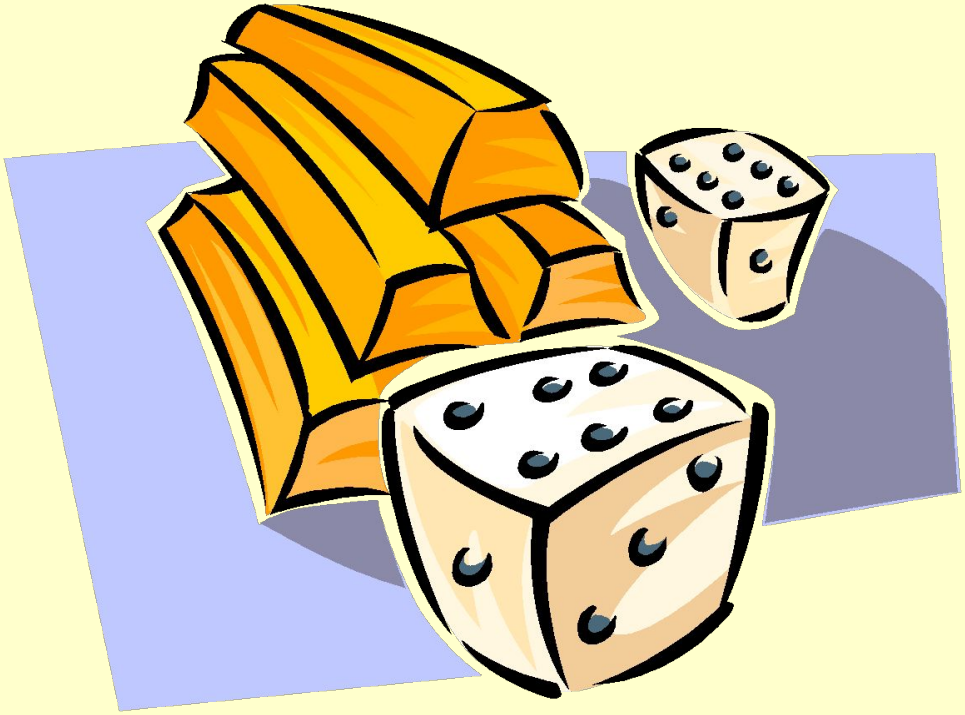


Построение регрессионных моделей и решение задачи предсказания

Два класса решаемых задач

X_{11}	X_{12}		…	X_{1m}
X_{21}	X_{22}		…	
. . .	. . .	X	. . .	. . .
…	…			…
X_{n1}				X_{nm}

Y_1
Y_2
Y . . .
…
Y_n

Методы : РГК, РЛС

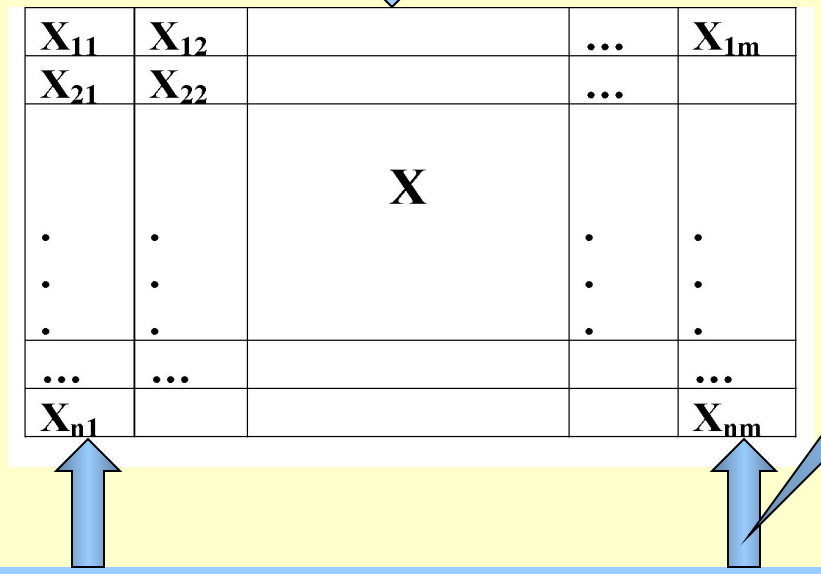
Задачи

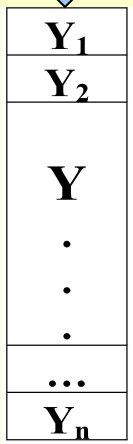
1. Построение модели Y(X)
2. Прогнозирование

Постановка задачи. Исходные данные

Независимые наблюдения
- предикторы

Зависимые переменные
- отклики



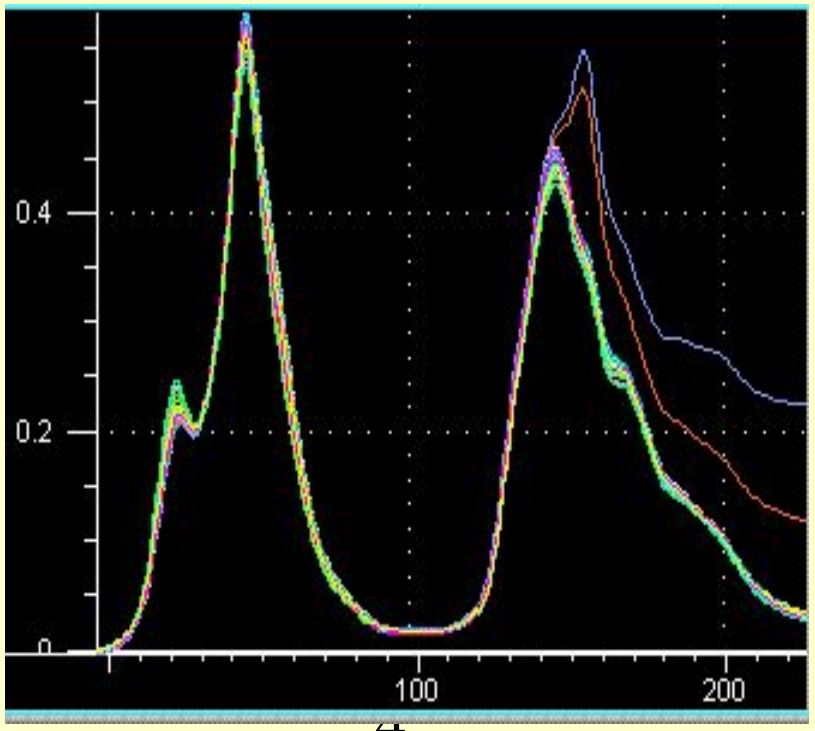


m -количество переменных
(факторов)

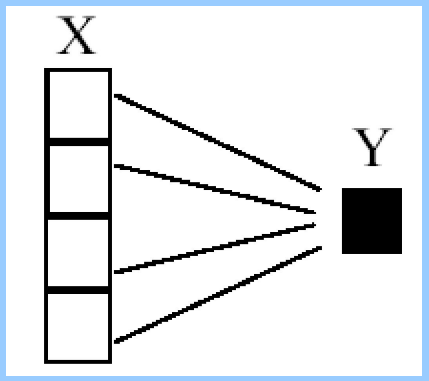
n –количество образцов
(наблюдений)

Цель исследования

1. Построить модель для известных наборов X и Y
2. Оценить возможности модели для предсказания неизвестных значений Y по новым значениям X.



Множественная регрессия.



$$y=Xb+f$$

$$y=b_0+b_1x_1+b_2x_2+..+b_mx_m+f$$

$$\hat{b}=(X^TX)^{-1}X^Ty$$

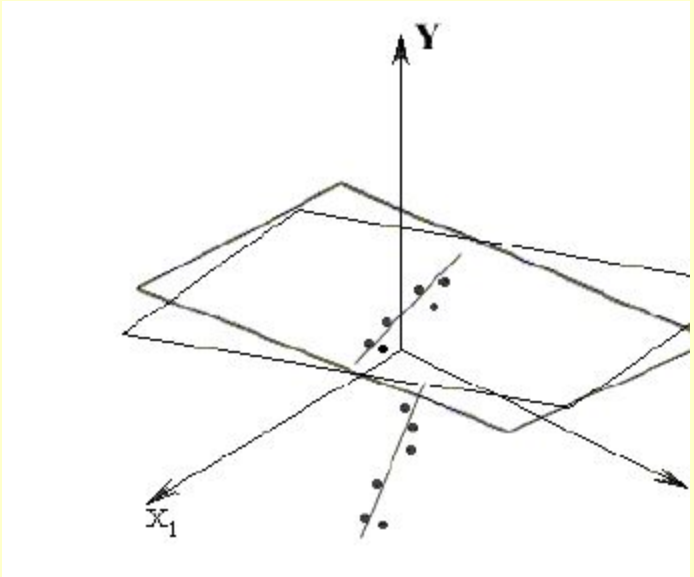
Проверка основных статистических гипотез об уравнении регрессии, его коэффициентах и прогнозируемых значениях откликов.

