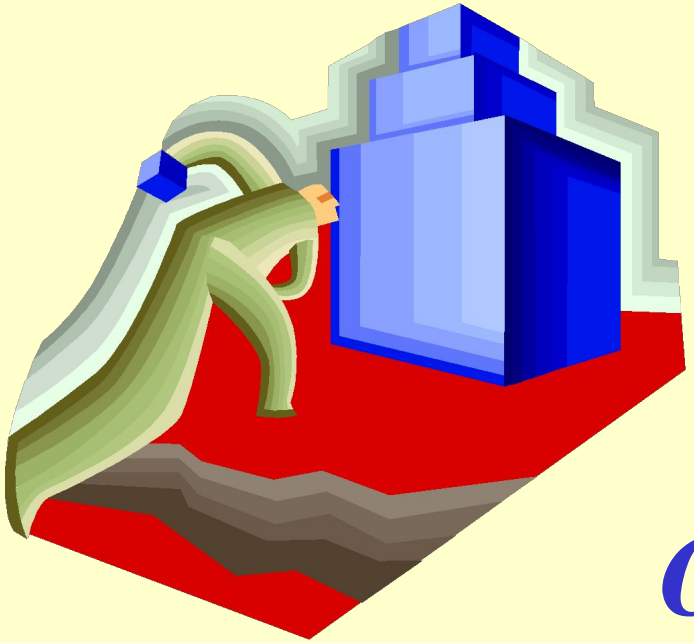


Проекционные методы. Основные понятия и примеры

Институт химической физики РАН, Москва

Родионова Оксана Евгеньевна

План лекции

1. **Ведение**
   - Немного истории
   - Природа многомерна
   - Пример - многомерный статистический контроль процессов
   - Два подхода к анализу данных
2. **Идеи, заложенные в проекционном подходе**
   - Данные – какие они бывают
   - Классы решаемых задач
   - **3. Метод главных компонент, основные понятия и примеры**

Метод наименьших квадратов

(простейший случай)

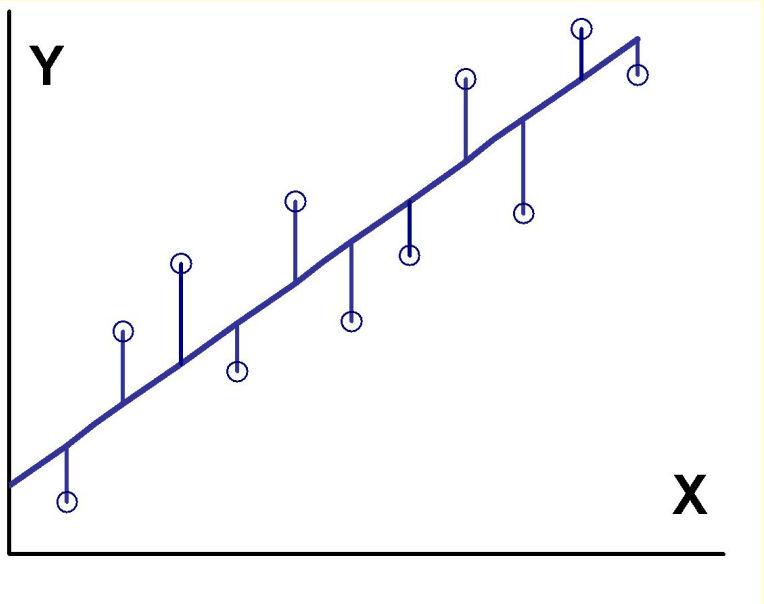
$$y = \alpha + \beta x$$

x_1	y_1
x_2	y_2
.	.
.	.
.	.
…	…
x_n	y_n

$$e_i$$

$$a \cong \alpha$$

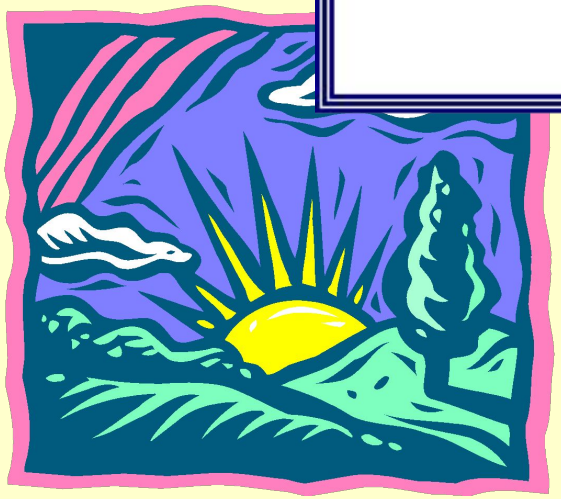
$$b \cong \beta$$



$$\sum_{i=1}^{n} (y_i - a - bx_i)^2 = \min$$

Многомерные данные

Исследуемое свойство <u>очень редко</u> зависит только от одной переменной



Контроль производственного процесса

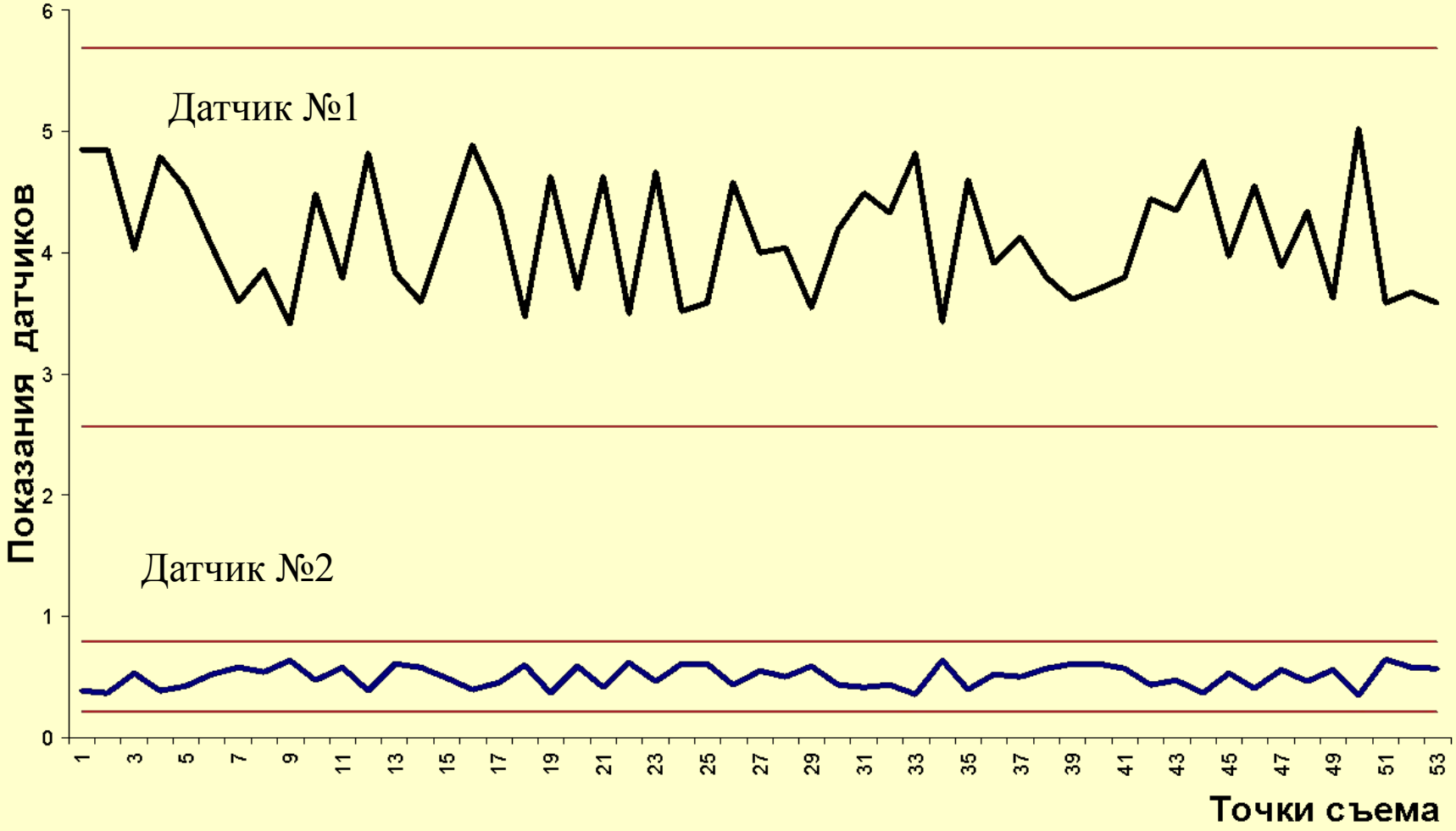
	X1	X2	X3	X4	X5	X6	X7	X8	X9	● ● ●	X17
s1	-1.19E-01	7.28E-01	-2.15E-02	5.22E-01	7.06E-04	7.32E-01	3.10E-04	-6.13E-04	-5.92E-05		9.74E-03
s2	-1.37E-01	7.28E-01	-2.89E-02	6.08E-01	7.09E-04	7.02E-01	6.58E-04	-1.22E-03	-1.49E-04		1.01E-02
s3	2.51E-02	-9.15E-02	6.73E-03	-1.13E-01	-9.07E-05	-7.58E-02	-2.29E-04	4.10E-04	5.65E-05		-1.43E-03
s4	-1.14E-01	6.70E-01	-2.18E-02	5.04E-01	6.50E-04	6.65E-01	3.83E-04	-7.34E-04	-7.96E-05		9.07E-03
s5	-7.93E-02	4.14E-01	-1.69E-02	3.51E-01	4.04E-04	3.98E-01	3.96E-04	-7.35E-04	-9.05E-05		5.78E-03
s6	1.51E-02	-6.38E-02	3.74E-03	-6.75E-02	-6.28E-05	-5.67E-02	-1.15E-04	2.07E-04	2.78E-05		-9.49E-04
s7	7.44E-02	-5.24E-01	1.11E-02	-3.24E-01	-5.06E-04	-5.45E-01	-1.73E-05	7.92E-05	-1.07E-05		-6.79E-03
s8	3.65E-02	-2.66E-01	5.12E-03	-1.59E-01	-2.56E-04	-2.78E-01	1.43E-05	-3.95E-07	-1.14E-05		-3.42E-03
s9	1.36E-01	-7.06E-01	2.89E-02	-6.01E-01	-6.88E-04	-6.77E-01	-6.83E-04	1.26E-03	1.56E-04		-9.86E-03
s10	-2.74E-02	3.60E-01	1.82E-03	1.12E-01	3.42E-04	4.12E-01	-4.31E-04	7.24E-04	1.22E-04		4.18E-03
s11	7.47E-02	-3.31E-01	1.80E-02	-3.34E-01	-3.25E-04	-2.99E-01	-5.30E-04	9.62E-04	1.28E-04		-4.84E-03
s12	-1.17E-01	7.02E-01	-2.16E-02	5.13E-01	6.81E-04	7.03E-01	3.40E-04	-6.63E-04	-6.76E-05		9.44E-03
s13	1.06E-01	-2.82E-01	3.23E-02	-4.82E-01	-2.85E-04	-1.87E-01	-1.25E-03	2.21E-03	3.14E-04		-4.99E-03
s14	7.39E-02	-5.28E-01	1.07E-02	-3.21E-01	-5.09E-04	-5.50E-01	2.49E-06	4.48E-05	-1.59E-05		-6.81E-03
s15	-9.87E-03	1.02E-01	-3.21E-04	4.17E-02	9.75E-05	1.13E-01	-8.29E-05	1.36E-04	2.44E-05		1.23E-03
s16	-1.06E-01	7.68E-01	-1.52E-02	4.62E-01	7.41E-04	8.03E-01	-2.54E-05	-2.68E-05	2.88E-05		9.90E-03
s17	-4.76E-02	2.66E-01	-9.52E-03	2.10E-01	2.59E-04	2.61E-01	1.92E-04	-3.61E-04	-4.19E-05		3.65E-03
	● ● ●										
s54	6.61E-02	-5.40E-01	7.19E-03	-2.85E-01	-5.19E-04	-5.78E-01	1.81E-04	-2.67E-04	-6.23E-05		-6.78E-03

