

Концепция скрытых (латентных) переменных в химическом анализе.

Часть 1. Качественный анализ

Померанцев
Алексей Леонидович



*Институт химической
физики РАН им Семенова*



*Российское хемометрическое
общество*

13.03.07

Лекция в МГУ

1

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

Хеометрика

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	образец/объект			2	2
...	0	0	переменная/свойство	1	3	1	1
...	1	1		1	1	1	1
...	0	1		1	4	1	1
...	0	1		1	5	1	3
...	0	1		1	1	1	1
...	1	0		1	1	2	4
...	0	1		2	1	1	1
Зонтов Ю.	1	0		4	6	0	1
...	0	1		1	2	1	1
...	0	0		1	3	1	2
...	1	1	1	1	1	2	
Померанцев А.	0	0	1	6	12	3	
...	1	1	1	2	2	1	
...	1	1	1	1	1	2	3
...	1	1	2	1	2	1	2

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

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

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 $\{a_{ij}, i = 1, \dots, I, j = 1, \dots, J\}$. Для приведенной в примере матрицы $I = 4, J = 3$ и $a_{23} = -7.5$.

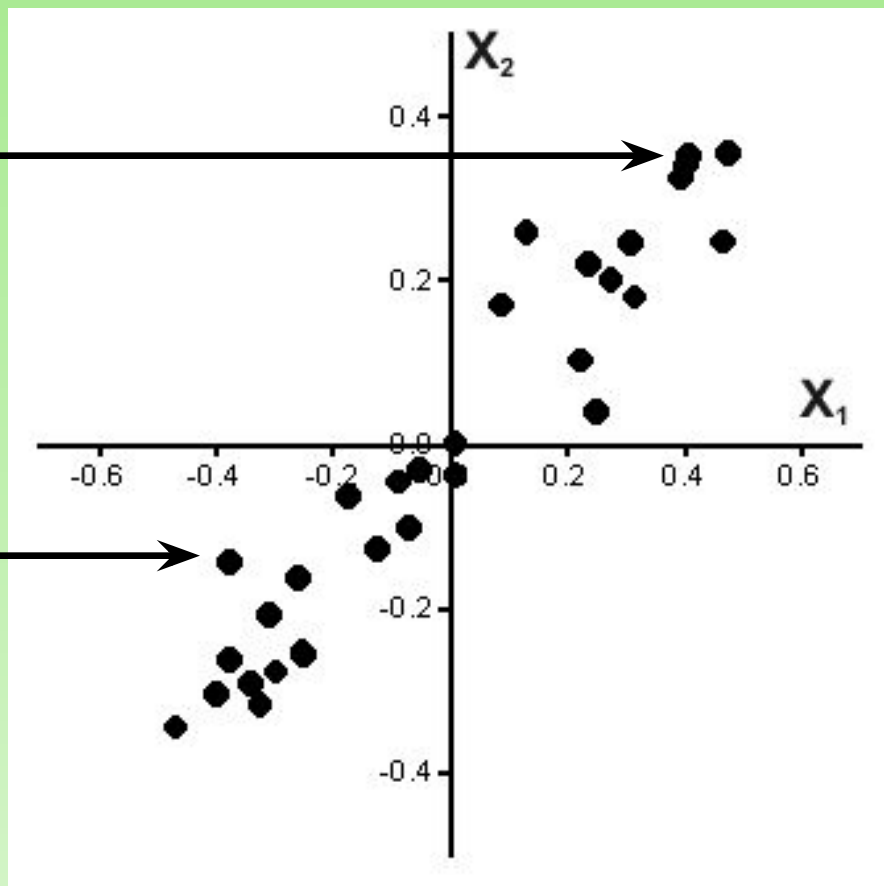
Пара чисел I и J называется размерностью матрицы и обозначается как $I \times J$. Примером матрицы в хемометрике может служить набор спектров, полученный для I образцов на J длинах волн.

1.2. Простейшие операции с матрицами

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

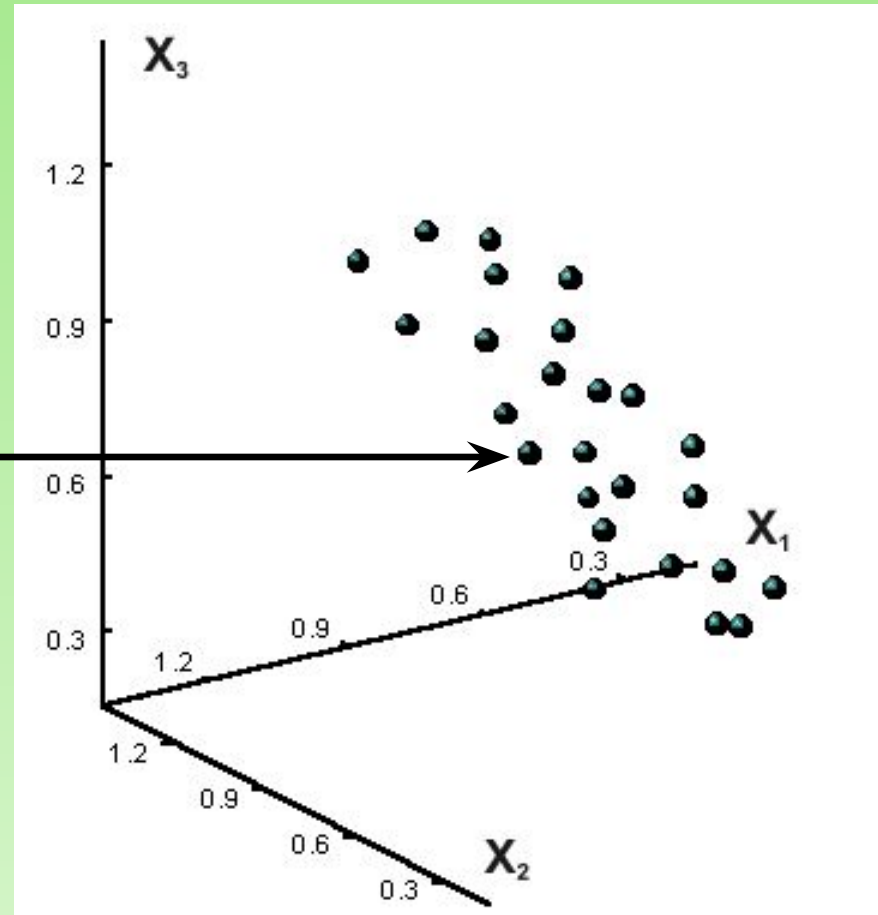
Графическое представление 2D-данных

	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

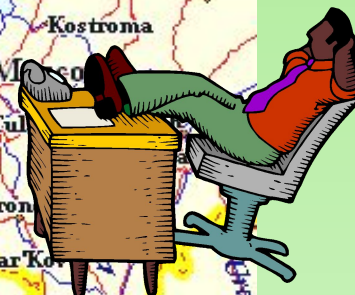


Графическое представление 3D-данных

	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



13.03.07

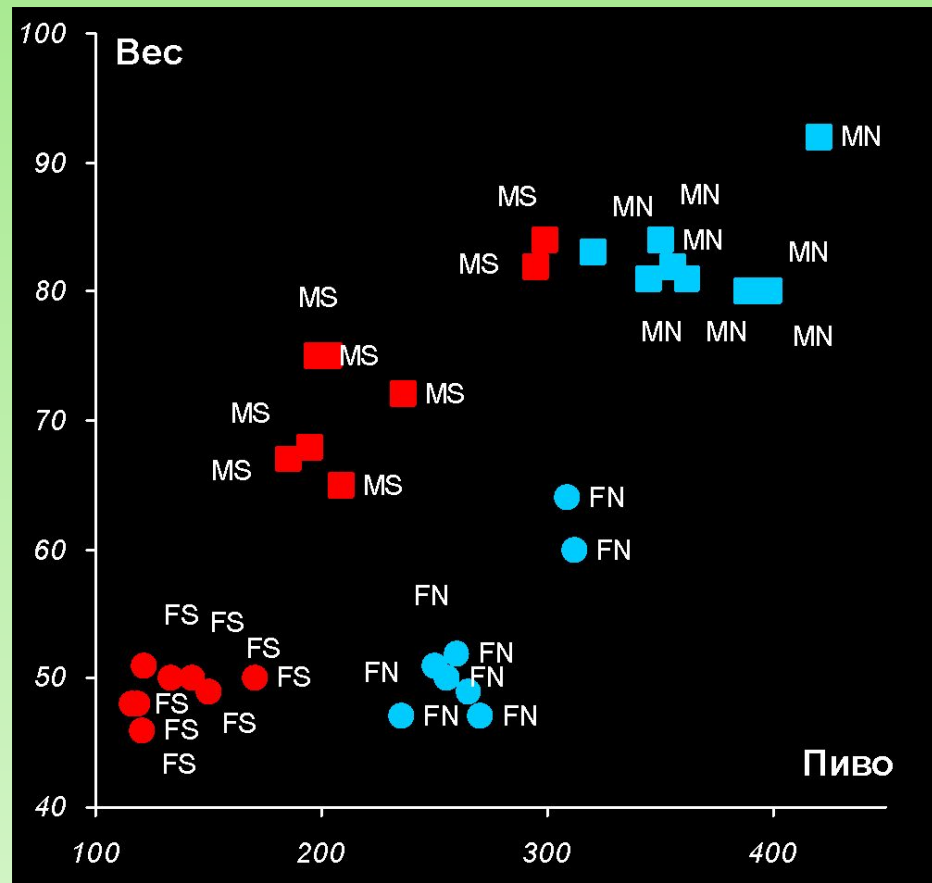
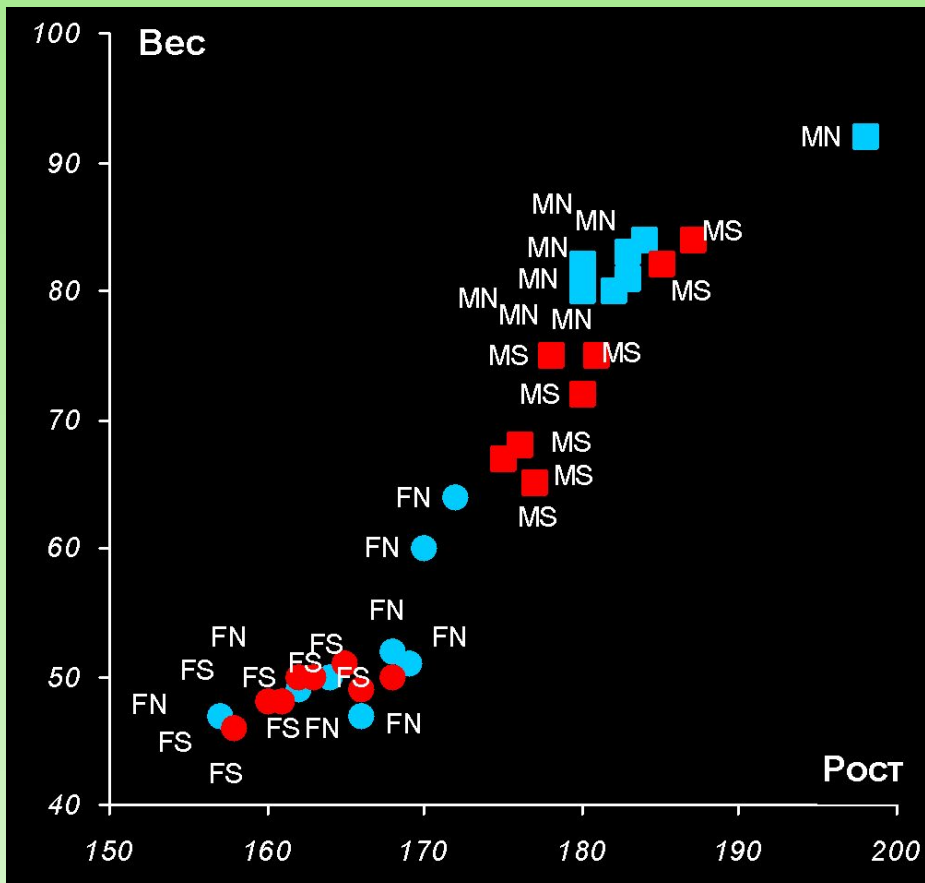
Лекция в МГУ

13

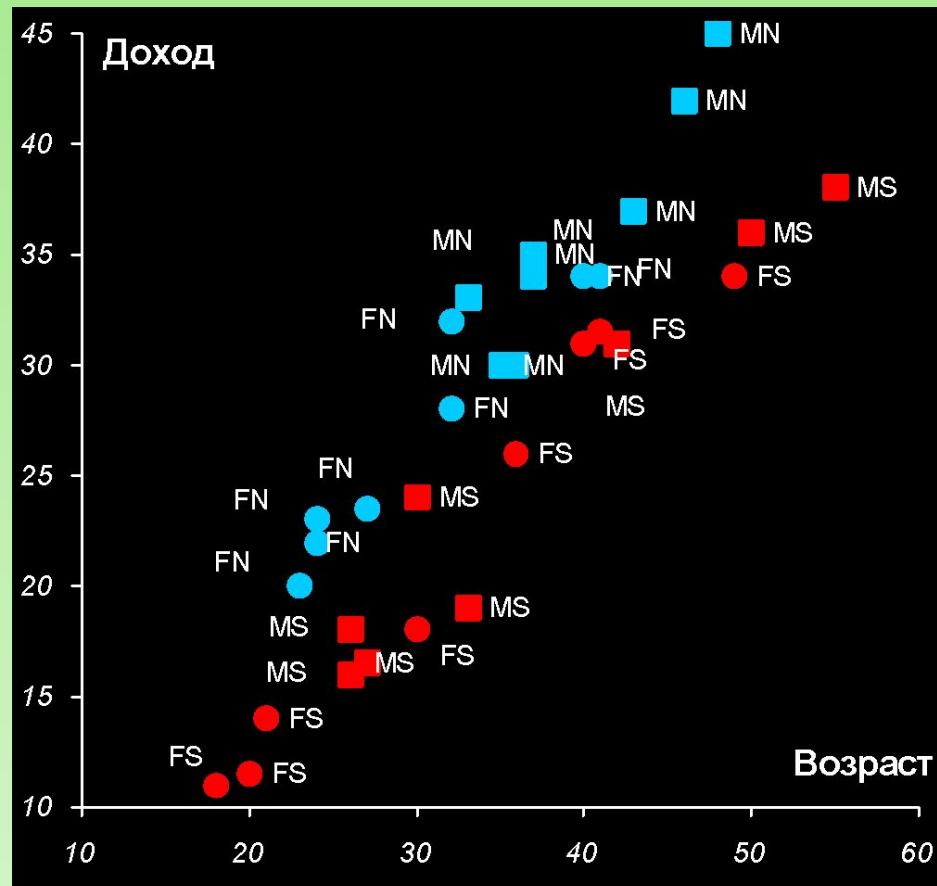
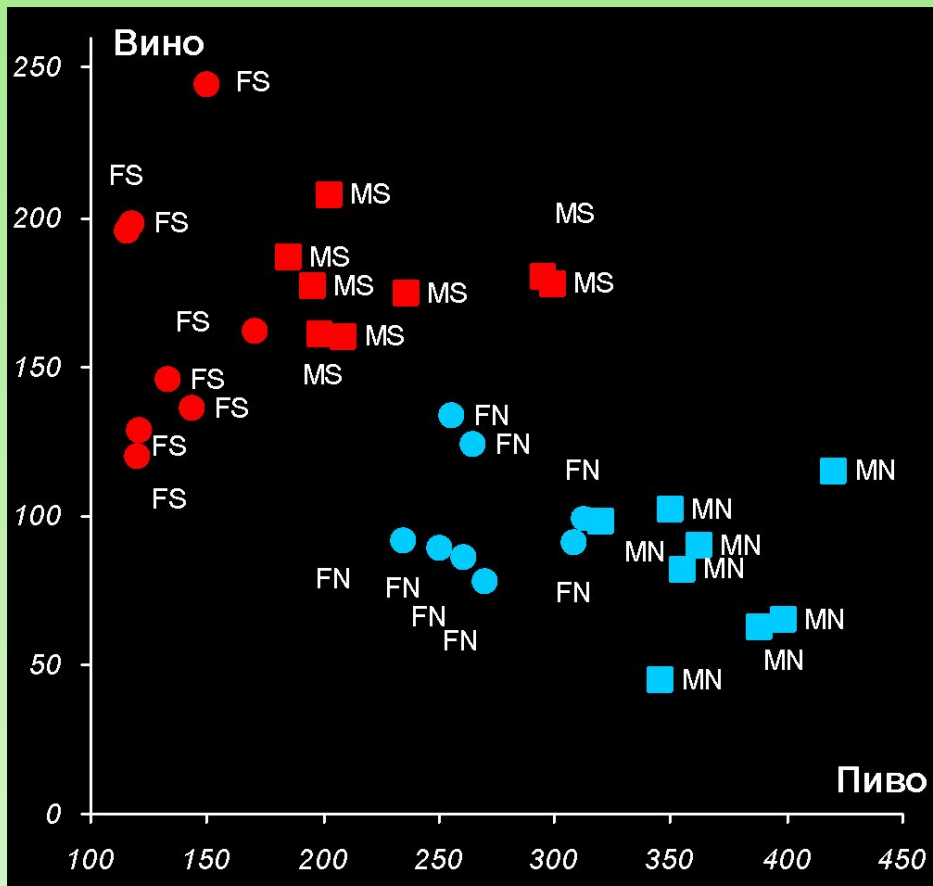
Таблица данных

