Концепция скрытых (латентных) переменных в химическом анализе. Часть 1. Качественный анализ

Померанцев
Алексей Леонидович

Институт химической физики РАН им Семенова

Российское хемометрическое общество

Вопрос 1 . Кто я ?

Еще студент

Уже не студент

Вопрос 2. Моя направленность

Экспериментатор

Теоретик

Я ошибся комнатой

Вопрос 3. Специальность

- Химик
- Физик
- Биолог
- Математик

Вопрос 4. Область

- Аналитическая химия
- Органическая химия
- Физическая химия
- Не занимаюсь ничем подобным

Вопрос 5. Специализация

Электрохимия

Хроматография

Спектроскопия

Масспектрометрия

Хемометрика

Вопрос 6. Знаю (читал) ...

1 журнал по аналитической химии

2 журнала по АХ

4 журнала по АХ

больше 6 журналов по АХ

Вопрос 7. Хемометрика это ...

статистика в аналитической химии

метрология применительно к АХ

то, что делают хемометрики

понятия не имею ?!

Данные о нас

Персона	Студент	Экс/Теор	Наука	Область	Специал	Журнал	Хемом
Иванов И.	1	1	1	1	1	1	1
Петров П.	1	1	1	1	1	1	1
Сидоров С.	1	1	1	1	1	2	2
...	0	0	1	1	3	1	1
...	1	1	1	1	1	1	1
...	0	1	1	1	4	1	1
...	0	1	1	1	5	1	3
...	0	1	1	1	1	1	1
...	1	0	1	1	1	2	4
...	0	1	1	2	1	1	1
Зонтов Ю.	1	0	1	4	6	0	1
...	0	1	1	1	2	1	1
...	0	0	1	1	3	1	2
...	1	1	1	1	1	1	2
Померанцев А.	0	0	3	1	6	12	3
...	1	1	1	2	2	1	1
...	1	1	1	1	1	2	3
...	1	1	2	1	2	1	2

образец/объект

переменная/свойство

Матрица данных

Матрицы и векторы: учебник

Chemometrics.Ru: учебники || А.Л. Померанцев. Матрицы и векторы - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://www.chemometrics.ru/materials/textbooks/matrix/ Go Links

1.1 Матрицы

Матрицей называется прямоугольная таблица чисел, например

$$\mathbf{A}=\begin{vmatrix} 1.2 & -5.3 & 0.25 \\ 10.2 & 1.5 & -7.5 \\ 2.3 & -1.2 & 5.6 \\ 4.5 & -0.8 & 9.5 \end{vmatrix}$$

Матрицы обозначаются заглавными полужирными буквами (A), а их элементы — соответствующими строчными буквами с индексами, т.е. a_{ij}. Первый индекс нумерует строки, а второй — столбцы. В хемометрике принято обозначать максимальное значение индекса той же буквой, что и сам индекс, но заглавной. Поэтому матрицу A можно также записать как $\{ a_{ij}, i = 1,..., I; j = 1,...,J\}$. Для приведенной в примере матрицы $I = 4$, $J = 3$ и $a_{23} = -7.5$.

Пара чисел I и J называется размерностью матрицы и обознается как $I \times J$. Примером матрицы в хемометрике может служить набор спектров, полученный для I образцов на J длинах волн.

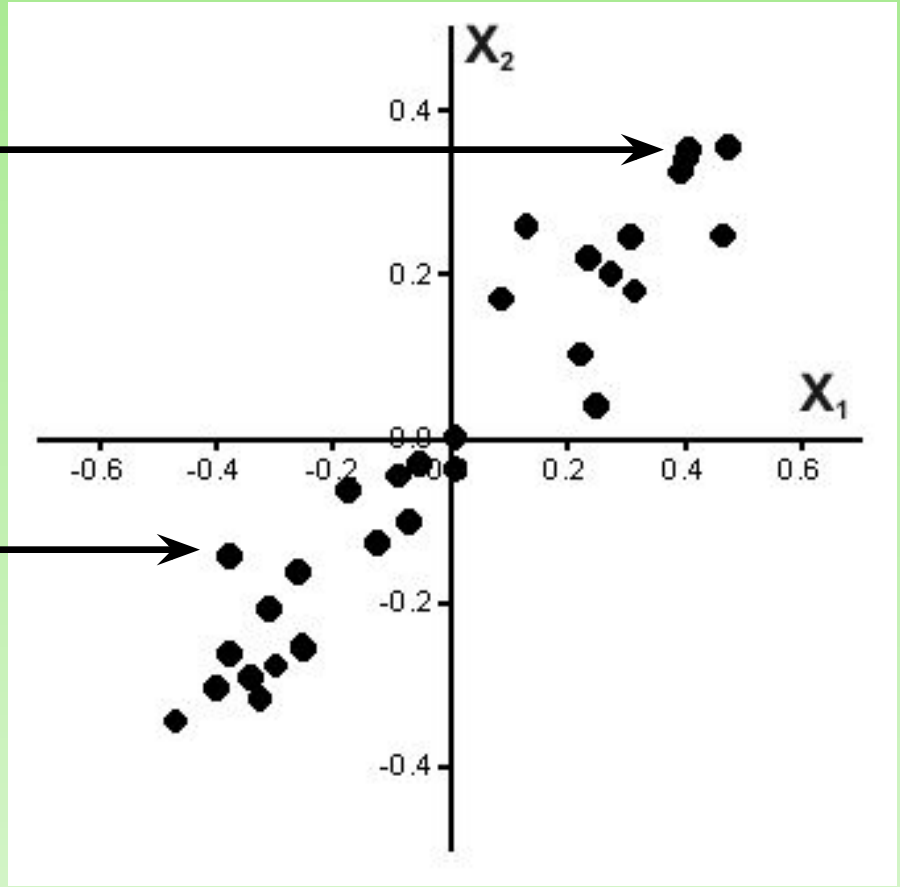
1.2. Простейшие операции с матрицами

Матрицы можно умножать на числа. При этом каждый элемент умножается на это число. Например —

Internet

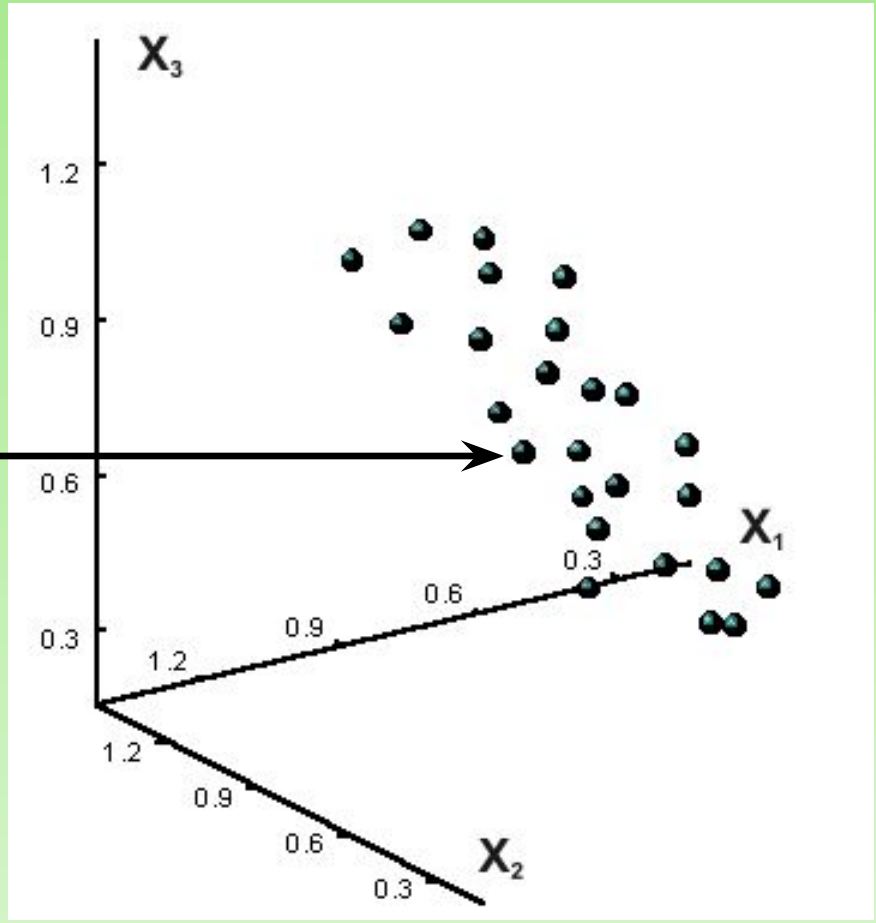
Графическое представление 2D-данных

	X_1	X_2
1	0.407	0.353
2	0.475	0.355
3	-0.088	-0.045
4	0.394	0.325
5	0.274	0.202
6	0.131	0.258
7	-0.053	-0.031
8	-0.124	-0.128
9	-0.469	-0.344
10	0.088	0.171
11	-0.261	-0.162
12	0.401	0.341
13	-0.376	-0.143
14	-0.251	-0.255

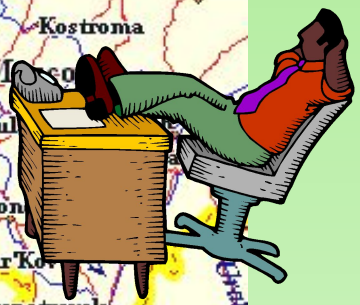


Графическое представление 3D-данных

	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



Почти наши данные



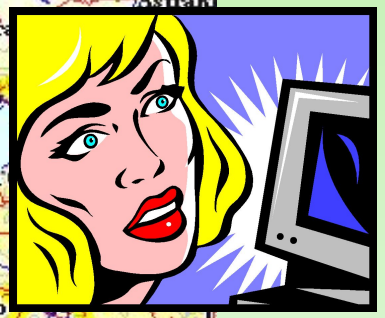
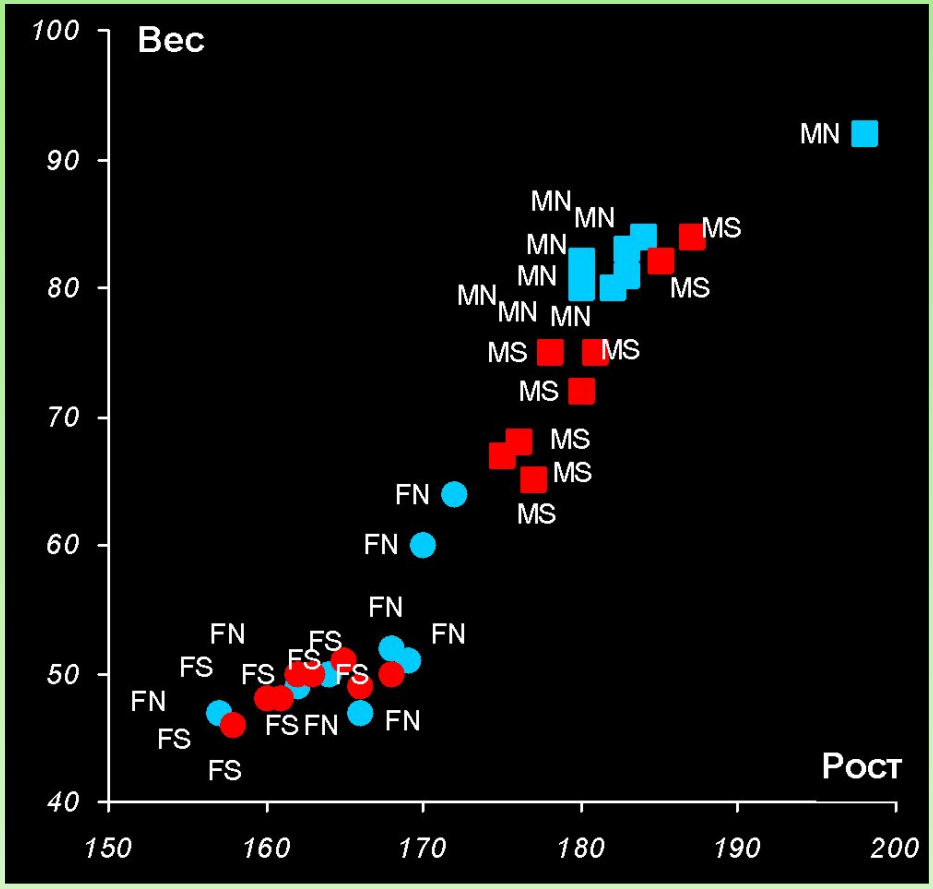
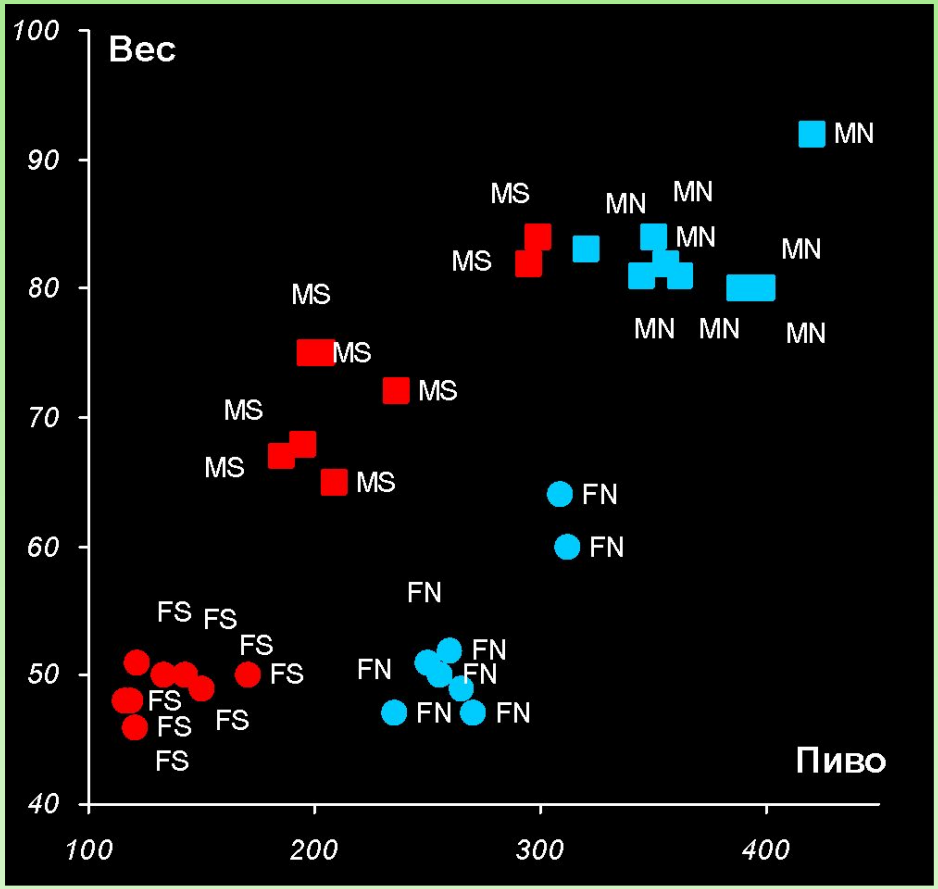


Таблица данных

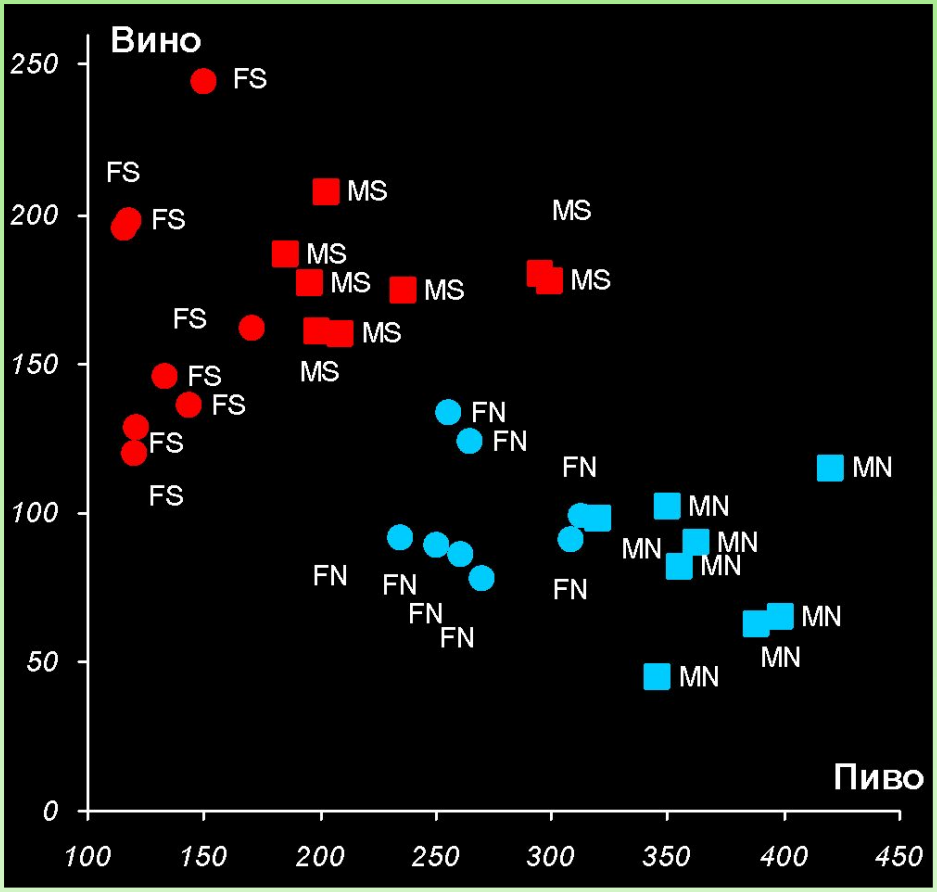
Персона №	Пол M/F	Регион S/N	Рост см	Вес кг	Волосы S/L	Обувь размер	Возраст лет	Доход K€/год	Пиво л/год	Вино л/год	Сила баллы	IQ баллы
1	M	N	198	92	S	48	48	45	420	115	98	100
2	M	N	184	84	S	44	33	33	350	102	92	130
3	M	N	183	83	S	44	37	34	320	98	91	127
4	M	N	182	80	S	42	35	30	398	65	85	140
5	M	N	180	80	S	43	36	30	388	63	84	129
6	M	N	183	81	S	42	37	35	345	45	90	105
7	M	N	180	82	S	44	43	37	355	82	88	109
8	M	N	180	81	S	44	46	42	362	90	86	113
9	M	S	185	82	S	45	26	16	295	180	92	109
10	M	S	187	84	S	46	27	17	299	178	95	119
11	M	S	177	65	S	41	26	18	209	160	86	120
12	M	S	180	72	S	43	33	19	236	175	85	115
13	M	S	181	75	S	43	42	31	198	161	83	105
14	M	S	176	68	S	42	50	36	195	177	82	96
15	M	S	175	67	L	42	55	38	185	187	80	105
16	M	S	178	75	S	42	30	24	203	208	81	118
17	F	N	166	47	S	36	32	28	270	78	75	112
18	F	N	170	60	L	38	23	20	312	99	81	110
19	F	N	172	64	L	39	24	22	308	91	82	102
20	F	N	169	51	L	36	24	23	250	89	78	98
21	F	N	168	52	L	37	27	24	260	86	78	100
22	F	N	157	47	L	36	32	32	235	92	70	127
23	F	N	164	50	L	38	41	34	255	134	76	101
24	F	N	162	49	L	37	40	34	265	124	75	108
25	F	S	168	50	L	37	49	34	170	162	76	135
26	F	S	166	49	L	36	21	14	150	245	75	123
27	F	S	158	46	L	34	30	18	120	120	70	119
28	F	S	163	50	L	36	18	11	143	136	75	102
29	F	S	162	50	L	36	20	12	133	146	74	132
30	F	S	165	51	L	36	36	26	121	129	76	126
31	F	S	161	48	L	35	41	32	116	196	75	120
32	F	S	160	48	L	35	40	31	118	198	74	129