Цель исследования

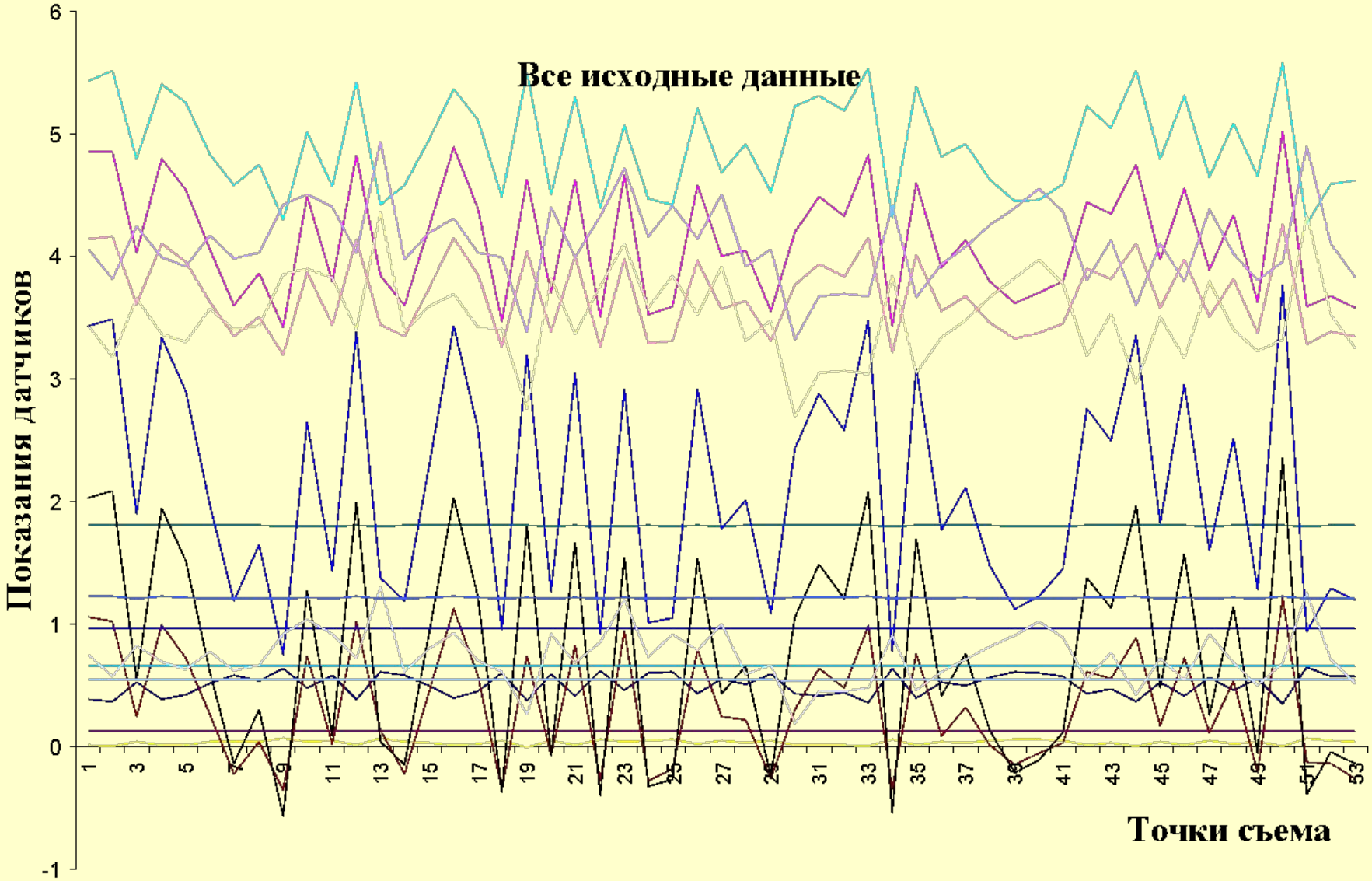
Контроль производства

Возможность воздействовать на процесс для его стабилизации

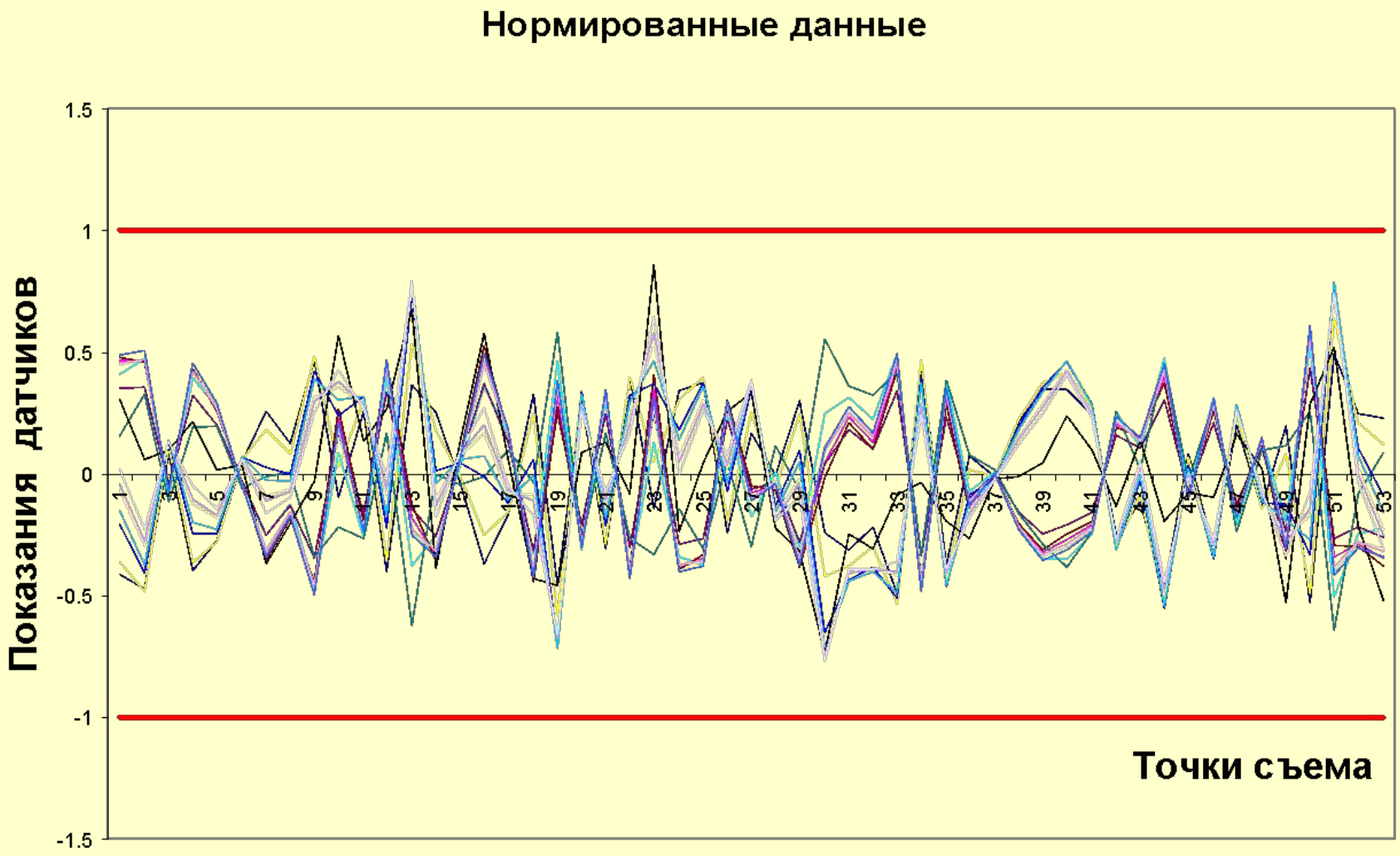
Контроль производственного процесса



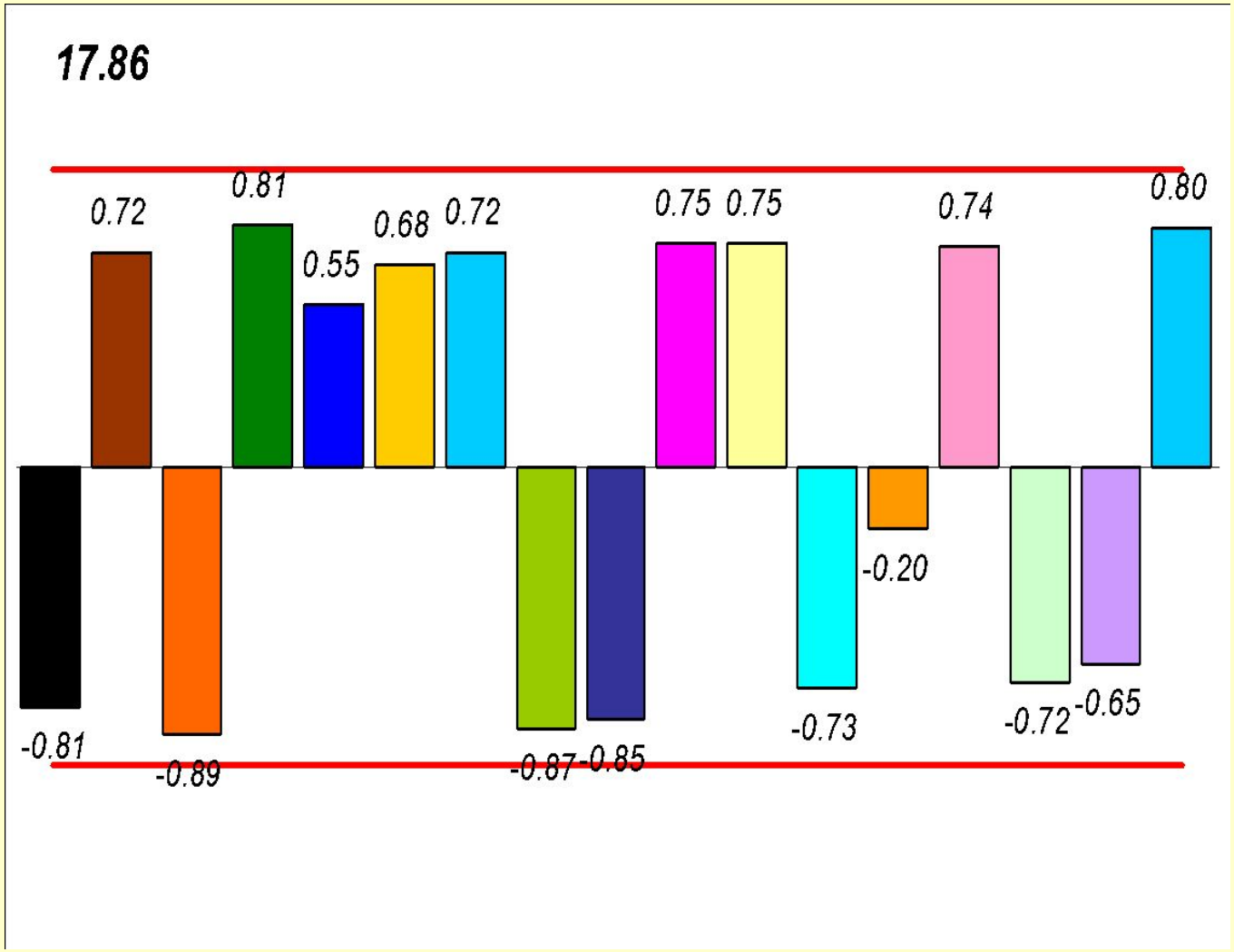
Контроль производственного процесса



Контроль производственного процесса



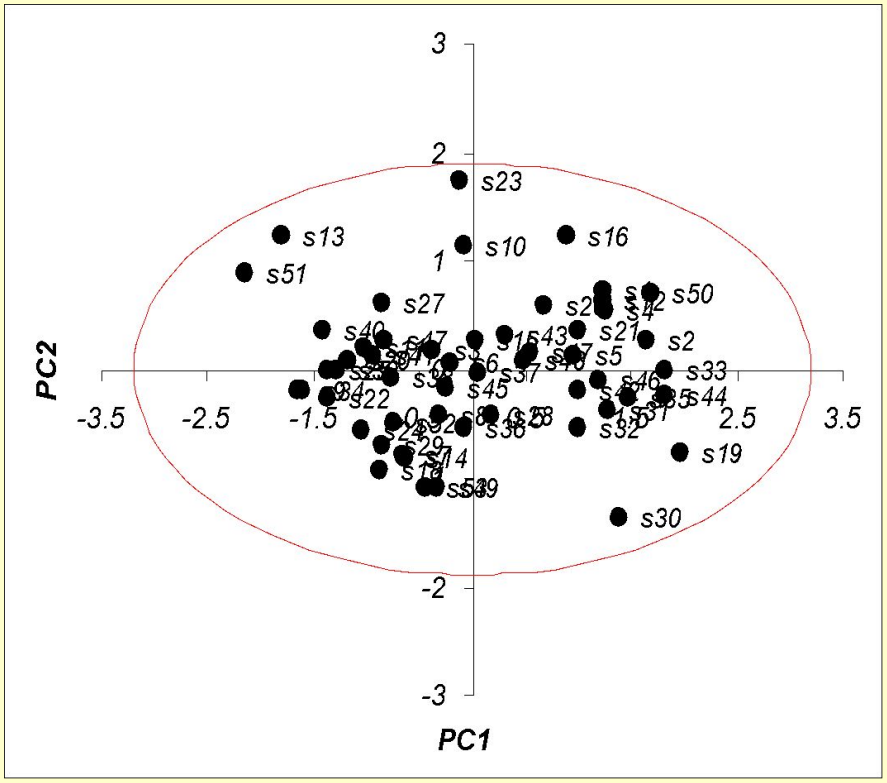
Контроль производственного процесса



On

Контроль производственного процесса

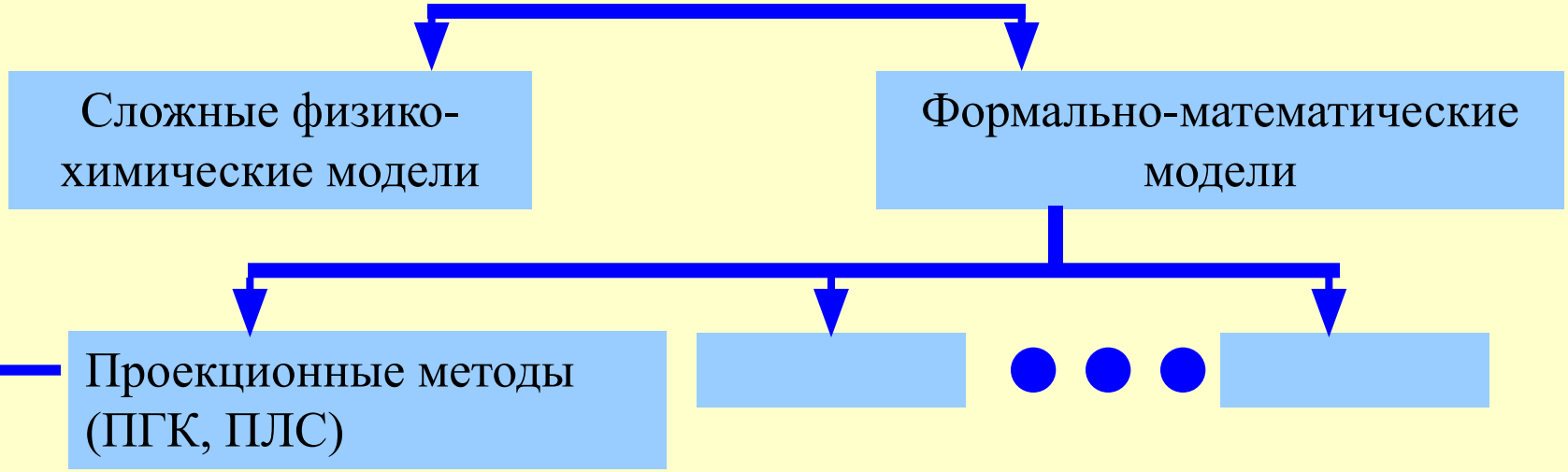
Точки съема



Моделирование производилось на основе анализа измерений и внутренних связей присущих этому набору данных