Персона №	Пол M/F	Регион S/N	Рост см	Вес кг	Волосы S/L	Обувь размер	Возраст лет	Доход К€/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

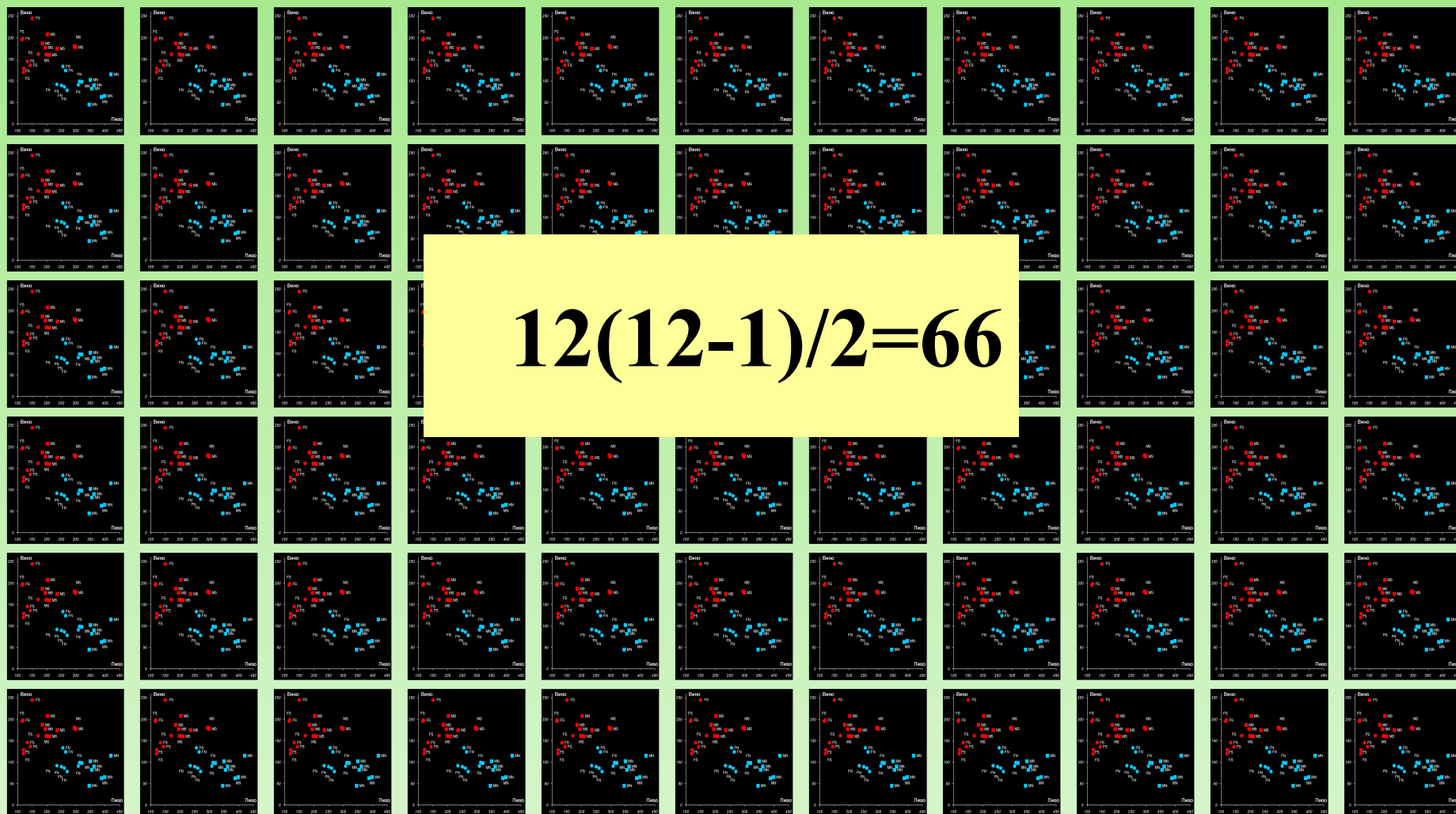
Корреляции 1



Корреляции 2



Все возможные корреляции



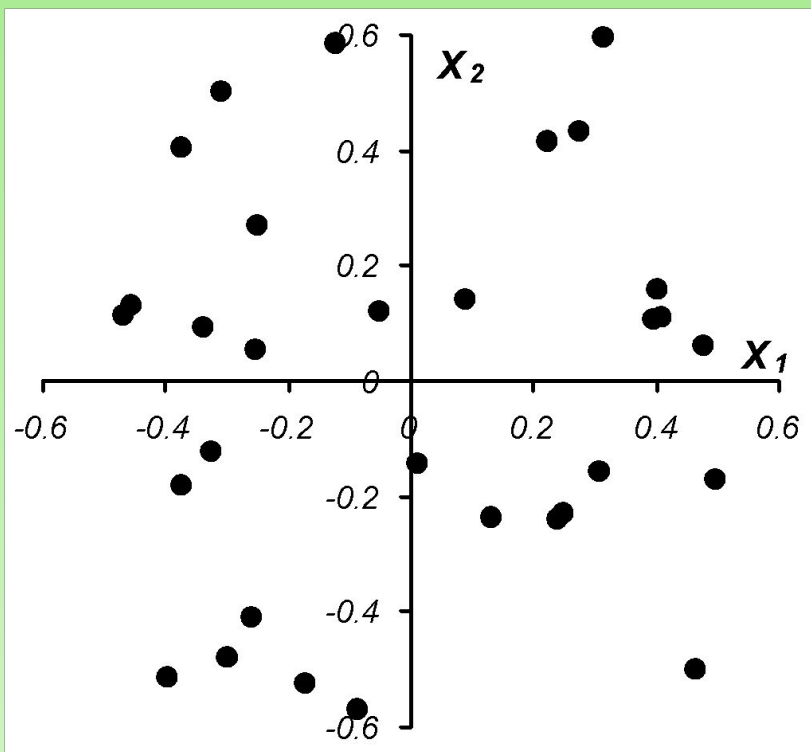
13.03.07

Лекция в МГУ

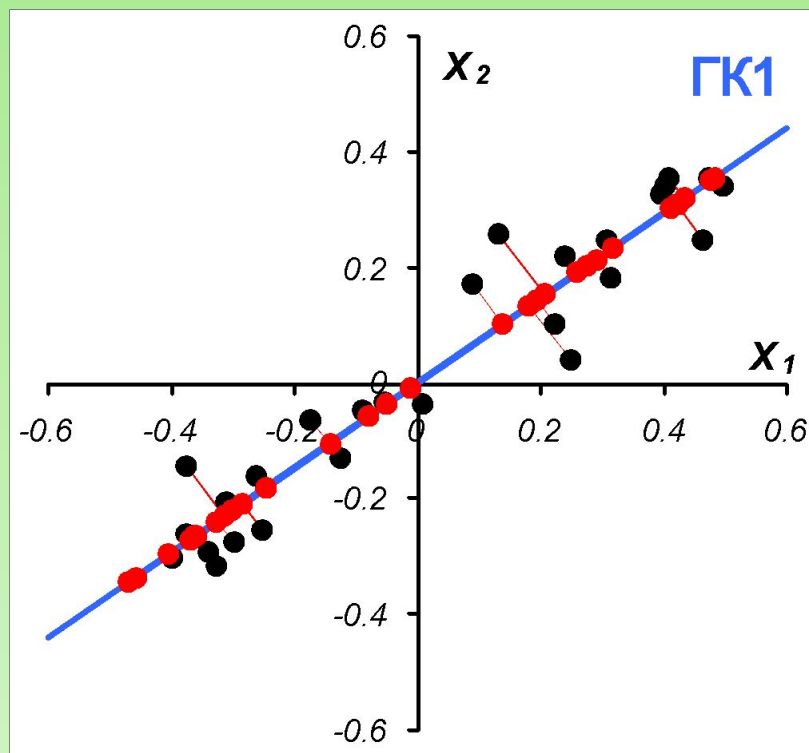
17

Главные направления и проекции

Данные без структуры



Данные со скрытой структурой



$$X_2 = aX_1 + E$$

Проекция на подпространство

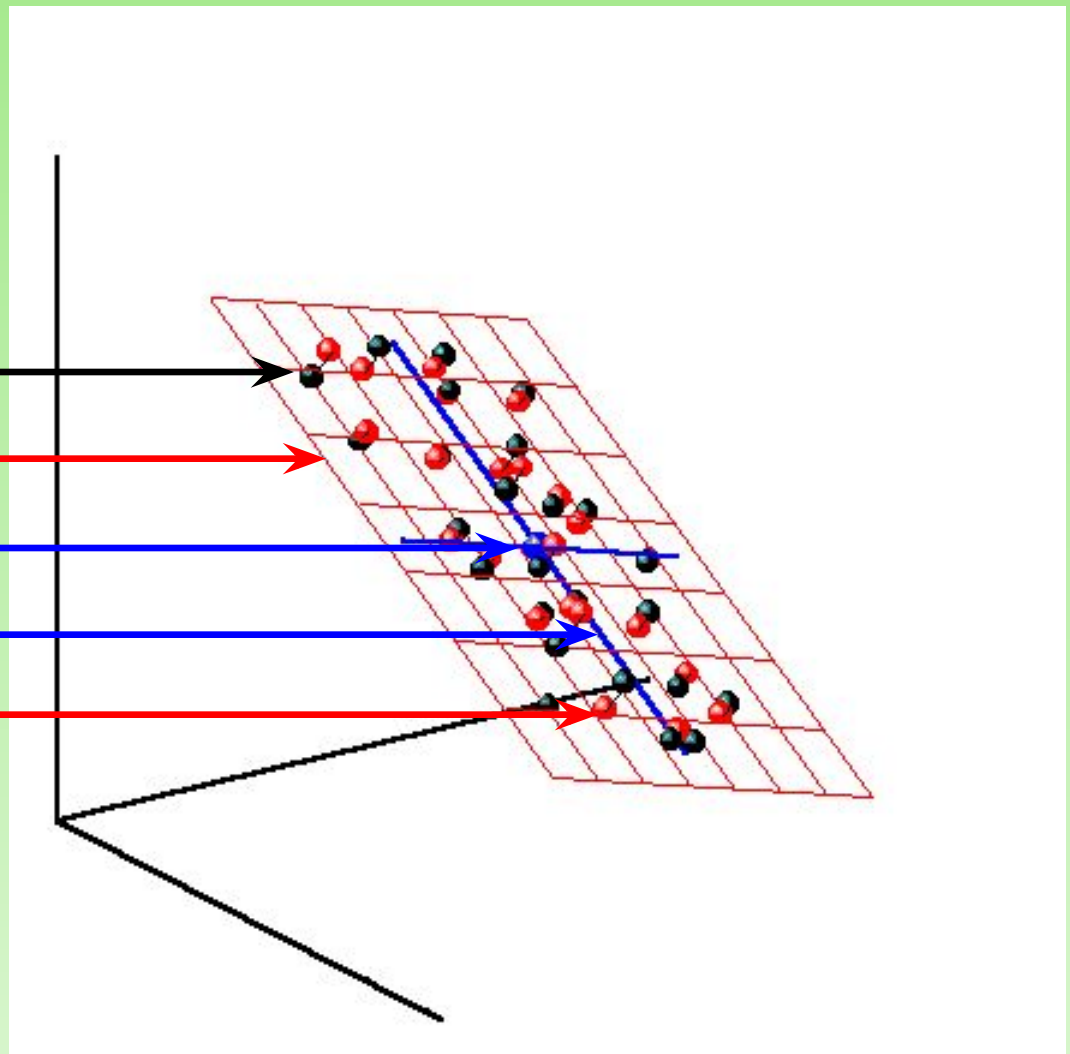
Исходные данные

Подпространство

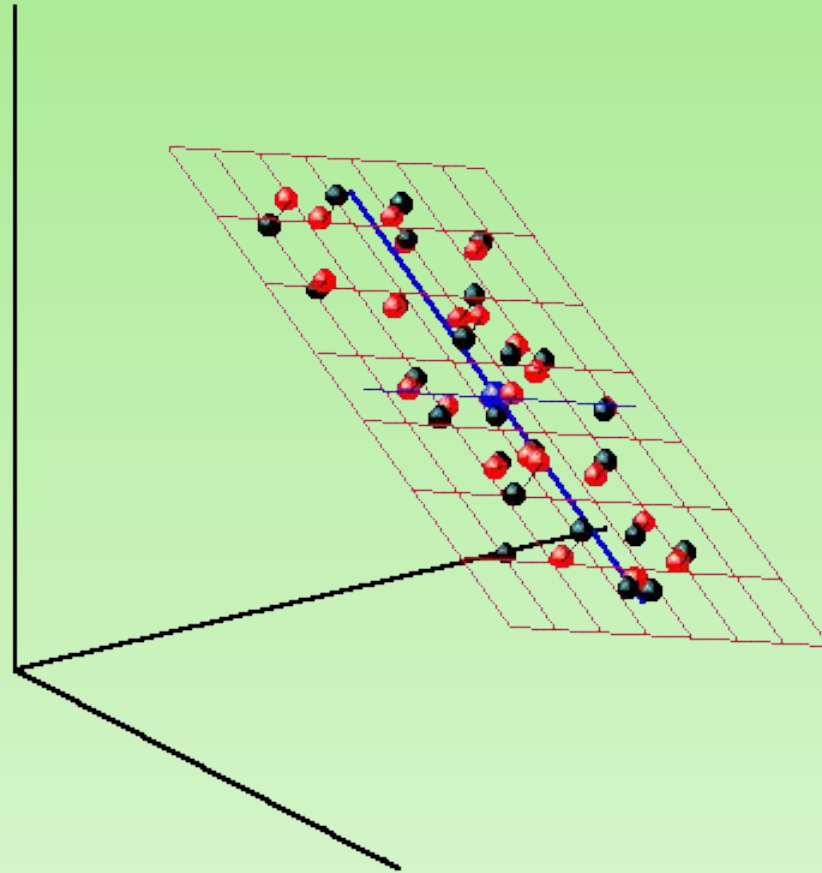
Центр данных

Главные компоненты

Проекции данных



Представление данных в подпространстве



Метод главных компонент: МГК (РСА)