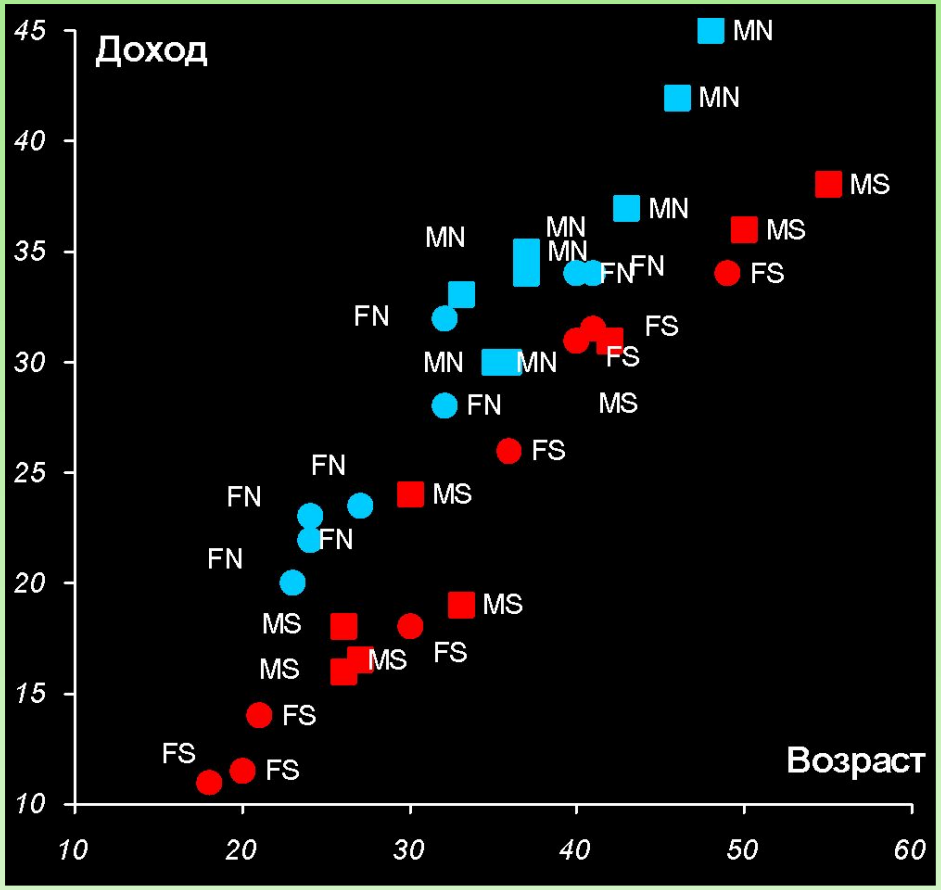
Корреляции 1



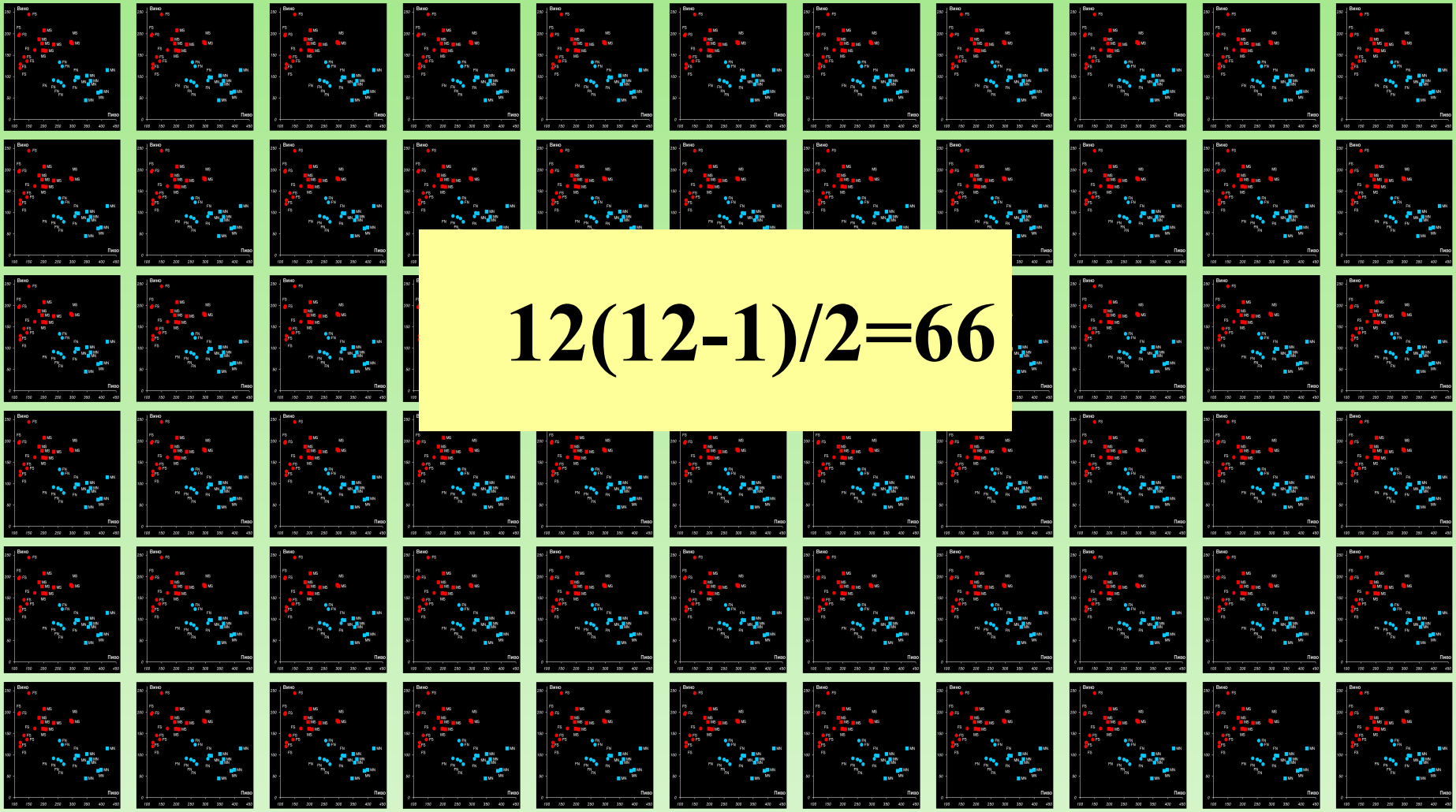


Корреляции 2





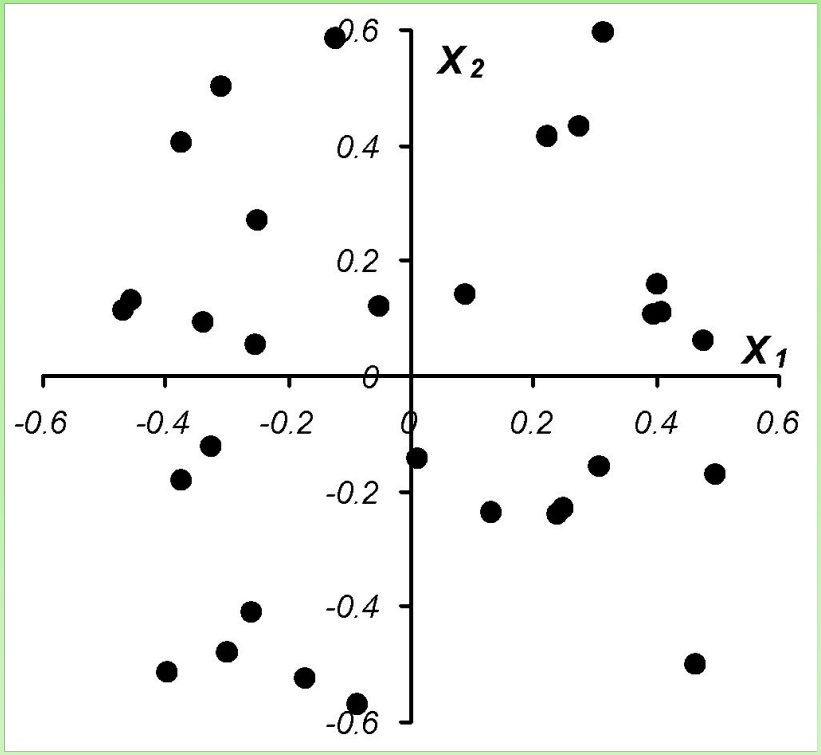
Все возможные корреляции



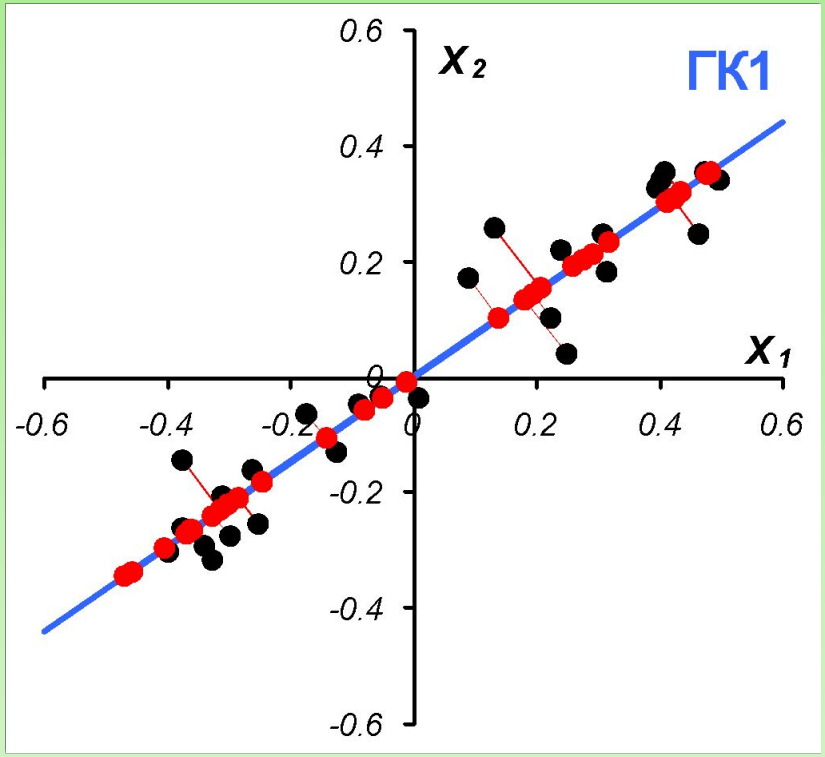
12(12-1)/2=66

Главные направления и проекции

Данные без структуры

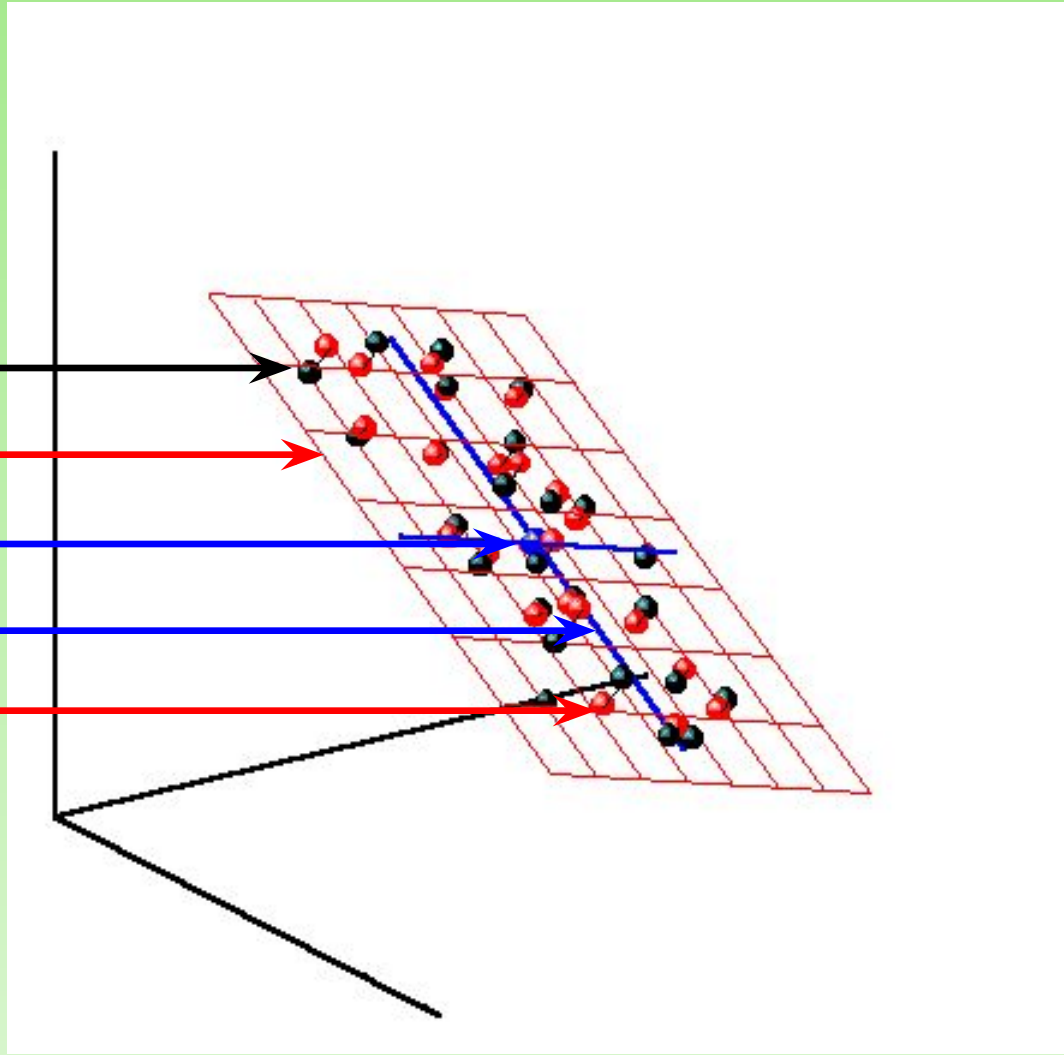


Данные со скрытой структурой

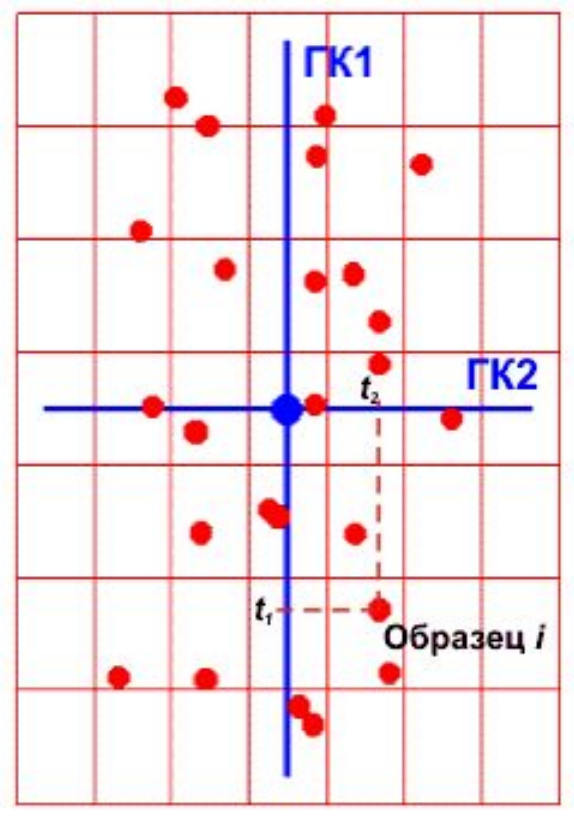


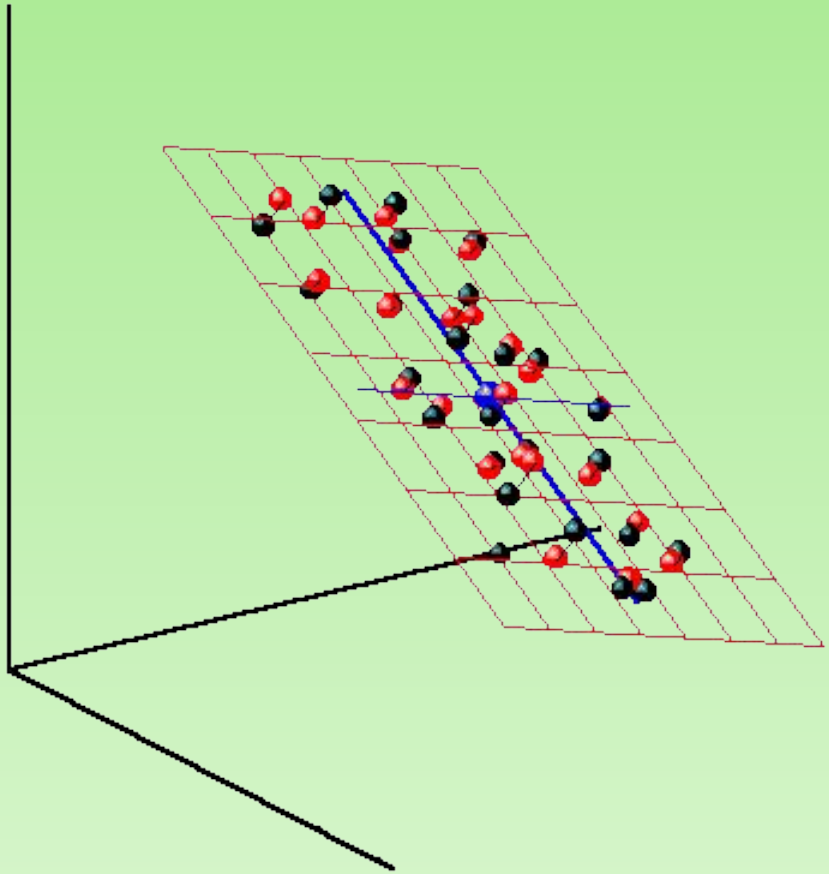
$$X_2 = aX_1 + E$$

Проекция на подпространство

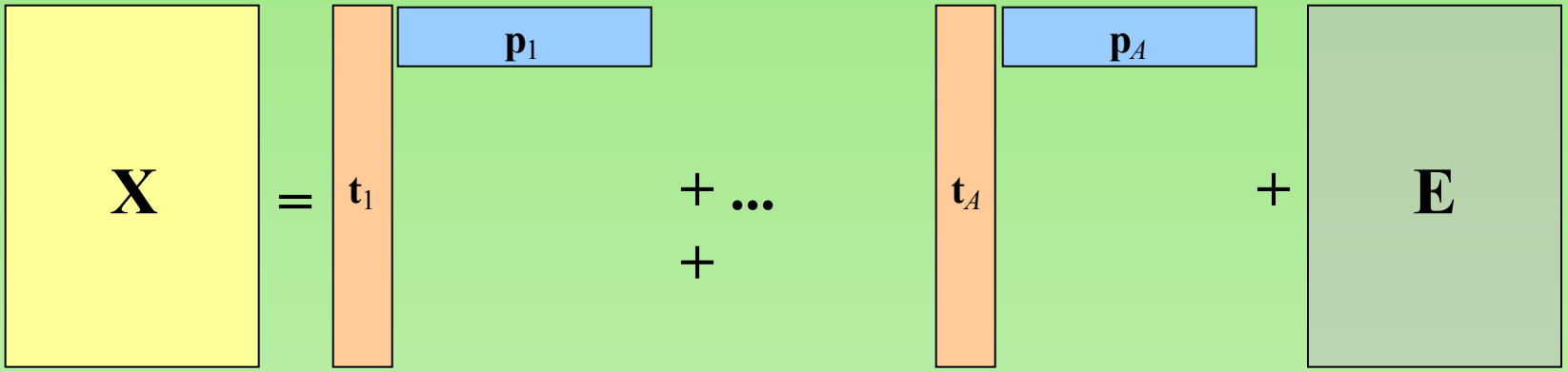


Представление данных в подпространстве





Метод главных компонент: МГК (PCA)



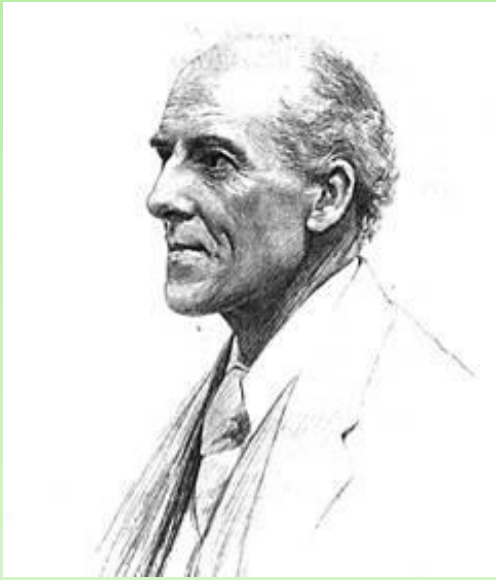
X - матрица *данных,*

E - матрица *погрешностей, обе* ($I \times J$)

T - матрица *счетов*: ($I \times A$),

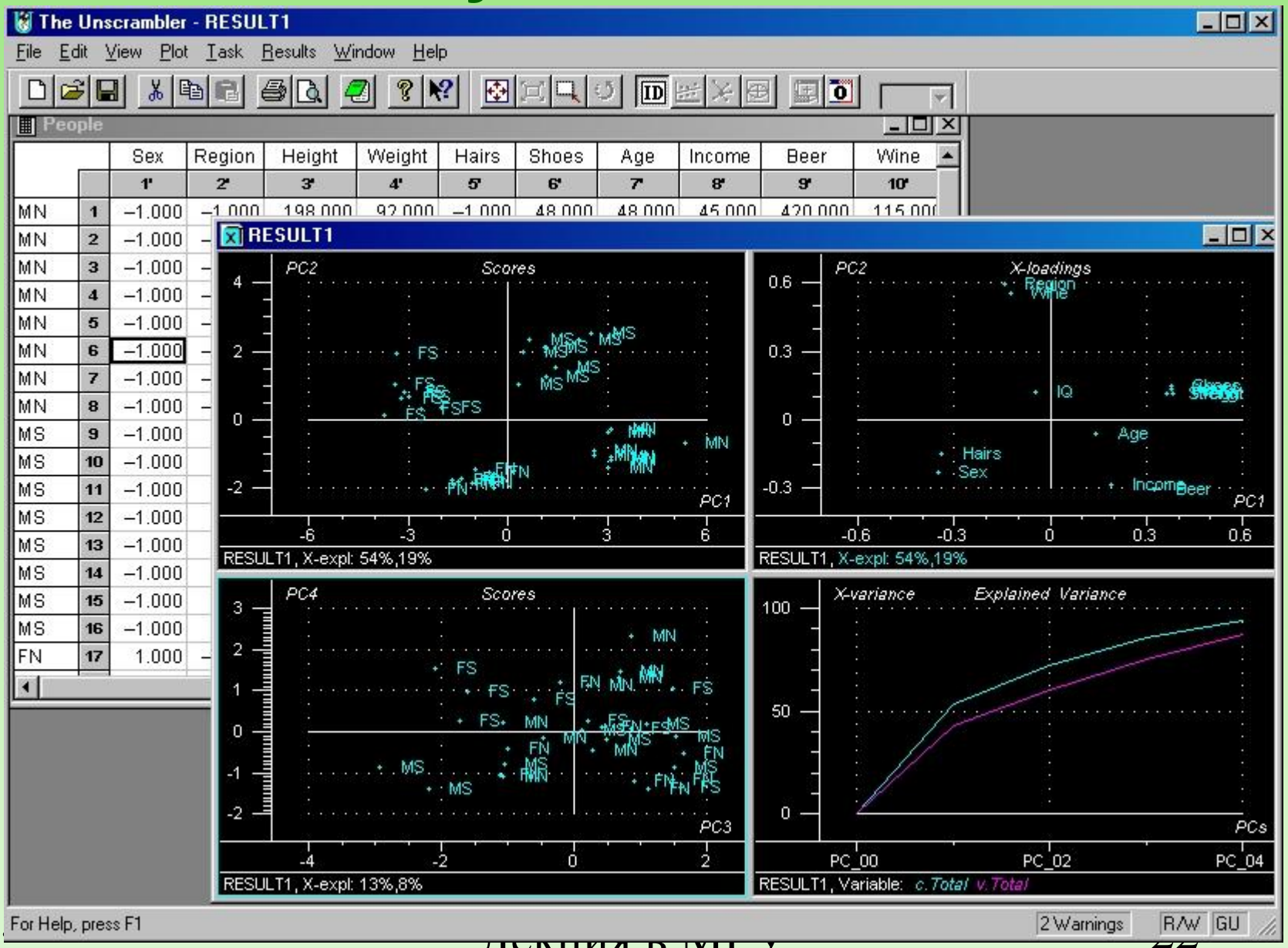
P - матрица *нагрузок*: ($A \times J$)

A - число главных компонент ($A << J$)