Не применялись содержательные физико-химические модели

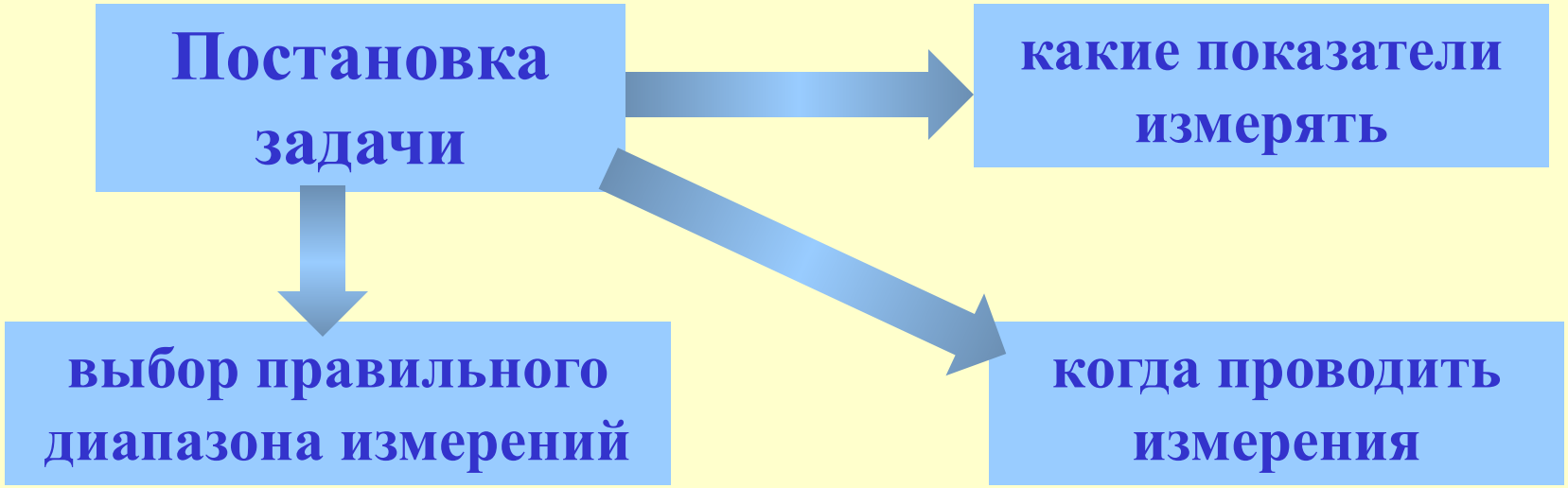
Моделирование многомерных данных (процессов или явлений)



1. Позволяют работать с достаточно большими массивами данных
2. Существенно понижают размерность изучаемой системы
3. Анализируют и позволяют выделить латентные структуры данных
4. Позволяют отделять содержательную часть от шума

Содержательная составляющая задачи.

Никакие многомерные методы не помогут, если данные не содержат полезной информации об изучаемом свойстве