Сложности

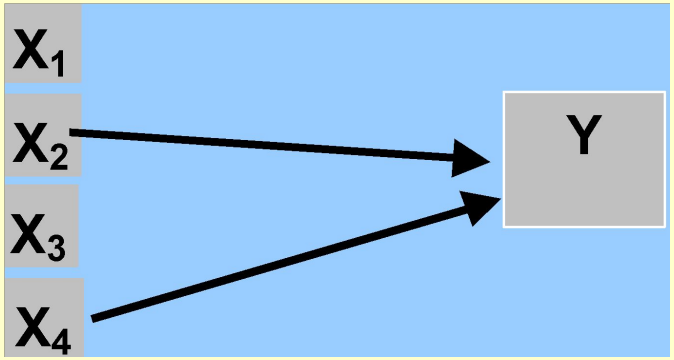
1. Количество переменных больше, чем количество образцов

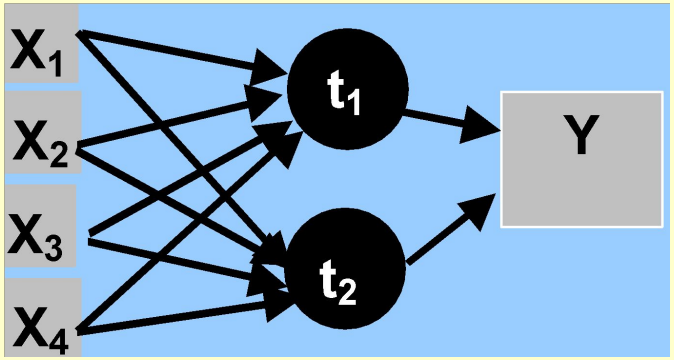
2. Наличие связей между переменными в X -матрице

Коллинеарность



Коллинеарность означает, что между переменными, составляющими матрицу X, существует взаимная корреляция, т.е. они в некоторой степени линейно зависимы между собой, например X1=f (X2, X3, …, Xn)

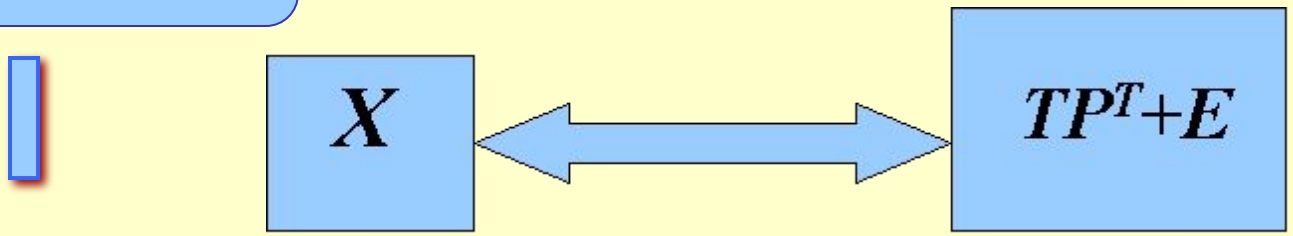


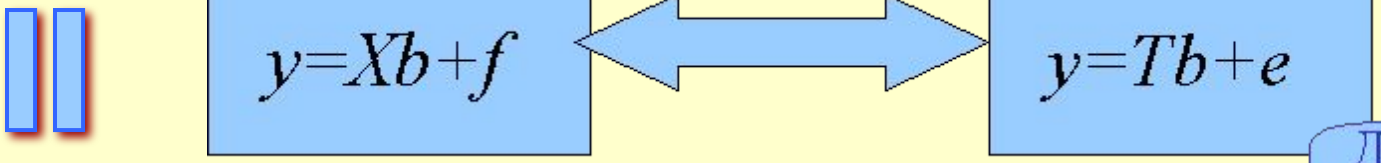


Регрессия на главные компоненты (РГК)

Для «нужного» количества ГК

Двухэтапная процедура РГК

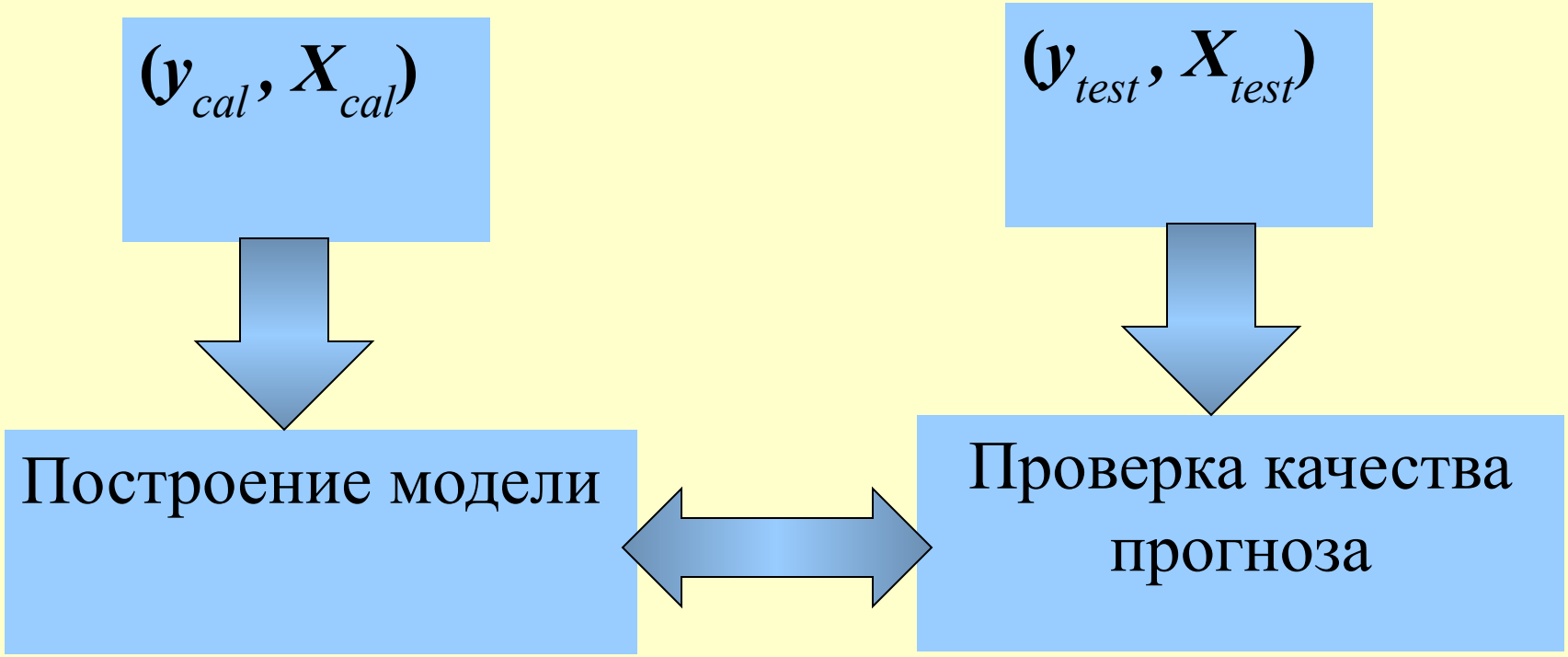




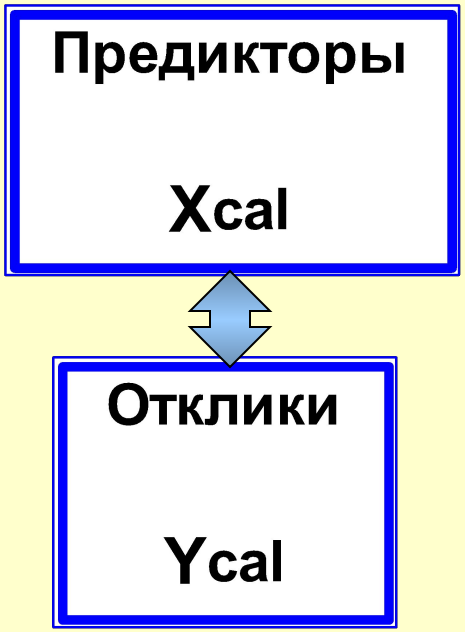
Диагональная матрица

$$\hat{b} = (X^T X)^{-1} X^T y \quad \Leftrightarrow \quad \hat{b} = P(T^T T)^{-1} P^T X^T y$$