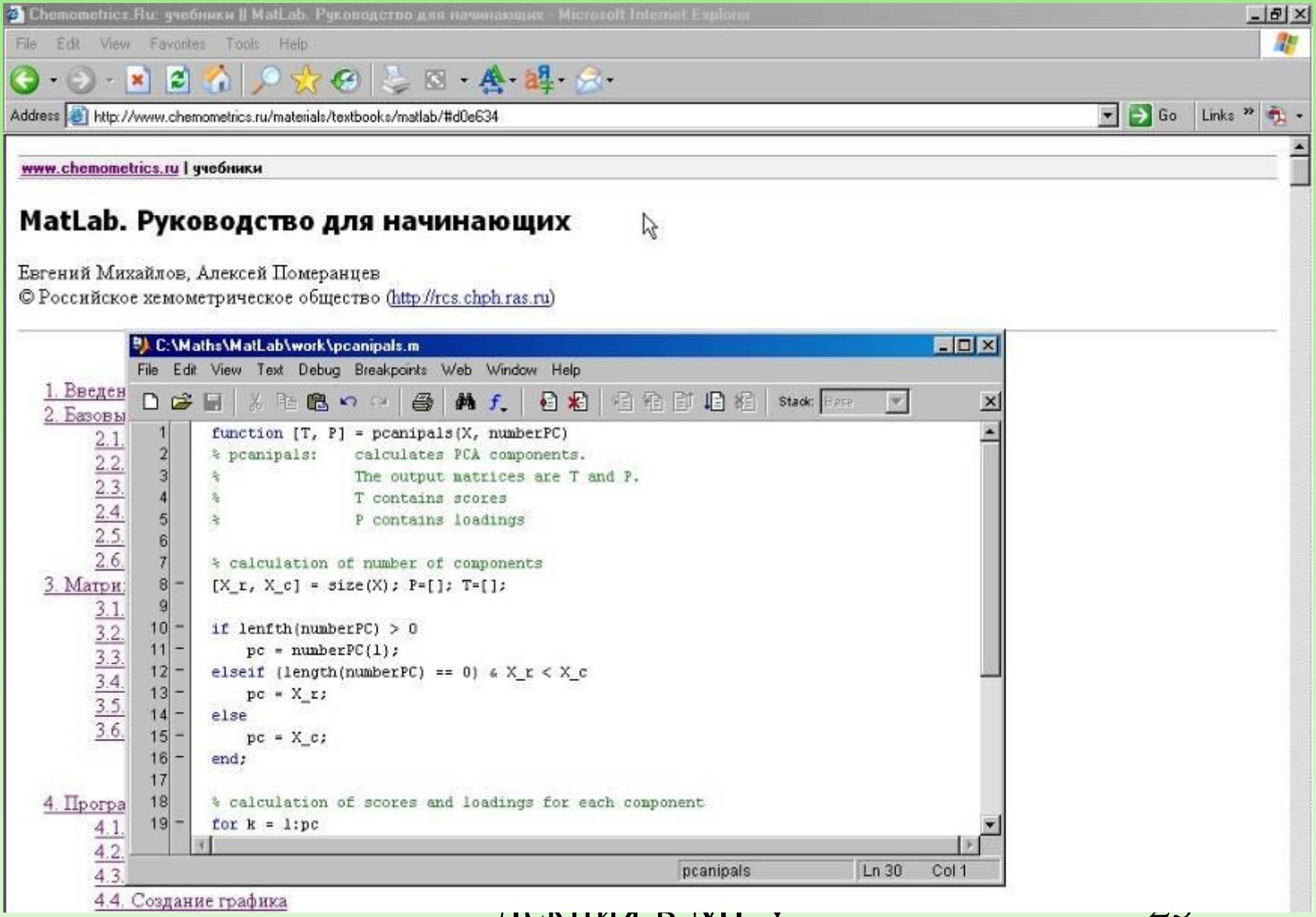


Karl Pearson, 1901

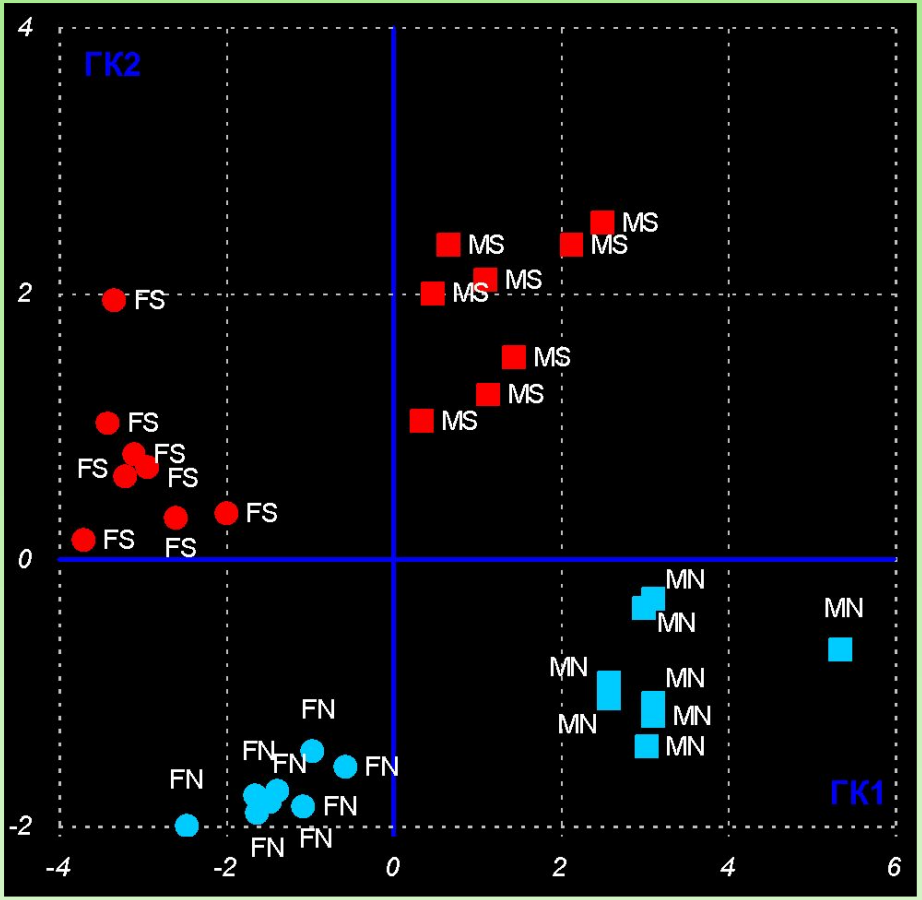
Unscrambler® by CAMO ASA

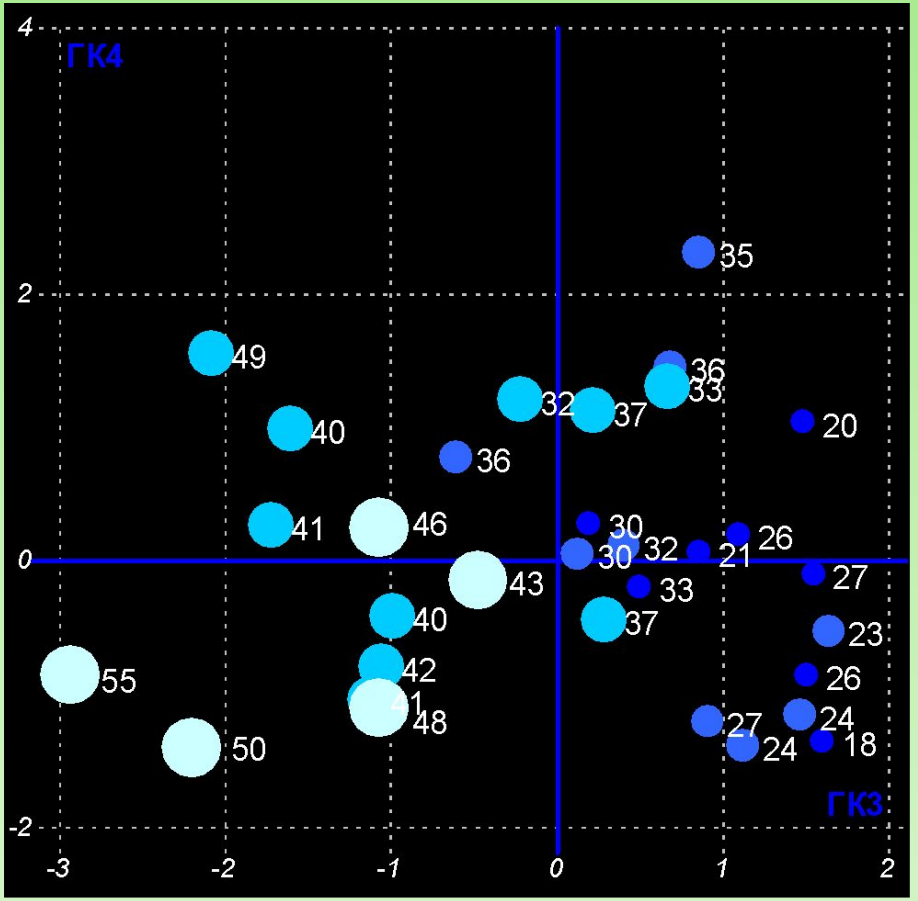


Matlab

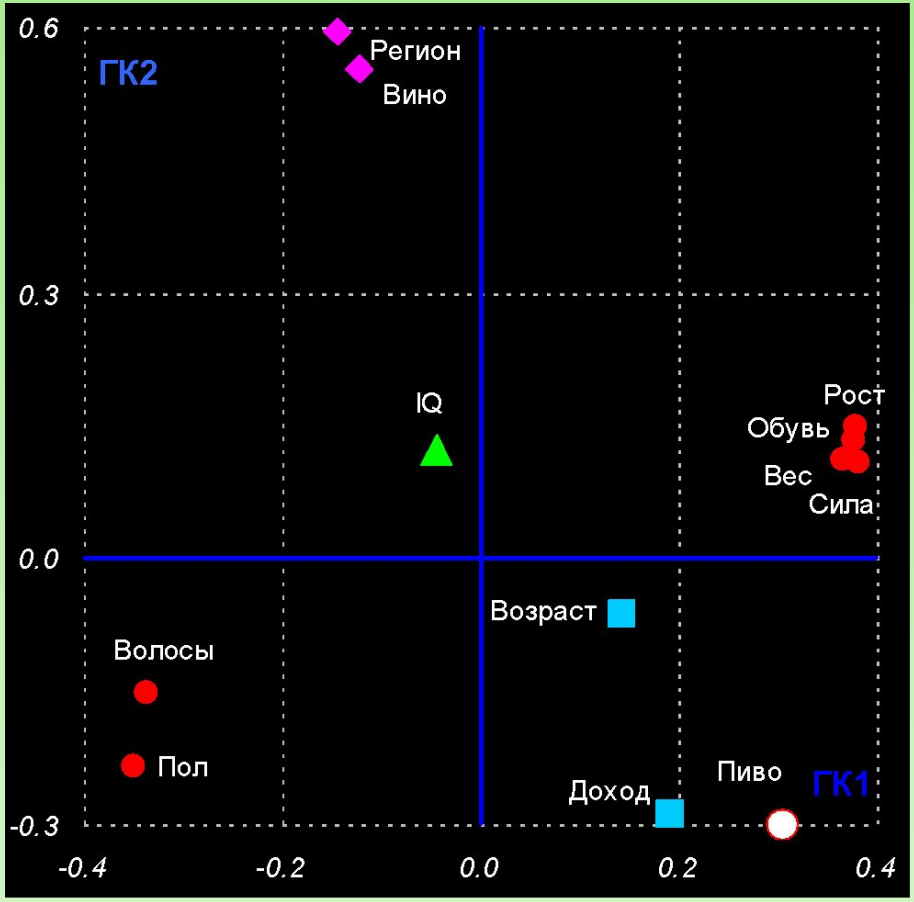

Chemometrics.Ru: учебники || MatLab. Руководство для начинающих - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.chemometrics.ru/materials/textbooks/matlab/#d0e634
Go Links
www.chemometrics.ru | учебники
MatLab. Руководство для начинающих
Евгений Михайлов, Алексей Померанцев
© Российское хемометрическое общество (http://rcs.chph.ras.ru)
1. Введен
2. Базовы
2.1.
2.2.
2.3.
2.4.
2.5.
2.6.
3. Матри
3.1.
3.2.
3.3.
3.4.
3.5.
3.6.
4. Програ
4.1.
4.2.
4.3.
4.4. Создание графика
C:\Maths\MatLab\work\pcanipals.m
File Edit View Text Debug Breakpoints Web Window Help
Stack: Base
function [T, P] = pcanipals(X, numberPC)
% pcanipals: calculates PCA components.
% The output matrices are T and P.
% T contains scores
% P contains loadings
% calculation of number of components
[X_r, X_c] = size(X); P=[]; T=[];
if lenfth(numberPC) > 0
pc = numberPC(1);
elseif (length(numberPC) == 0) & X_r < X_c
pc = X_r;
else
pc = X_c;
end;
% calculation of scores and loadings for each component
for k = 1:pc
pcanipals
Ln 30 Col 1

