МАШИННОЕ ОБУЧЕНИЕ И АНАЛИЗ ДАННЫХ

(Machine Learning and Data Mining)

Н. Ю. Золотых

http://www.uic.unn.ru/~zny/ml

Лекция 4

Метод главных компонент

4.1. Историческая справка и определения

Метод главных компонент (Principal Component Analysis, PCA) — один из основных, «классических», методов понижения размерности.

Sylvester J.J. On the reduction of a bilinear quantic of the nth order to the form of a sum of n products by a double orthogonal substitution. Messenger of Mathematics, 19, 42–46 (1889)

Pearson C. On lines and planes of closest fit to systems of points in space. Phil. Mag., Series B., 2 (11), 559–572 (1901)

Джеймс Джозеф Сильвестр (1814–1897) — английский математик. Несколько лет работал в США. Основал «American Journal of Mathematics».

Карл Пирсон (1857–1936) — английский математик, статистик и биолог. Основатель математической статистики.

$$x^{(1)}, x^{(2)}, \ldots, x^{(N)} \in \mathbf{R}^d.$$

Рассмотрим следующие задачи:

1) Аппроксимировать данные линейными многообразиями меньшей размерности: найти линейное многообразие заданной размерности $k < d$, сумма квадратов расстояний до которого минимальна.

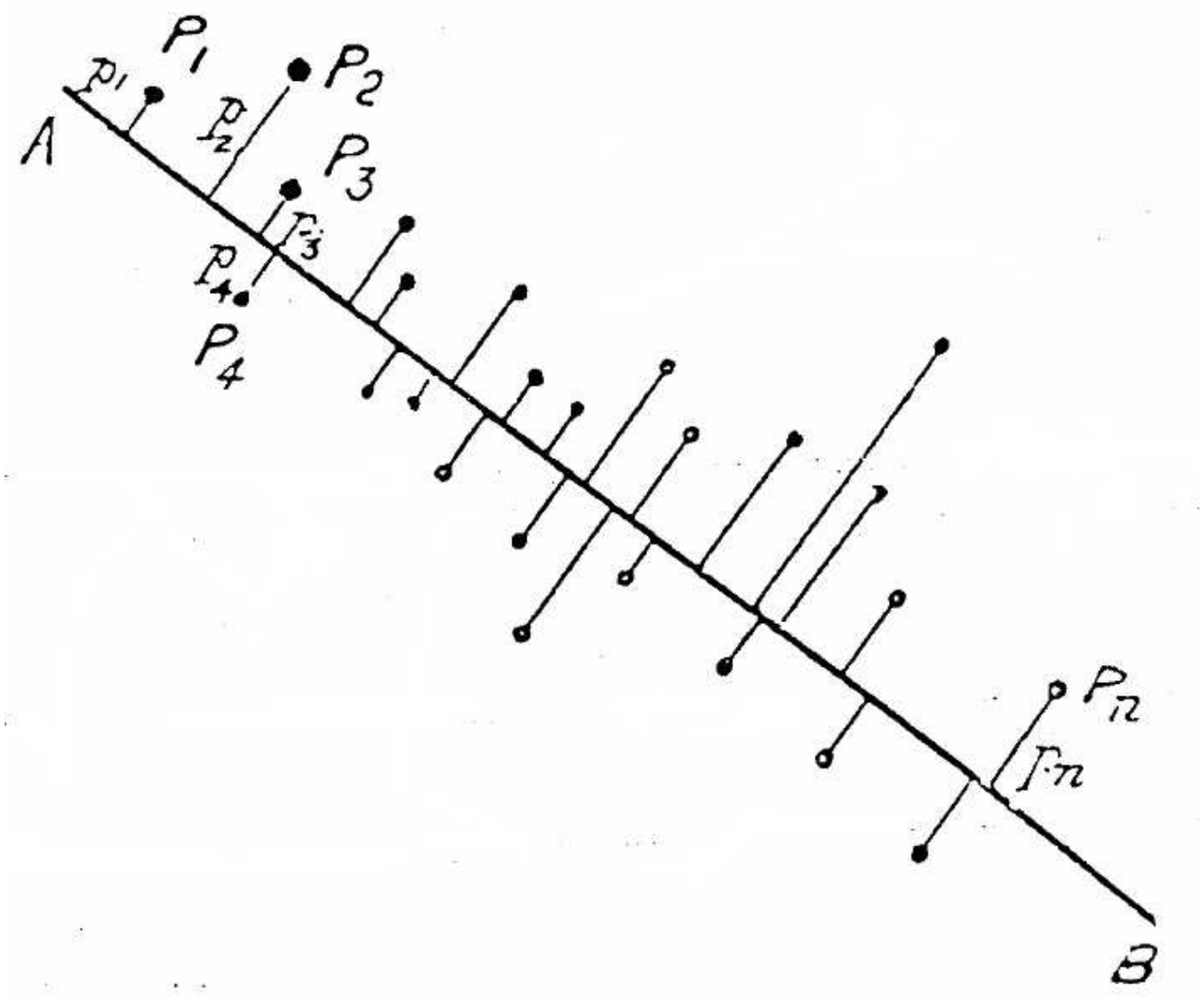
2) Найти подпространство заданной размерности, в ортогональной проекции на которое разброс (рассеивание) (выборочная дисперсия для $k = 1$) максимален.