Моделирование – хемометрический подход



Обучающий набор данных



1. Набор должен быть достаточно большим
2. Должны охватывать всю будущую совокупность
3. Измерения **X,** по возможности, должны быть несложными

Измеренные референтным методом

Планирование эксперимента

Теория пробоотбора

Построение модели

1. (y_{cal}, X_{cal}) → Модель

2. X_{cal} + Модель → $\hat{y}_{cal}$

3. $(y^i_{cal} - \hat{y}^i_{cal})^2$ → Ошибка моделирования

4. $\frac{\sum(\hat{y}_{cal} - y_{cal})^2}{n}$ → Калибровочная остаточная дисперсия

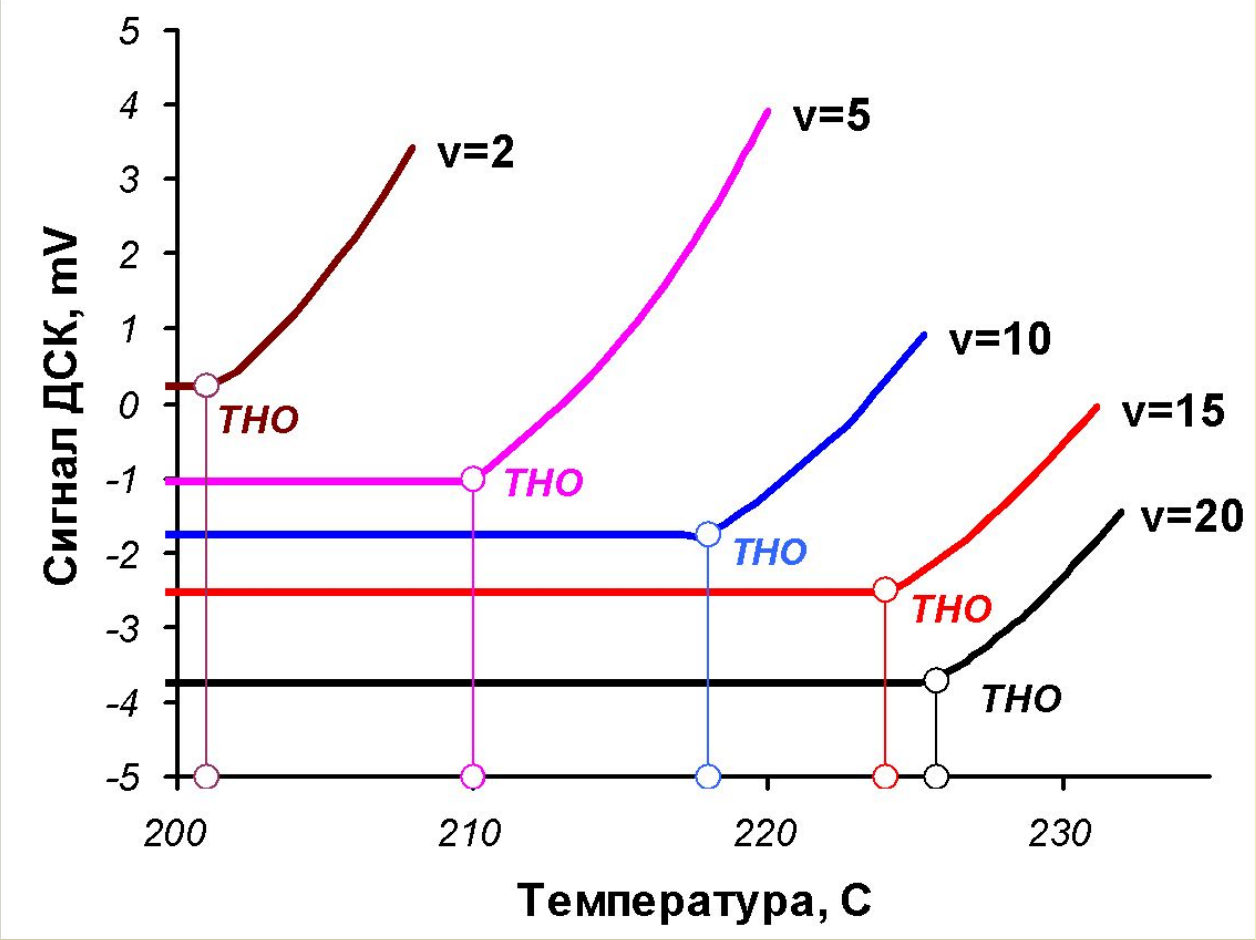
5. $RMSEC = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i,cal} - y_{cal})^2}{n}}$ → Стандартная ошибка калибровки

Оценка антиоксидантов методом ДСК

Объект	**Антиоксиданты в ПП**
Цель	**Оценка эффективности АО**
Y- измерения	**Длительное термостарение**
X- измерения	**Температура начала окисления**
Эксперимент	**Дифференц. калориметрия**
Обработка	**Регрессия на главные компоненты**

ДСК эксперимент

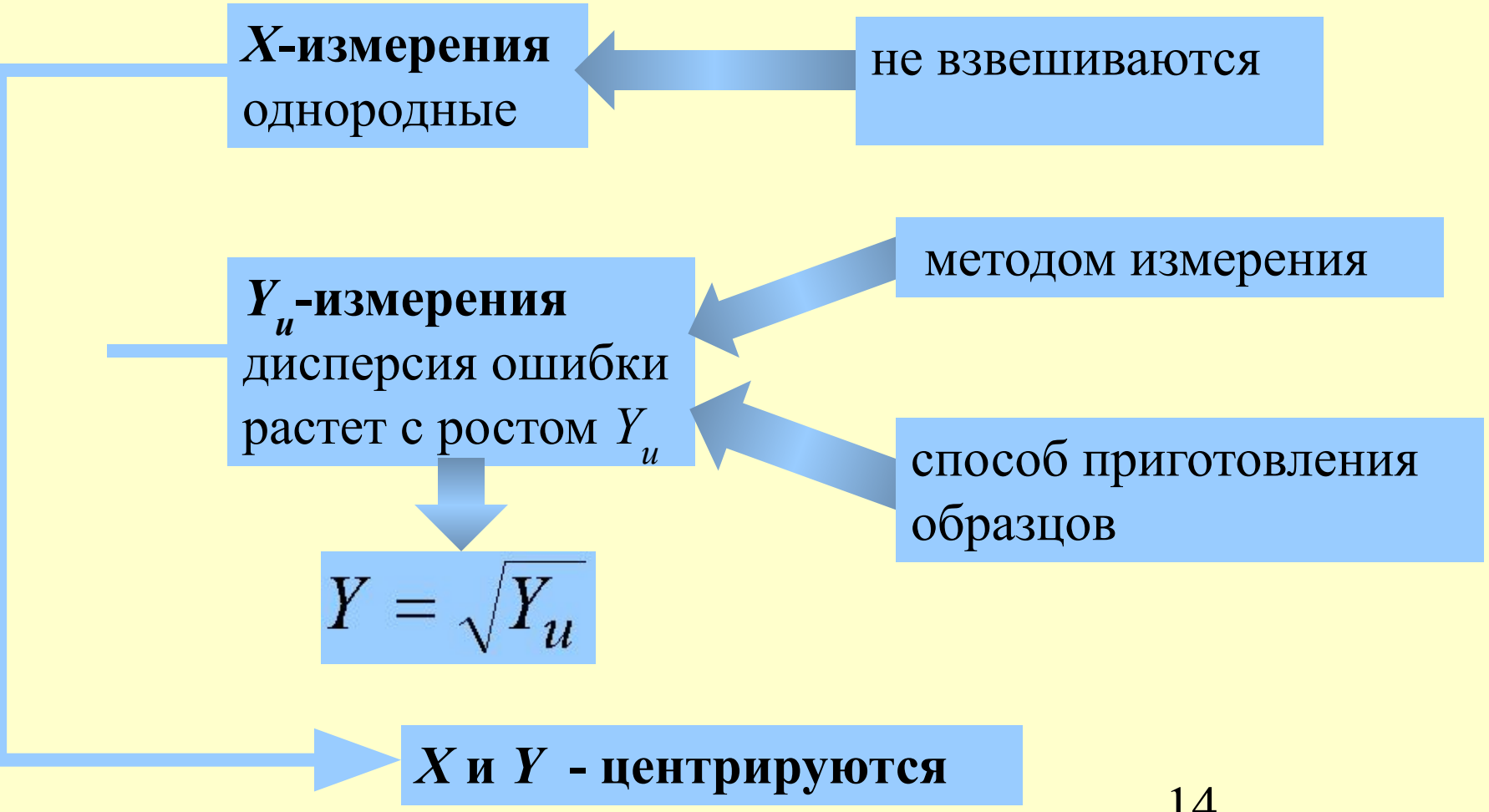
Оценка температуры начала окисления (ТНО) при разных скоростях нагрева v