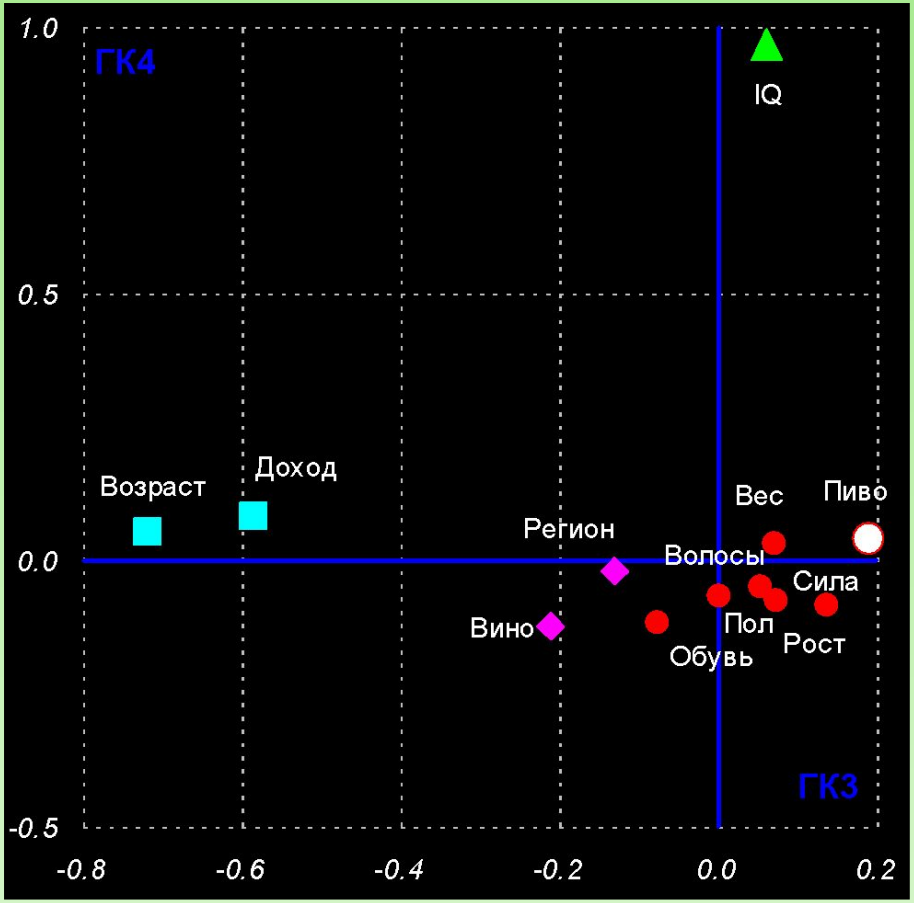
Люди: Графики счетов



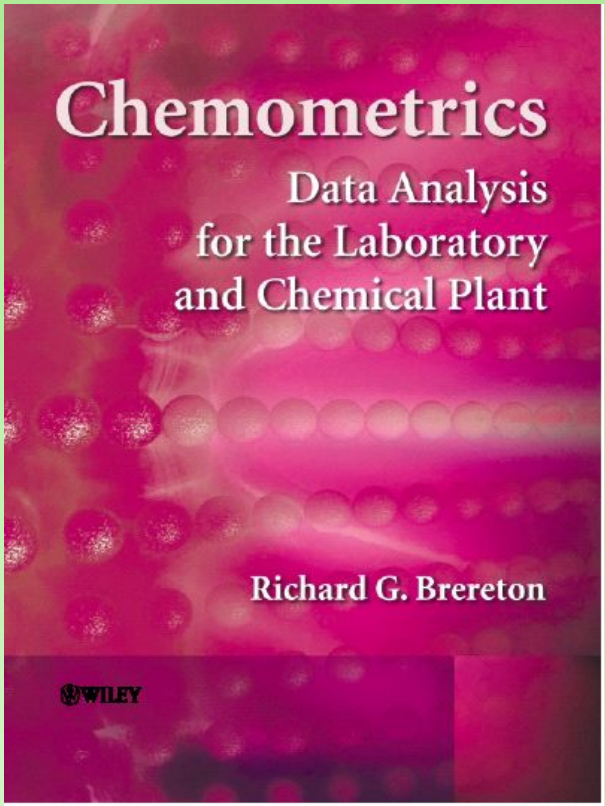


Люди: Графики нагрузок

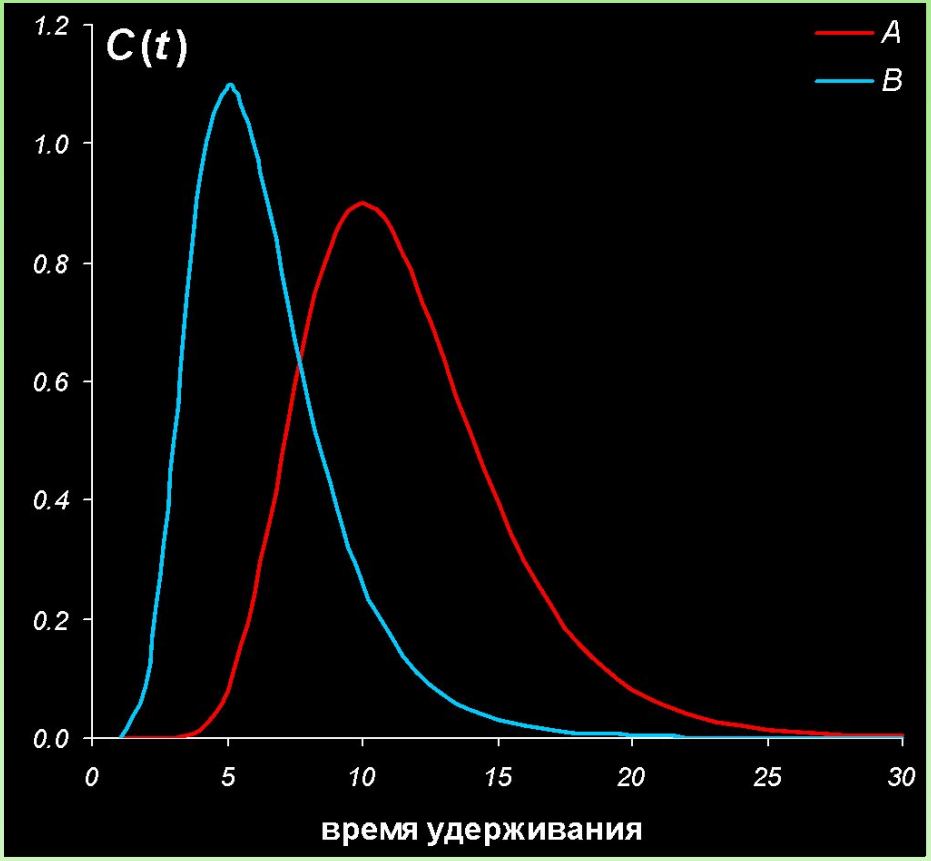


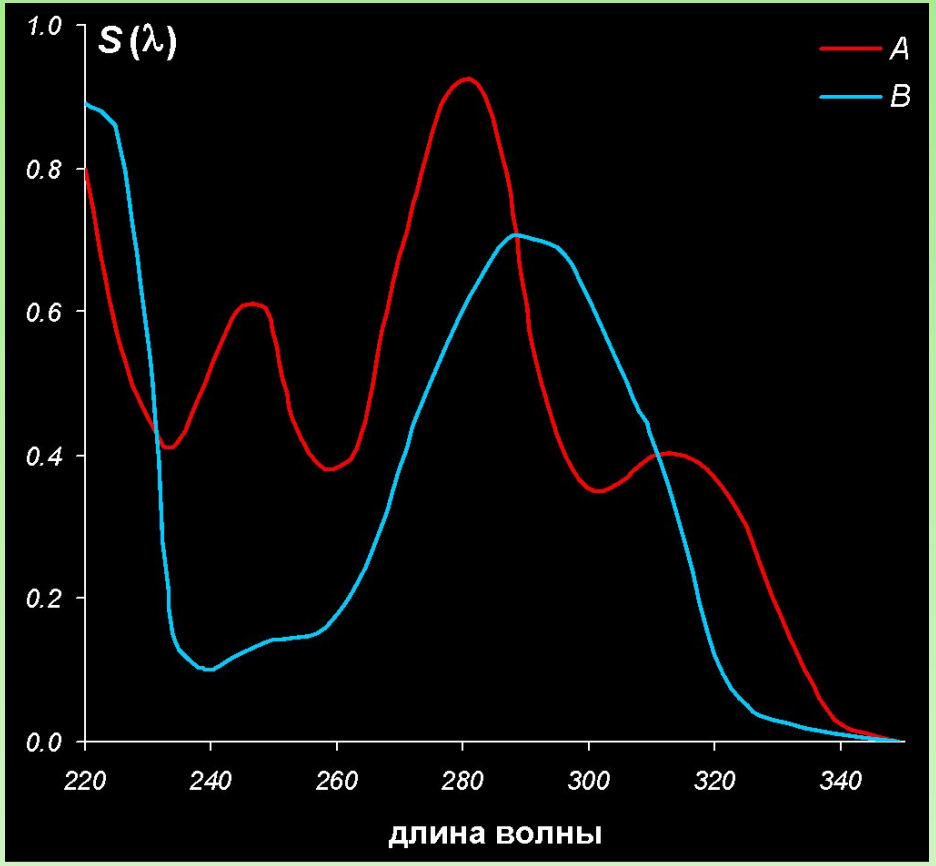


Разделение пиков в ВЭЖХ-ДАМ



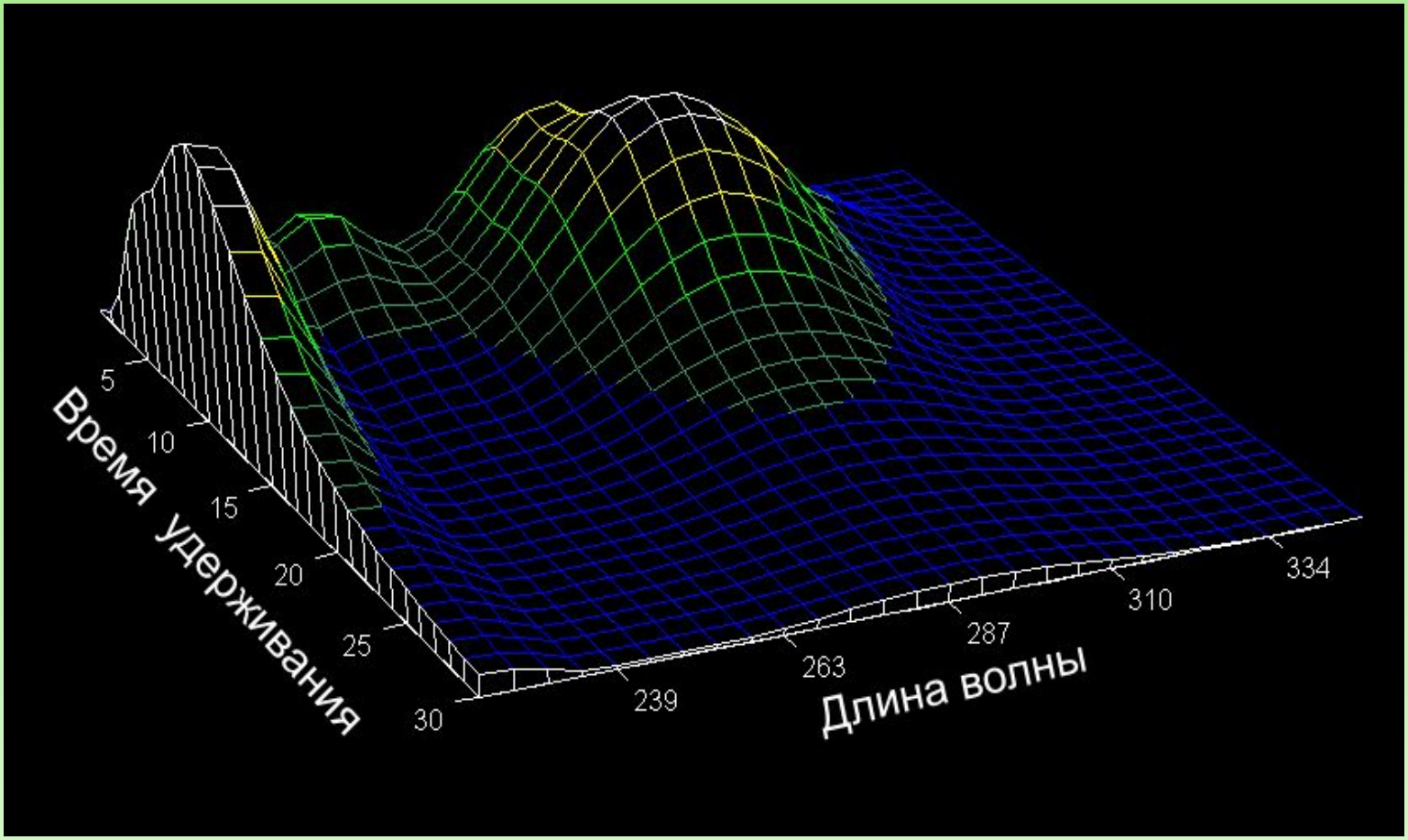
Свойства чистых веществ A и B



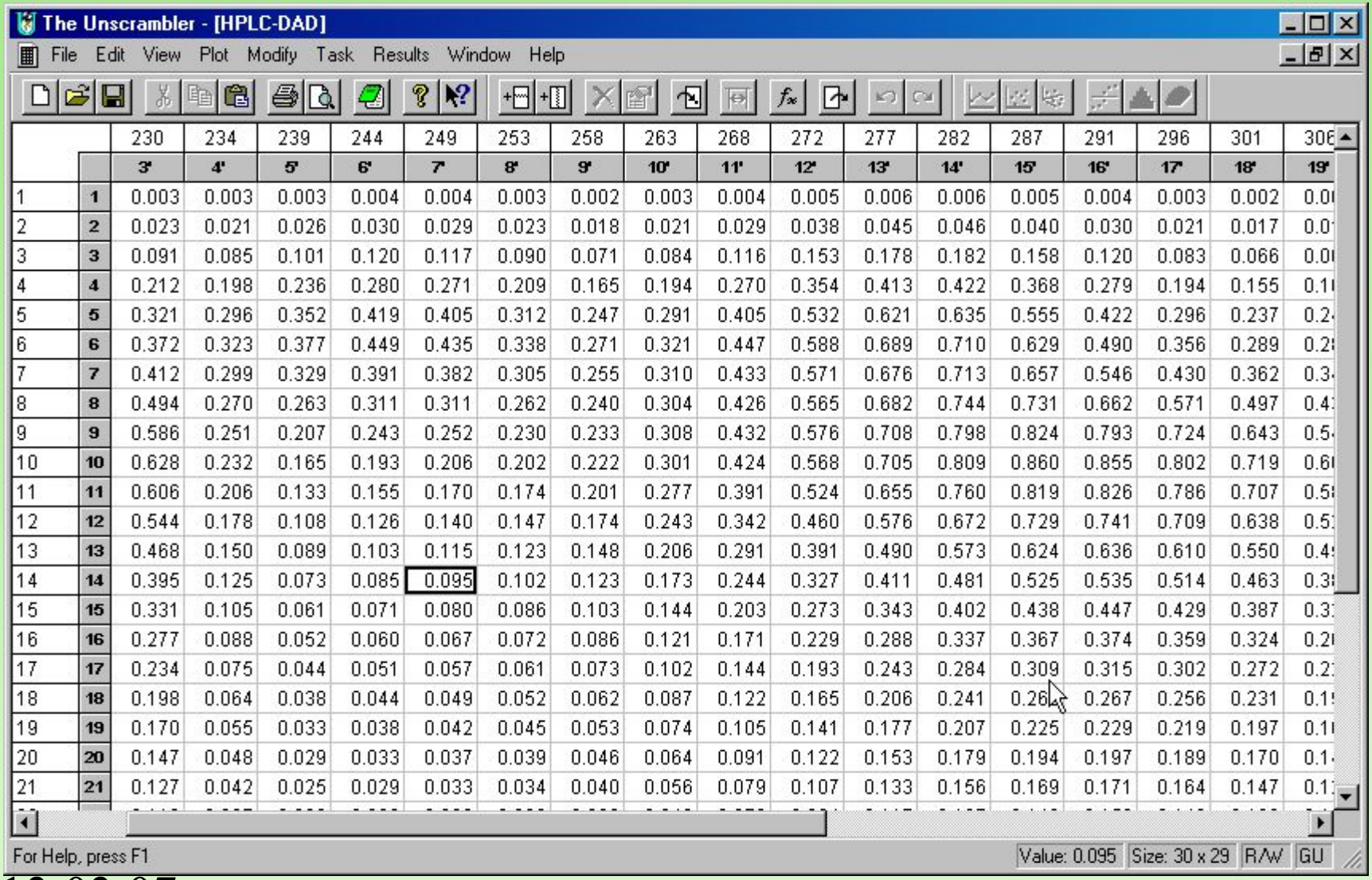


$$X=CS^{T}+E$$

Хроматограмма смеси

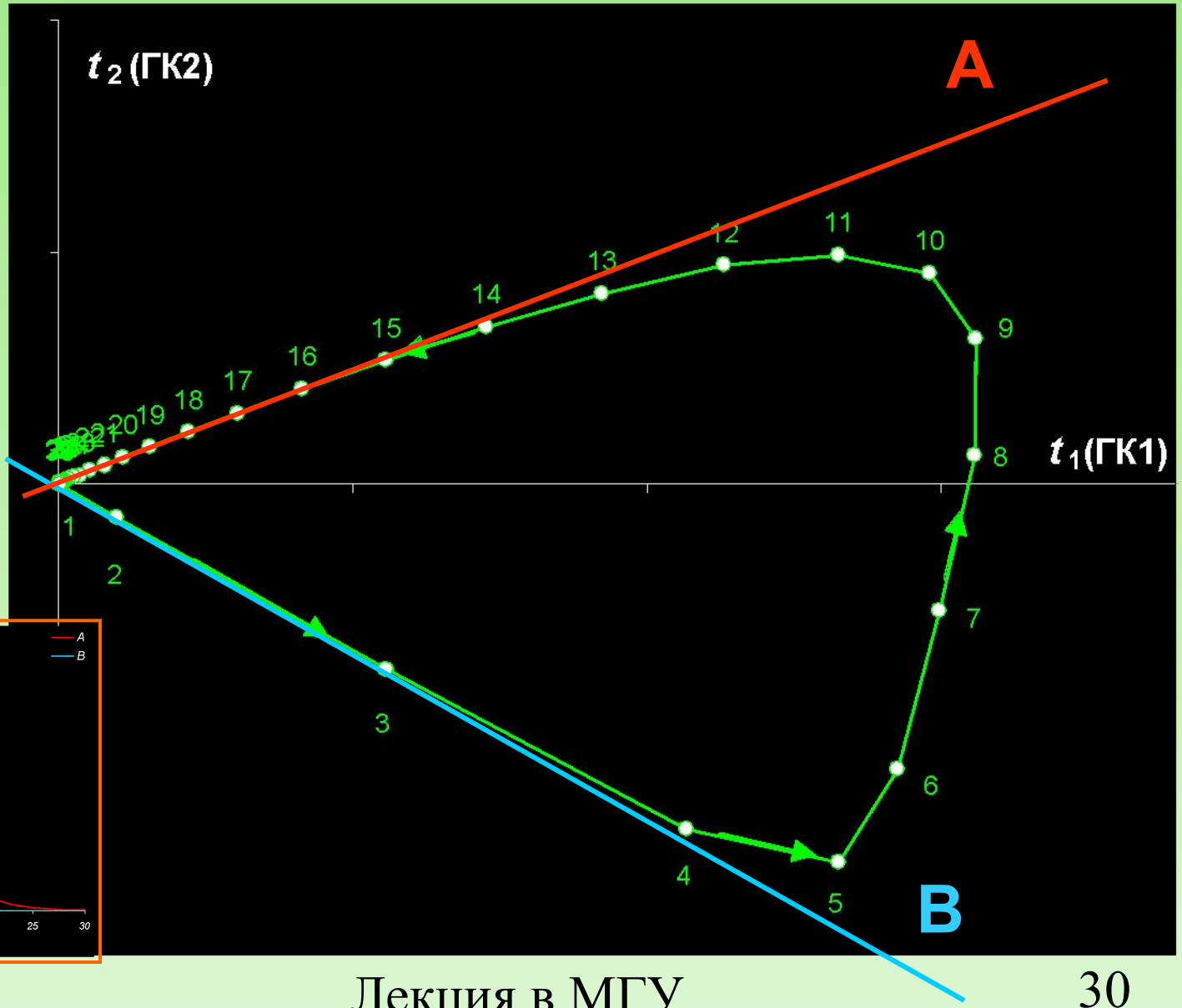


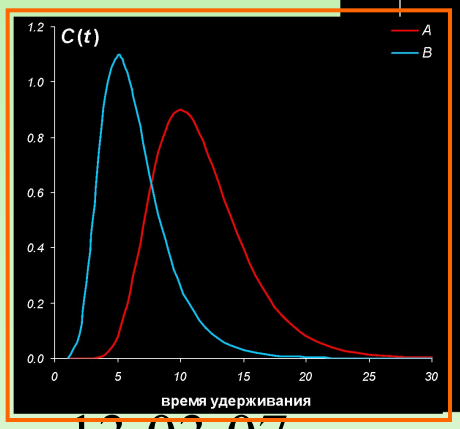
Данные в Unscrambler



		230	234	239	244	249	253	258	263	268	272	277	282	287	291	296	301	306
		3'	4'	5'	6'	7'	8'	9'	10'	11'	12'	13'	14'	15'	16'	17'	18'	19'
1	1	0.003	0.003	0.003	0.004	0.004	0.003	0.002	0.003	0.004	0.005	0.006	0.006	0.005	0.004	0.003	0.002	0.0
2	2	0.023	0.021	0.026	0.030	0.029	0.023	0.018	0.021	0.029	0.038	0.045	0.046	0.040	0.030	0.021	0.017	0.0
3	3	0.091	0.085	0.101	0.120	0.117	0.090	0.071	0.084	0.116	0.153	0.178	0.182	0.158	0.120	0.083	0.066	0.0
4	4	0.212	0.198	0.236	0.280	0.271	0.209	0.165	0.194	0.270	0.354	0.413	0.422	0.368	0.279	0.194	0.155	0.1
5	5	0.321	0.296	0.352	0.419	0.405	0.312	0.247	0.291	0.405	0.532	0.621	0.635	0.555	0.422	0.296	0.237	0.2
6	6	0.372	0.323	0.377	0.449	0.435	0.338	0.271	0.321	0.447	0.588	0.689	0.710	0.629	0.490	0.356	0.289	0.2
7	7	0.412	0.299	0.329	0.391	0.382	0.305	0.255	0.310	0.433	0.571	0.676	0.713	0.657	0.546	0.430	0.362	0.3
8	8	0.494	0.270	0.263	0.311	0.311	0.262	0.240	0.304	0.426	0.565	0.682	0.744	0.731	0.662	0.571	0.497	0.4
9	9	0.586	0.251	0.207	0.243	0.252	0.230	0.233	0.308	0.432	0.576	0.708	0.798	0.824	0.793	0.724	0.643	0.5