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1 + \dots + \mathbf{t}_A \mathbf{p}_A + \mathbf{E}$$

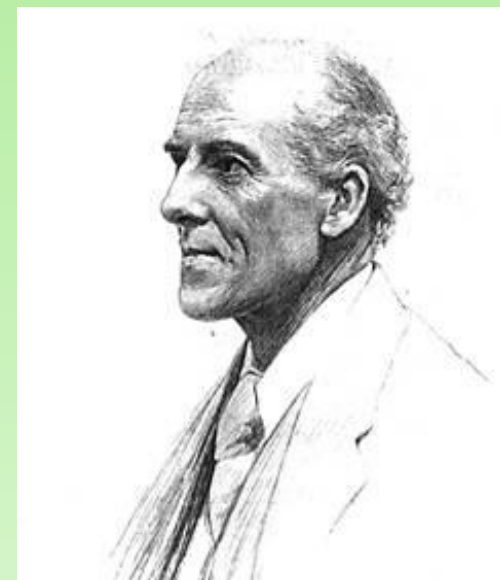
\mathbf{X} - матрица данных,

\mathbf{E} - матрица погрешностей, обе $(I \times J)$

\mathbf{T} - матрица счетов: $(I \times A)$,

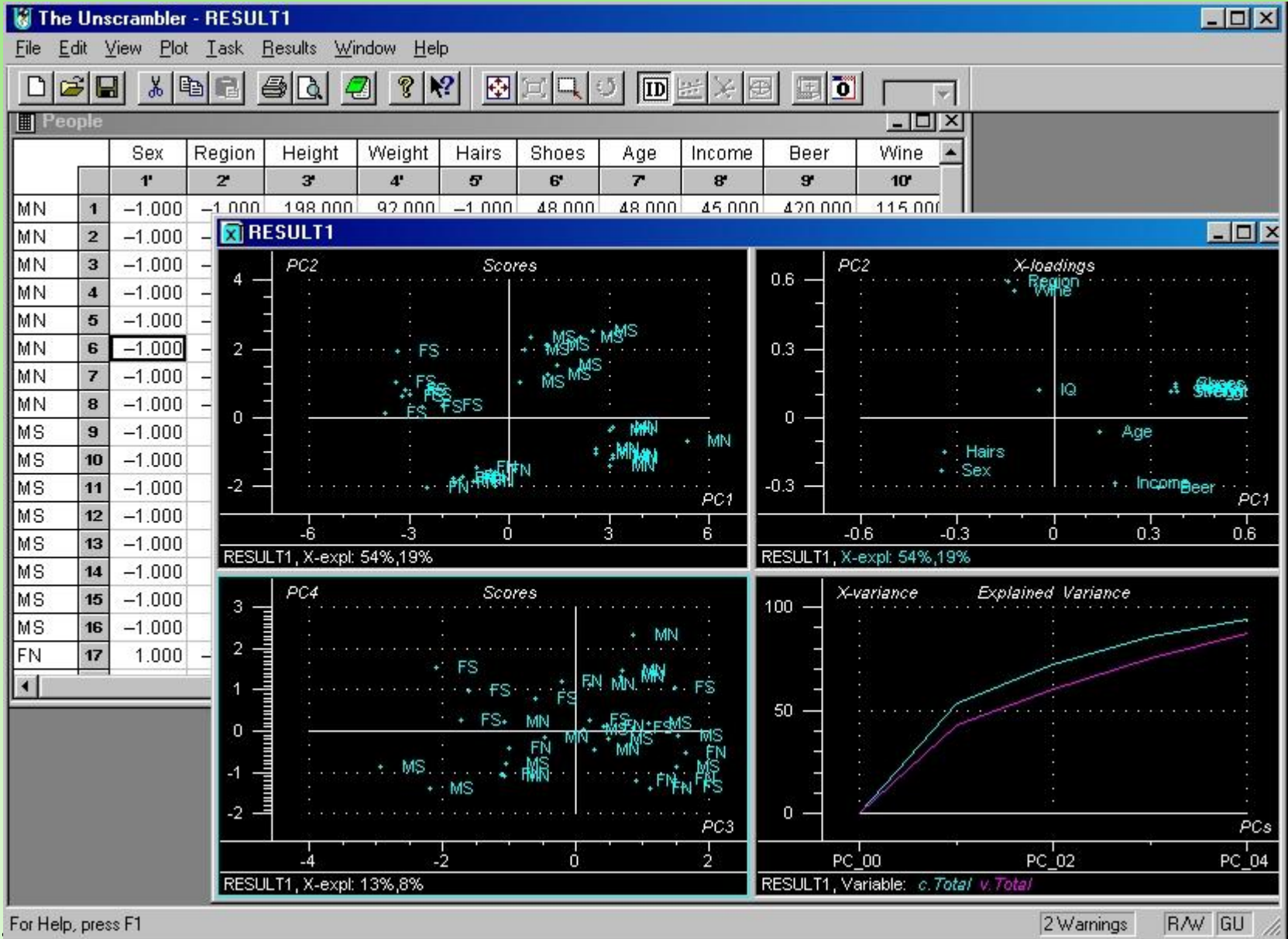
\mathbf{P} - матрица нагрузок: $(A \times J)$

A - число главных компонент ($A \ll J$)

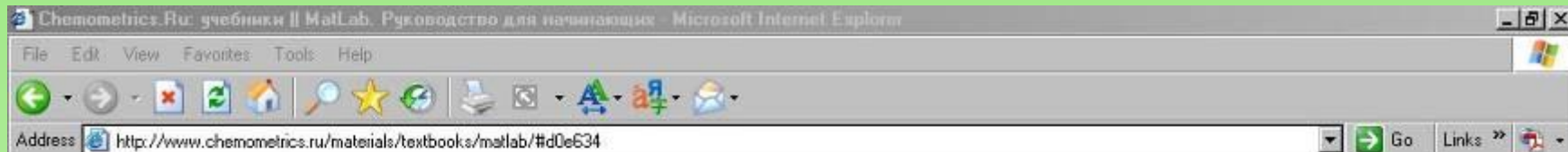


Karl Pearson, 1901

Unscrambler® by CAMO ASA



Matlab



www.chemometrics.ru | учебники

MatLab. Руководство для начинающих

Евгений Михайлов, Алексей Померанцев

© Российское хемометрическое общество (<http://rcs.chph.ras.ru>)

[1. Введение](#)

[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](#)

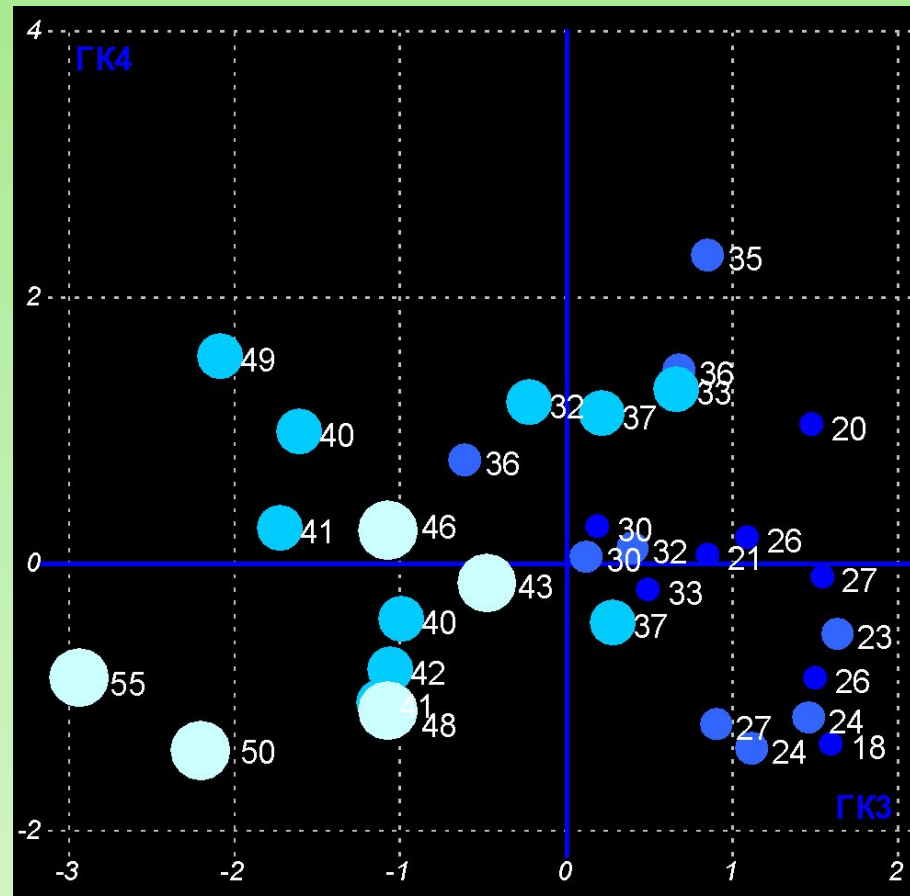
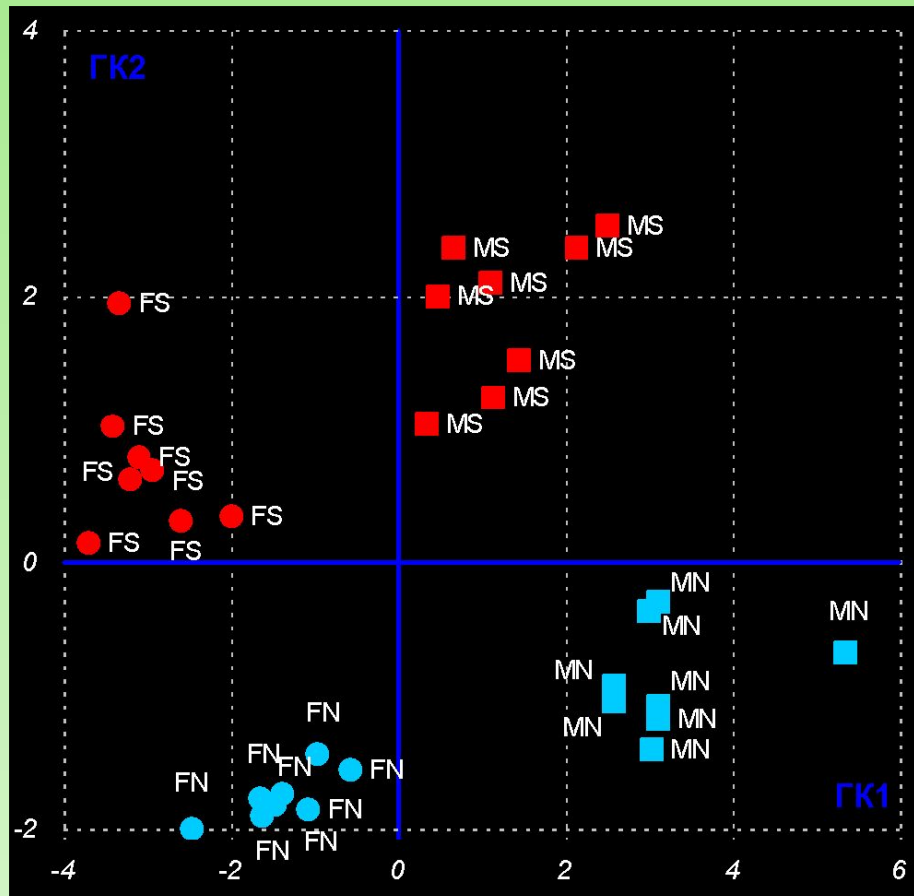
[4.1.](#)

[4.2.](#)

[4.3.](#)

[4.4. Создание графика](#)