ДСК данные и референтные данные

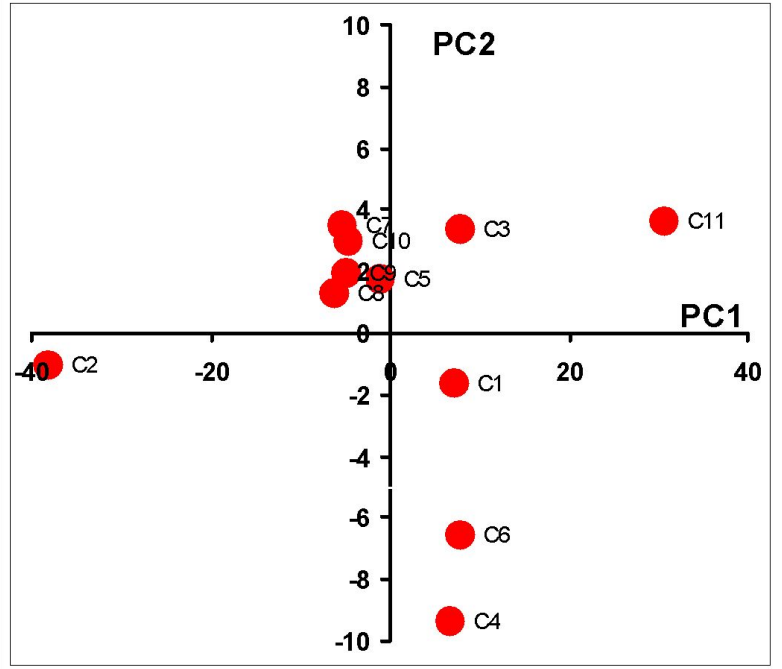
Образцы		Время старения в печке (дни)	ТНО (С) для разных скоростей нагрева (град/мин)				
			2	5	10	15	20
калибровка	C1	6	193.0	200.0	207.1	210.1	209.1
	C2	1	173.6	179.2	181.7	190.9	193.2
	C3	2	192.5	203.5	204.4	208.5	212.9
	C4	18	194.0	197.7	209.7	212.8	202.0
	C5	3	193.4	192.7	199.1	207.9	209.2
	C6	15	194.0	197.7	209.7	212.8	205.3
	C7	1.5	185.8	193.1	199.0	205.2	209.7
	C8	2.5	185.8	193.1	199.0	205.2	207.1
	C9	3	186.0	192.1	197.0	211.3	207.0
	C10	3	186.0	192.1	197.0	211.0	208.2
	C11	5	203.0	208.5	216.5	222.9	222.0
контроль	T1	0.5	185.0	191.7	197.0	197.2	211.2
	T2	17	194.0	197.7	209.7	212.8	203.1
	T3	8	186.8	191.0	208.2	205.1	205.1
	T4	5	203.9	213.9	220.2	221.4	227.2

Предварительная обработка данных.

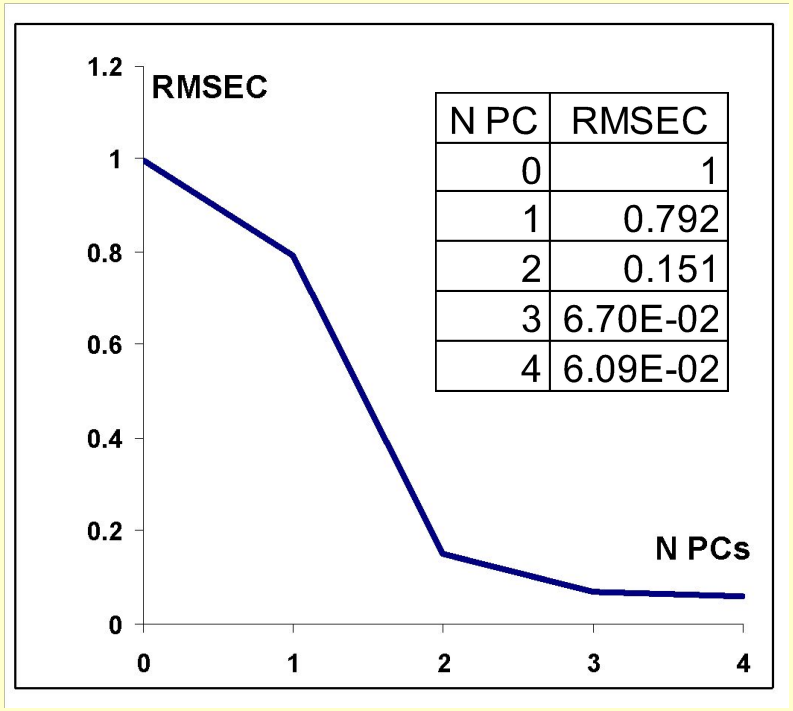


Метод главных компонент в примере с АО

График счетов (ГК1-ГК2)



Стандартная ошибка калибровки



N PC	RMSEC
0	1
1	0.792
2	0.151
3	6.70E-02
4	6.09E-02

ГК1-ГК2: объясняют 96% структуры X и 97 % структуры Y

Тестовый набор данных

Предикторы

Xtest

Отклики

Ytest

1. Набор должен быть достаточно большим
2. Должны охватывать всю будущую совокупность
3. Не должны быть «***слишком***» похож на калибровочный набор

Измеренные референтны методом

Используются только для оценки ошибки предсказания

Моделирование – стадия проверки

6. (y_{test}, X_{test}) → Используются для проверки качества прогноза

7. X_{test} + Модель → $\hat{y}_{test}$

8. $(y^i_{test} - \hat{y}^i_{test})^2$ → Ошибка прогнозирования

9. $\dfrac{\sum_{i=1}^{n}(y^i_{test} - \hat{y}^i_{test})^2}{n}$ → Проверочная дисперсия

10. $RMSEP = \sqrt{\dfrac{\sum_{i=1}^{n}(y^i_{test} - \hat{y}^i_{test})^2}{n}}$ → Стандартная ошибка прогноза

Способы проверки

Проверка на тестовом наборе ⟷ Самый надежный способ

Перекрестная проверка ⟷ Используется тогда, когда нельзя собрать тестовый массив

Проверка корректировкой размахом ⟷

1. Самый быстрый и самый грубый способ
2. Не использует тестовый массив

Перекрестная проверка

Тестовый набор отсутствует

(y_{test}, X_{test})

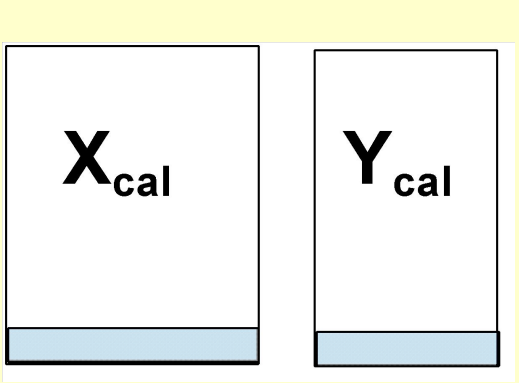
Самый медленный способ проверки и не всегда надежный

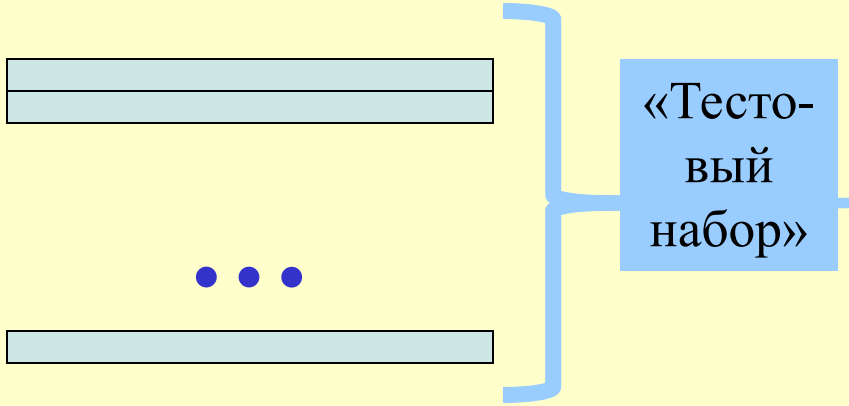
Моделируют тестовый набор используя калибровочный