10	10	0.628	0.232	0.165	0.193	0.206	0.202	0.222	0.301	0.424	0.568	0.705	0.809	0.860	0.855	0.802	0.719	0.6
11	11	0.606	0.206	0.133	0.155	0.170	0.174	0.201	0.277	0.391	0.524	0.655	0.760	0.819	0.826	0.786	0.707	0.5
12	12	0.544	0.178	0.108	0.126	0.140	0.147	0.174	0.243	0.342	0.460	0.576	0.672	0.729	0.741	0.709	0.638	0.5
13	13	0.468	0.150	0.089	0.103	0.115	0.123	0.148	0.206	0.291	0.391	0.490	0.573	0.624	0.636	0.610	0.550	0.4
14	14	0.395	0.125	0.073	0.085	0.095	0.102	0.123	0.173	0.244	0.327	0.411	0.481	0.525	0.535	0.514	0.463	0.3
15	15	0.331	0.105	0.061	0.071	0.080	0.086	0.103	0.144	0.203	0.273	0.343	0.402	0.438	0.447	0.429	0.387	0.3
16	16	0.277	0.088	0.052	0.060	0.067	0.072	0.086	0.121	0.171	0.229	0.288	0.337	0.367	0.374	0.359	0.324	0.2
17	17	0.234	0.075	0.044	0.051	0.057	0.061	0.073	0.102	0.144	0.193	0.243	0.284	0.309	0.315	0.302	0.272	0.2
18	18	0.198	0.064	0.038	0.044	0.049	0.052	0.062	0.087	0.122	0.165	0.206	0.241	0.26	0.267	0.256	0.231	0.1
19	19	0.170	0.055	0.033	0.038	0.042	0.045	0.053	0.074	0.105	0.141	0.177	0.207	0.225	0.229	0.219	0.197	0.1
20	20	0.147	0.048	0.029	0.033	0.037	0.039	0.046	0.064	0.091	0.122	0.153	0.179	0.194	0.197	0.189	0.170	0.1
21	21	0.127	0.042	0.025	0.029	0.033	0.034	0.040	0.056	0.079	0.107	0.133	0.156	0.169	0.171	0.164	0.147	0.1

График счетов

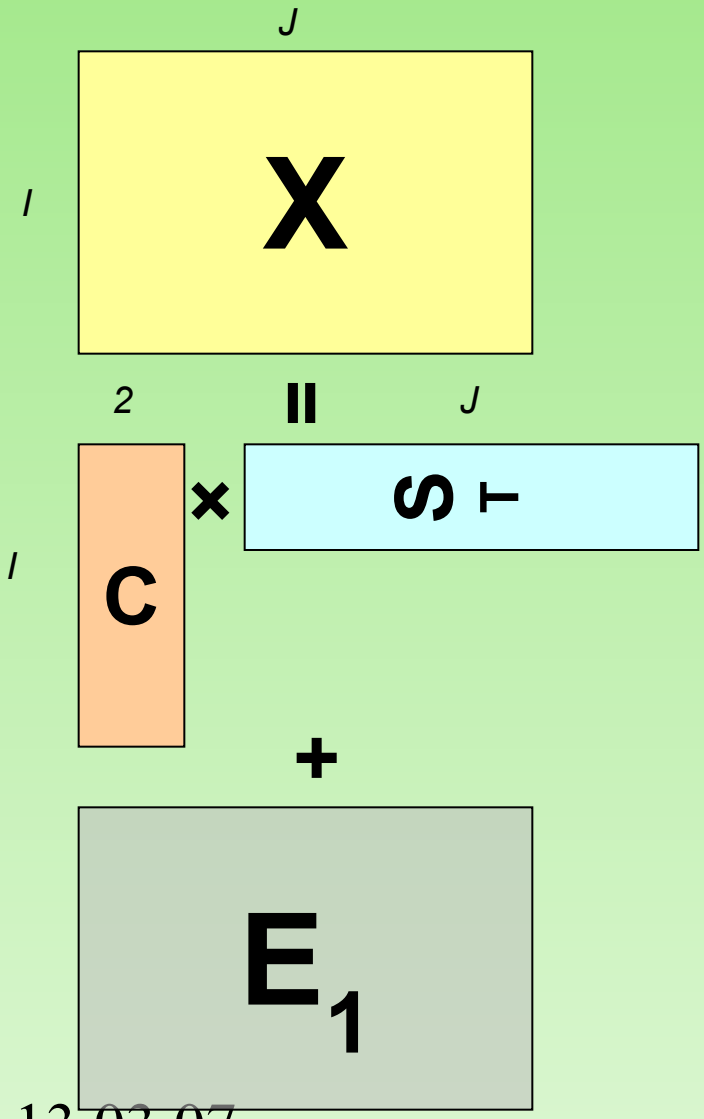


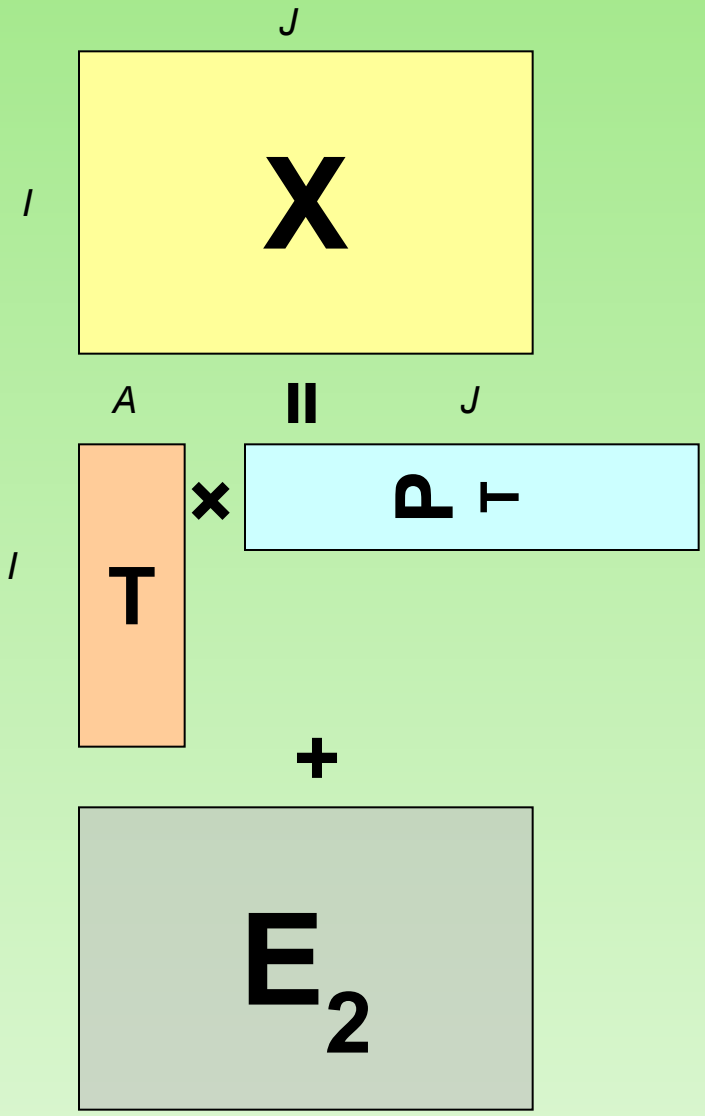


Выводы из графика счетов

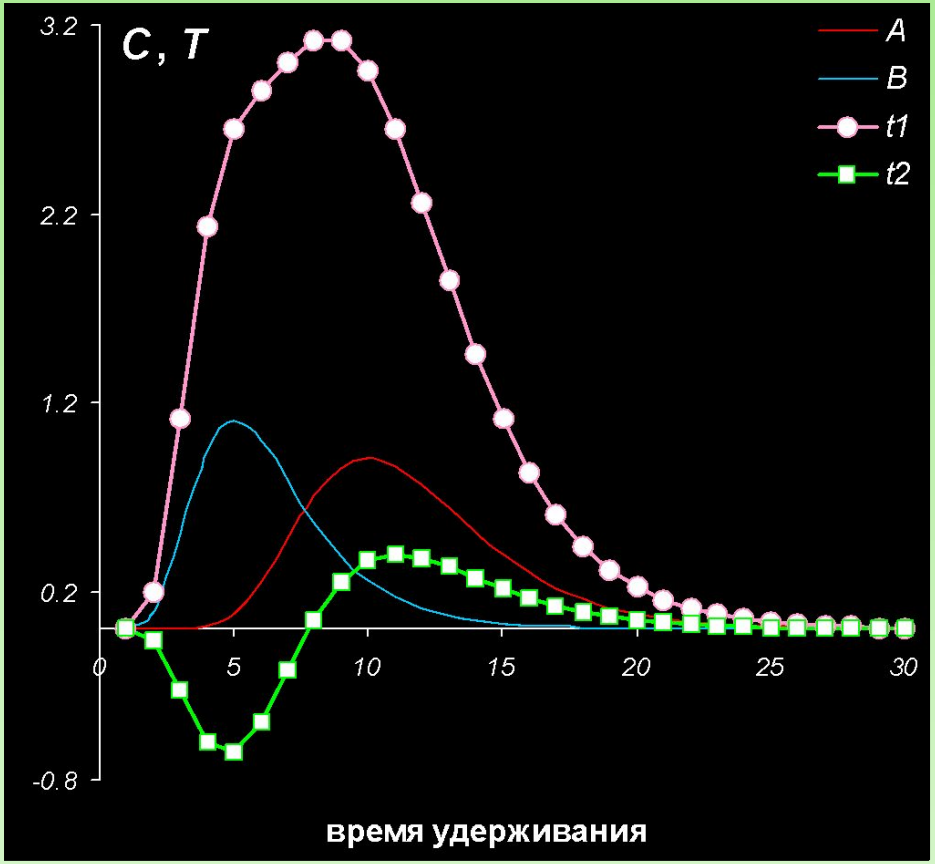
1. линейные участки = чистые компоненты
2. кривые участки = коэлюция
3. ближе к началу = меньше интенсивность
4. число поворотов = число чистых компонент