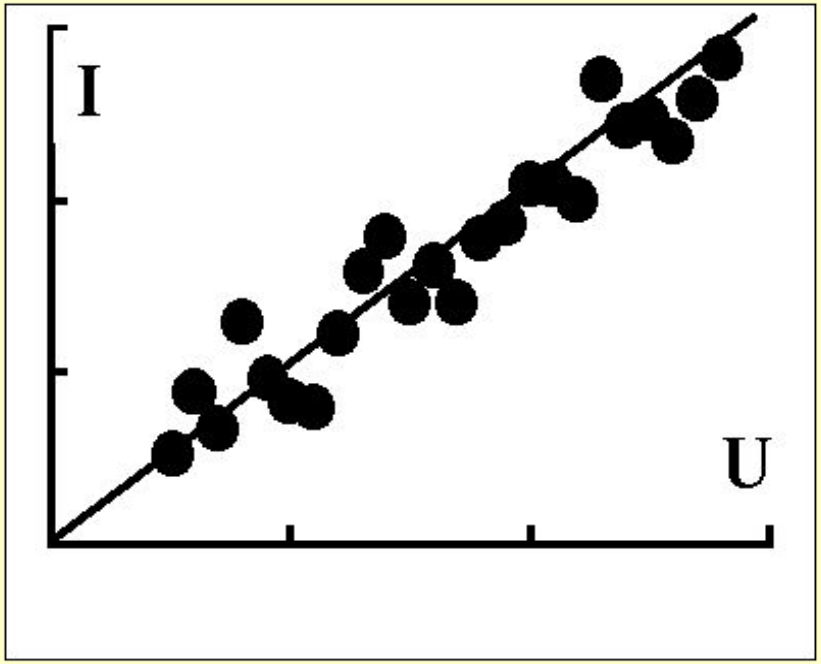
Данные

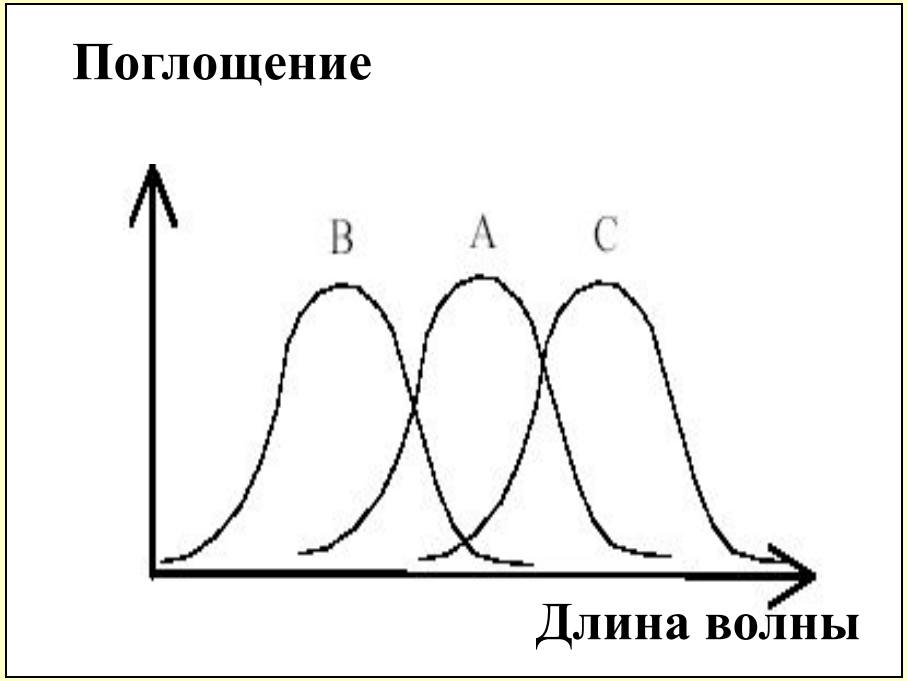
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
s1	-1.19E-01	7.28E-01	-2.15E-02	5.22E-01	7.06E-04	7.32E-01	3.10E-04	-6.13E-04	-5.92E-05	1.28E+00	1.32E+00	-7.03E-02	1.23E-03	4.67E-01	-4.21E-02	1.44E-02	9.74E-03
s2	-1.37E-01	7.28E-01	-2.89E-02	6.08E-01	7.09E-04	7.02E-01	6.58E-04	-1.22E-03	-1.49E-04	1.35E+00	1.39E+00	-3.27E-01	2.48E-04	4.84E-01	-2.94E-01	-1.62E-01	1.01E-02
s3	2.51E-02	-9.15E-02	6.73E-03	-1.13E-01	-9.07E-05	-7.58E-02	-2.29E-04	4.10E-04	5.65E-05	-1.96E-01	-2.02E-01	1.49E-01	3.83E-04	-6.80E-02	1.43E-01	9.43E-02	-1.43E-03
s4	-1.14E-01	6.70E-01	-2.18E-02	5.04E-01	6.50E-04	6.65E-01	3.83E-04	-7.34E-04	-7.96E-05	1.20E+00	1.24E+00	-1.36E-01	8.59E-04	4.34E-01	-1.09E-01	-3.60E-02	9.07E-03
s5	-7.93E-02	4.14E-01	-1.69E-02	3.51E-01	4.04E-04	3.98E-01	3.96E-04	-7.35E-04	-9.05E-05	7.71E-01	7.94E-01	-2.02E-01	7.80E-05	2.76E-01	-1.83E-01	-1.03E-01	5.78E-03
s6	1.51E-02	-6.38E-02	3.74E-03	-6.75E-02	-6.28E-05	-5.67E-02	-1.15E-04	2.07E-04	2.78E-05	-1.29E-01	-1.33E-01	7.08E-02	1.40E-04	-4.52E-02	6.71E-02	4.31E-02	-9.49E-04
s7	7.44E-02	-5.24E-01	1.11E-02	-3.24E-01	-5.06E-04	-5.45E-01	-1.73E-05	7.92E-05	-1.07E-05	-8.87E-01	-9.13E-01	-1.02E-01	-1.47E-03	-3.26E-01	-1.19E-01	-1.15E-01	-6.79E-03
s8	3.65E-02	-2.66E-01	5.12E-03	-1.59E-01	-2.56E-04	-2.78E-01	1.43E-05	-3.95E-07	-1.14E-05	-4.46E-01	-4.59E-01	-6.86E-02	-8.12E-04	-1.64E-01	-7.71E-02	-6.99E-02	-3.42E-03
s9	1.36E-01	-7.06E-01	2.89E-02	-6.01E-01	-6.88E-04	-6.77E-01	-6.83E-04	1.26E-03	1.56E-04	-1.31E+00	-1.35E+00	3.50E-01	-1.12E-04	-4.71E-01	3.18E-01	1.80E-01	-9.86E-03
s10	-2.74E-02	3.60E-01	1.82E-03	1.12E-01	3.42E-04	4.12E-01	-4.31E-04	7.24E-04	1.22E-04	5.29E-01	5.43E-01	3.97E-01	2.27E-03	2.02E-01	4.03E-01	3.03E-01	4.18E-03
s11	7.47E-02	-3.31E-01	1.80E-02	-3.34E-01	-3.25E-04	-2.99E-01	-5.30E-04	9.62E-04	1.28E-04	-6.54E-01	-6.74E-01	3.20E-01	5.44E-04	-2.31E-01	3.01E-01	1.91E-01	-4.84E-03
s12	-1.17E-01															-6.77E-03	9.44E-03
s13	1.06E-01															5.65E-01	-4.99E-03
s14	7.39E-02															-1.26E-01	-6.81E-03
s15	-9.87E-03															6.61E-02	1.23E-03
s16	-1.06E-01															1.94E-01	9.90E-03
s17	-4.76E-02															-3.44E-02	3.65E-03
s18	9.54E-02															-1.20E-01	-8.53E-03
s19	-1.32E-01															-4.71E-01	7.72E-03
s20	8.91E-02															1.82E-01	-6.07E-03
s21	-8.91E-02															-4.98E-02	6.92E-03
s22	1.15E-01															1.19E-01	-8.60E-03
s23	-4.08E-02															4.63E-01	6.30E-03
s24	9.92E-02															7.74E-04	-8.06E-03
s25	1.08E-01															1.85E-01	-7.60E-03
s26	-6.95E-02															5.55E-02	6.00E-03
s27	4.90E-02															2.72E-01	-2.23E-03
s28	-1.65E-03															-1.37E-01	-7.53E-04
s29	8.73E-02	-5.70E-01	1.46E-02	-3.82E-01	-5.52E-04	-5.83E-01	-1.37E-04	2.94E-04	1.91E-05	-9.87E-01	-1.02E+00	-2.30E-02	-1.27E-03	-3.61E-01	-4.36E-02	-6.48E-02	-7.52E-03
s30	-6.93E-02	6.98E-02	-2.51E-02	3.19E-01	7.94E-05	-2.22E-02	1.11E-03	-1.96E-03	-2.85E-04	3.19E-01	3.31E-01	-8.06E-01	-2.95E-03	9.79E-02	-7.88E-01	-5.47E-01	2.10E-03
s31	-8.99E-02	3.66E-01	-2.28E-02	4.03E-01	3.61E-04	3.20E-01	7.21E-04	-1.30E-03	-1.76E-04	7.48E-01	7.72E-01	-4.52E-01	-9.80E-04	2.63E-01	-4.30E-01	-2.79E-01	5.51E-03
s32	-6.32E-02	2.05E-01	-1.79E-02	2.85E-01	2.05E-04	1.59E-01	6.44E-04	-1.15E-03	-1.61E-04	4.63E-01	4.78E-01	-4.31E-01	-1.23E-03	1.59E-01	-4.15E-01	-2.78E-01	3.35E-03
s33	-1.42E-01	6.98E-01	-3.20E-02	6.33E-01	6.83E-04	6.57E-01	8.32E-04	-1.53E-03	-1.95E-04	1.33E+00	1.37E+00	-4.62E-01	-3.32E-04	4.74E-01	-4.28E-01	-2.57E-01	9.92E-03
s34	1.32E-01	-6.89E-01	2.81E-02	-5.85E-01	-6.72E-04	-6.62E-01	-6.61E-04	1.23E-03	1.51E-04	-1.28E+00	-1.32E+00	3.38E-01	-1.27E-04	-4.60E-01	3.06E-01	1.73E-01	-9.62E-03
s35	-1.08E-01	4.80E-01	-2.61E-02	4.84E-01	4.72E-04	4.35E-01	7.66E-04	-1.39E-03	-1.84E-04	9.49E-01	9.79E-01	-4.61E-01	-7.76E-04	3.36E-01	-4.34E-01	-2.75E-01	7.03E-03
s36	2.13E-02	-2.11E-01	9.91E-04	-9.02E-02	-2.02E-04	-2.34E-01	1.57E-04	-2.55E-04	-4.66E-05	-3.28E-01	-3.38E-01	-1.62E-01	-1.06E-03	-1.24E-01	-1.67E-01	-1.29E-01	-2.56E-03
s37	-2.39E-03	4.55E-03	-7.90E-04	1.09E-02	4.75E-06	1.94E-03	3.28E-05	-5.80E-05	-8.35E-06	1.36E-02	1.41E-02	-2.32E-02	-7.97E-05	4.43E-03	-2.26E-02	-1.55E-02	9.42E-05
s38	6.29E-02	-3.26E-01	1.35E-02	-2.79E-01	-3.18E-04	-3.13E-01	-3.21E-04	5.94E-04	7.36E-05	-6.09E-01	-6.27E-01	1.66E-01	-3.74E-05	-2.18E-01	1.51E-01	8.57E-02	-4.56E-03
s39	1.02E-01	-5.03E-01	2.26E-02	-4.51E-01	-4.91E-04	-4.74E-01	-5.80E-04	1.07E-03	1.35E-04	-9.53E-01	-9.83E-01	3.19E-01	1.85E-04	-3.40E-01	2.94E-01	1.75E-01	-7.12E-03
s40	1.00E-01	-4.22E-01	2.50E-02	-4.49E-01	-4.16E-04	-3.75E-01	-7.68E-04	1.39E-03	1.86E-04	-8.52E-01	-8.79E-01	4.75E-01	9.52E-04	-3.00E-01	4.51E-01	2.90E-01	-6.29E-03
s41	7.05E-02	-3.24E-01	1.66E-02	-3.14E-01	-3.17E-04	-2.97E-01	-4.70E-04	8.56E-04	1.12E-04	-6.31E-01	-6.51E-01	2.77E-01	3.94E-04	-2.24E-01	2.60E-01	1.62E-01	-4.69E-03
s42	-7.27E-02	3.21E-01	-1.76E-02	3.24E-01	3.15E-04	2.90E-01	5.17E-04	-9.38E-04	-1.25E-04	6.35E-01	6.55E-01	-3.12E-01	-5.35E-04	2.25E-01	-2.94E-01	-1.87E-01	4.70E-03
s43	-3.34E-02	2.26E-01	-5.31E-03	1.46E-01	2.18E-04	2.33E-01	3.24E-05	-7.79E-05	-1.84E-06	3.87E-01	3.98E-01	2.54E-02	5.64E-04	1.42E-01	3.33E-02	3.69E-02	2.96E-03
s44	-1.36E-01	6.25E-01	-3.21E-02	6.08E-01	6.13E-04	5.73E-01	9.11E-04	-1.66E-03	-2.18E-04	1.22E+00	1.26E+00	-5.37E-01	-7.70E-04	4.32E-01	-5.04E-01	-3.15E-01	9.06E-03
s46	2.41E-02	-1.53E-01	4.17E-03	-1.06E-01	-1.48E-04	-1.55E-01	-4.92E-05	1.01E-04	8.29E-06	-2.67E-01	-2.75E-01	2.90E-03	-3.05E-04	-9.74E-02	-2.79E-03	-1.12E-02	-2.03E-03
s47	-9.11E-02	4.29E-01	-2.11E-02	4.06E-01	4.21E-04	3.98E-01	5.77E-04	-1.05E-03	-1.37E-04	8.29E-01	8.55E-01	-3.33E-01	-3.91E-04	2.95E-01	-3.11E-01	-1.91E-01	6.17E-03
s49	5.84E-02	-2.37E-01	1.49E-02	-2.62E-01	-2.34E-04	-2.07E-01	-4.71E-04	8.49E-04	1.15E-04	-4.85E-01	-5.00E-01	2.96E-01	6.47E-04	-1.70E-01	2.82E-01	1.83E-01	-3.57E-03
s50	-4.05E-02	2.15E-01	-8.50E-03	1.79E-01	2.10E-04	2.08E-01	1.93E-04	-3.58E-04	-4.36E-05	3.98E-01	4.10E-01	-9.53E-02	7.91E-05	1.43E-01	-8.56E-02	-4.69E-02	2.99E-03
s51	5.69E-02	-4.96E-01	5.12E-03	-2.44E-01	-4.76E-04	-5.37E-01	2.36E-04	-3.68E-04	-7.53E-05	-7.94E-01	-8.16E-01	-2.82E-01	-2.11E-03	-2.96E-01	-2.95E-01	-2.36E-01	-6.15E-03
s52	-1.54E-01	9.00E-01	-2.94E-02	6.78E-01	8.74E-04	8.93E-01	5.19E-04	-9.96E-04	-1.08E-04	1.61E+00	1.66E+00	-1.87E-01	1.14E-03	5.84E-01	-1.51E-01	-5.12E-02	1.22E-02
s54	1.44E-01	-5.35E-01	3.82E-02	-6.45E-01	-5.30E-04	-4.48E-01	-1.28E-03	2.30E-03	3.16E-04	-1.13E+00	-1.17E+00	8.31E-01	2.08E-03	-3.95E-01	7.95E-01	5.24E-01	-8.29E-03
s55	7.20E-02	-4.51E-01	1.27E-02	-3.16E-01	-4.37E-04	-4.56E-01	-1.63E-04	3.29E-04	2.92E-05	-7.91E-01	-8.14E-01	2.23E-02	-8.46E-04	-2.88E-01	5.29E-03	-2.35E-02	-6.00E-03
s56	6.61E-02	-5.40E-01	7.19E-03	-2.85E-01	-5.19E-04	-5.78E-01	1.81E-04	-2.67E-04	-6.23E-05	-8.78E-01	-9.03E-01	-2.51E-01	-2.08E-03	-3.26E-01	-2.67E-01	-2.19E-01	-6.78E-03

1. **Количественные и качественные**
2. **Управляемые и неуправляемые**
3. **Прямые измерения и косвенные**

Данные

Данные → **Структура** + **Шум**





Два класса решаемых задач

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

Метод- МГК

Задачи

1. Анализ структуры, поиск латентных переменных
2. Классификация и дискриминация

Y_{11}	Y_{12}	…	Y_{1k}
Y_{21}	Y_{22}	…	
.	.	.	.
.	.	.	.
.	.	.	.
…	…		…
Y_{n1}			Y_{nk}

Методы : РГК, ПЛС

Задачи

1. Построение модели Y(X)
2. Прогнозирование

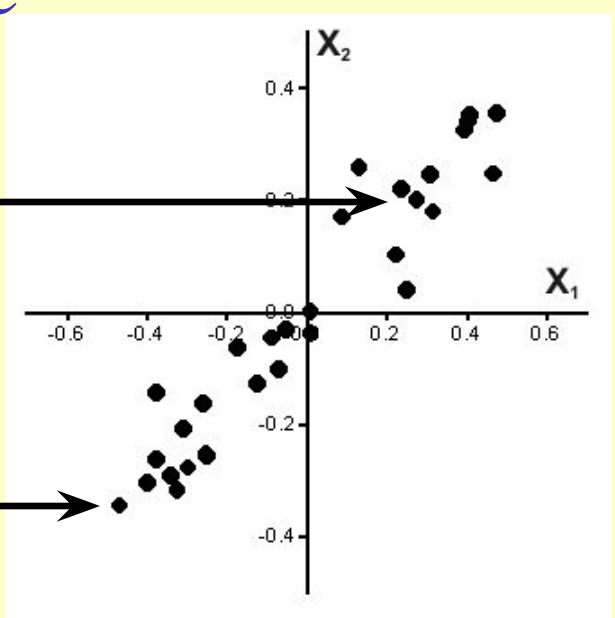
n – количество образцов

m – количество переменных (факторов)

Проекционные методы

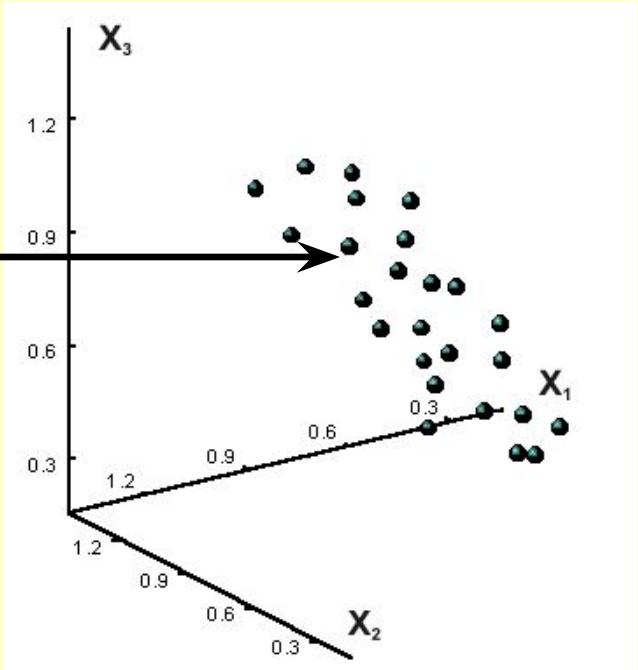
P=2

	X_1	X_2
1	0.407	0.353
2	0.475	0.355
3	0.274	0.202
4	0.394	0.325
5	-0.088	-0.045
6	-0.053	-0.031
7	-0.253	-0.253
8	-0.124	-0.128
9	-0.251	-0.255
10	0.088	0.171
11	-0.261	-0.162
12	0.401	0.341
13	-0.469	-0.344
14	-0.376	-0.143



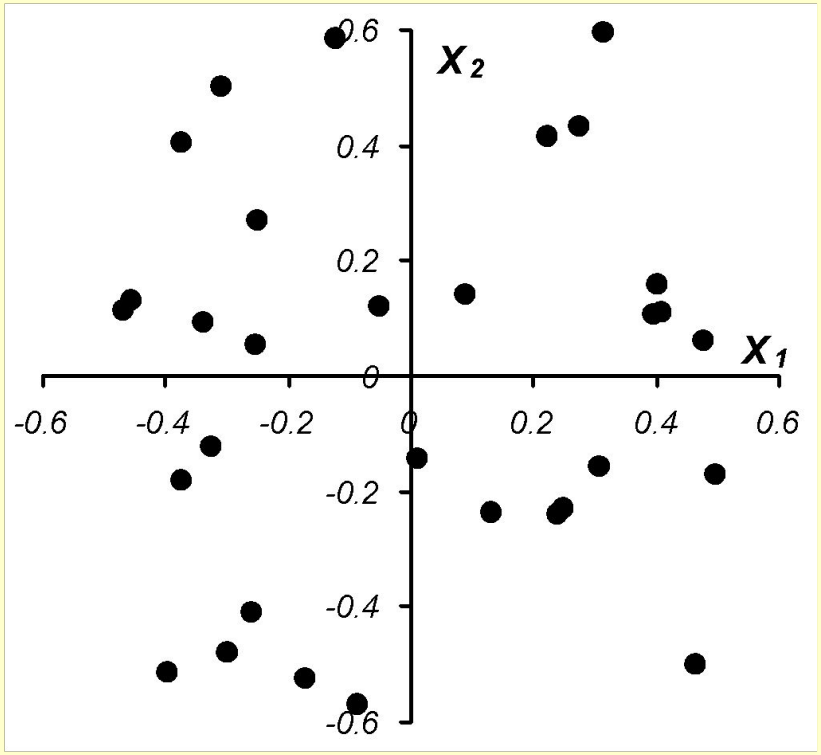
P=3

	X_1	X_2	X_3
1	0.631	0.421	0.504
2	0.663	0.537	0.510
3	0.544	0.825	0.637
4	0.662	0.954	0.736
5	0.581	1.178	0.866
6	0.758	0.338	0.482
7	0.679	0.611	0.634
8	0.644	0.870	0.744
9	0.713	1.030	0.756
10	0.748	1.166	0.914
11	0.787	0.372	0.482
12	0.820	0.635	0.678
13	0.773	0.831	0.676
14	0.735	0.964	0.861

