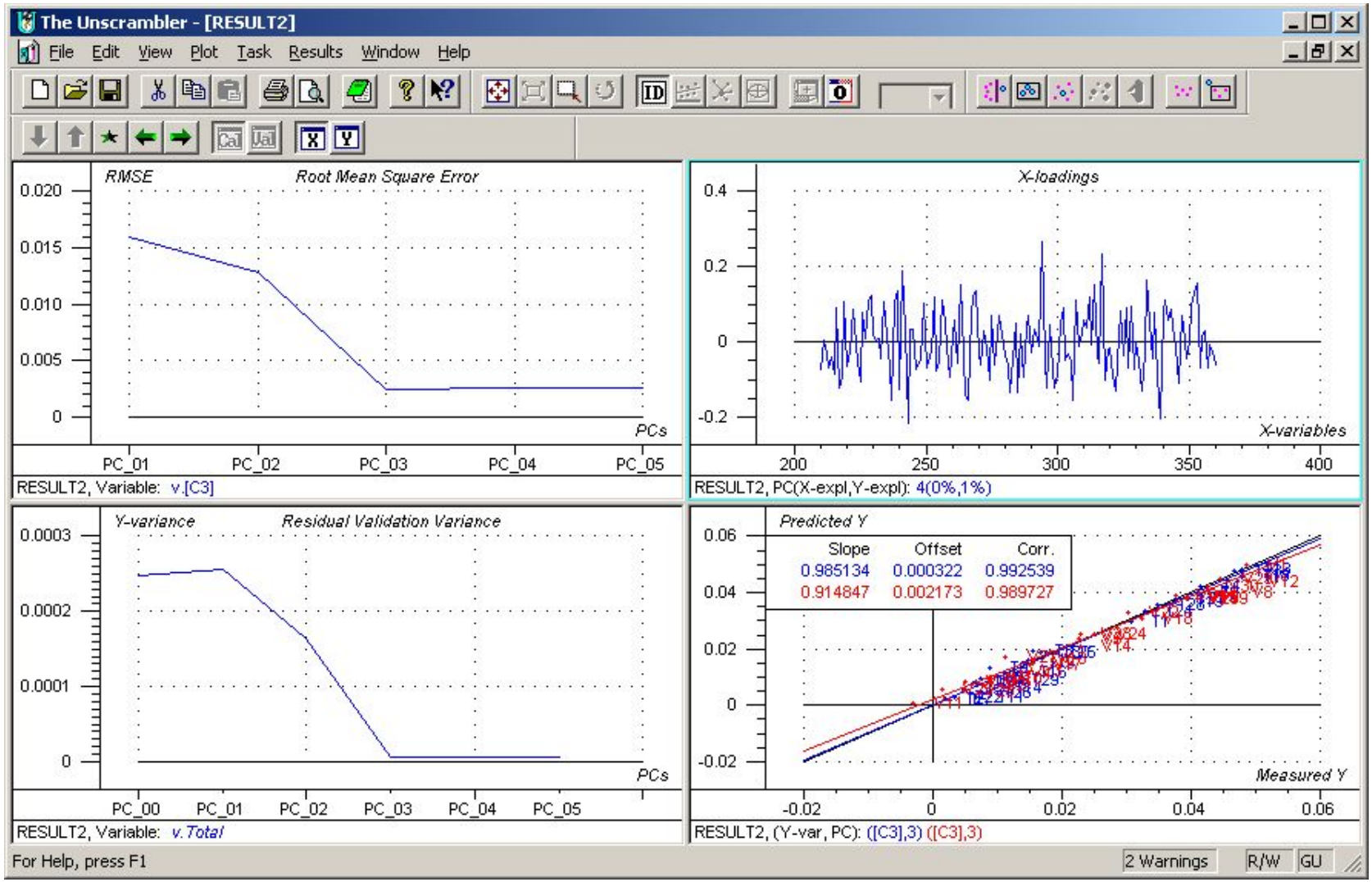


Интерпретация ПЛС-модели: выбросы (Octane)

График T - U как средство детекции выбросов (outliers)



Проверка ПЛС-моделей



Сравнение моделей (Simdata)

Сравнение моделей калибровки трехкомпонентной смеси ПАУ (Simdata)

	МЛР (MLR)	РГК (PCR)	ПЛС1-Р (PLS1-R)	ПЛС2-Р (PLS2-R)
[C1]	0.1312	0.0576	0.0575	0.0575
[C2]	0.0527	0.0241	0.0245	0.0245
[C3]	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 [C3] - неудовлетворительные