3) Найти подпространство заданной размерности, в ортогональной проекции на которое среднеквадратичное расстояние между каждой парой точек максимально.

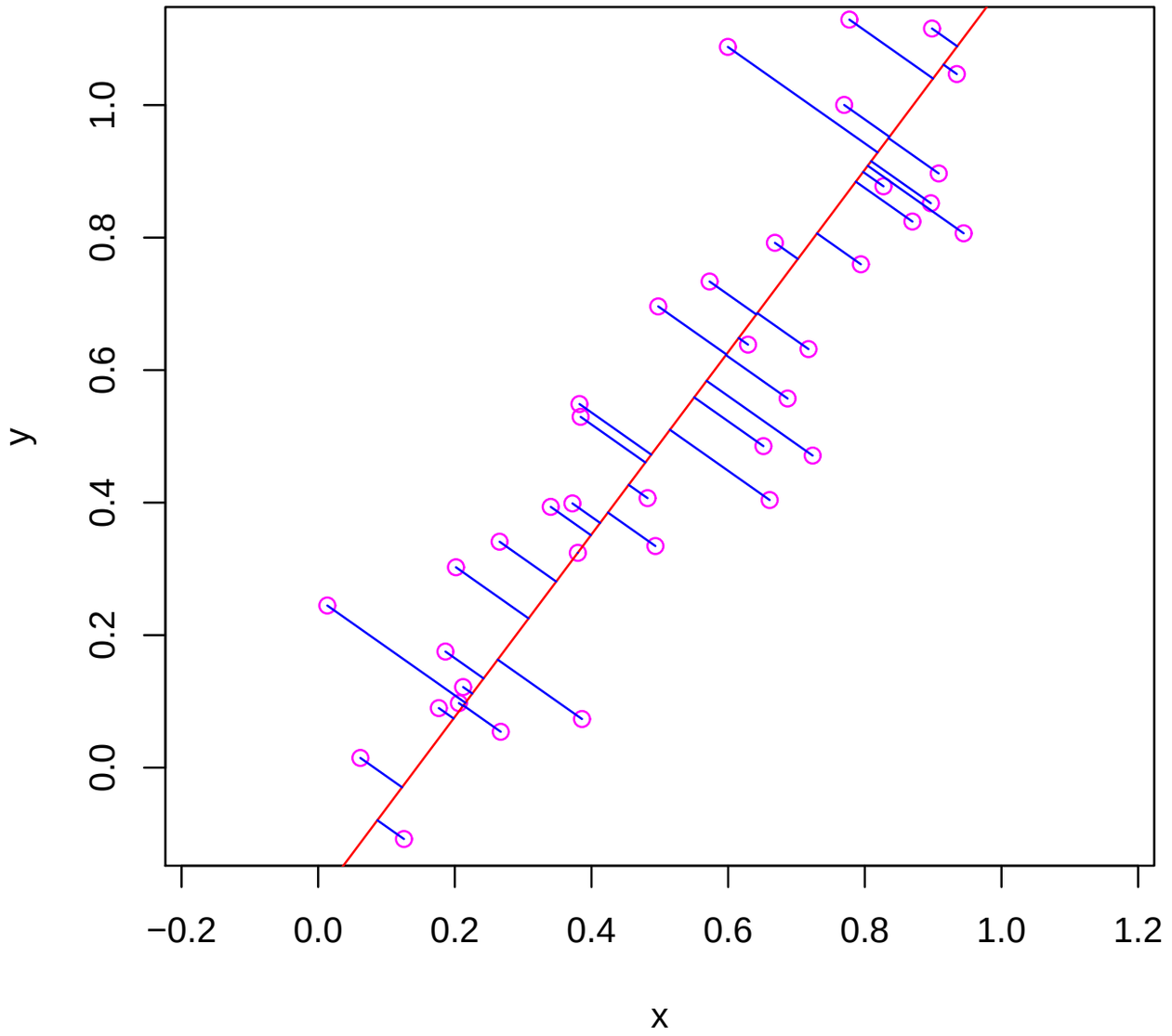
Все эти задачи имеют одно и то же решение — PCA.