(y_{cal}, X_{cal})

Создают как бы «тестовый массив»

Полная перекрестная проверка





Модель 1 | Модель 2 | . . . | Модель N

Модель

Проверка корректировкой размахом

«Быстрый»

«Грубый»

Требует построения лишь одной модели

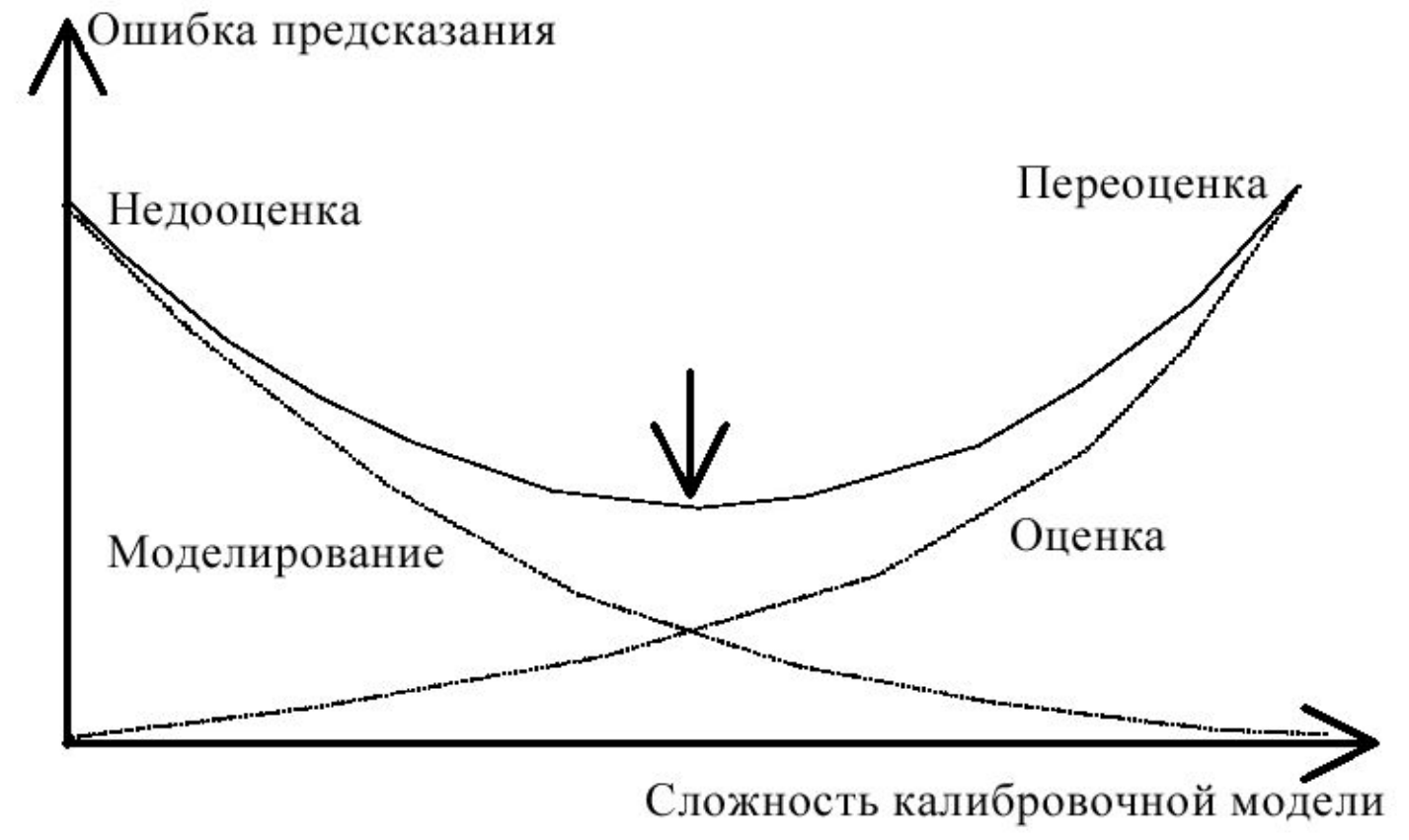
Ошибка предсказания всегда оценивается слишком оптимистично