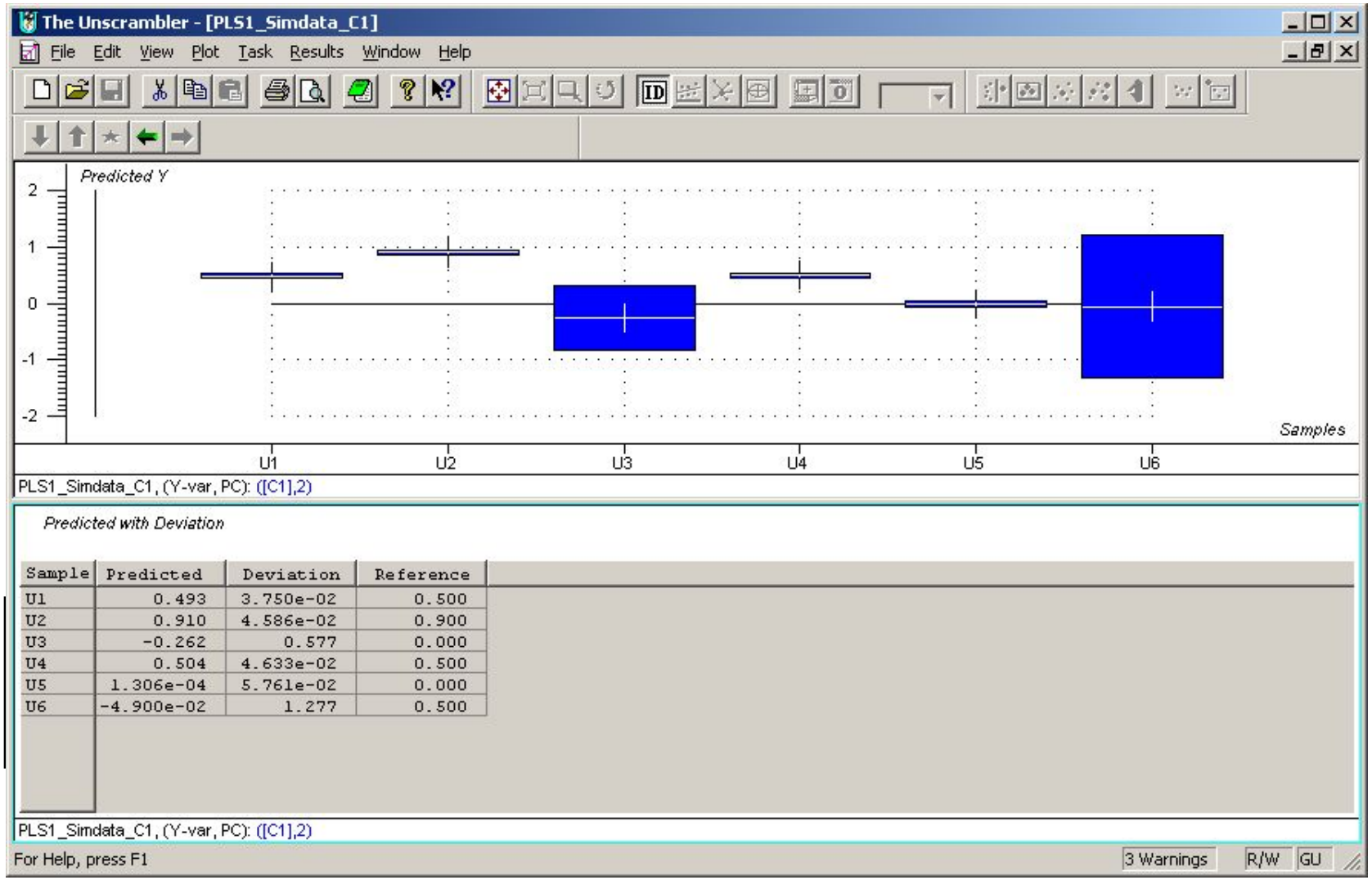
Сравнение методов калибровки

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

Предсказание: диагностика соответствия новых образцов

- с построением калибровочной модели проблемы еще не кончаются
- возможность выявления образцов, несоответствующих данной регрессионной модели является одним из преимуществ многомерного подхода в калибровке
- Deviation - эмпирический параметр, характеризующий меру соответствия нового образца калибровочной модели
- рассмотрим наш пример…

Диагностика предсказания (Simdata)



Принципы построения «хорошей» калибровки

- правильно приготовить (собрать) образцы
- визуально изучить данные, если необходимо, применить предварительную обработку данных (preprocessing)
- 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

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, обносить калибровку
- 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