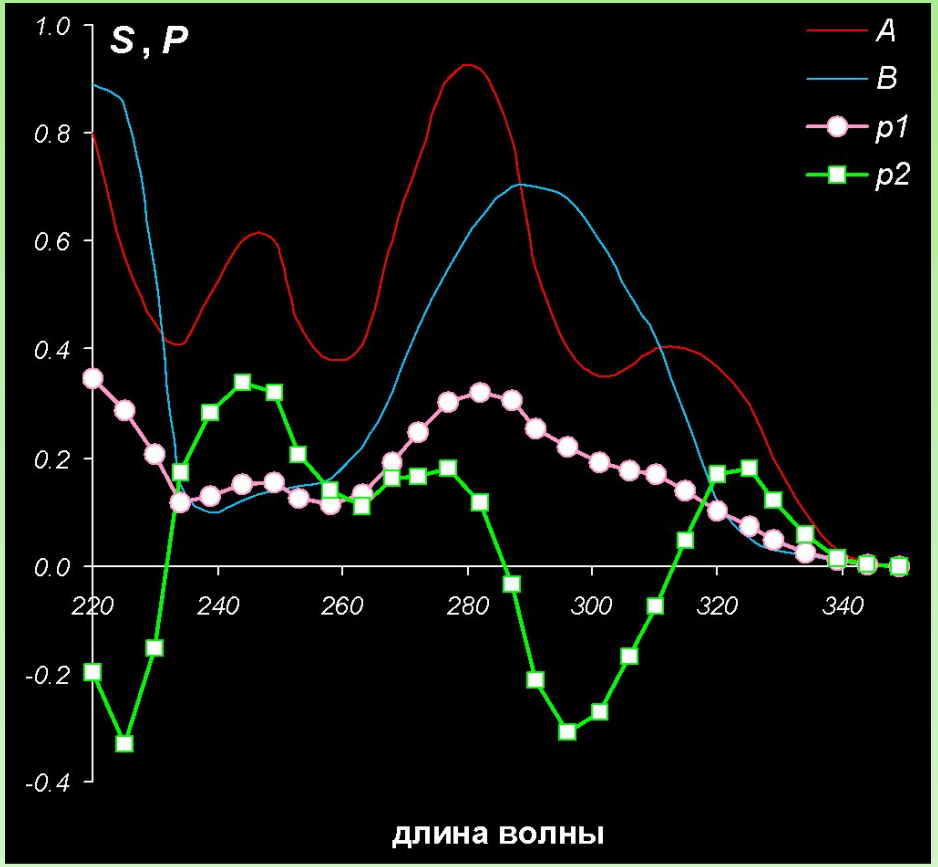
Факторный анализ и анализ ГК





Счета и нагрузки





Прокрустово преобразование

$$X \approx CS^T$$

$$X \approx TP^T$$

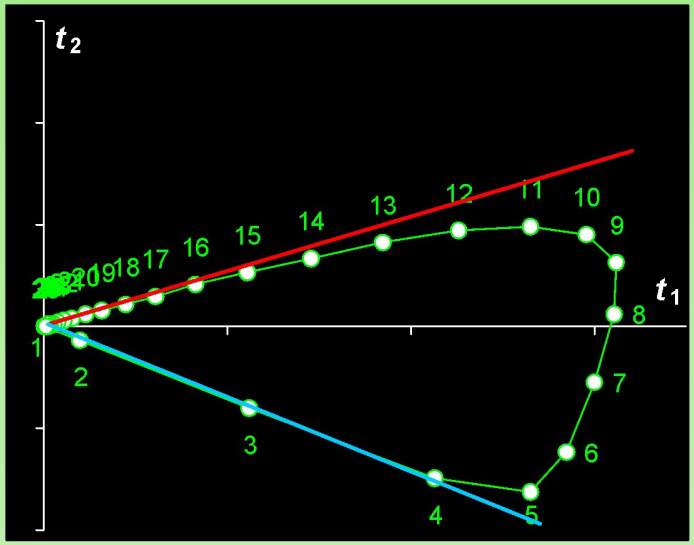
$$I = RR^T = \text{единичная}$$

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

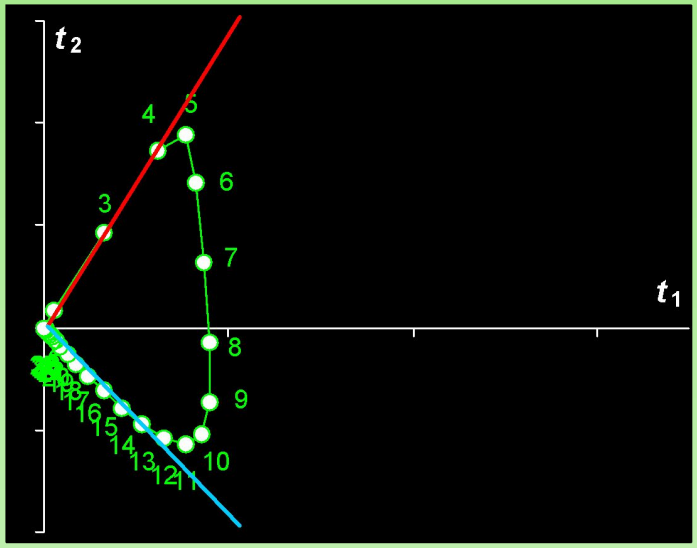
$$C \approx TR \qquad S \approx PR$$

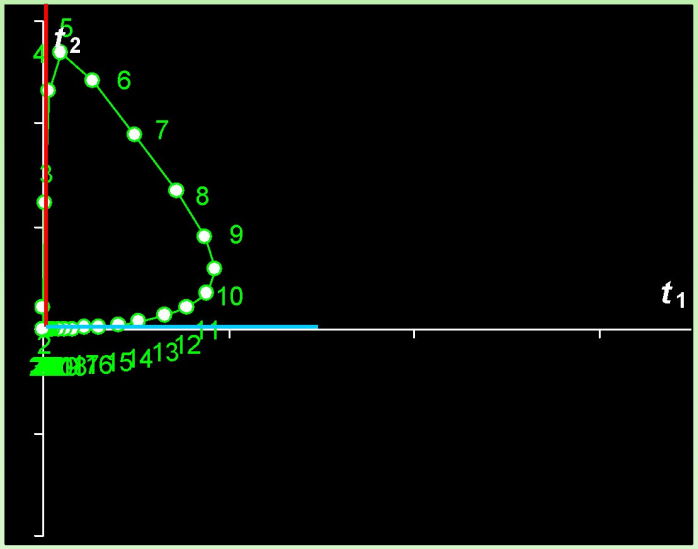
$$R = R_{stratch} \times R_{rotation}$$

Преобразование счетов



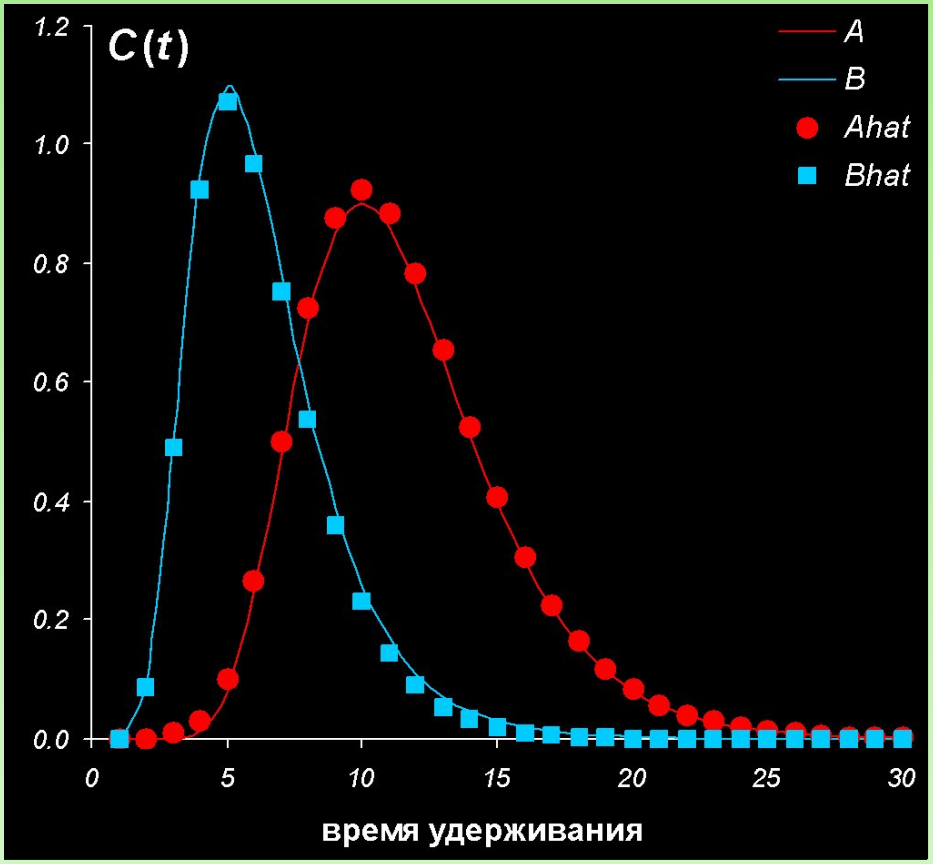
СЖАТИЕ

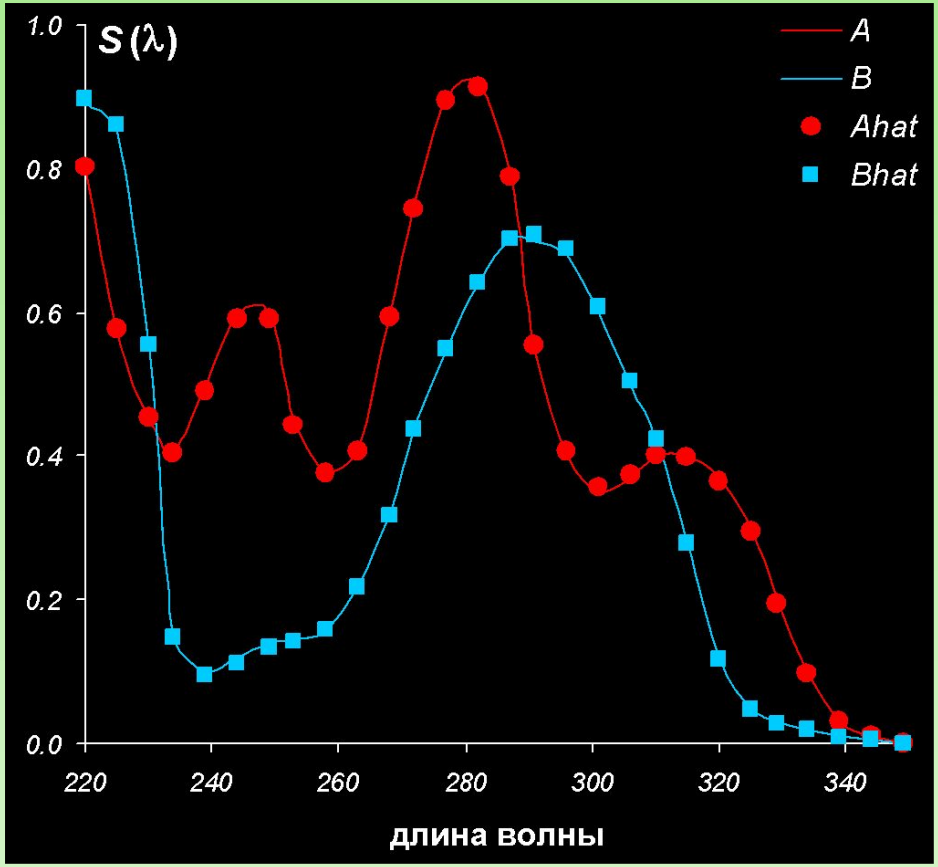




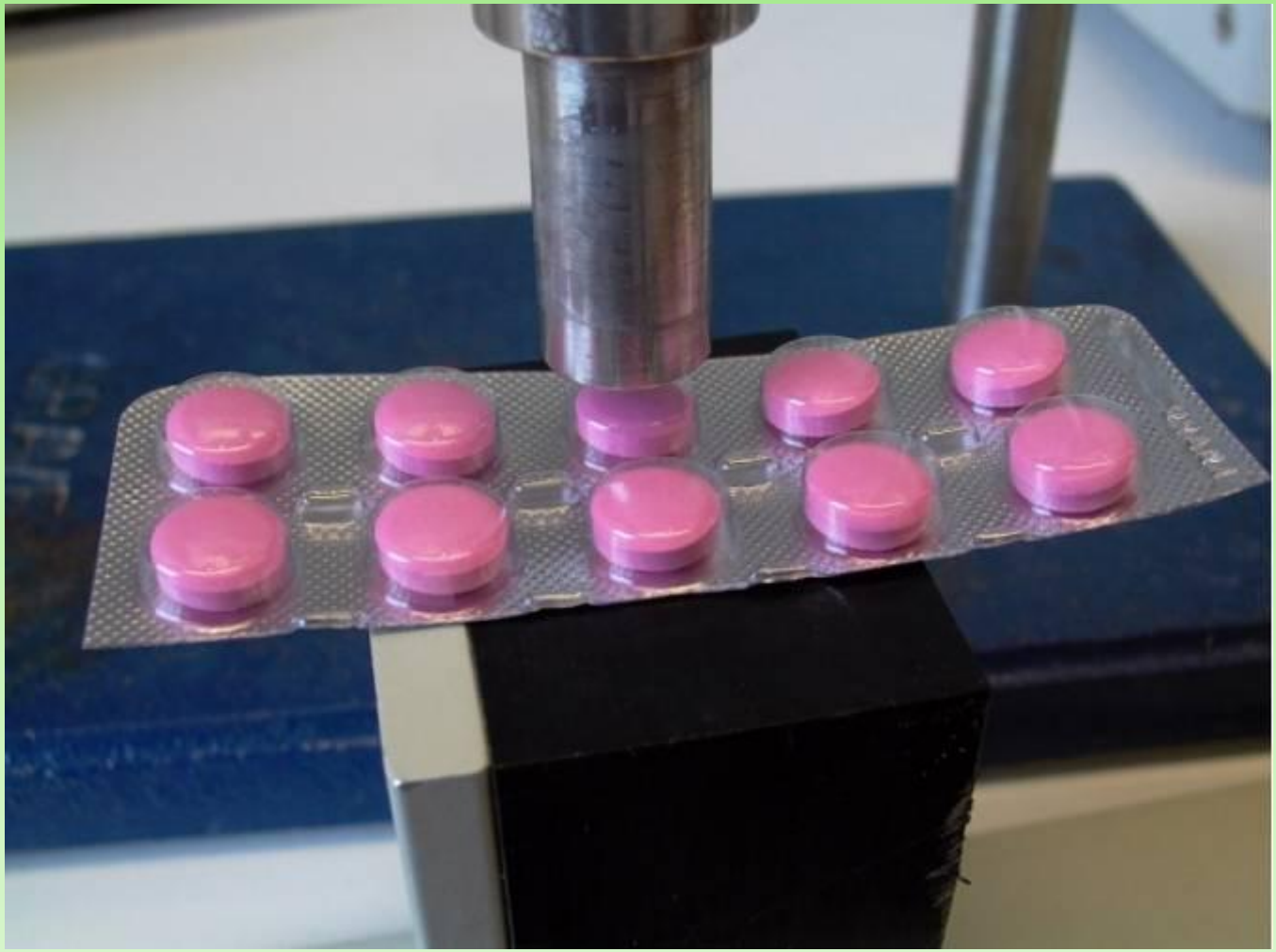
ВРАЩЕНИЕ