Люди: Графики счетов

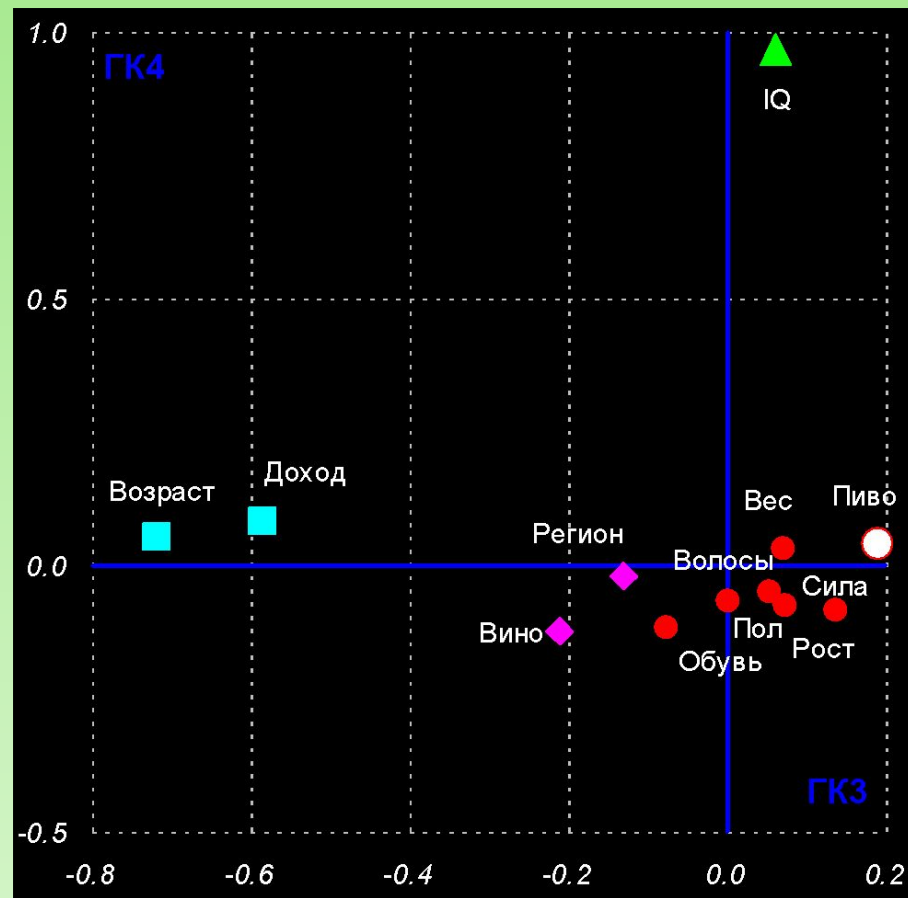
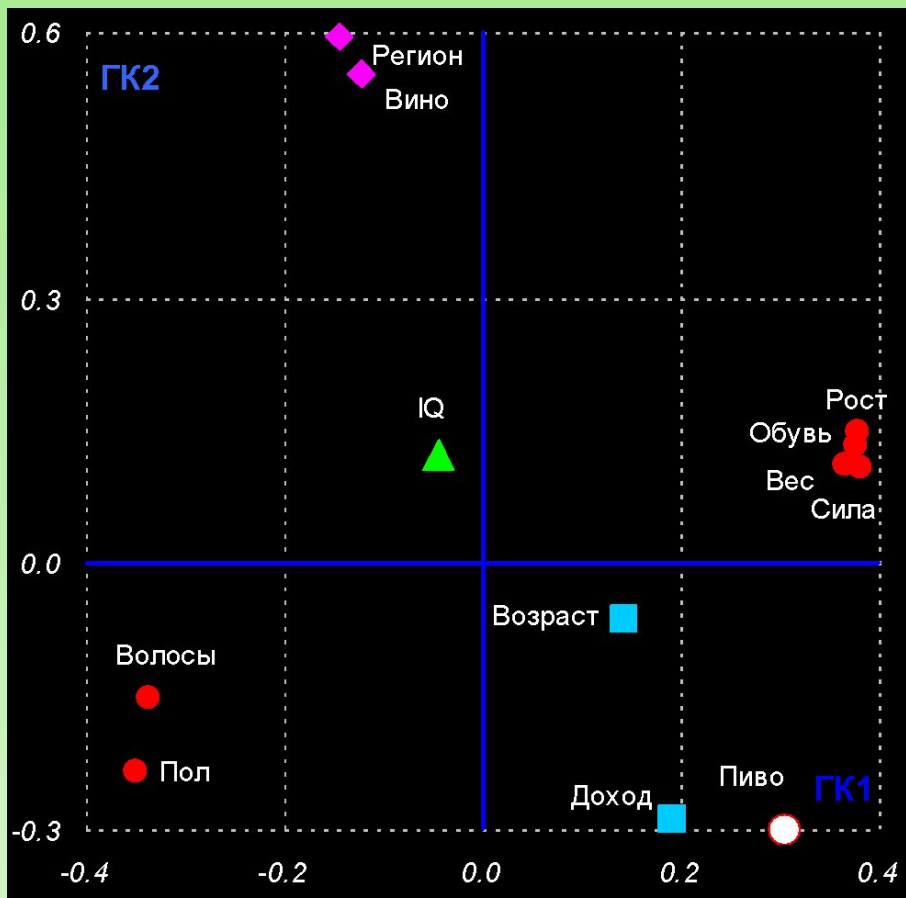


13.03.07

Лекция в МГУ

24

Люди: Графики нагрузок

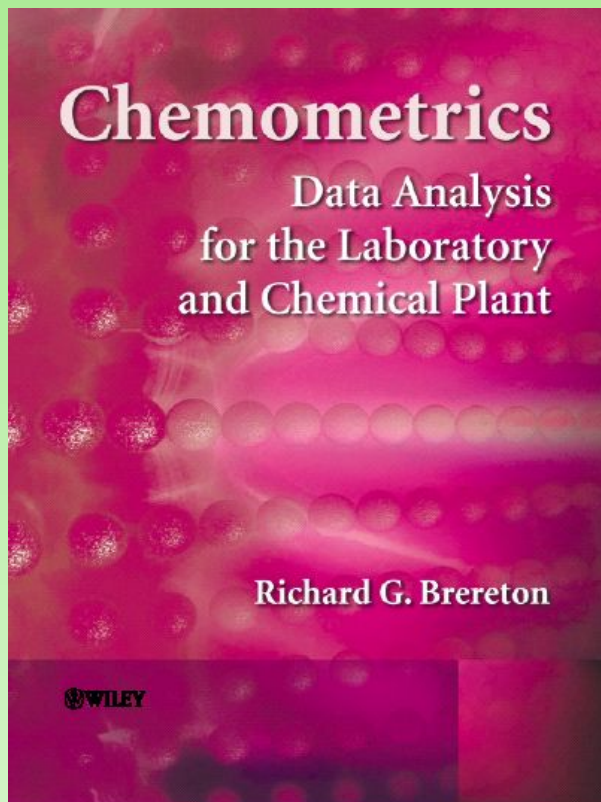


13.03.07

Лекция в МГУ

25

Разделение пиков в ВЭЖХ-ДАМ

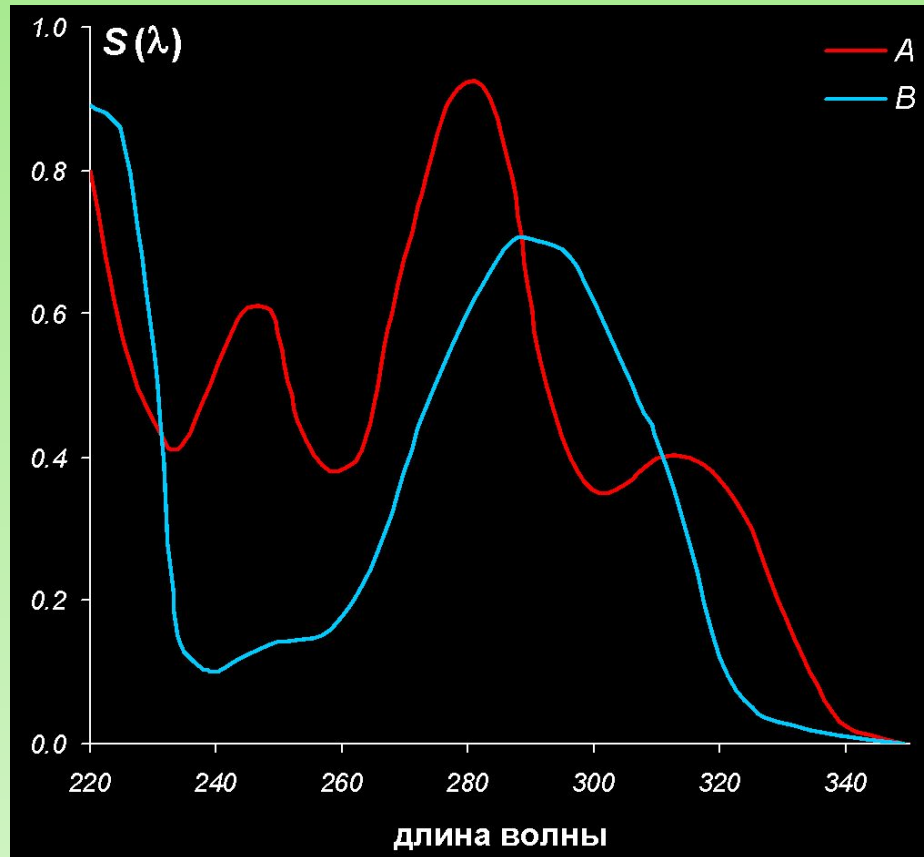
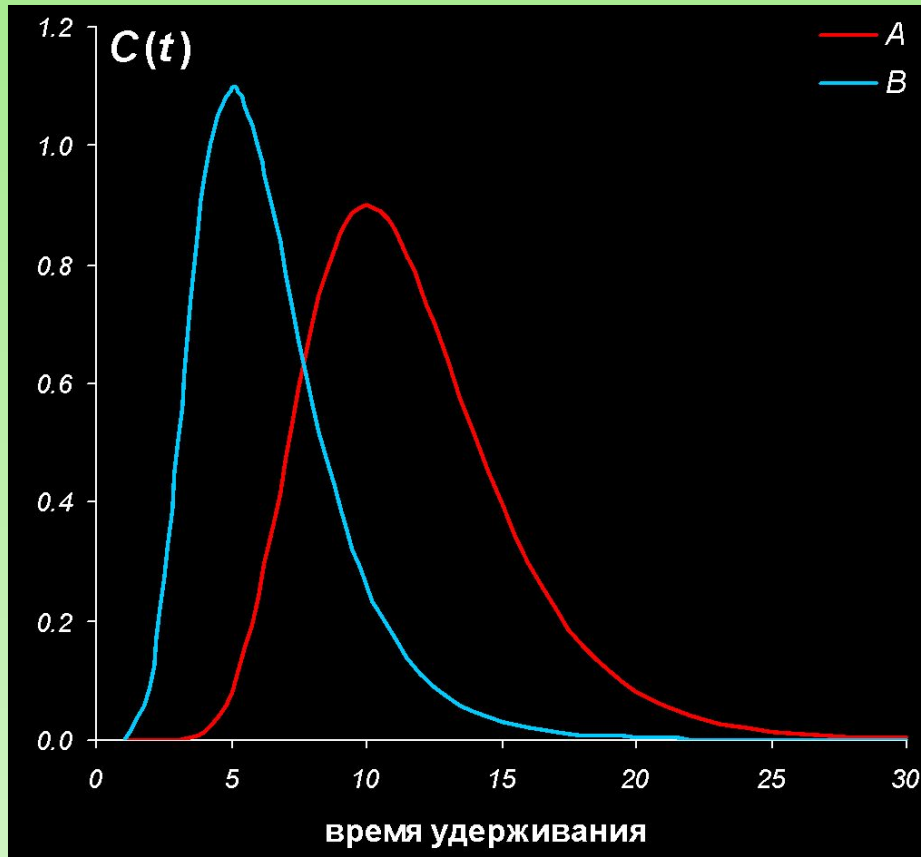


13.03.07

Лекция в МГУ

26

Свойства чистых веществ А и В



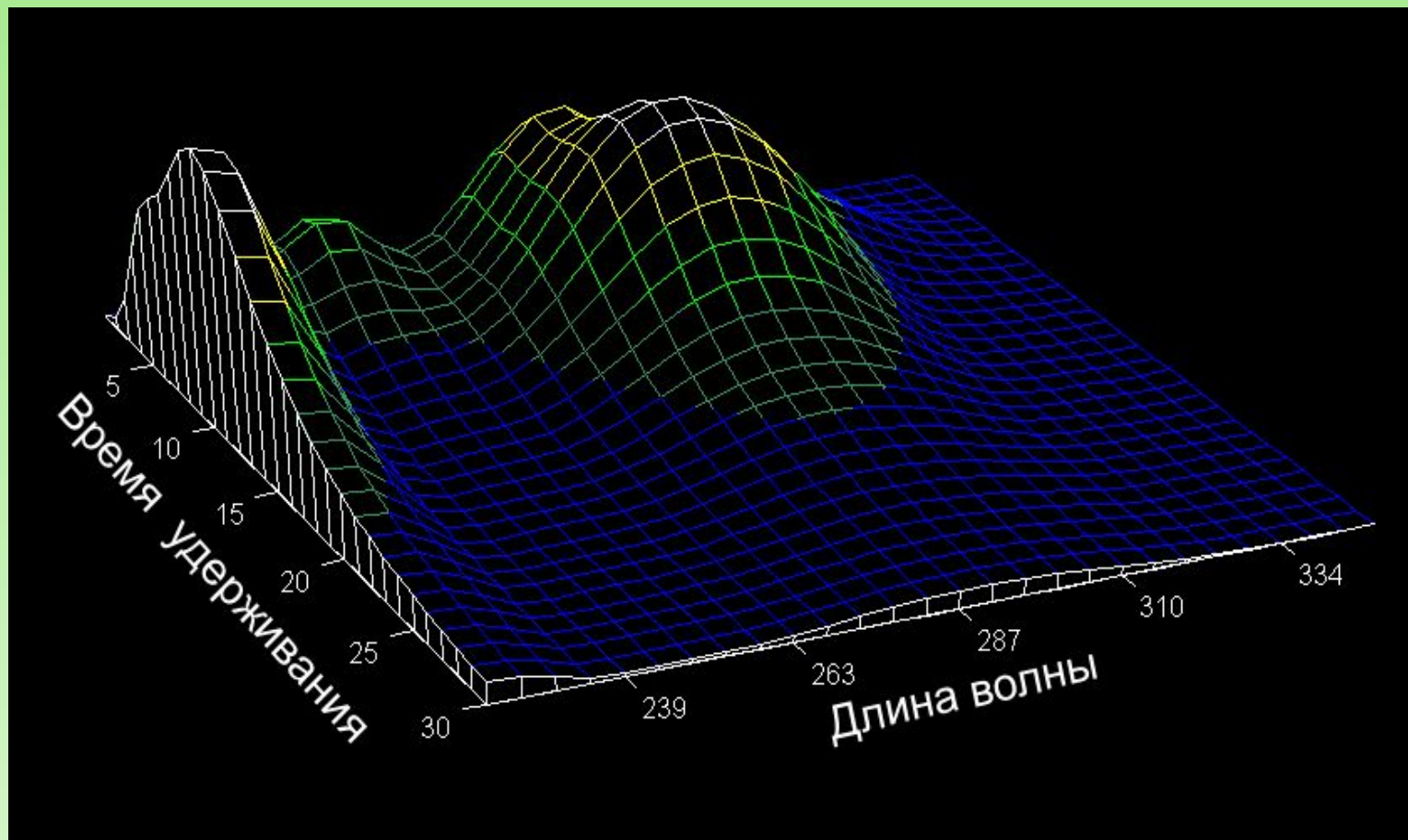
$$X = CS^T + E$$

13.03.07

Лекция в МГУ

27

Хроматограмма смеси



Данные в Unscrambler

The Unscrambler - [HPLC-DAD]

File Edit View Plot Modify Task Results Window Help

		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	0.001
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	0.013
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	0.050
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	0.116
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	0.178
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	0.220
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	0.293
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	0.418
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	0.564
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	0.640
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	0.628
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	0.559
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	0.471
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	0.384
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	0.308
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	0.245
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	0.193
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	0.261	0.267	0.256	0.231	0.152
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	0.118
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	0.091
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	0.062