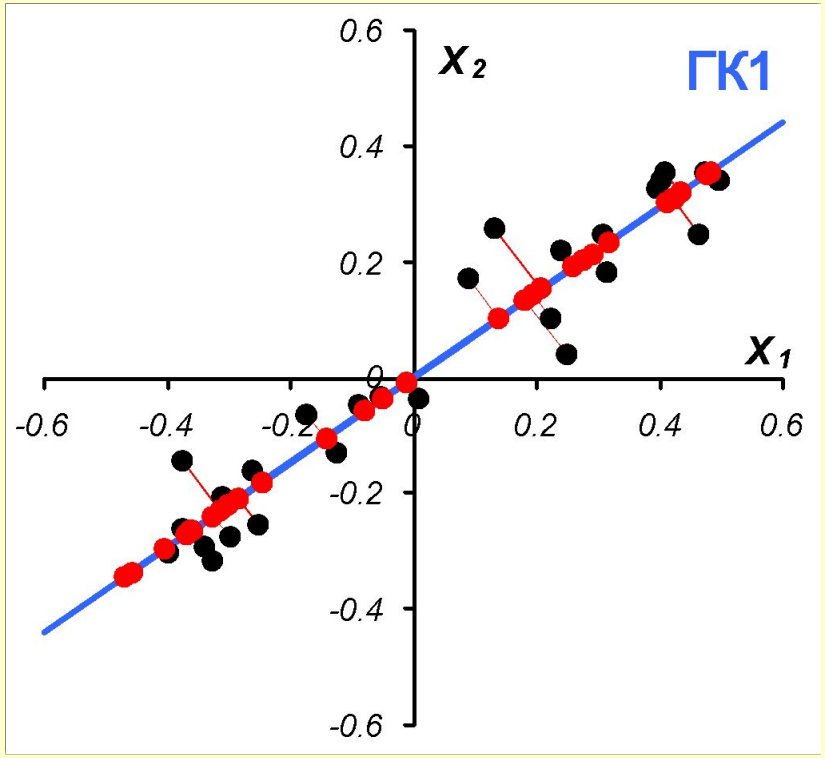


Проекционные методы

Данные без структуры

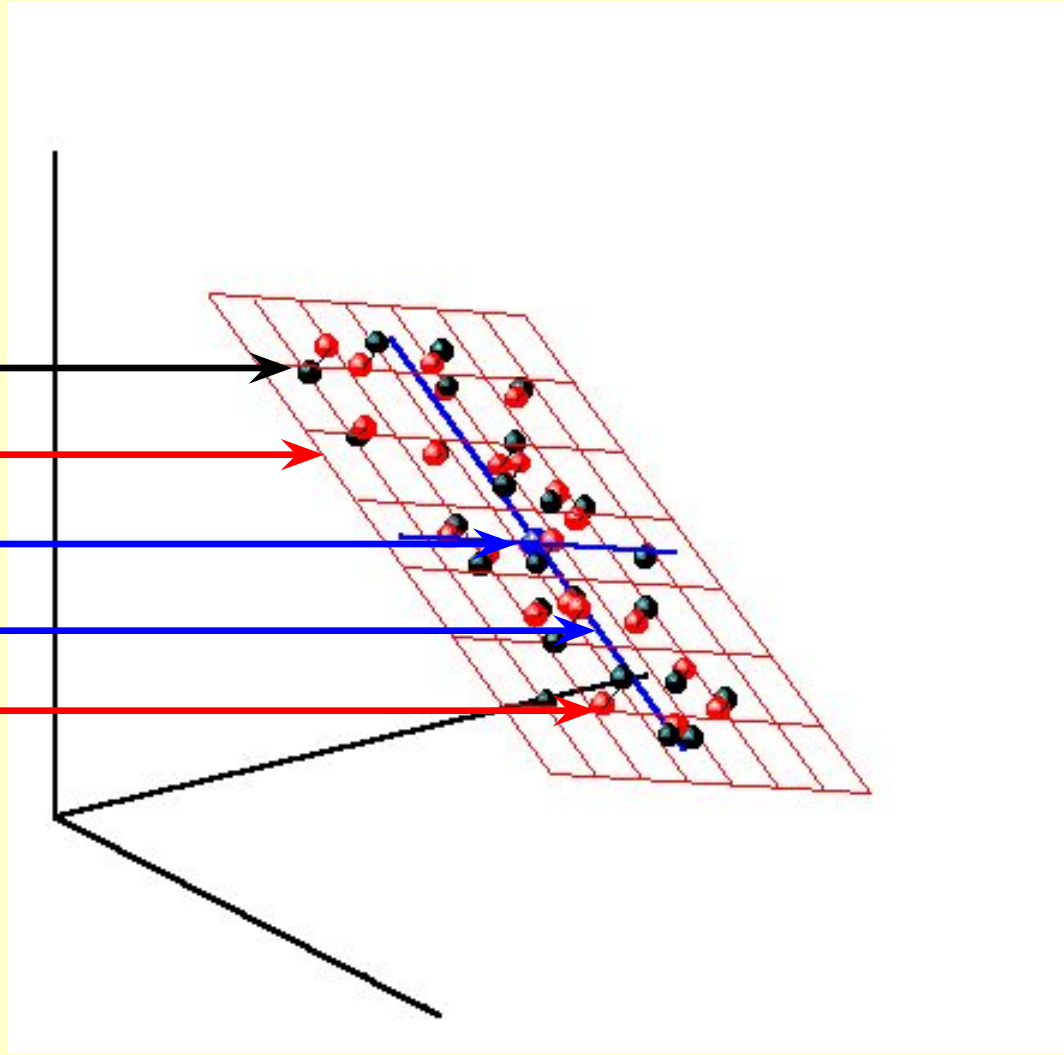


Данные со скрытой структурой

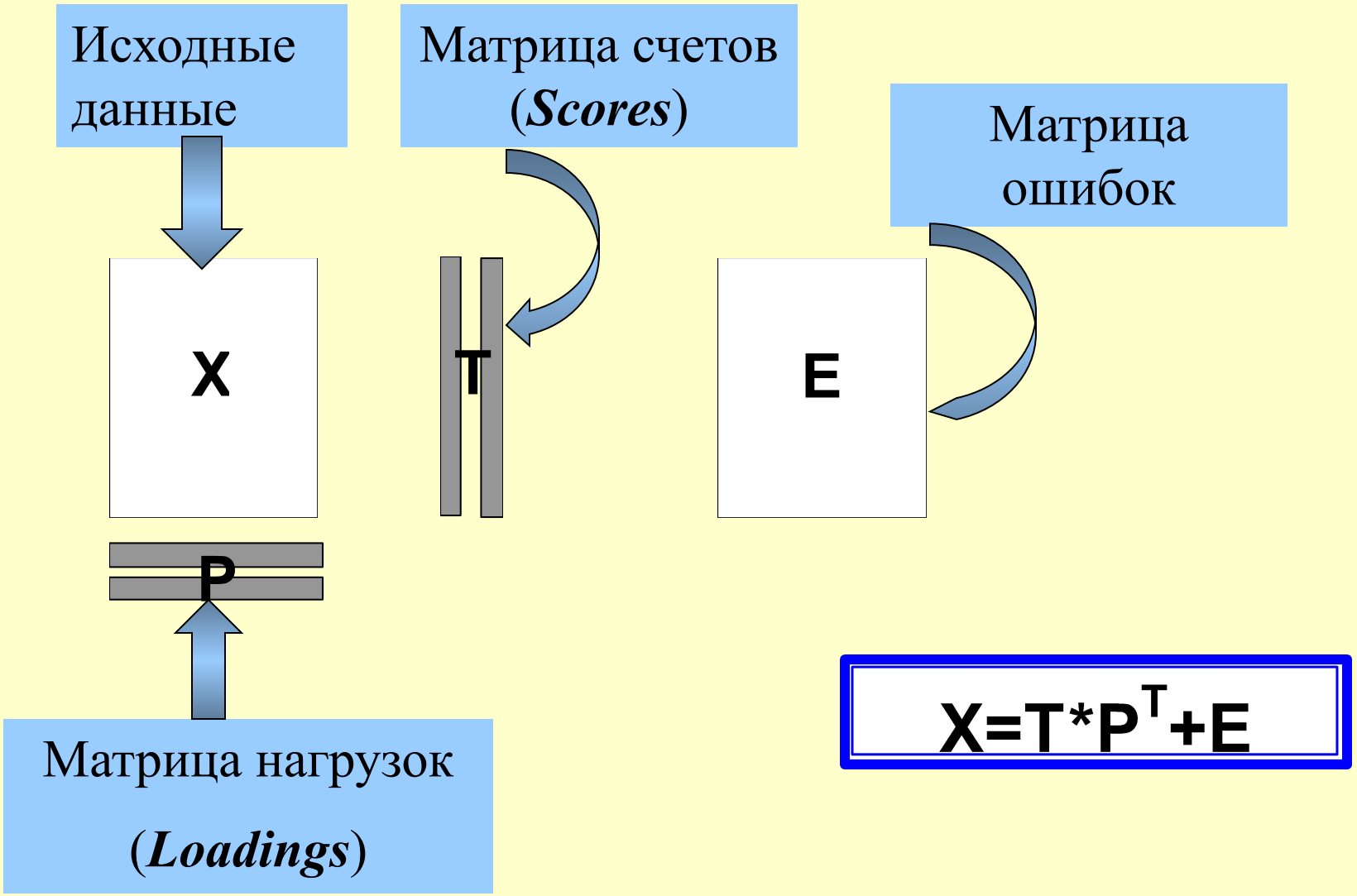


$$X_2 = aX_1 + E$$

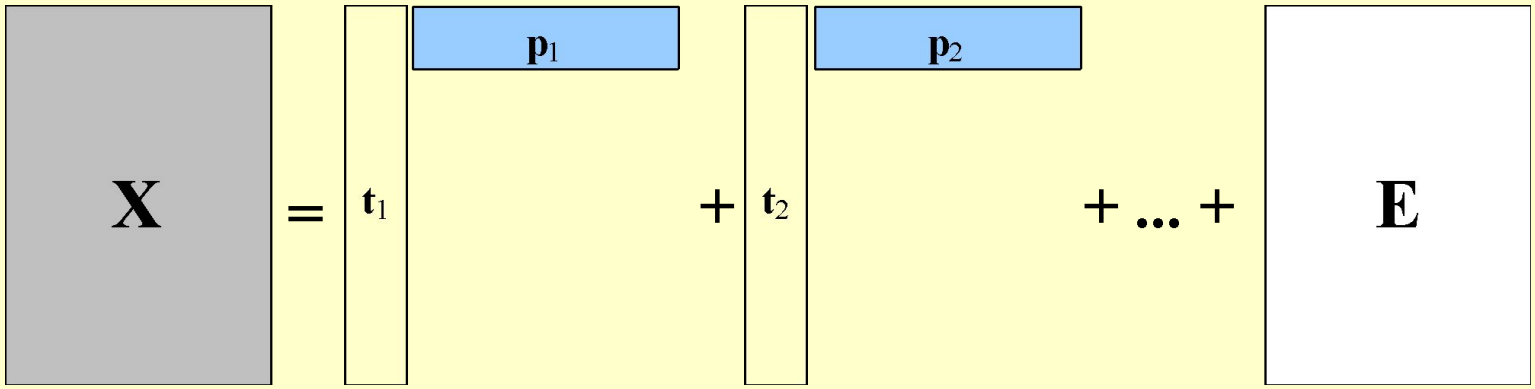
Проекция на подпространство



Метод главных компонент (PCA)



Метод главных компонент



$$\mathbf{t}=\mathbf{Xp} \Leftarrow \max|\mathbf{Xp}|^2 \text{ при условии } |\mathbf{p}|=1 \Leftrightarrow \mathbf{X}^t\mathbf{Xp} = \lambda\mathbf{p} \; ; \; \mathbf{t}^T\mathbf{t}=\lambda$$

X - матрица *данных,* **E** - матрица *ошибок, обе* ($n \times p$)