Результат Прокрустова преобразования





Классификация поддельных лекарств



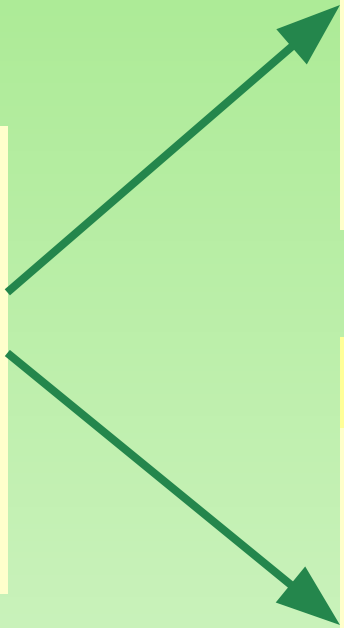
Измерения: БИК спектроскопия

Образцы

Подлинники:
11 серий по 5 таблеток

Подделки:
4 серии по 5 таблеток

Всего
15×5= 75 образцов



Обучающий набор

Подлинники:
8 серий по 5 таблеток

Всего:
8×5= 40 образцов

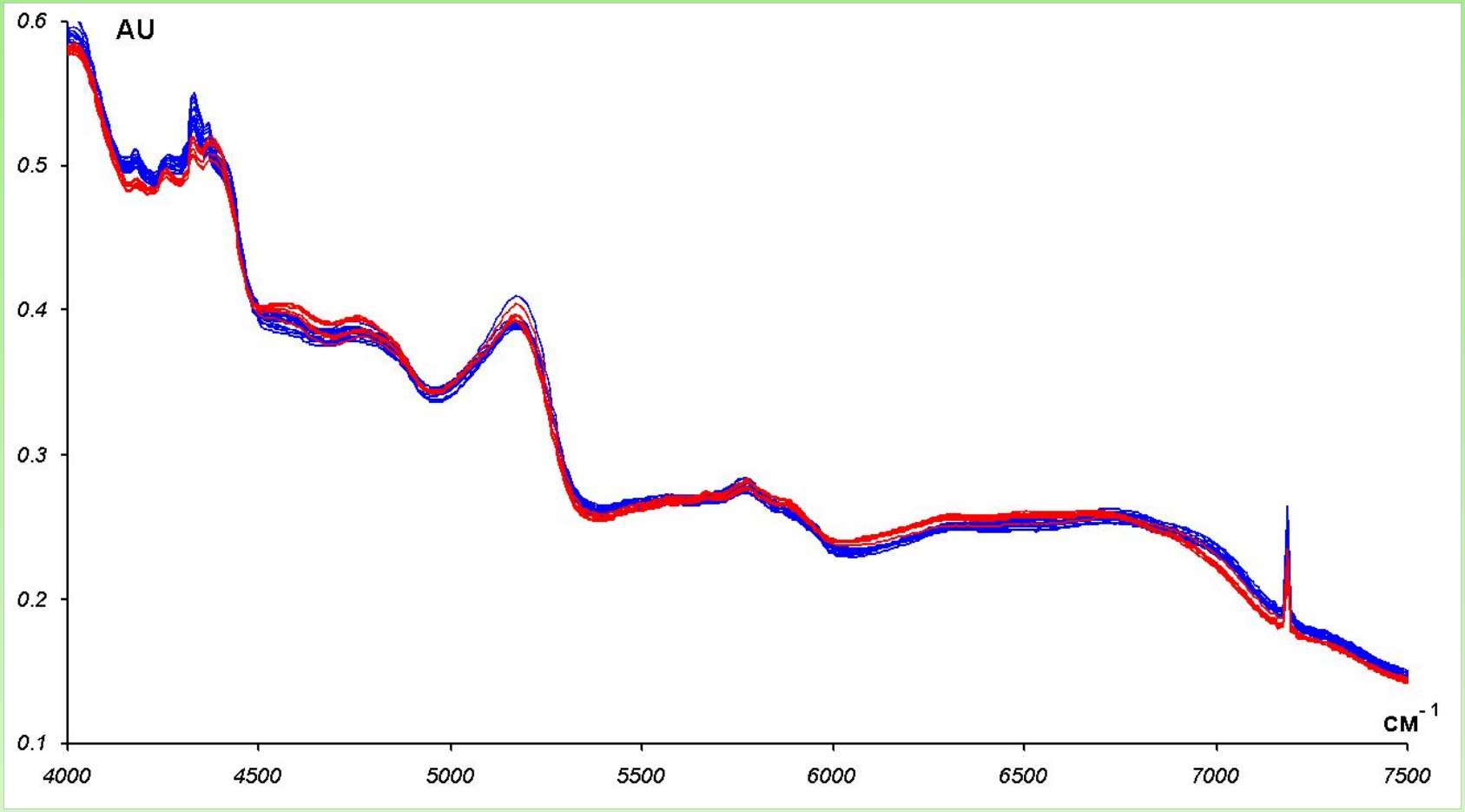
Проверочный набор

Подлинники:
3 серии по 5 таблеток

Подделки:
4 серии по 5 таблеток

Всего:
7×5= 35 образцов

БИК спектры мезима



Настоящие **Фальшивые**

Данные: БИК спектры

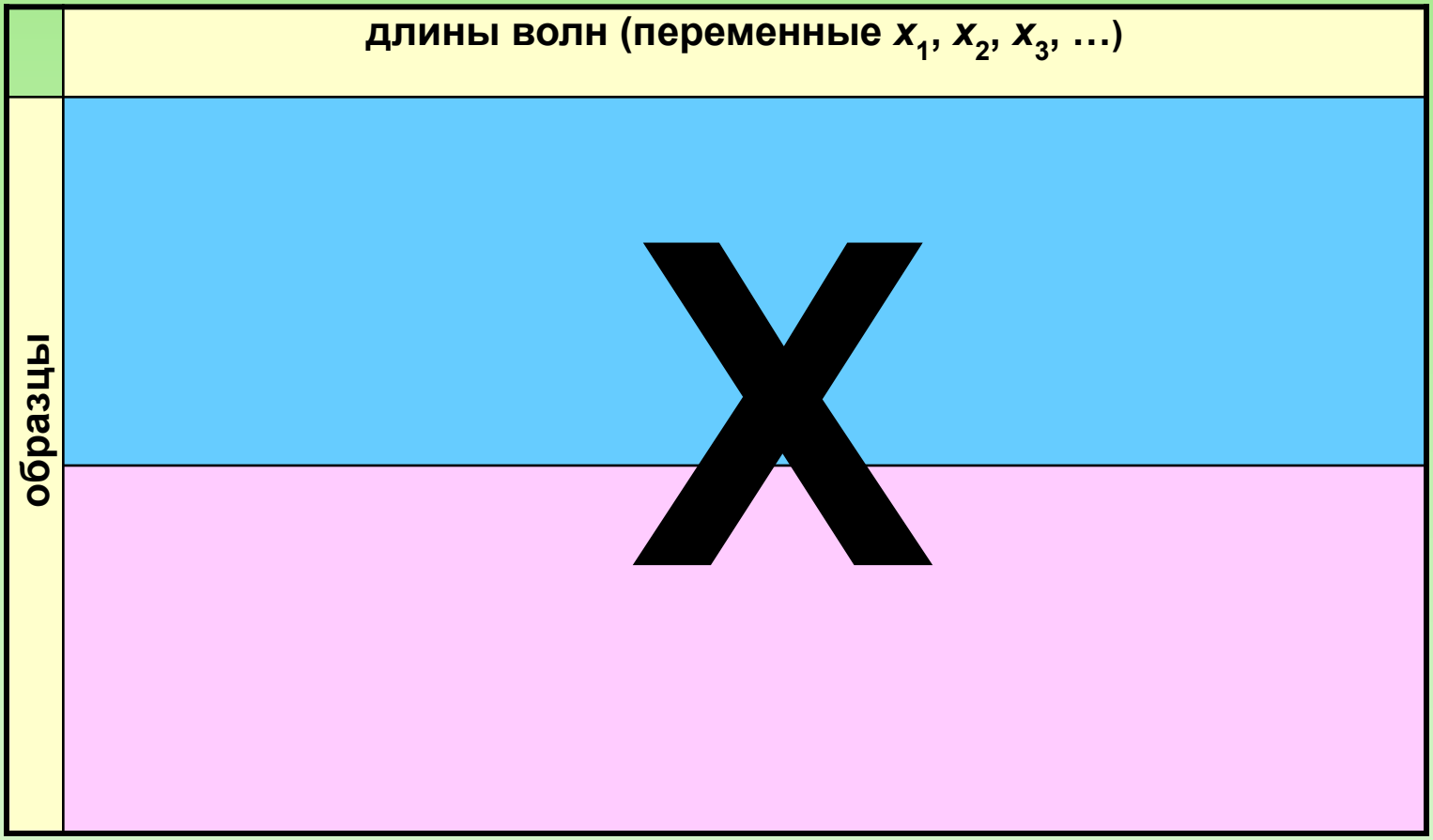
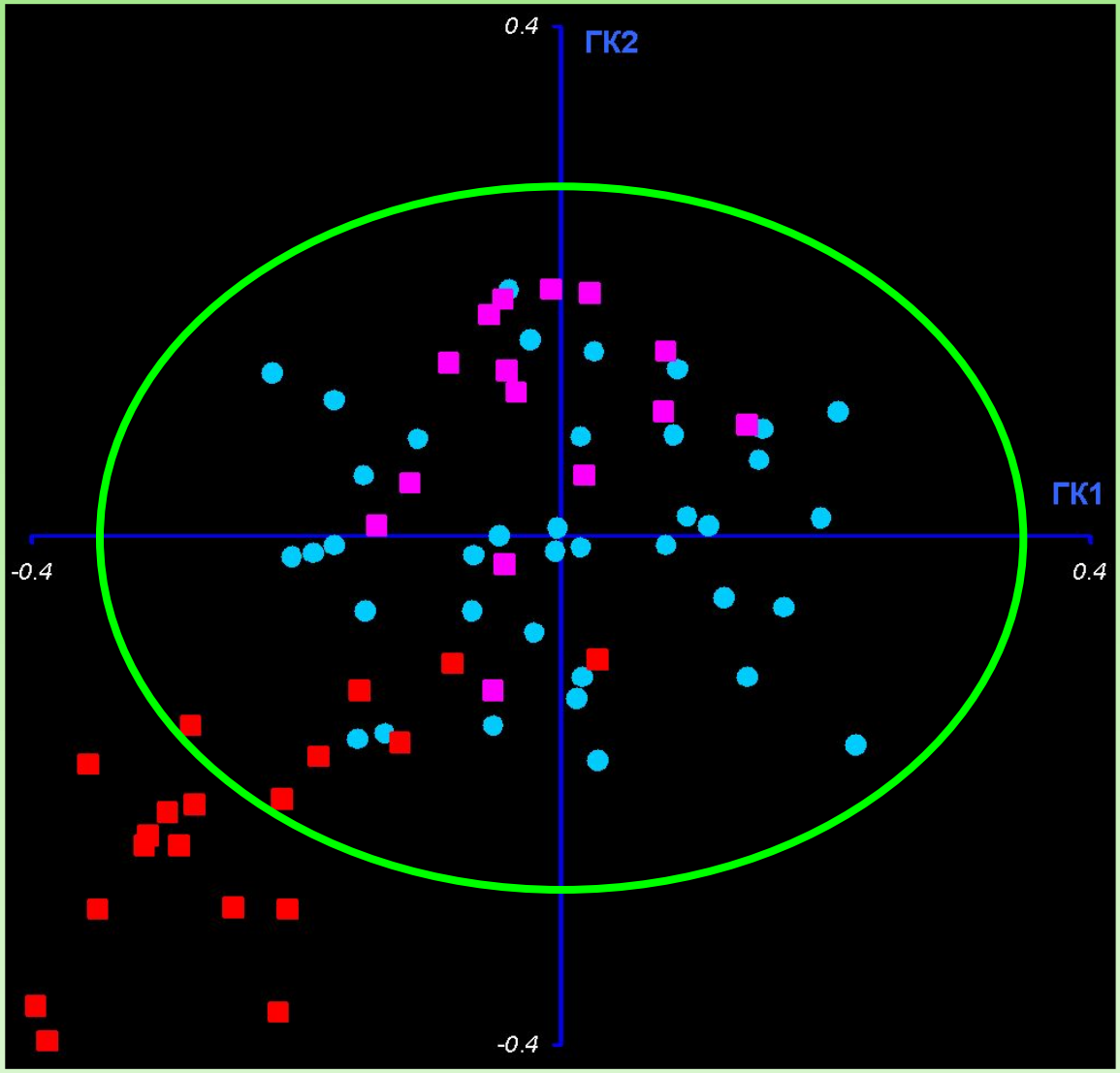