Размах – это мера влияния образца на модель

$$h_i = \frac{1}{n} + \sum_{a=1}^{A} \frac{t_{ia}^2}{t_a^T t_a}$$

$$f_{ij}^{corrected} = \frac{f_{ij}}{1 - h_i}$$

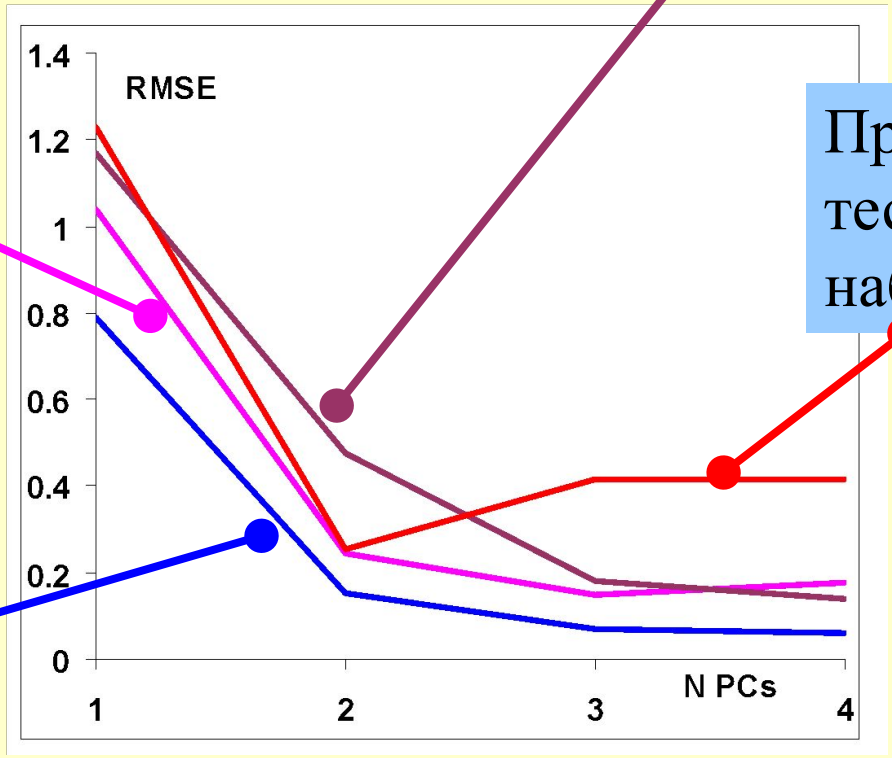
Сколько выбрать главных компонент



Ошибка моделирования и ошибка предсказания

Перекрестная проверка

Проверка корректировкой размахом

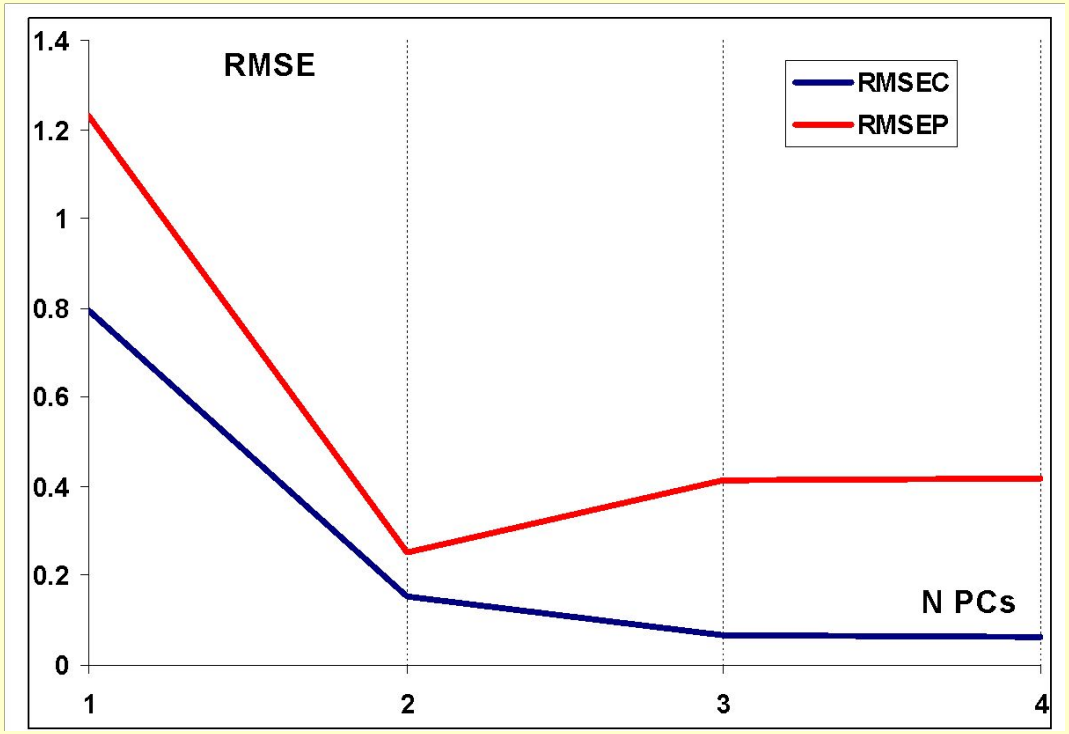


Проверка на тестовом наборе

Ошибка моделирования не зависит от вида проверки

Количество ГК для АО примера

N PCs	RMSEC	RMSEP
1	0.792	1.228
2	0.151	0.253
3	6.70E-02	0.414
4	6.09E-02	0.417



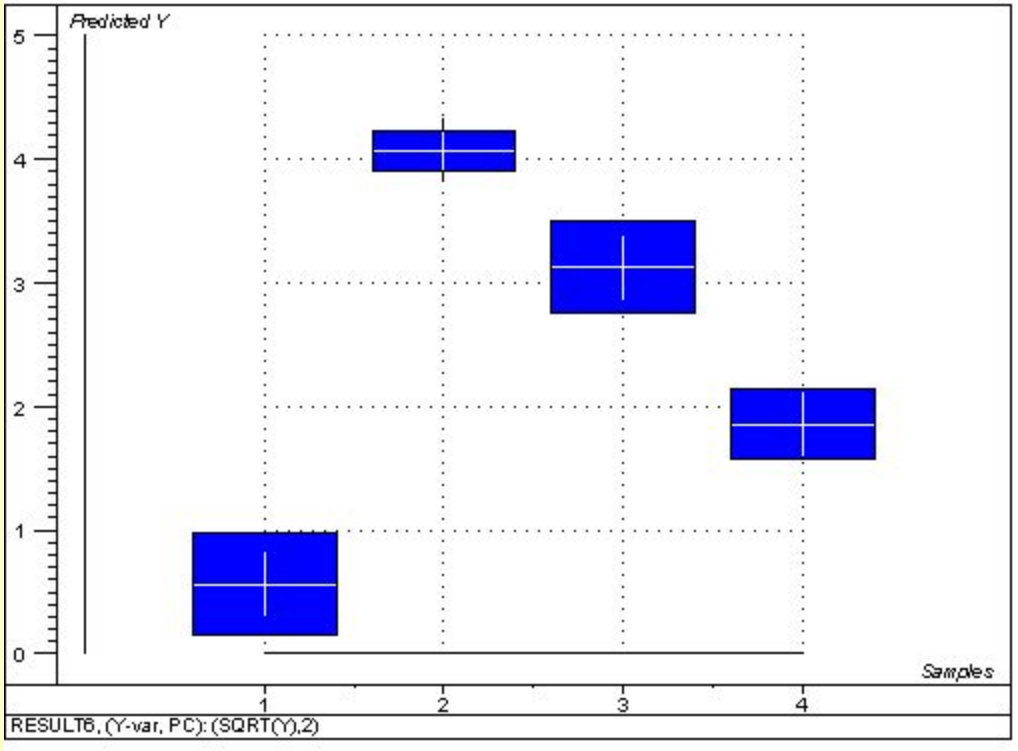
2 главные компоненты

Прогноз эффективности АО

Образец	Пред-ние	Откл-ие	Изм-ние
Ts1	0.564	0.407	0.707
Ts2	4.072	0.16	4.123
Ts3	3.125	0.371	2.828
Ts4	1.856	0.287	2.236

RMSEP = 0.253

Yпред=Y±2*RMSEP



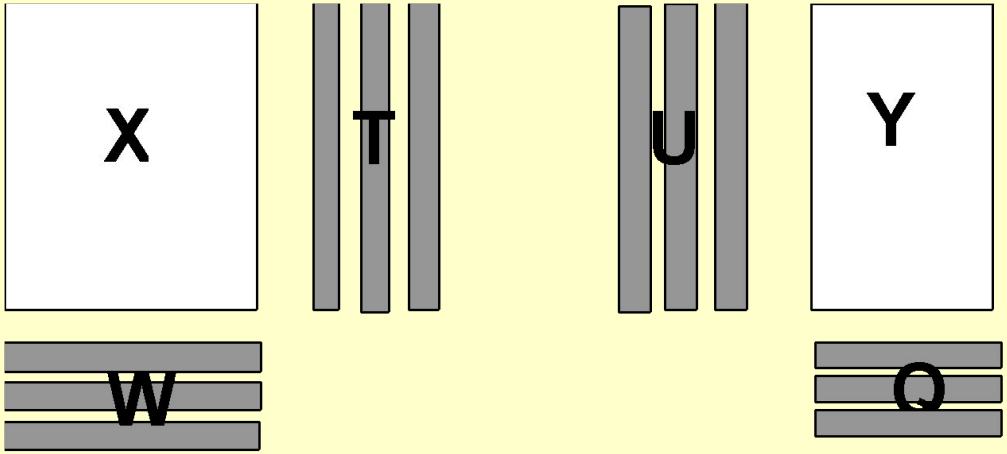
Слабость РГК

РГК – мощное средство борьбы с мультиколлинеарностью в матрице X

РГК –двухэтапный метод = **Декомпозиция *X* по МГК** + **МЛР**

Эта декомпозиция не учитывает связи между *X* и *Y*

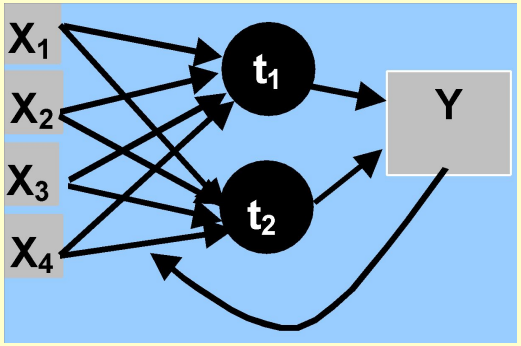
Регрессия на латентные структуры (ПЛС - регрессия)