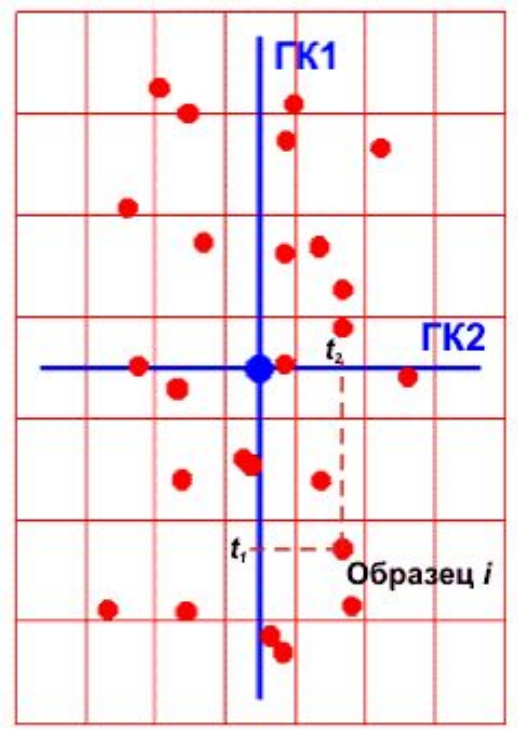
T - матрица *счетов*: ($n \times k$), **P** - матрица *нагрузок*: ($k \times p$)

k - число главных компонент ($k<<p$)

Karl Pearson, 1901

Матрица счетов T (scores)

$$X = T * P^T + E$$



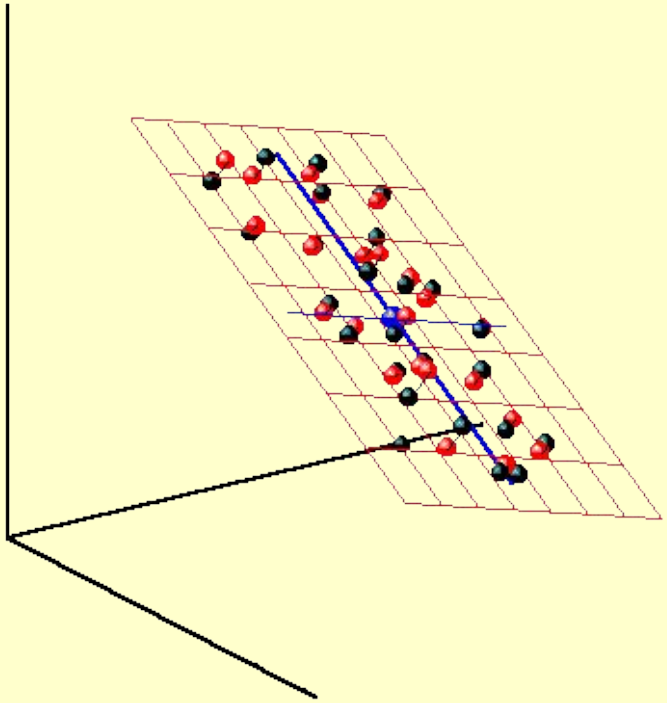
Строка – координаты одного объекта в новой системе координат

Столбец – проекция всех объектов на одну ось главных компонент

t_{11}	t_{12}
t_{21}	t_{22}
.	.
.	.
.	.
…	…
t_{n1}	t_{n2}

Матрица нагрузок P (loadings)

$$X=T*P^{T}+E$$



p_{11}	p_{12}		…	p_{1m}
p_{21}	p_{22}		…	p_{2m}

P^T- матрица перехода из пространства ***X*** в пространство главных компонент

Остатки E

$$X = T * P^T + E$$

e_{11}	e_{12}		…	e_{1m}
e_{21}	e_{22}		…	
.	.	**E**	.	.
.	.		.	.
.	.		.	.
…	…			…
e_{n1}				e_{nm}

$$e_i^2 = \sum_{k=1}^{m} e_{ik}^2$$

e_i - определяет расстояние от исходного объекта до подпространства главных компонент

$$e_{tot}^2 = \sum_{i=1}^{n} e_i^2$$

совокупная ошибка для всех объектов

матрица ***E*** имеет ту же структуру что и ***X***

$E_0, E_1, \ldots$

E_0 – ошибка при ***0***-м ГК, т.е. центрированная матрица ***X***

Математическое обеспечение

Специальные пакеты

UNSCRAMBLER

SIMCA

PLS -ToolBox для MatLab …

Стандартные статистические пакеты

STATISTICA

SPSS

SAS …

Использование стандартного мат. обеспечения для написания проекционных процедур

MatLab

Excel+ VBA+.dll

Пример. Демографические данные

		Height	Weight	Hairleng	Shoesize	Age	Income	Beer	Wine	Sex	Swim	A
		1	2	3	4	5	6	7	8	9	10	1
MA							0e+04	420.0000	115.0000	–1.0000	98.0000	–1.
MA							0e+04	350.0000	102.0000	–1.0000	92.0000	–1.
MA							0e+04	320.0000	98.0000	–1.0000	91.0000	–1.
FA							0e+04	270.0000	78.0000	1.0000	75.0000	–1.
FA							0e+04	312.0000	99.0000	1.0000	81.0000	–1.
FA	6	172.0000	64.0000	1.0000	39.0000	24.0000	2.2000e+04	308.0000	91.0000	1.0000	82.0000	–1.
MA	7	182.0000	80.0000	–1.0000	42.0000	35.0000	3.0000e+04	398.0000	65.0000	–1.0000	85.0000	–1.
MA	8	180.0000	80.000								00	–1.
FA	9	169.0000	51.000								00	–1.
FA	10	168.0000	52.000								00	–1.
MA	11	183.0000	81.000								00	–1.
FA	12	157.0000	47.000								00	–1.
FA	13	164.0000	50.000								00	–1.
FA	14	162.0000	49.000								00	–1.
MA	15	180.0000	82.000								00	–1.
MA	16	180.0000	81.000								00	–1.
MB	17	185.0000	82.000								00	1.
MB	18	187.0000	84.000								00	1.
FB	19	168.0000	50.000								00	1.
FB	20	166.0000	49.000								00	1.
FB	21	158.0000	46.000								00	1.
MB	22	177.0000	65.000								00	1.
MB	23	180.0000	72.000								00	1.
MB	24	181.0000	76.000								00	1.

Количество объектов (n) = 32

Количество переменных (m) = 12

Рост (*Height*)	в сантиметрах
Вес (*Weight*)	в килограммах
Длина волос (*Hairleng*)	короткие: -1; длинные: +1
Размер обуви (*Shoesize*)	Европейский стандарт
Возраст (*Age*)	в годах
Доход (*Income*)	в евро
Потребление пива (*Beer*)	литров в год
Потребление вина (*Wine*)	литров в год
Пол (*Sex*)	мужской: -1; женский: +1
Способность плавать (*Swim*)	индекс, основанный на 500 м дистанции
Место жительства (*A/B*)	A: -1 (Скандинавия); B: +1 (Средиземноморье)
Коэффициент интеллекта (*IQ*)	Стандартный евр. тест

Предварительная обработка данных

Цель – преобразование исходных данных в форму, наиболее удобную для анализа.

Автошкалирование = Центрирование относительно среднего + Взвешивание

$$x_{ik}^{scaled} = x_{ik} \frac{1}{SDev}$$

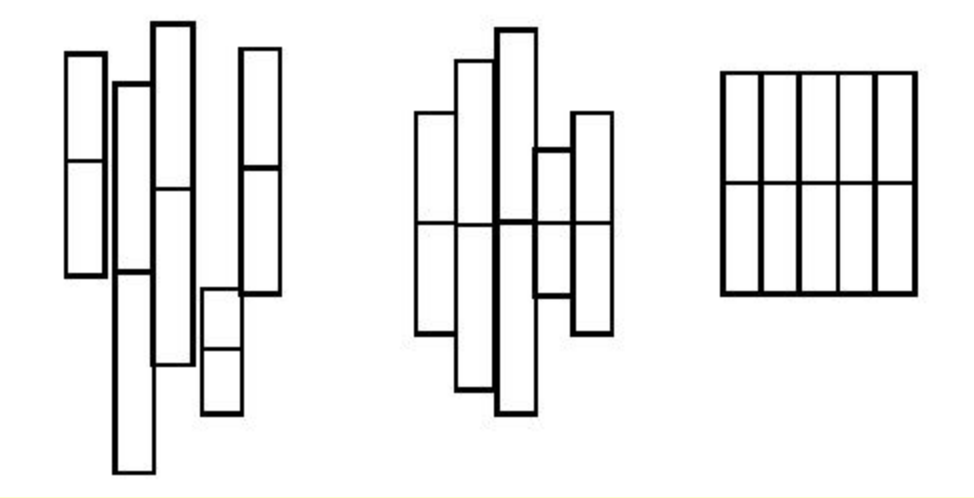
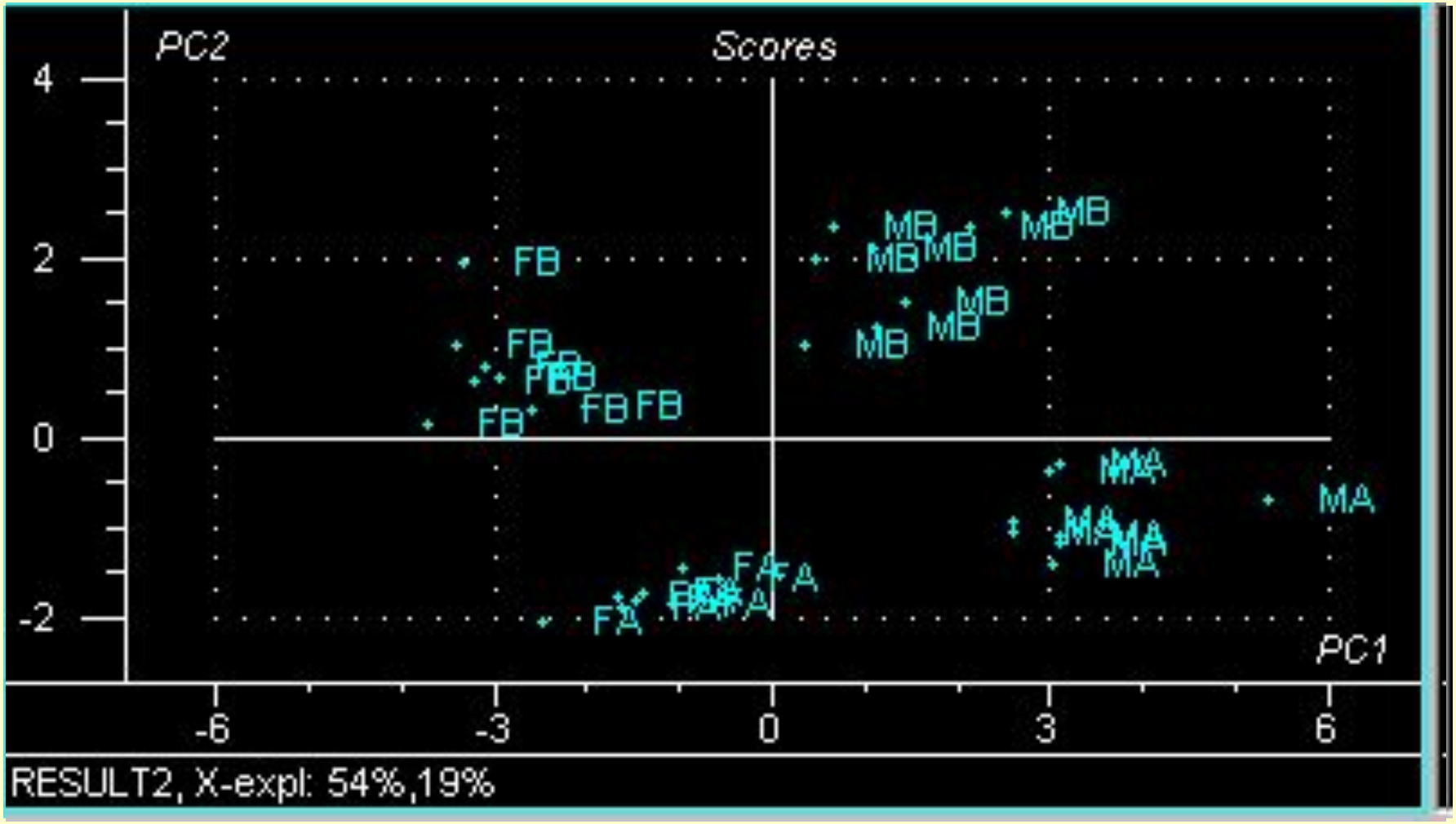
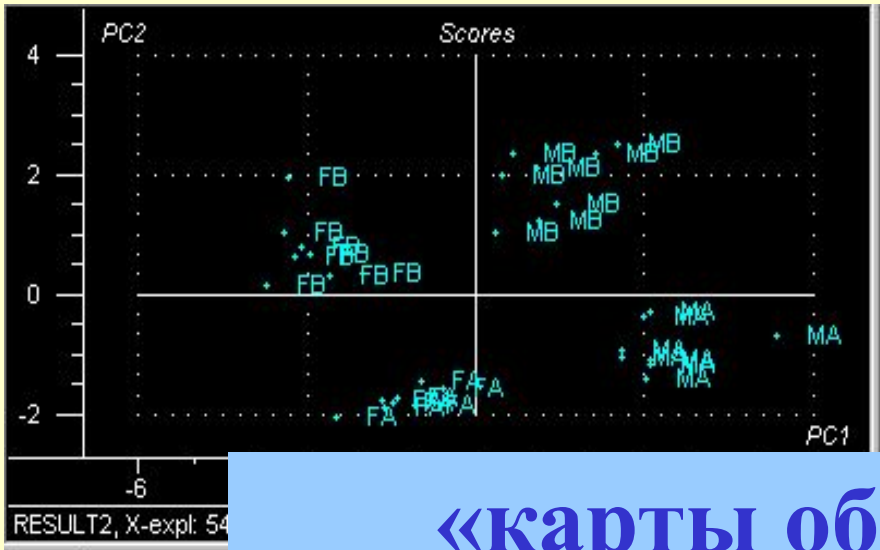