График МГК счетов



Обучающий набор

Подлинники:
8 серий по 5 таблеток

Всего:
8×5= 40 образцов

Проверочный набор

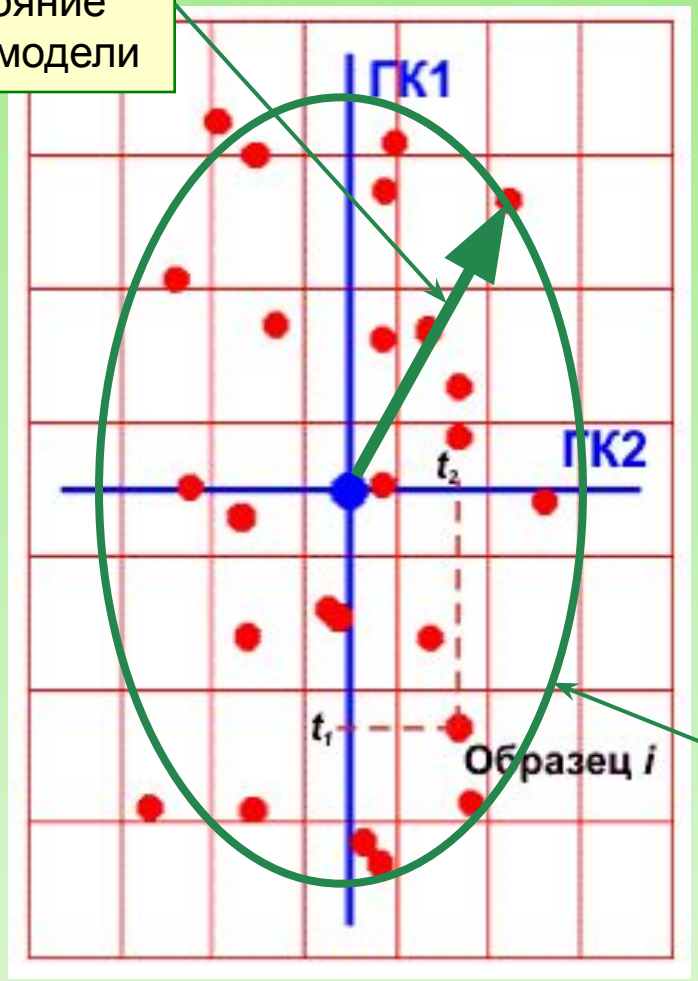
Подлинники:
3 серии по 5 таблеток

Подделки:
4 серии по 5 таблеток

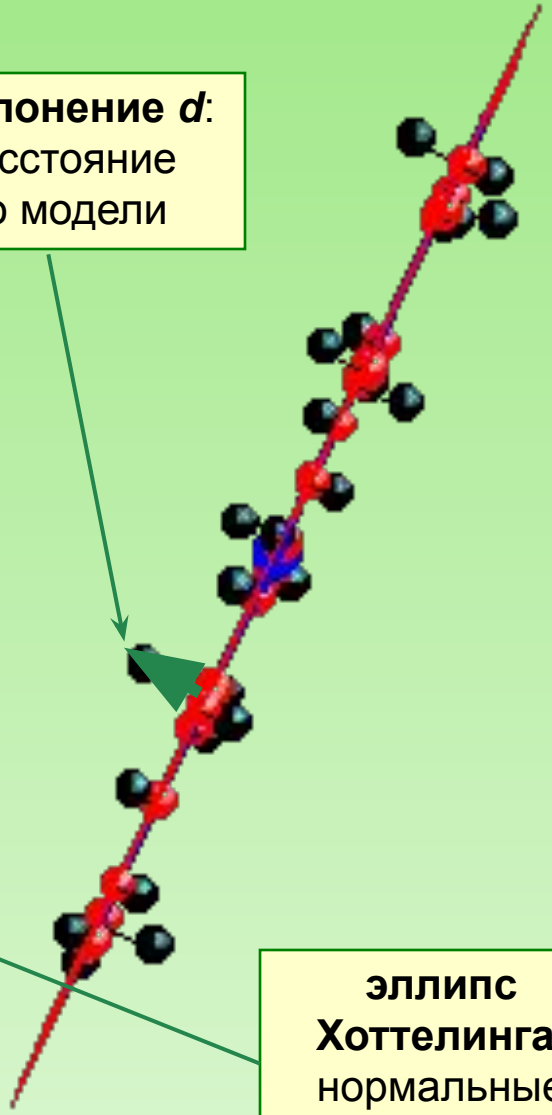
Всего:
7×5= 35 образцов

Размах и отклонение

Размах *h*:
расстояние
внутри модели

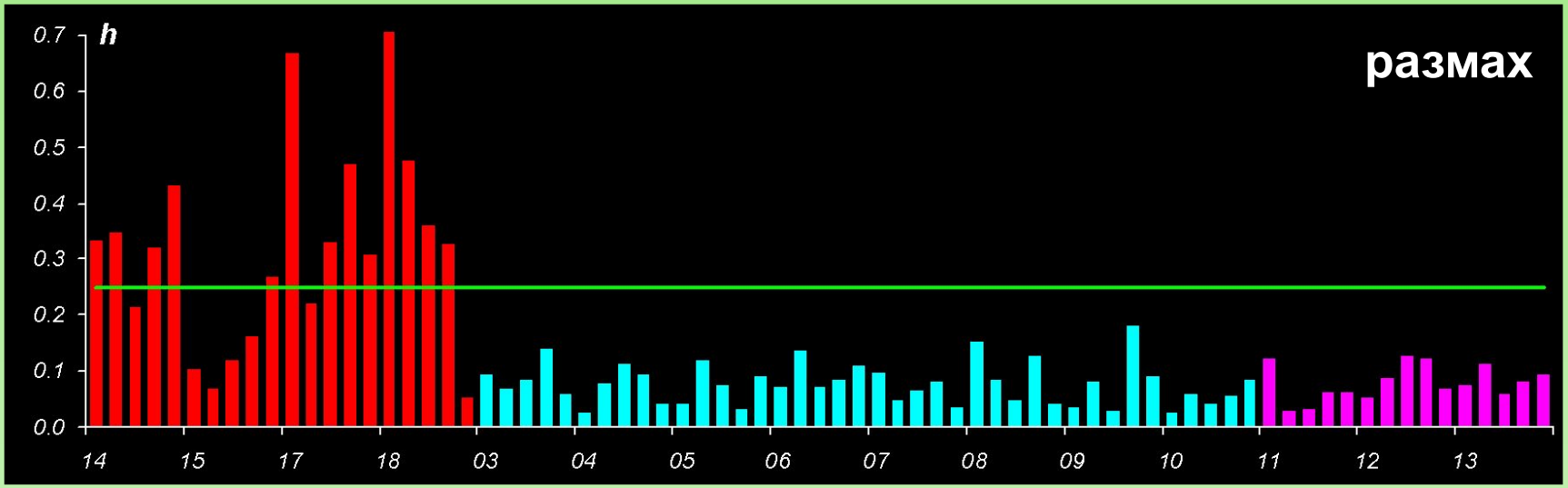


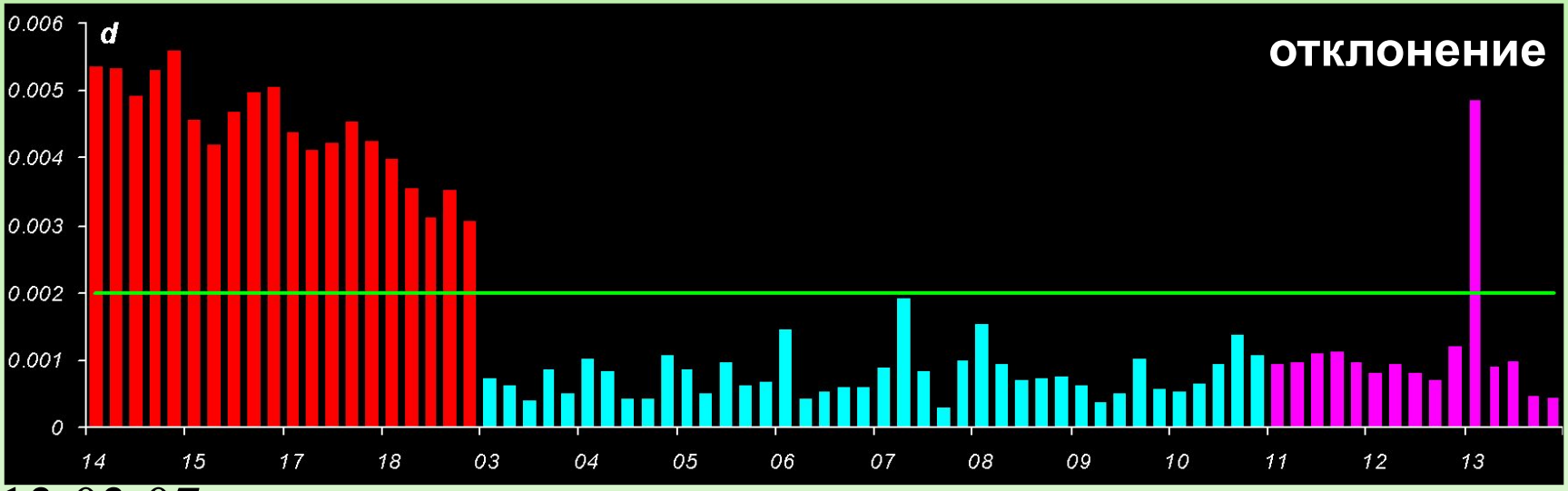
Отклонение *d*:
расстояние
до модели



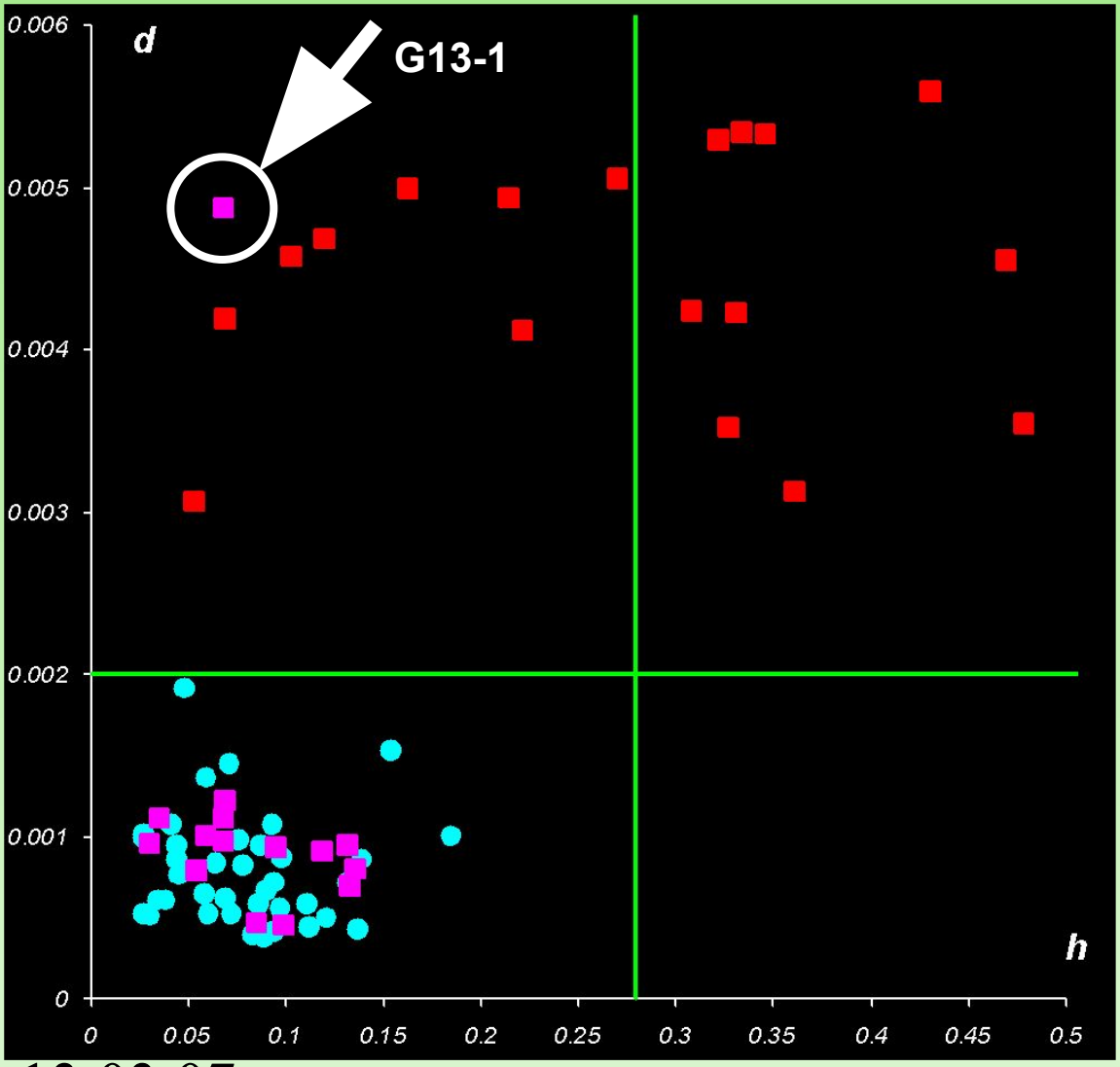
**эллипс
Хоттелинга**:
нормальные
образцы

Размах и отклонение для мезима





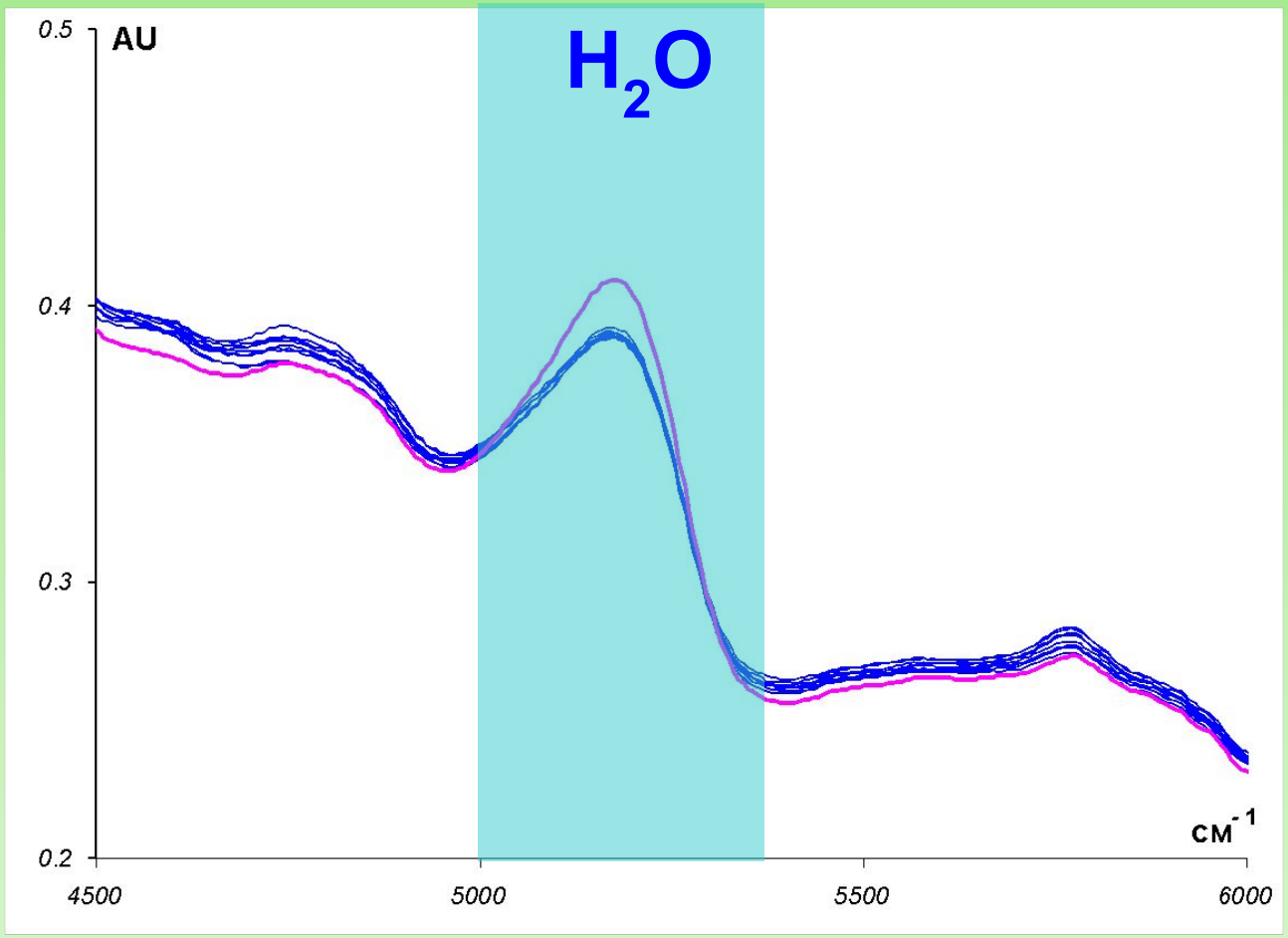
SIMCA



Soft
Independent
Modeling of
Class
Analogy

формальное
независимое
моделирование
аналогий
классов

Выброс G13-1



Хемометрика: два определения

Дедуктивное

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

сайт Российского хемометрического общества

Индуктивное

Хемометрика – это то, что делают хемометрики.

сайт Международного хемометрического общества

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

Svante Wold

Что делают хемометрики?

- контролируют производство полупроводников, аспирина и пива;
- исследуют причины разрушения старинных документов;
- проводят допинг контроль спортсменов;
- расшифровывают состав косметики Древнего Египта;
- локализуют месторождение золота в Швеции;
- идентифицируют подозреваемого в убийстве Анны Линд;
- проводят диагностику артрита и рака на ранних стадиях;
- исследуют органические субстанции в кометном веществе;
- исследуют кормовой рацион на свинооткормочных комплексах;
- проверяют, как рацион питания влияет на умственные способности;
- определяют фальшивые лекарства;
- и еще многое, многое другое

А если серьезно ....

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

Два «не» и три «да»

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

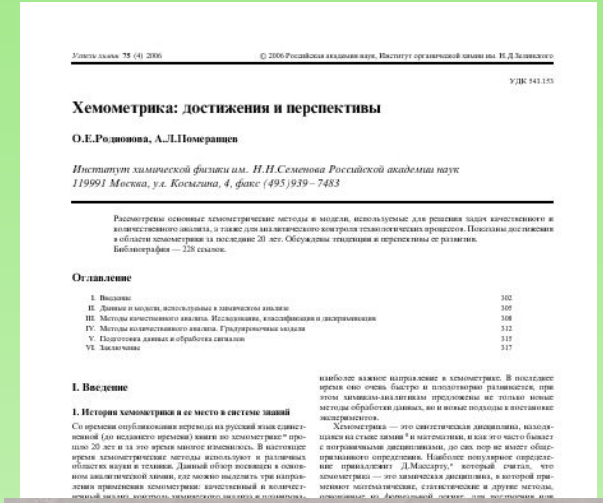
Хемометрика решает следующие задачи в области химии:

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

(2) как организовать и представить эту информацию,

(3) как получить данные, содержащие такую информацию.

Что можно почитать ...



Успехи химии 75 (4) 2006 © 2006 Российская академия наук. Институт органической химии им. Н.Д.Зелинского

УДК 543.153

Хемометрика: достижения и перспективы

О.Е.Родионова, А.Л.Померанцев

Институт химической физики им. Н.Н.Семенова Российской академии наук
119991 Москва, ул. Косыгина, 4, факс (495)939–7483

Рассмотрены основные хемометрические методы и модели, используемые для решения задач качественного и количественного анализа, а также для анализа и контроля технологических процессов. Показаны достижения в области хемометрики за последние 20 лет. Обсуждены тенденции и перспективы ее развития. Библиография — 228 ссылок.

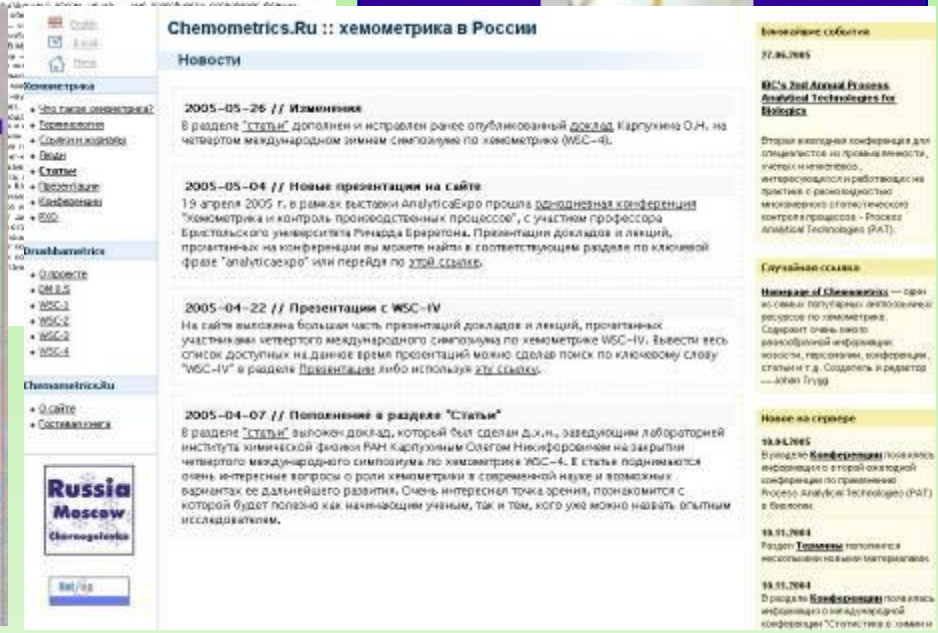
Оглавление

I. Введение

1. История хемометрики и ее место в системе знаний

Со времени опубликования перевода на русский язык единственной (до недавнего времени) книги по хемометрике* прошло 20 лет и за это время многое изменилось. В настоящее время хемометрические методы используют в различных областях науки и техники. Данный обзор посвящен в основном аналитической химии, где можно выделить три направления применения хемометрики: качественный и количественный анализ данных, управление экспериментом и планирование, наиболее важное направление в хемометрике. В последнее время оно очень быстро и плодотворно развивается, при этом химико-аналитическим предложены не только новые методы обработки данных, но и новые подходы в постановке экспериментов.

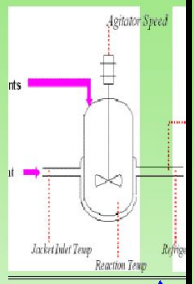
Хемометрика — это синтетическая дисциплина, находящаяся на стыке химии* и математики, и как это часто бывает с пограничными дисциплинами, до сих пор не имеет общепризнанного определения. Наиболее популярное определение принадлежит Д.Массарту,* который считал, что хемометрика — это химическая дисциплина, в которой применяют математические, статистические и другие методы,



Что осталось на потом ...

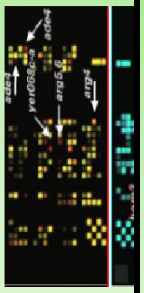
MSPC в фармацевтике

Реактор



10.02.05

Хемометрика в биологии

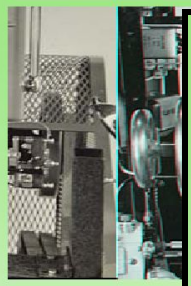


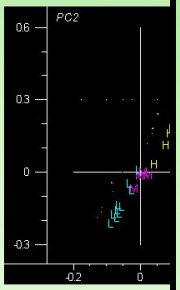
10.02.05

Электронные язык и нос

10.02.05

Определение качества бензина по ИК-спектру





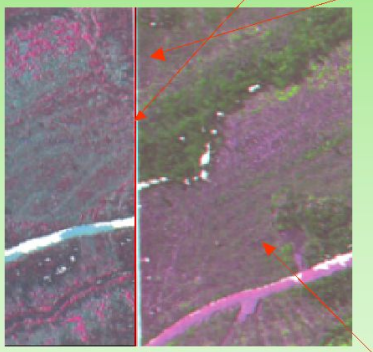
10.02.05

Исследование состояния лесов (Канада)

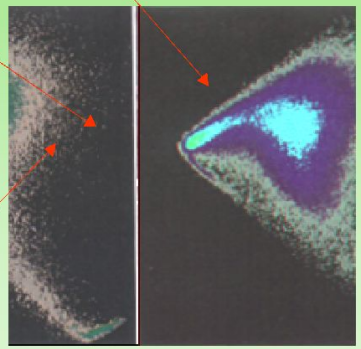
Новые посадки

Старые деревья

Исходный аэроснимок



Он же в пространстве ГК1



Область с высоким коэффициентом отражения

07.10.06 EVENT 30

Спасибо за внимание!