For Help, press F1

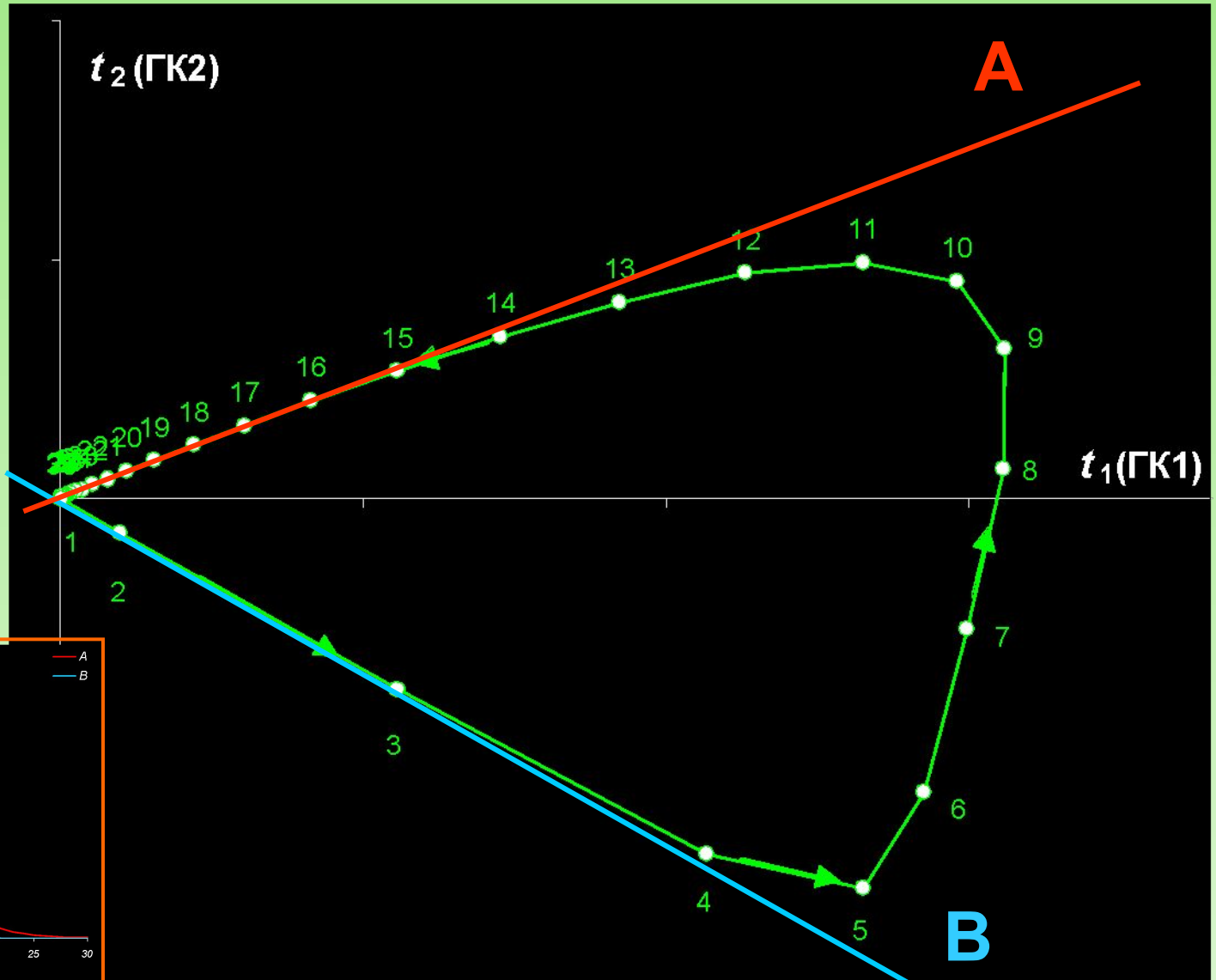
Value: 0.095 Size: 30 x 29 R/W GU

13.03.07

Лекция в МГУ

29

График счетов



13.03.07

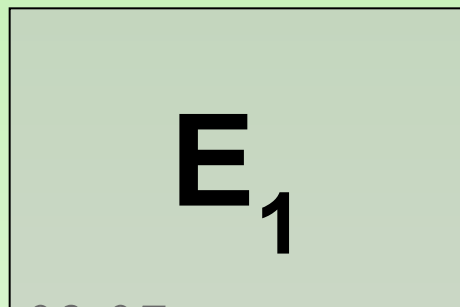
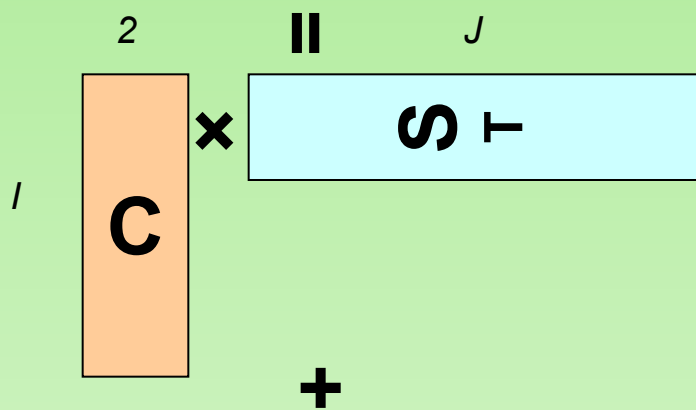
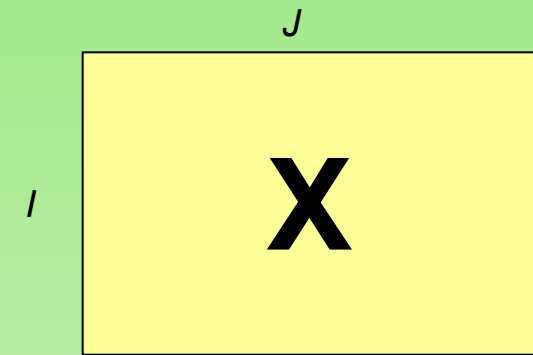
Лекция в МГУ

30

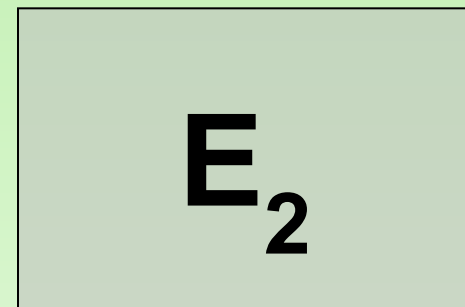
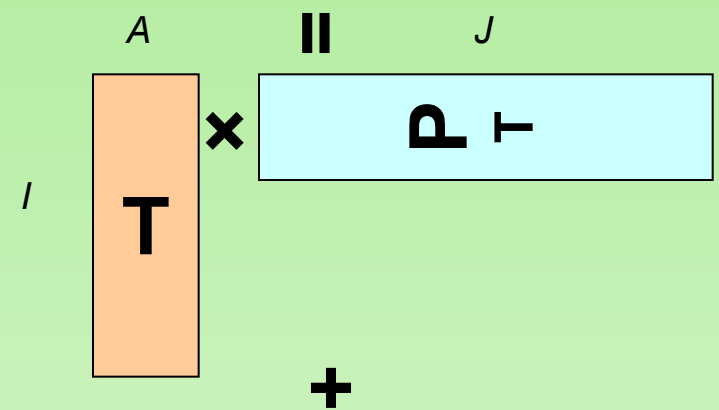
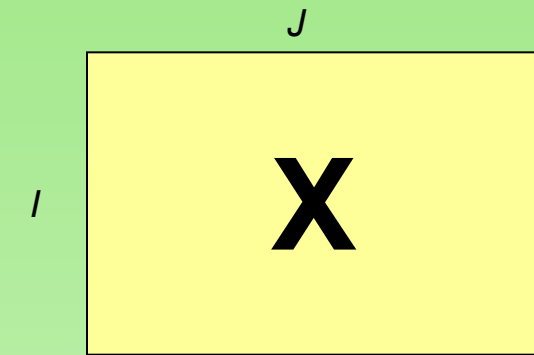
Выводы из графика счетов

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

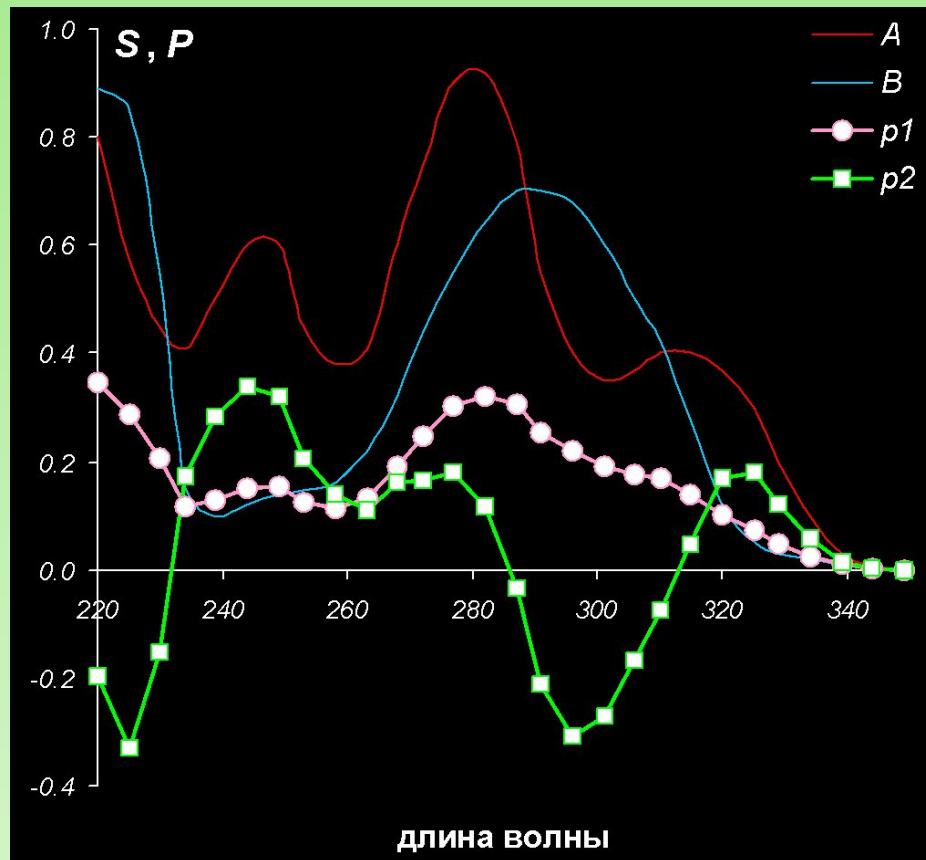
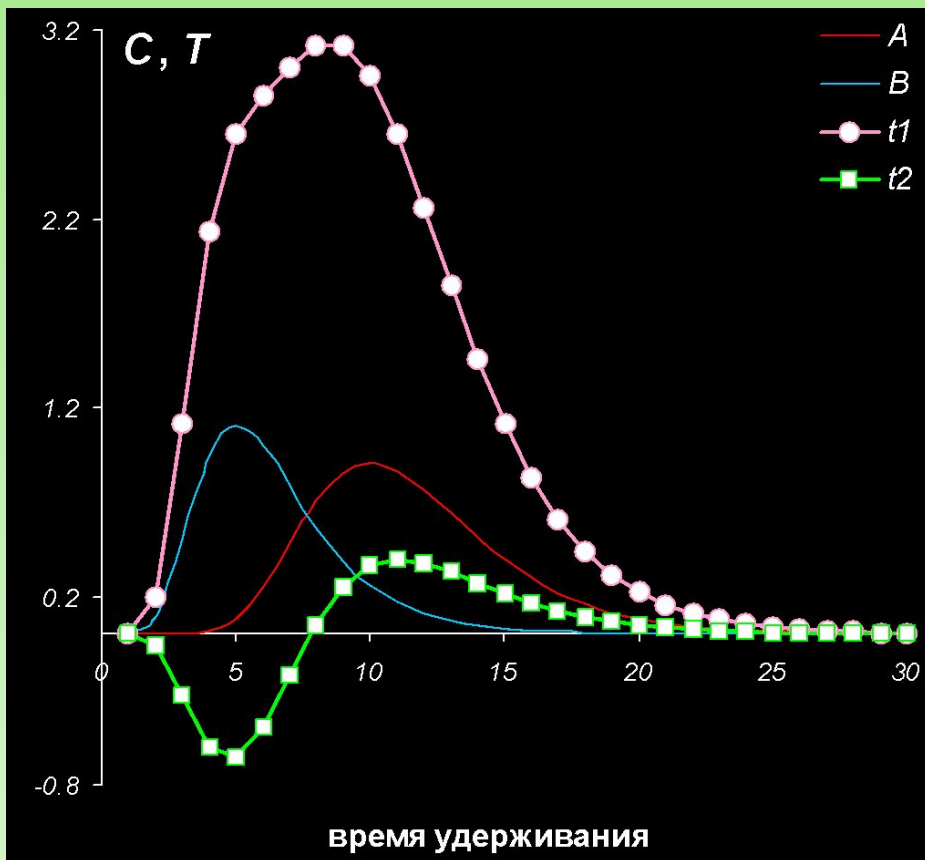
Факторный анализ и анализ ГК