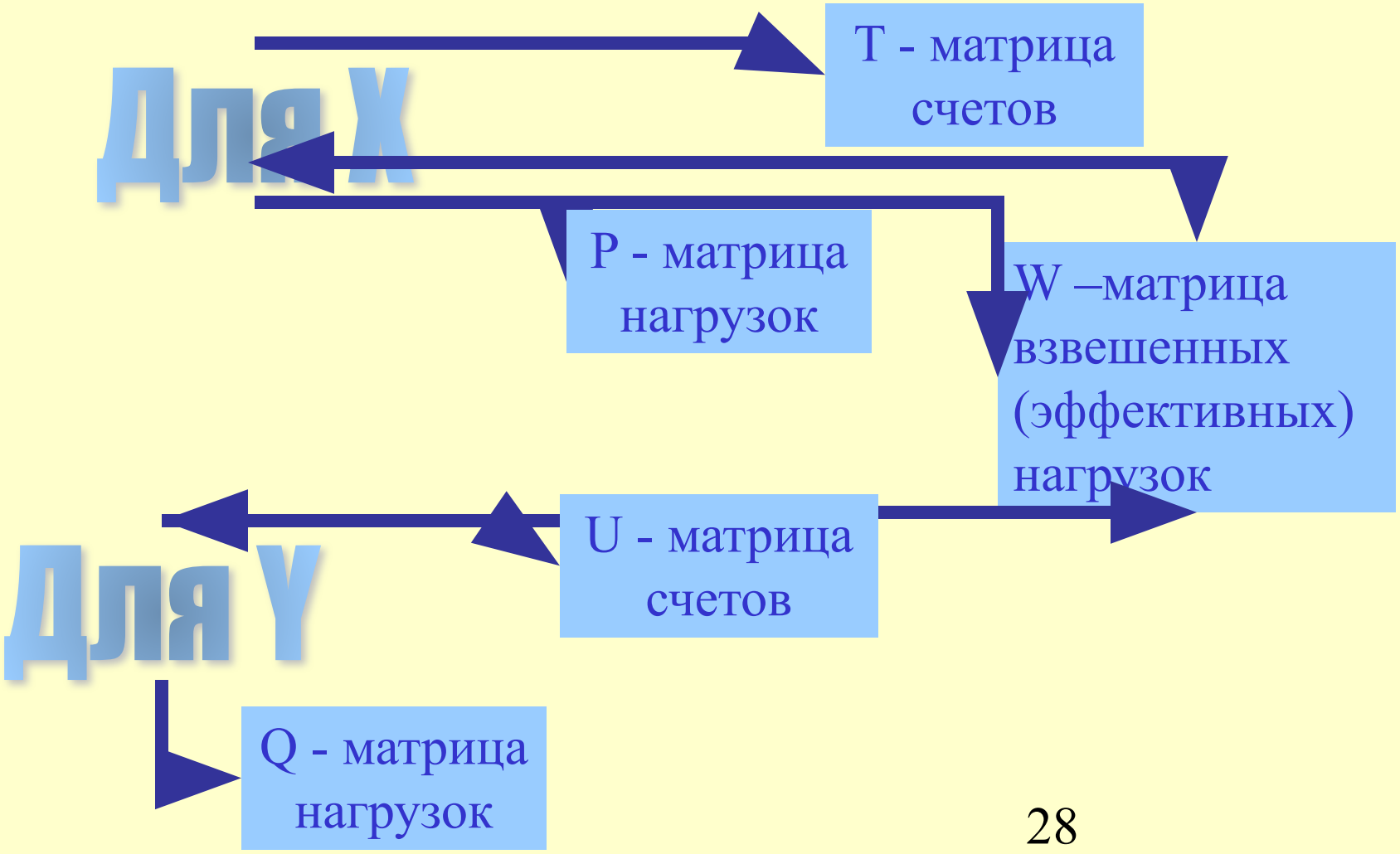
Схематическое представление

$$X = \sum_A TP^T + E$$

$$Y = \sum_A UQ^T + F$$

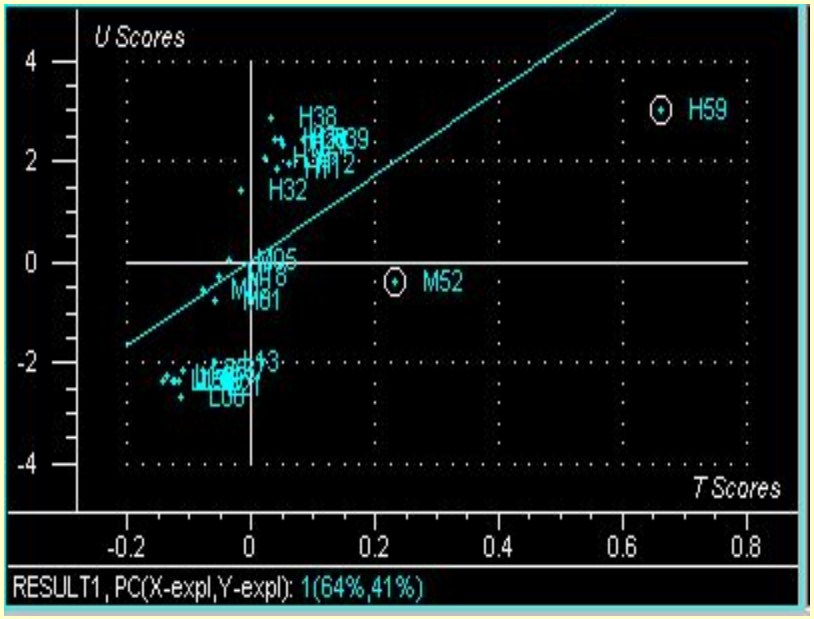


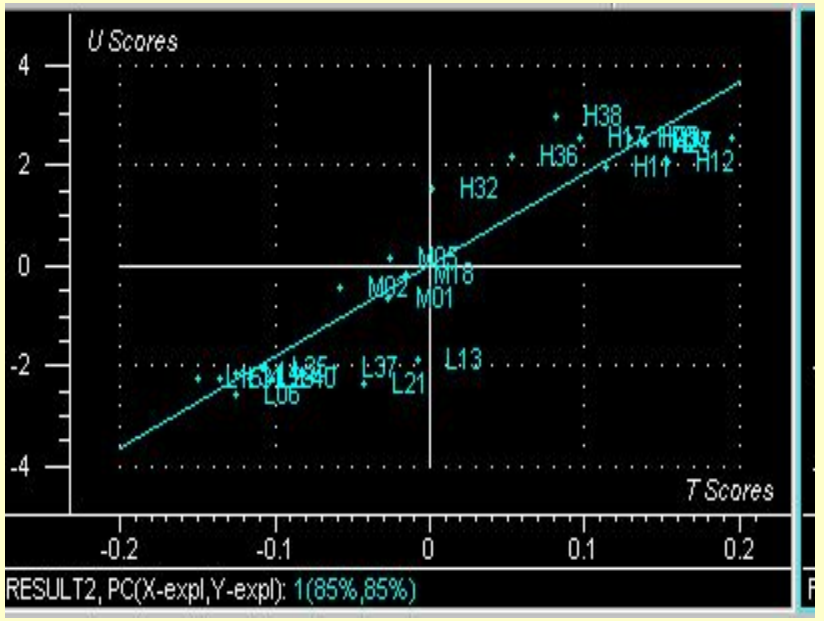
Интерпретация ПЛС-модели



Графике зависимости X-Y

U - T



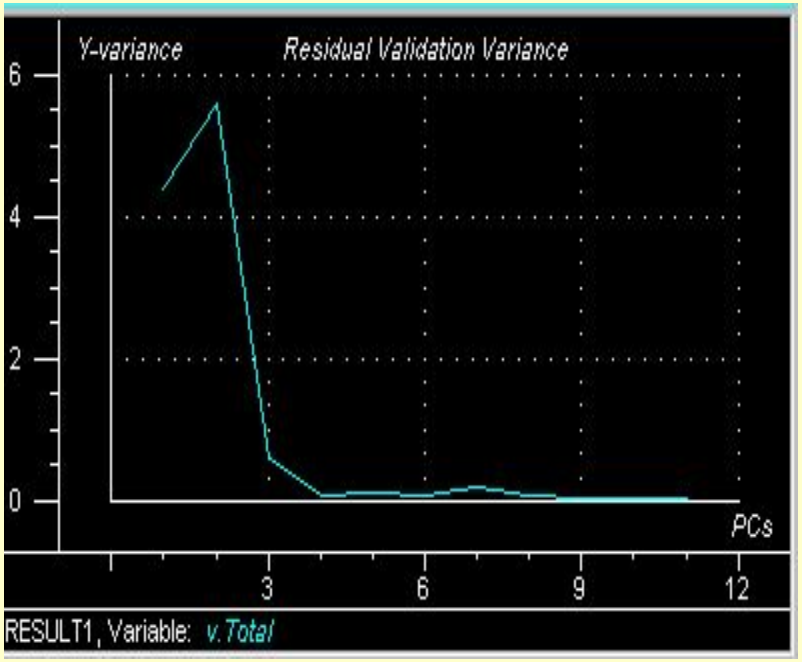


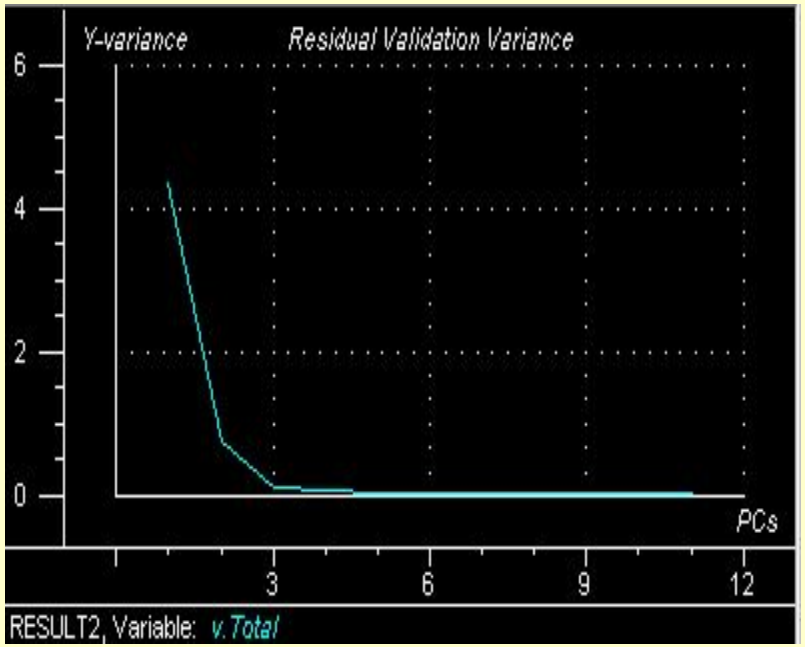
Данные содержат выбросы

Данные не содержат выбросы

График остаточной дисперсии

Остаточная дисперсия Y – количества ГК

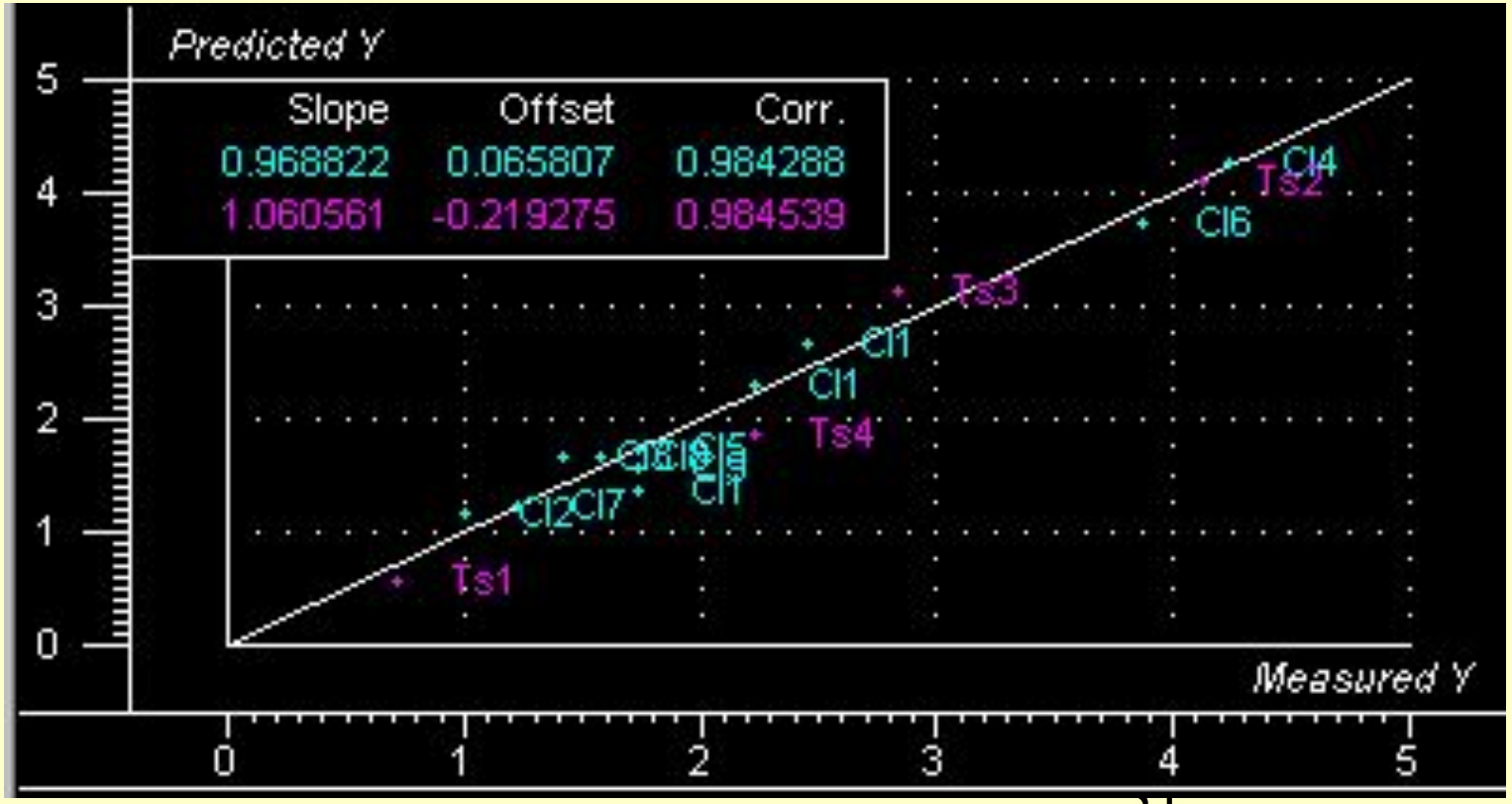




Для ПЛС-моделей дисперсия должна падать

Заключительный график

Предсказанные значения Y - измеренные значения Y



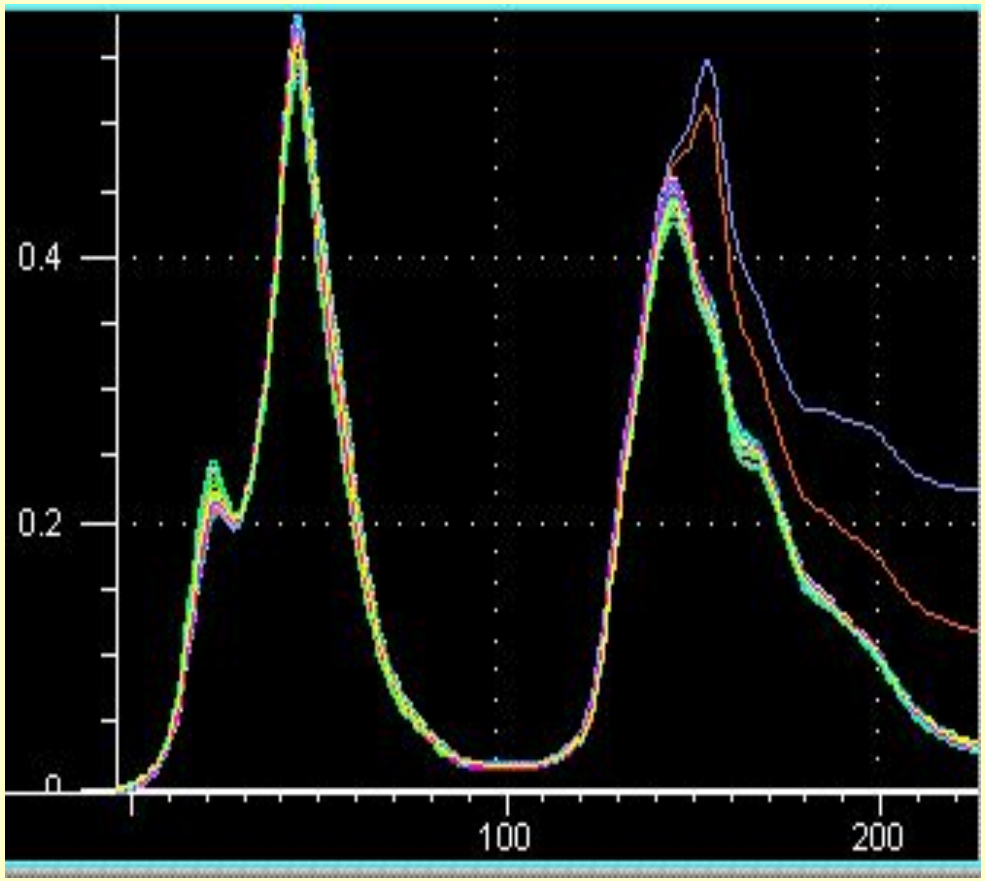
Определение октанового числа бензина по данным ИК-спектроскопии

Исходные данные

Обучающий массив = 26 образца

Прогнозный массив = 13 образцов

Количество переменных (длин волн) = 226 (1100 – 1550 nm)

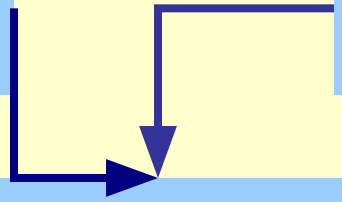


Выводы

Два основных проекционных регрессионных метода.

Регрессия на главные компоненты

Регрессия на латентные структуры.



1. Уменьшают размерность исследуемых данных
2. Позволяют проанализировать скрытые в данных закономерности

Выбор меньшего числа ГК дает более устойчивую модель

Проверка с помощью представительного тестового набора наиболее надежный способ оценки ошибки прогнозирования