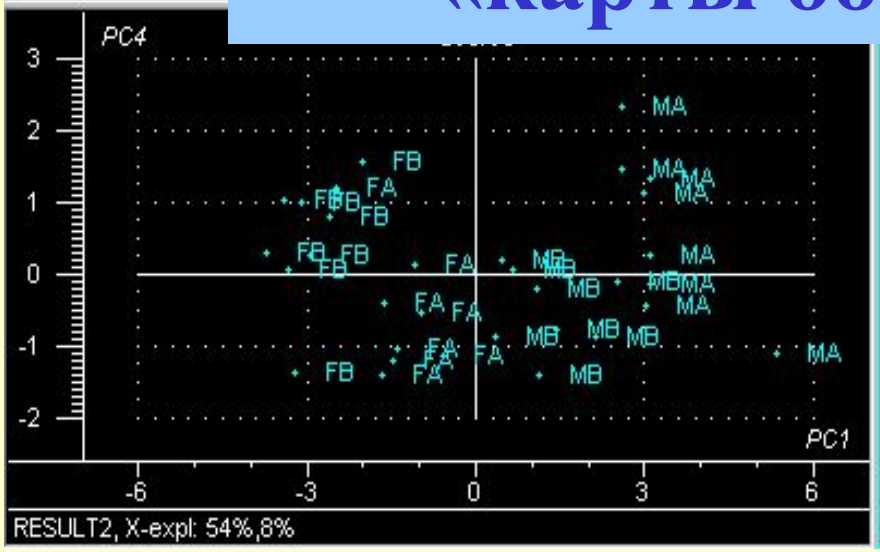
График счетов (ГК1-ГК2)



Графики счетов



«карты образцов»



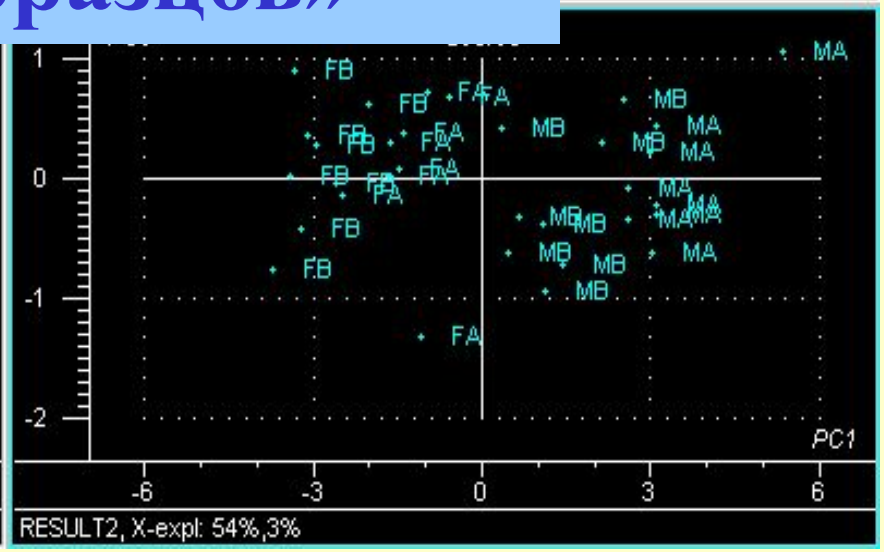
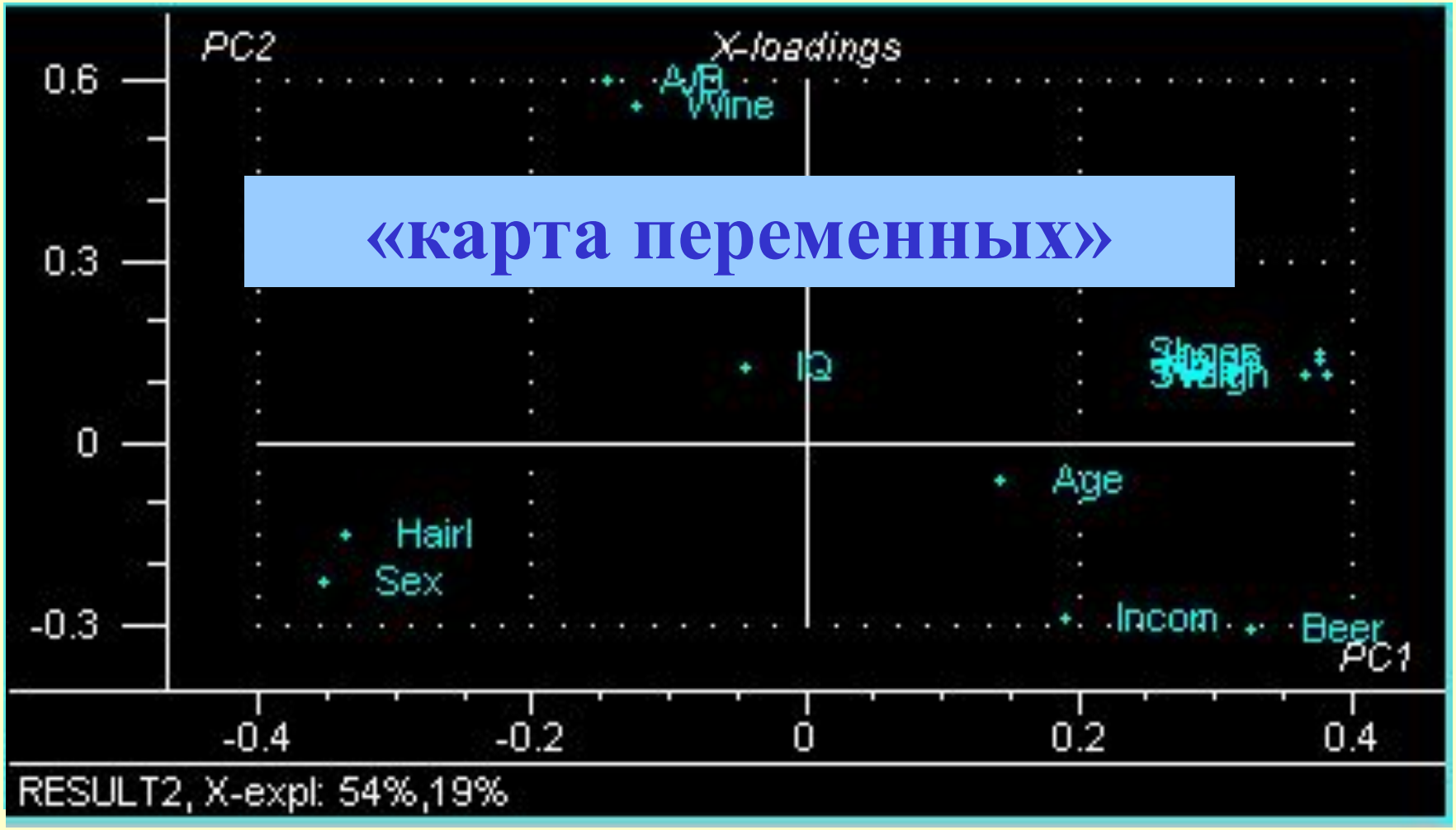


График нагрузок (ГК1-ГК2)



«карта переменных»

ГК1-ГК2 счета и нагрузки

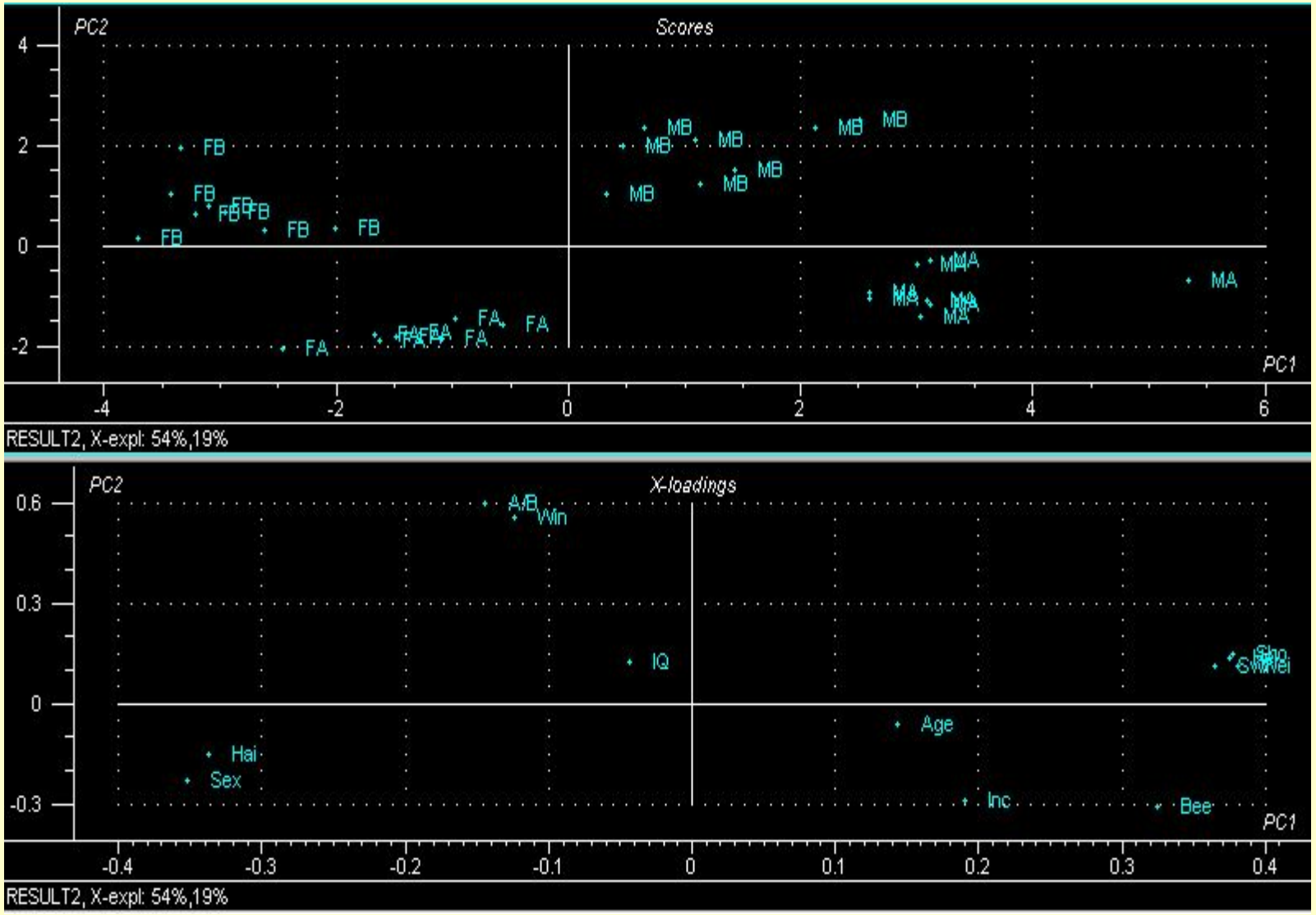
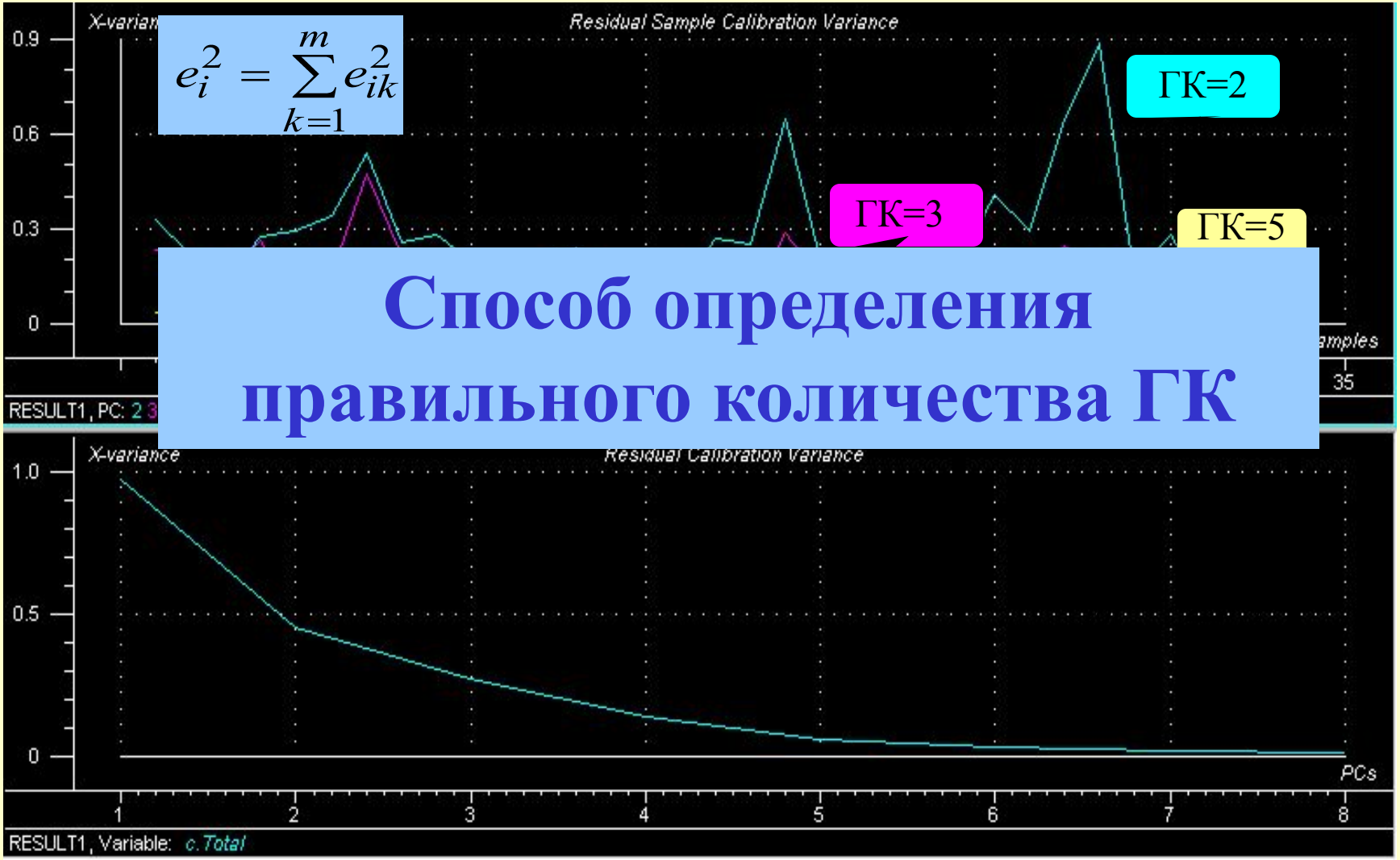