13.03.07



Счета и нагрузки



Прокрустово преобразование



$$X \approx CS^T$$

$$X \approx TP^T$$

$$I = RR^T = \text{единичная}$$

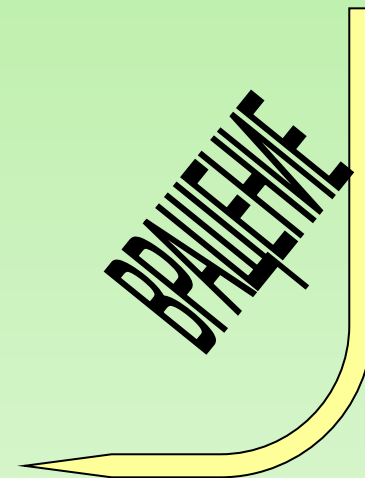
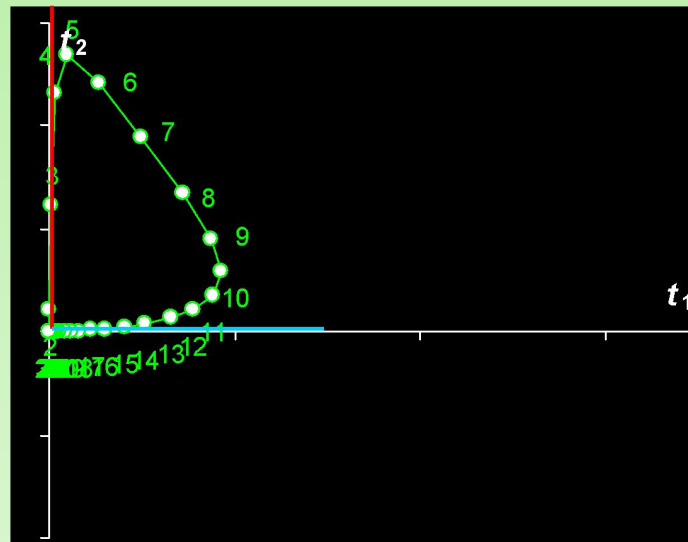
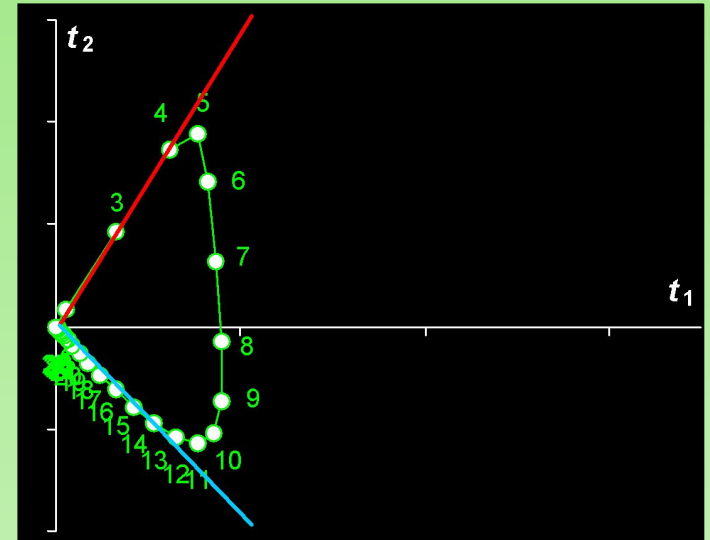
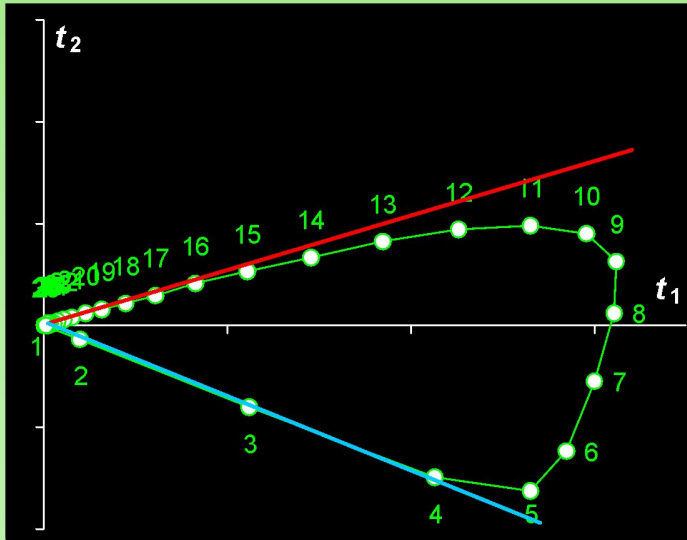
$$X \approx T(RR^T)P^T = (TR)(PR)^T$$

$$C \approx TR$$

$$S \approx PR$$

$$R = R_{\text{stretch}} \times R_{\text{rotation}}$$

Преобразование счетов

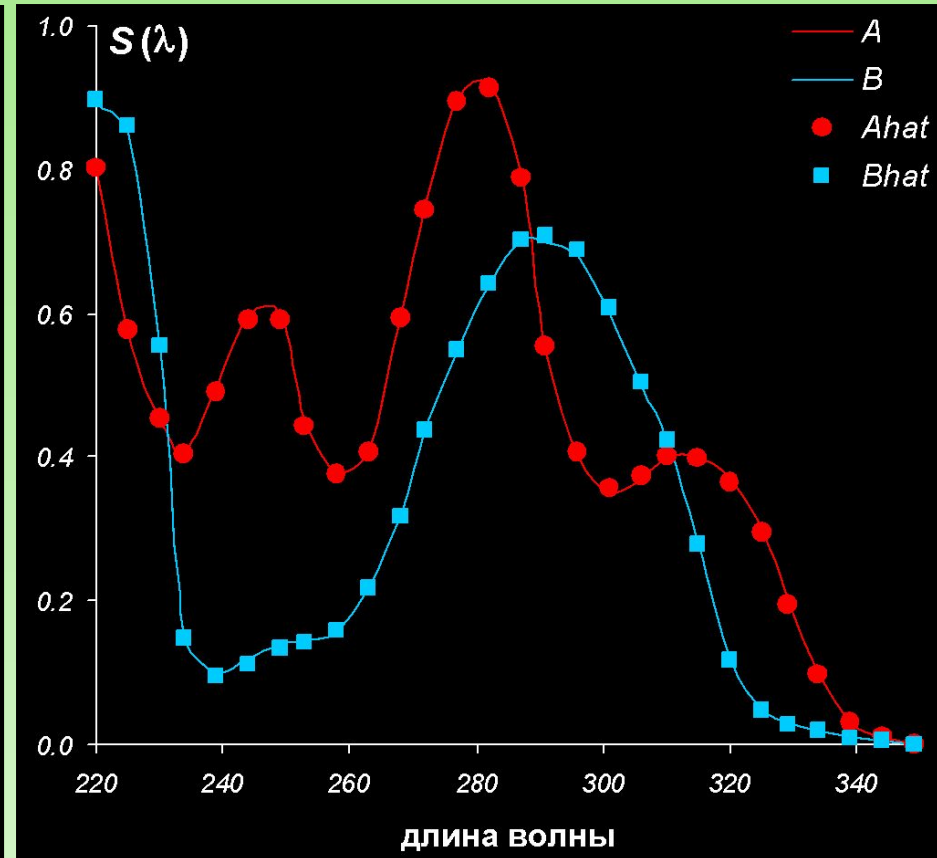
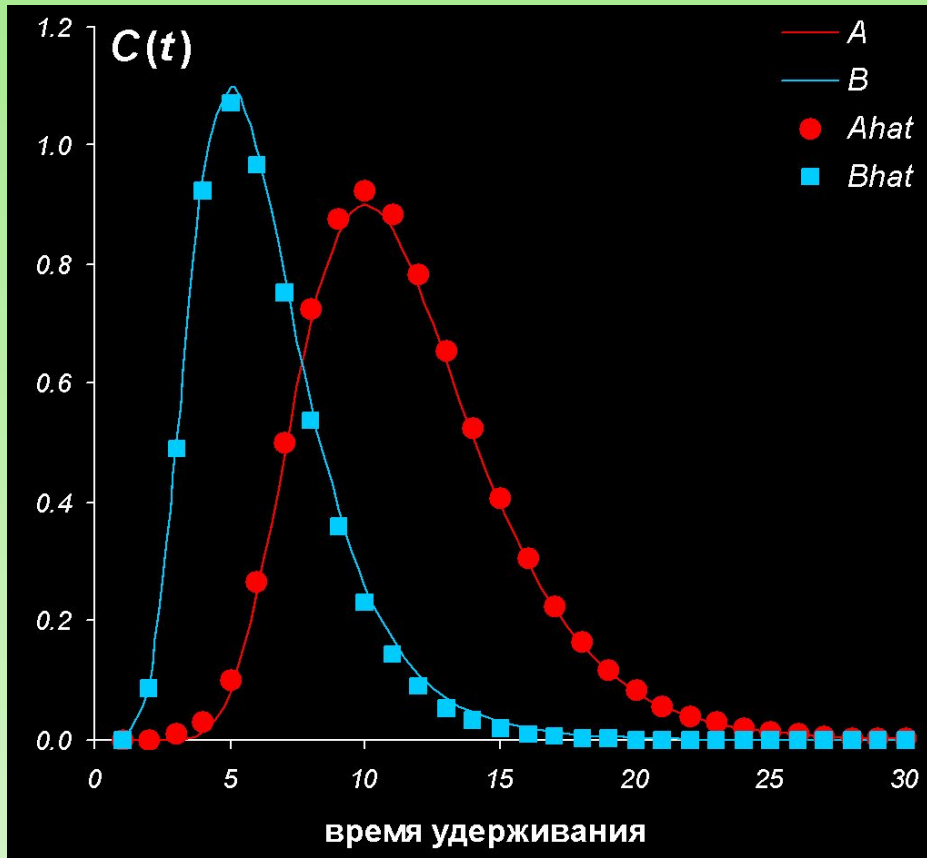


13.03.07

Лекция в МГУ

35

Результат Прокрустова преобразования



Классификация поддельных лекарств



13.03.07

Лекция в МГУ

37

Измерения: БИК спектроскопия



13.03.07

Лекция в МГУ

38


Образцы

Подлинники:
11 серий по 5 таблеток

Подделки:
4 серии по 5 таблеток


Всего
 $15 \times 5 = 75$ образцов


Обучающий набор

Подлинники: 
8 серий по 5 таблеток

Всего:
 $8 \times 5 = 40$ образцов

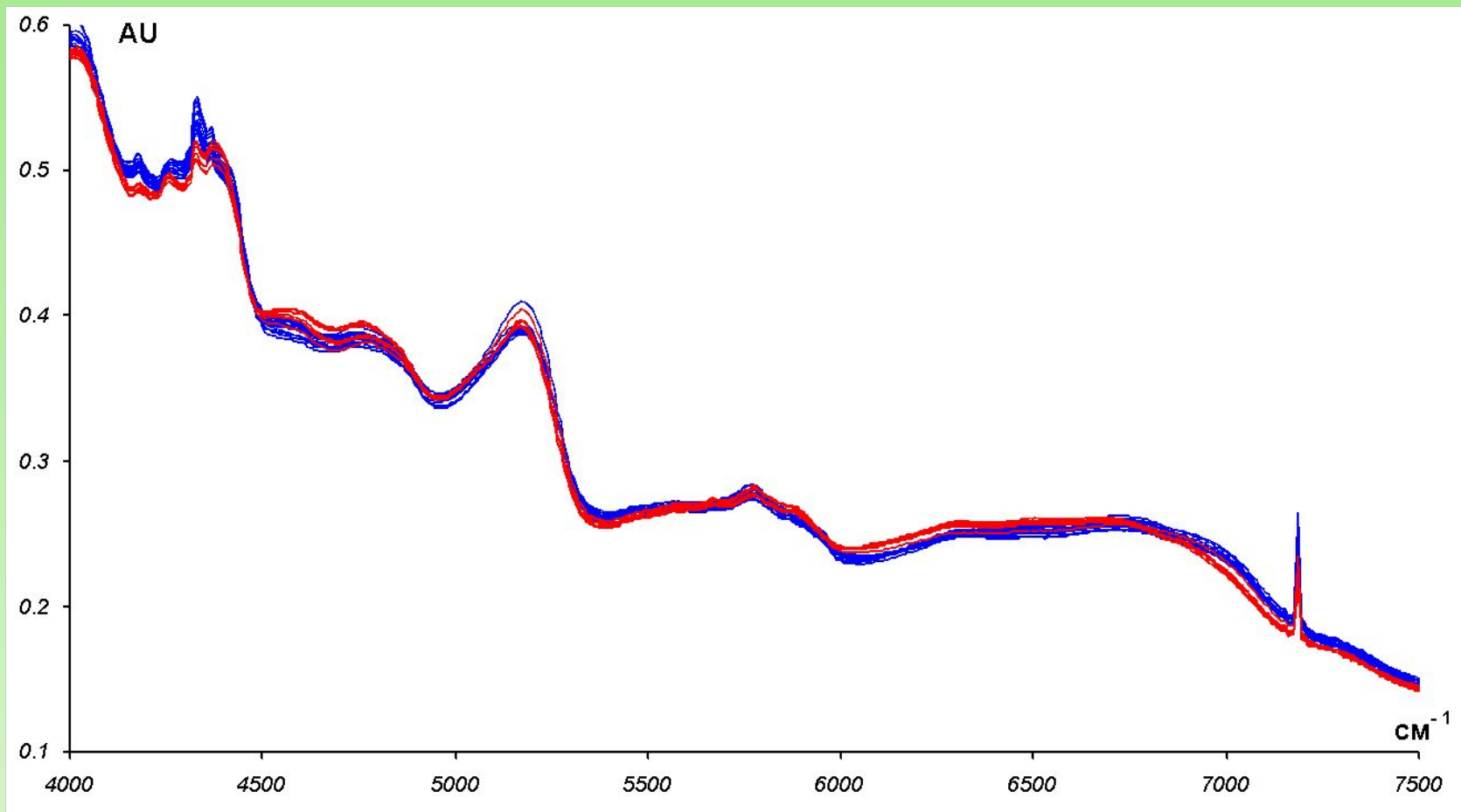
Проверочный набор

Подлинники: 
3 серии по 5 таблеток

Подделки: 
4 серии по 5 таблеток

Всего:
 $7 \times 5 = 35$ образцов

БИК спектры мезима



Настоящие Фальшивые

13.03.07

Лекция в МГУ

40

Данные: БИК спектры

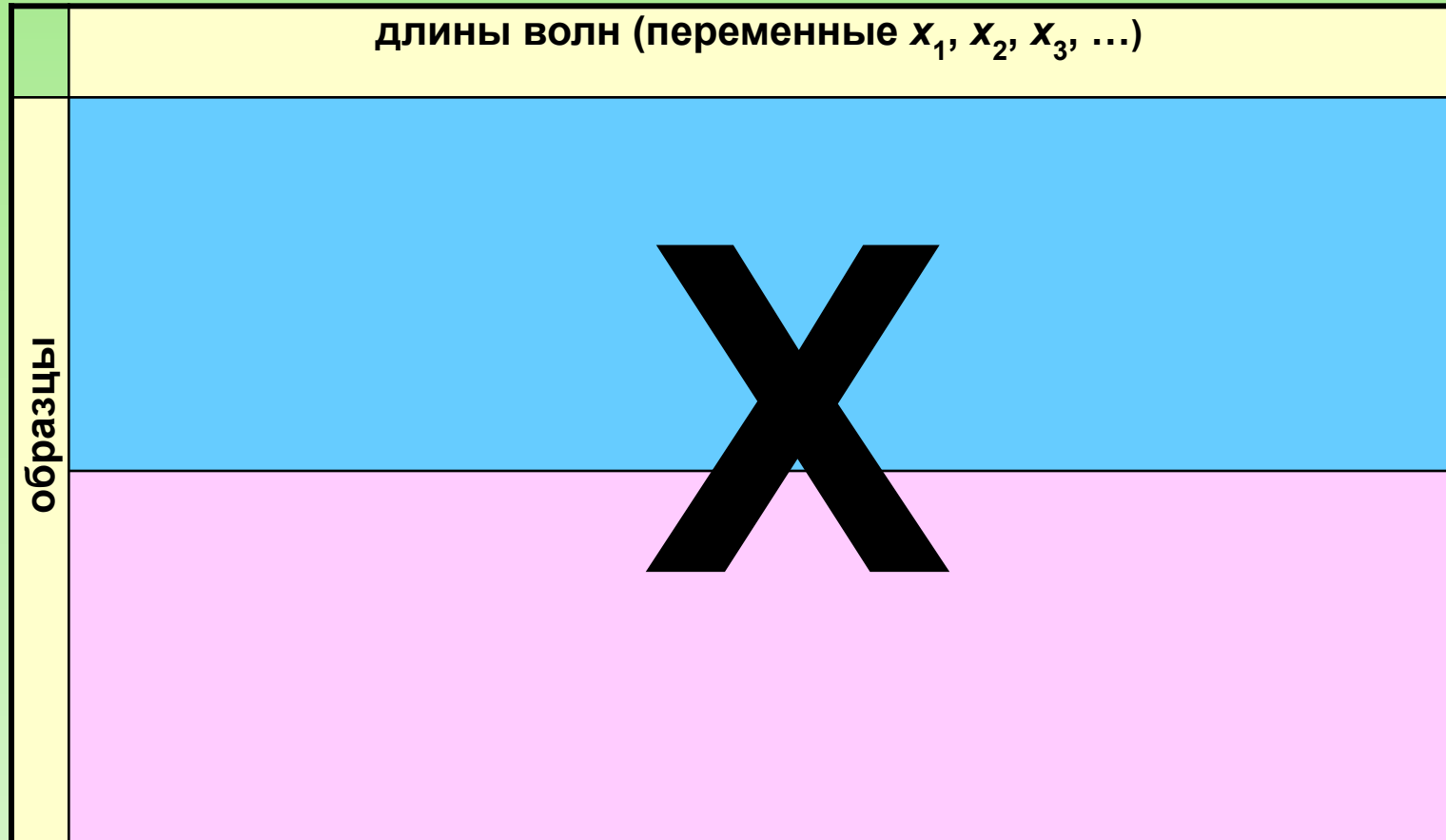
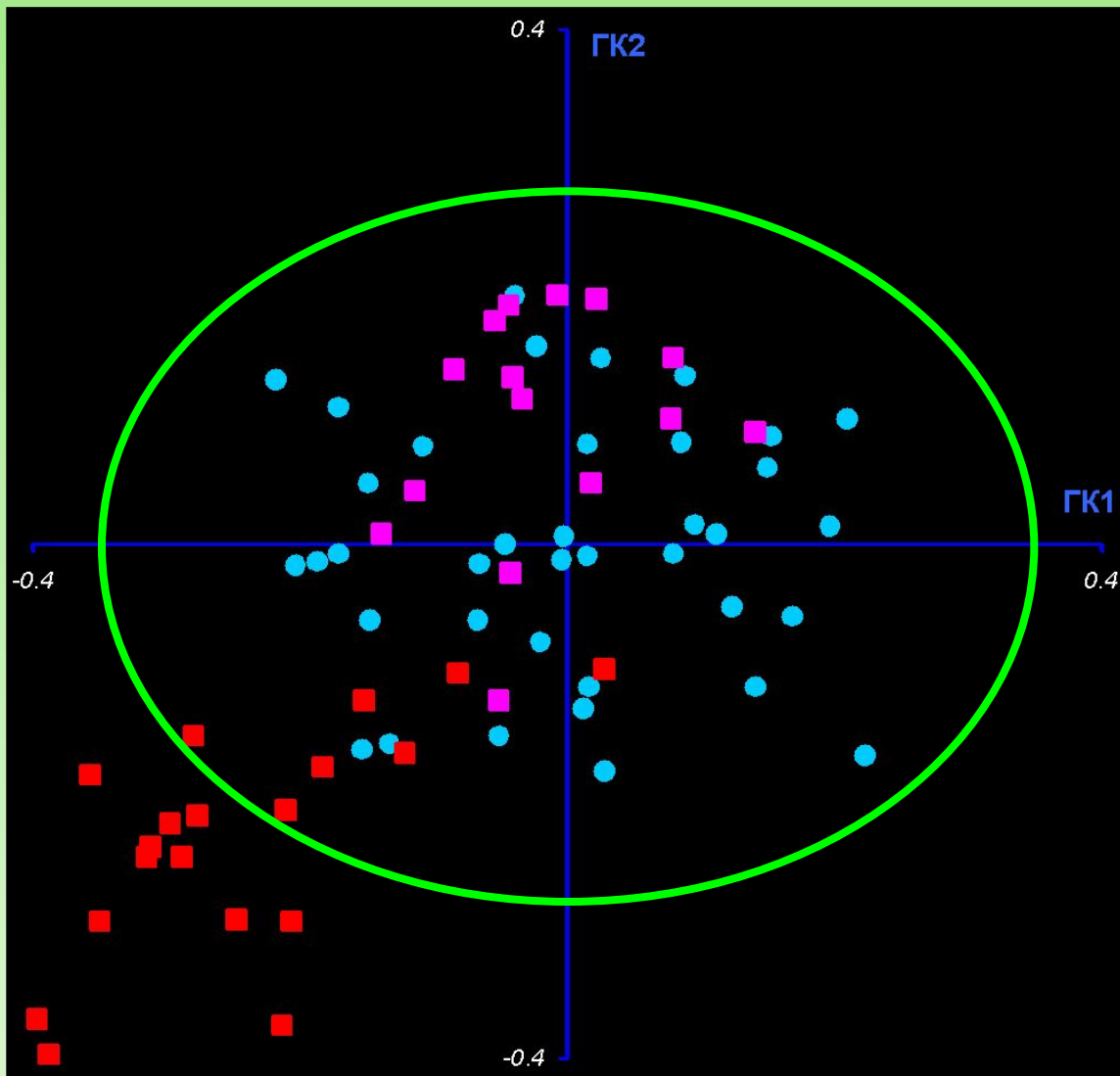


График МГК счетов



Обучающий набор

Подлинники: ●
8 серий по 5 таблеток

Всего:
 $8 \times 5 = 40$ образцов

Проверочный набор

Подлинники: ■
3 серии по 5 таблеток

Подделки: ■
4 серии по 5 таблеток

Всего:
 $7 \times 5 = 35$ образцов

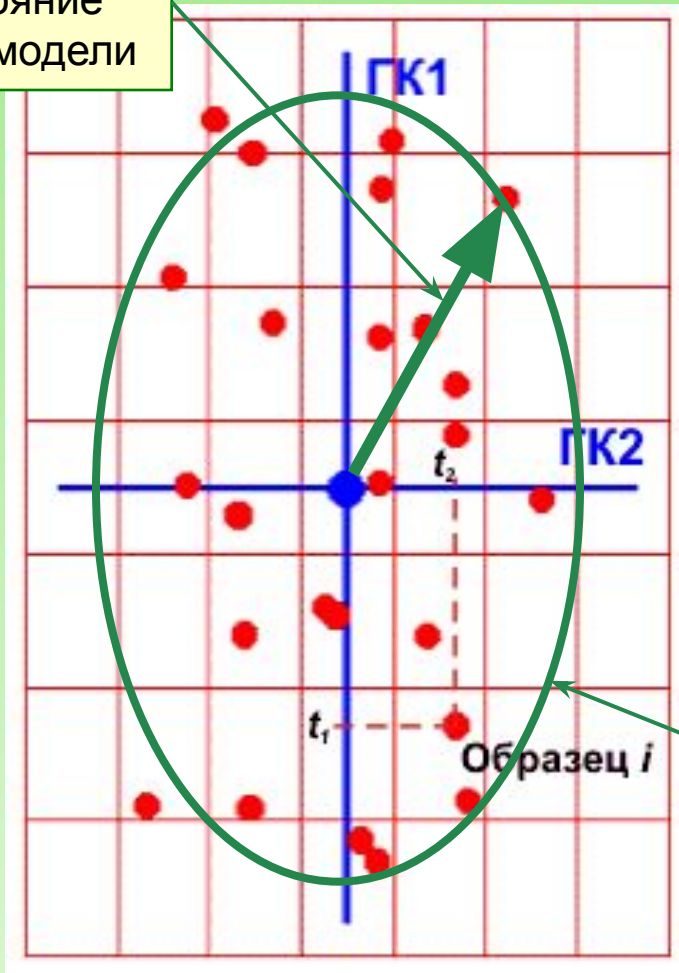
13.03.07

Лекция в МГУ

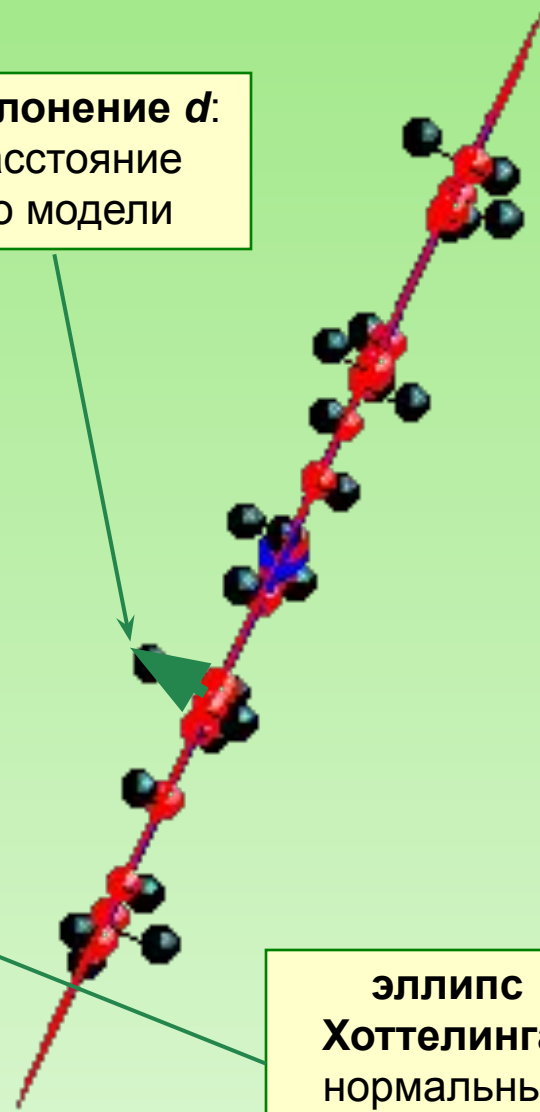
42

Размах и отклонение

Размах h :
расстояние
внутри модели



Отклонение d :
расстояние
до модели



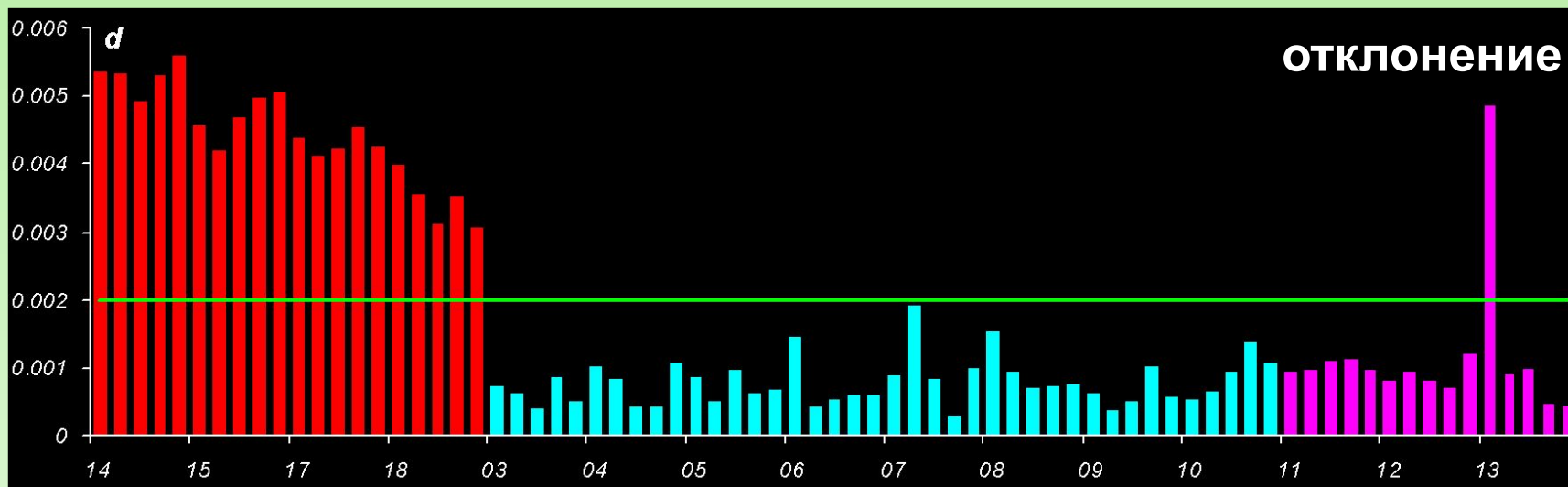
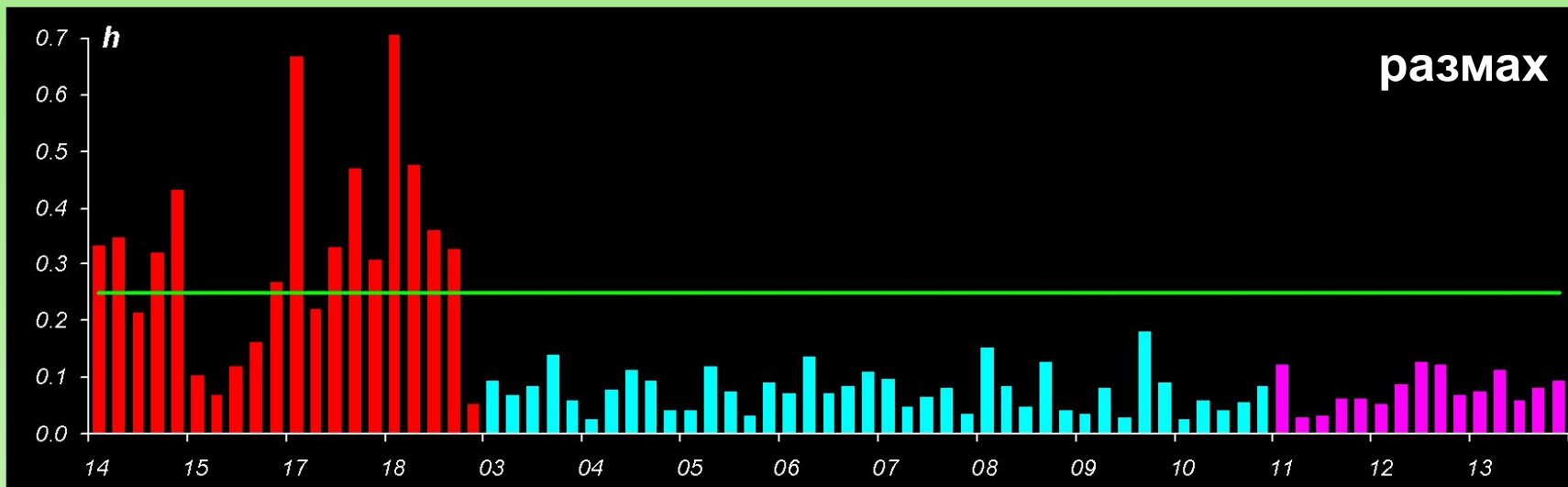
**эллипс
Хоттелинга:**
нормальные
образцы