График ошибок



Способ определения правильного количества ГК

Цели и «инструменты»

Основные цели МГК

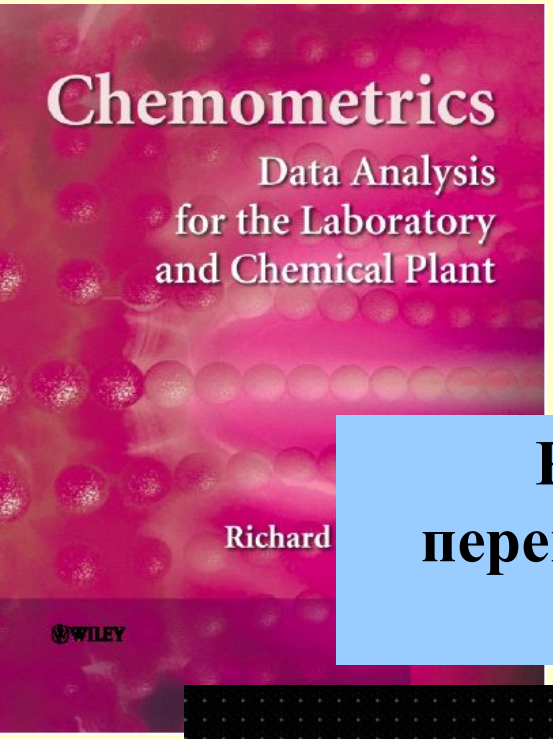
1. Представление объектов в пространстве, отражающем внутреннюю структуру изучаемых данных
2. Понижение размерности системы, отделение содержательной части от шума

Основные «инструменты»

1. Графики счетов – «карты образцов»
2. Графики нагрузок – «карты переменных»
3. Графики остатков – способ выбора количества ГК

Что может быть не так?

1. Данные не содержат необходимой информации
2. Использовано недостаточное количество ГК
3. Использовано излишние количество ГК
4. Не удалены выбросы
5. Удалены точки (псевдовыбросы) содержащие важную информацию
6. Недостаточный анализ графиков счетов/нагрузок
7. Использована только стандартная (машинная) диагностика, без содержательного анализа.
8. Использованы неверные методы предварительной обработки данных



Анализ смеси

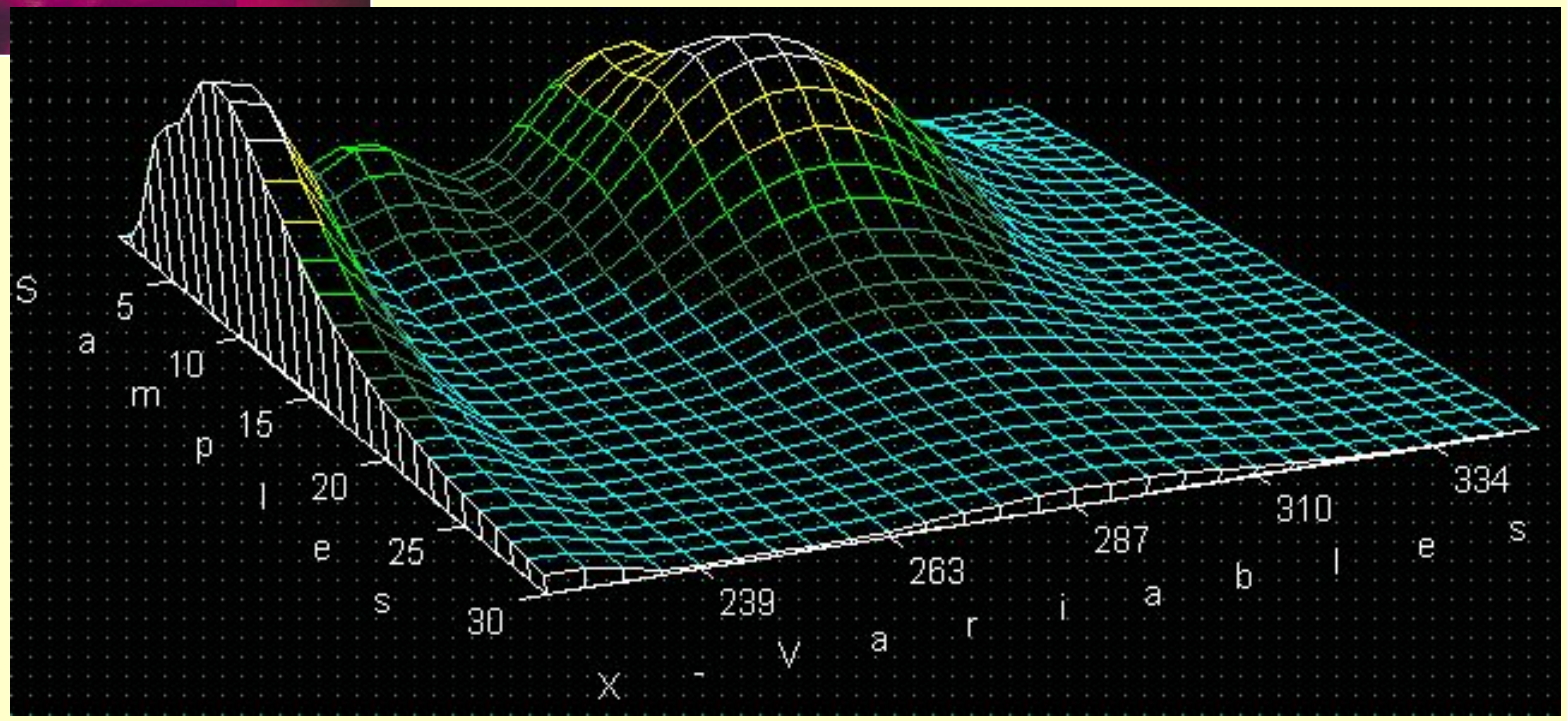
Разделение перекрывающихся пиков

n=30
(сек)

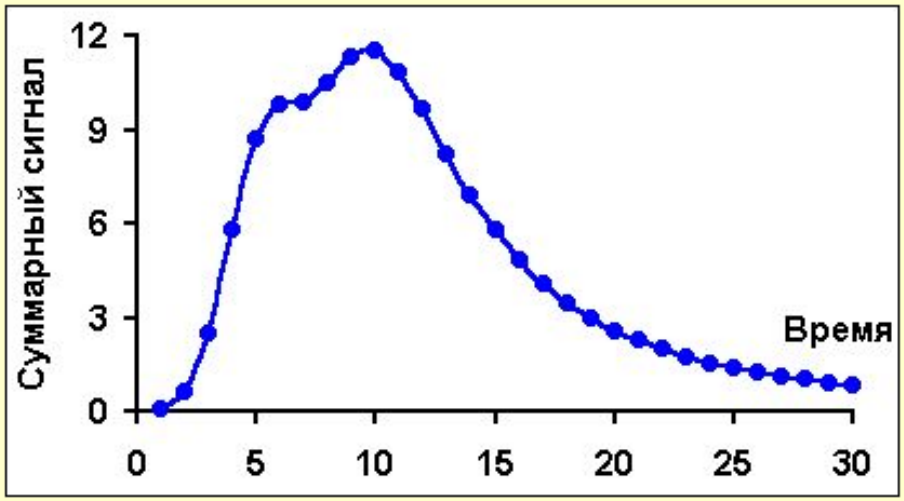
время

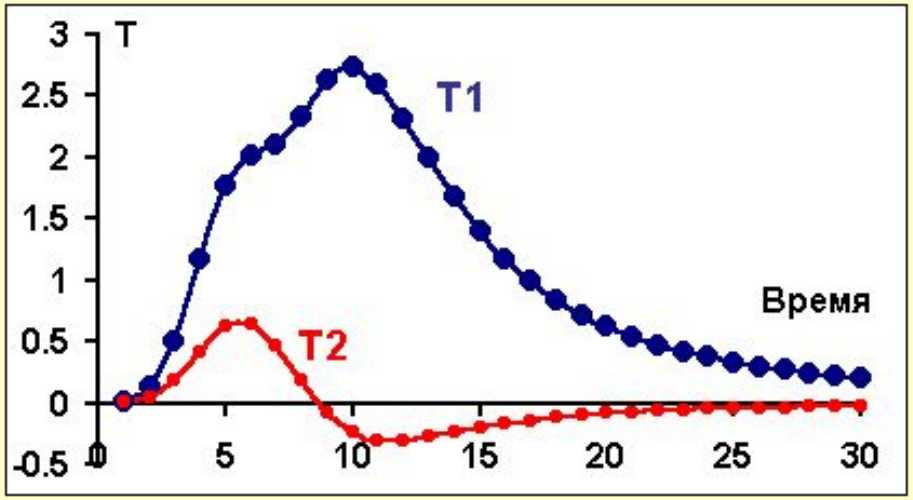
X
ВЖХ-ДМД

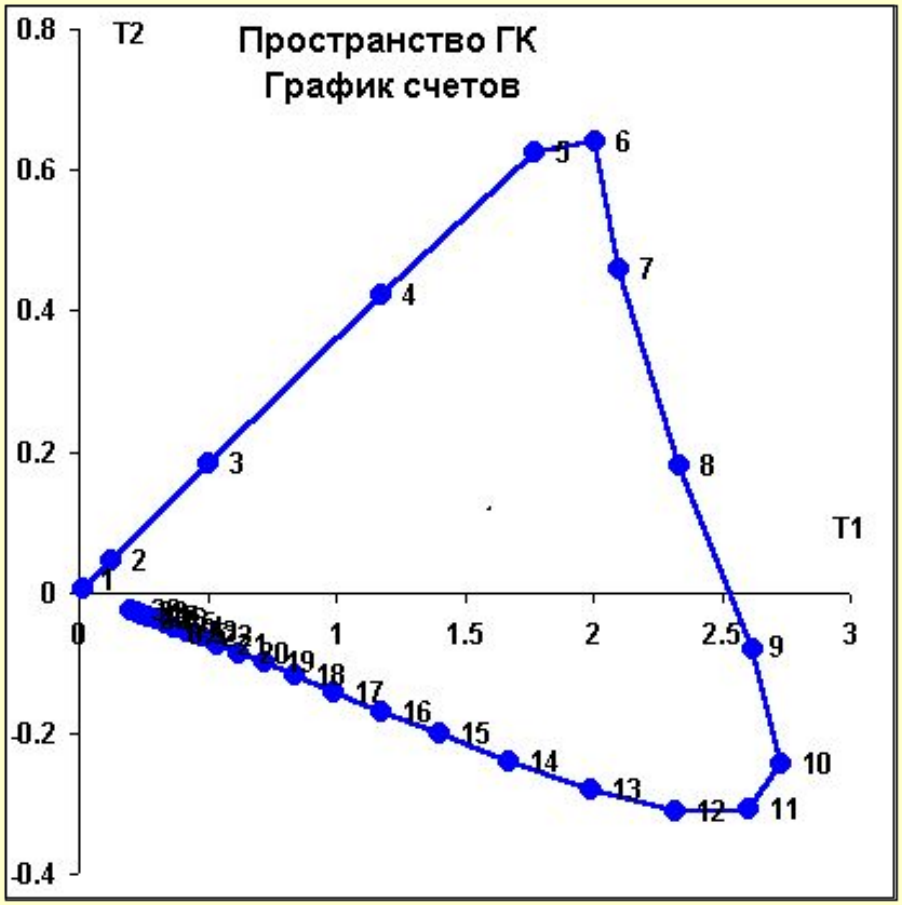
длина волны
m=28
(нм)



Разделение пиков







Продолжение - за компьютером