13.03.07

Лекция в МГУ

43

Размах и отклонение для мезима

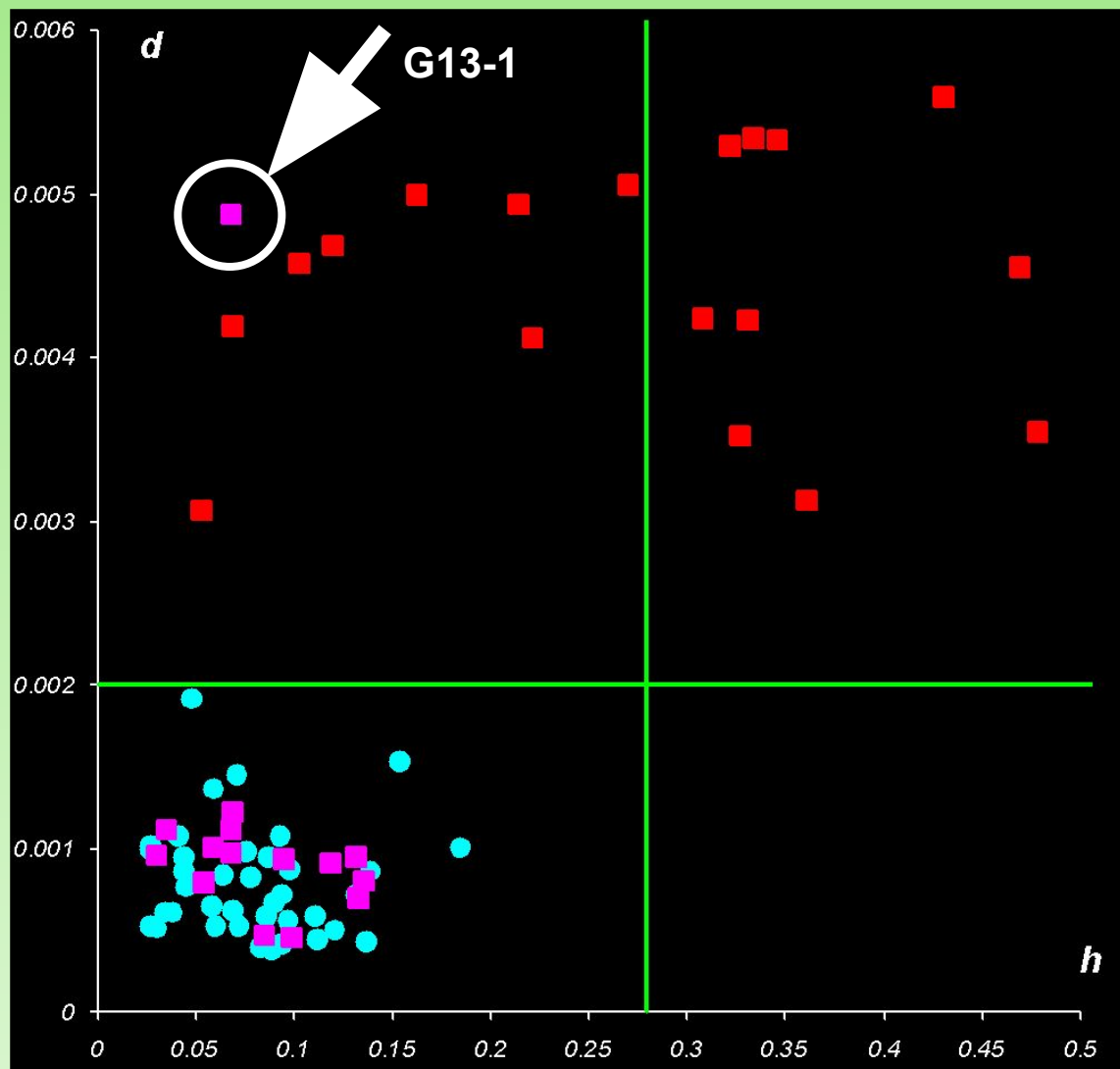


13.03.07

Лекция в МГУ

44

SIMCA



Soft

Independent

Modeling of

Class

Analogy

формальное

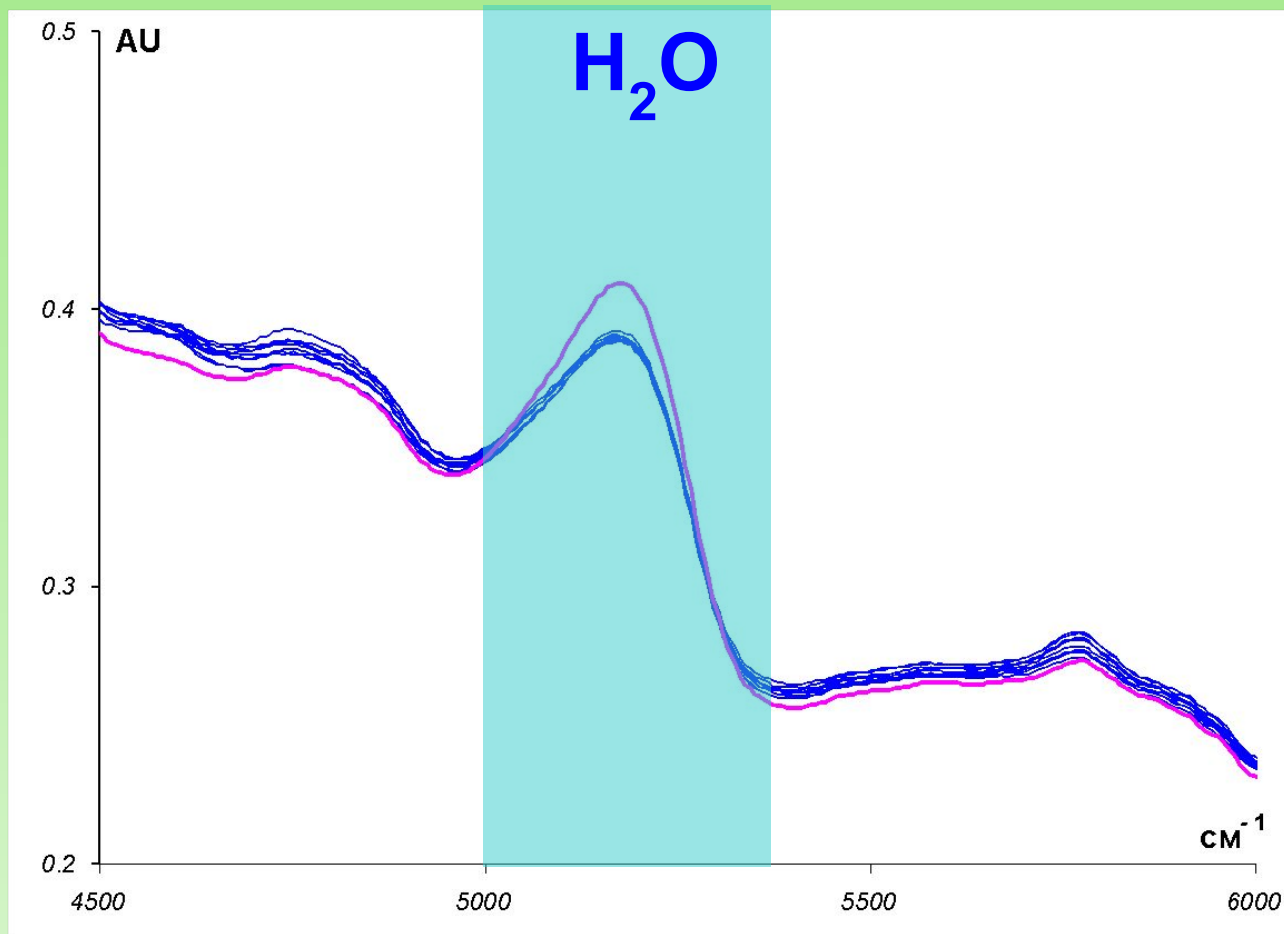
независимое

моделирование

аналогий

классов

Выброс G13-1



Хеометрика: два определения

Дедуктивное

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

сайт Российского хеометрического общества

Индуктивное

Хеометрика – это то, что делают хеометрики.

сайт Международного хеометрического общества

Хеометрики – 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** (зачастую 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. Хемометрия \neq химическая метрология
2. Хемометрия \neq статистика в химии

Хемометрия 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 ...

Химическая физика 75 (4) 2006 © 2006 Российская академия наук, Институт химической физики им. Н.Д. Зелинского

УДК 543.133

Хемометрика: достижения и перспективы

О.Е.Рожнова, А.Л.Померанцев

Институт химической физики им. Н.Д.Зелинского 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 надежного контроля технологических процессов. Показаны достижения в области хемометрики за последние 20 лет. Обсуждаются тенденции и перспективы ее развития.
Библиография — 228 ссылок.

Оглавление

I. Введение	102
II. Данные и модели, используемые в хемометрике	106
III. Методы качественного анализа. Базисные функции и дискриминанты	108
IV. Методы количественного анализа. Гравиметрические модели	112
V. Подготовка данных и обработка сигналов	115
VI. Заключение	117

I. Введение

1. История хемометрики и ее место в системе знаний

С момента опубликования перевода на русский язык единственного (до недавнего времени) книги по хемометрике [1] прошло 20 лет и за это время многое изменилось. В настоящее время хемометрические методы используются в различных областях науки и техники. Данный обзор посвящен в основном аналитической химии, где можно выделить три направления прикладной хемометрики: качественный и количественный анализ, контроль технологических процессов и оптимизация.

Наиболее важное направление в хемометрике. В последние годы очень часто баггер и оптимизацию разделились, при этом, применяя-аналитика, предложенные не только новые методы обработки данных, но и новые подходы к построению моделей.

Хемометрика — это аналитическая дисциплина, находящаяся на стыке химии и математики, и как это часто бывает с междисциплинарными дисциплинами, до сих пор не имеет общепринятого определения. Наиболее популярное определение принадлежит Д.Маскору, который считал, что хемометрика — это эмпирическая дисциплина, в которой применяются математические, статистические и другие методы, позволяющие на достоящий анализ, для оптимизации...



Ким Эсбенсен

АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ

Российское Хемометрическое Общество

18 мая 2005 г.

Надеемся, что Ким Эсбенсен Райнхольд изобретатель иерархических Третейской системы по анализу данных (HSC-3) поднимет тему Глобальной и Сильной аналитической химии.

[Посмотреть мероприятие](#) [Презентовать предложение](#)

20 апреля 2005 г.

18 апреля 2005 г. в рамках выставки AnalyticaWorld прошла однодневная конференция "Хемометрика и вопросы прикладной аналитической химии". Приглашаем профессора R. Venema (University of Bath), UK ("PAT - Технологии (методы) анализа процессов" и другие дисциплины можно посмотреть на сайте Хемометрика России).

Программа конференции

2 марта 2005 г.

Следующий номер семинара посвящен хемометрике "Современные методы анализа квантового диода" (WSC-3) пройдет 14-15 февраля 2006 г. в Санкт-Петербурге.

Темы семинара конференции PAT/MSPC, серия семинаров отсюда и далее на русском языке.

[Подробнее](#)

Chemometrics.Ru :: хемометрика в России

Новости

2005-05-26 // Изменения
В разделе "Статьи" дополнен и исправлен ранее опубликованный доклад Карпухина О.Н. на четвертом международном зимнем симпозиуме по хемометрике (WSC-4).

2005-05-04 // Новые презентации на сайте
19 апреля 2005 г. в рамках выставки AnalyticaExpo прошла однодневная конференция "Хемометрика и контроль производственных процессов", с участием профессора Хемометрики университета Реннара Бреттона. Презентации докладов и лекций, прочитанных на конференции вы можете найти в соответствующем разделе по ключевой фразе "analyticaexpo" или перейти по [этой ссылке](#).

2005-04-22 // Презентации с WSC-IV
На сайте выложена большая часть презентаций докладов и лекций, прочитанных участниками четвертого международного симпозиума по хемометрике WSC-IV. Вывести весь список доступных на данное время презентаций можно сделав поиск по ключевому слову "WSC-IV" в разделе [Презентации](#) либо используя [эту ссылку](#).

2005-04-07 // Пополнение в разделе "Статьи"
В разделе "Статьи" выложен доклад, который был сделан д.и.н., заведующим лабораторией института химической физики РАН Карпухиным Олегом Никифоровичем на закрытии четвертого международного симпозиума по хемометрике WSC-4. В статье поднимаются очень интересные вопросы о роли хемометрики в современной науке и возможных вариантах ее дальнейшего развития. Очень интересна точка зрения, познакомится с которой будет полезно как начинающим ученым, так и тем, кого уже можно назвать опытным исследователем.

Ближайшие события

27.06.2005

UK's 3rd Annual Process Analytical Technologies for Biotech

Открытая ежегодная конференция для специалистов из фармацевтической, пищевой и химической промышленности, интересующихся работой на практике в промышленности - Process Analytical Technologies (PAT).

Случайная ссылка

Homepage of Chemometrics — официальный популярный англоязычный ресурс по хемометрике. Содержит очень много разнообразной информации: новости, публикации, конференции, статьи и т.д. Создатель и редактор — John Turrill

Новости на странице

19.04.2005
В разделе **Презентации** появилась информация о второй ежегодной конференции по применению Process Analytical Technology (PAT) в биотехнике.

18.01.2004
Российский семинар посвящен в основном анализу материалов.

19.01.2004
В разделе **Презентации** появилась информация о международном симпозиуме по хемометрике "Статистика в химии"

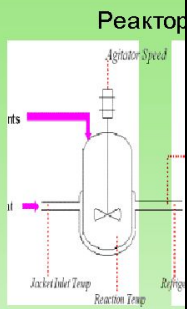
13.03.07

Лекция в МГУ

51

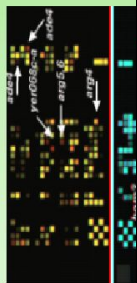
Что осталось на потом ...

MSPC в фармацевтике



10.02.05

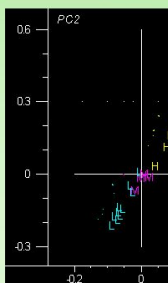
Хеометрика в биологии



10.02.05

Электронные язык и нос

Определение качества бензина по ИК-спектру



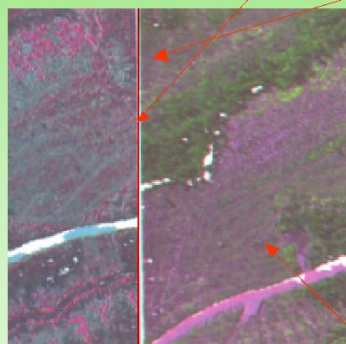
10.02.05

Исследование состояния лесов (Канада)

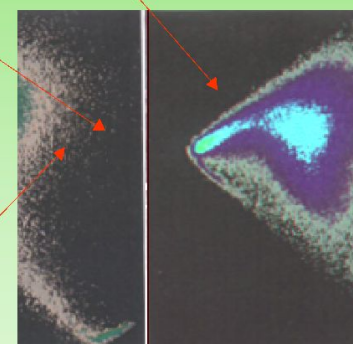
Новые посадки

Старые деревья

Исходный аэроснимок



Он же в пространстве ГК1



Область с высоким коэффициентом отражения

07.10.06

EVENT

30

13.03.07

Лекция в МГУ

52

Спасибо за внимание!



13.03.07

Лекция в МГУ

53