Многочисленные *обобщения* (в том числе нелинейные): хорошо приспособленный базис, метод главных кривых и многообразий, поиск наилучшей проекции (Projection Pursuit), самоорганизующиеся карты Кохонена, автоэнкодеры и т. д.

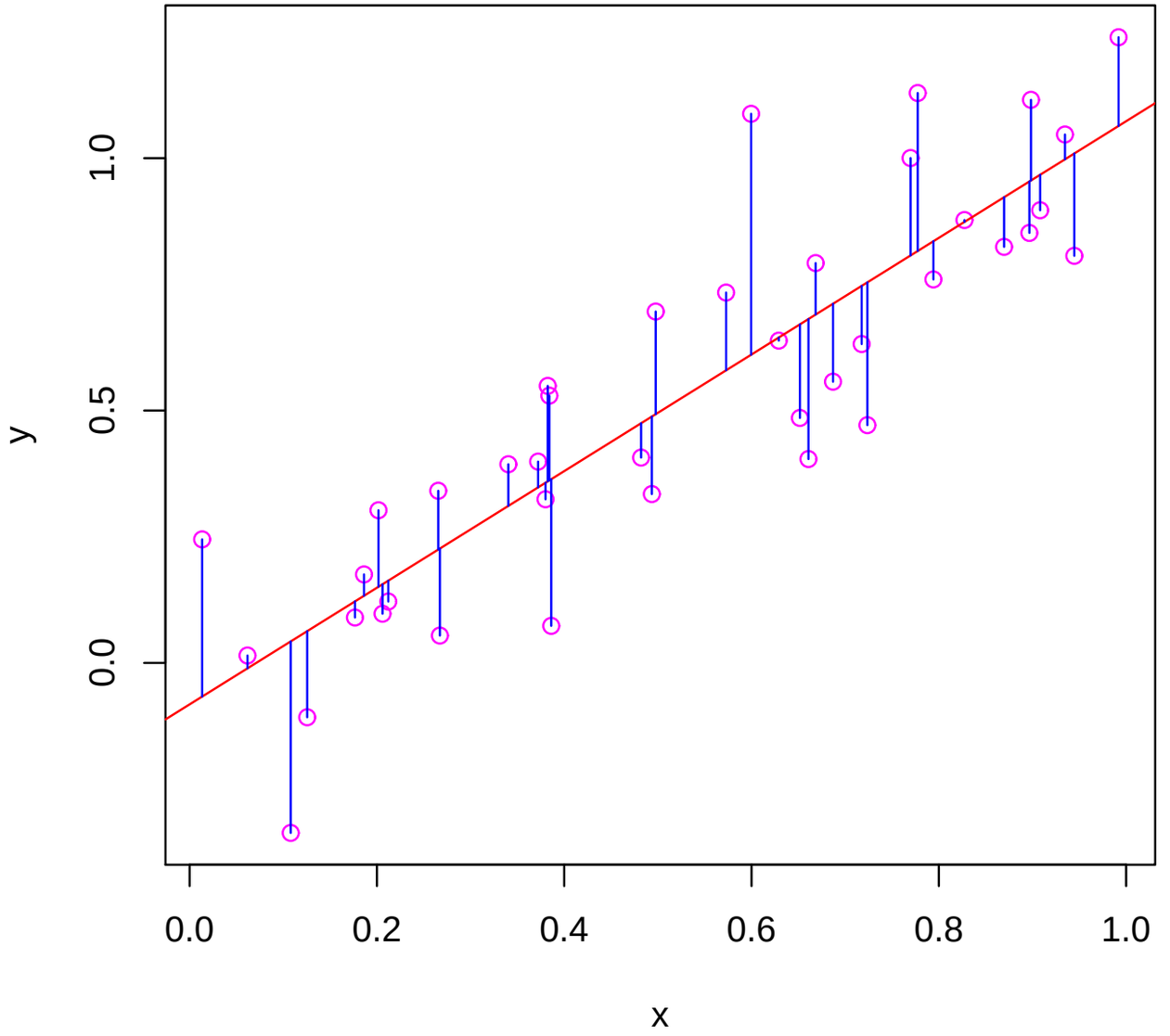
Рисунок из оригинальной статьи Пирсона



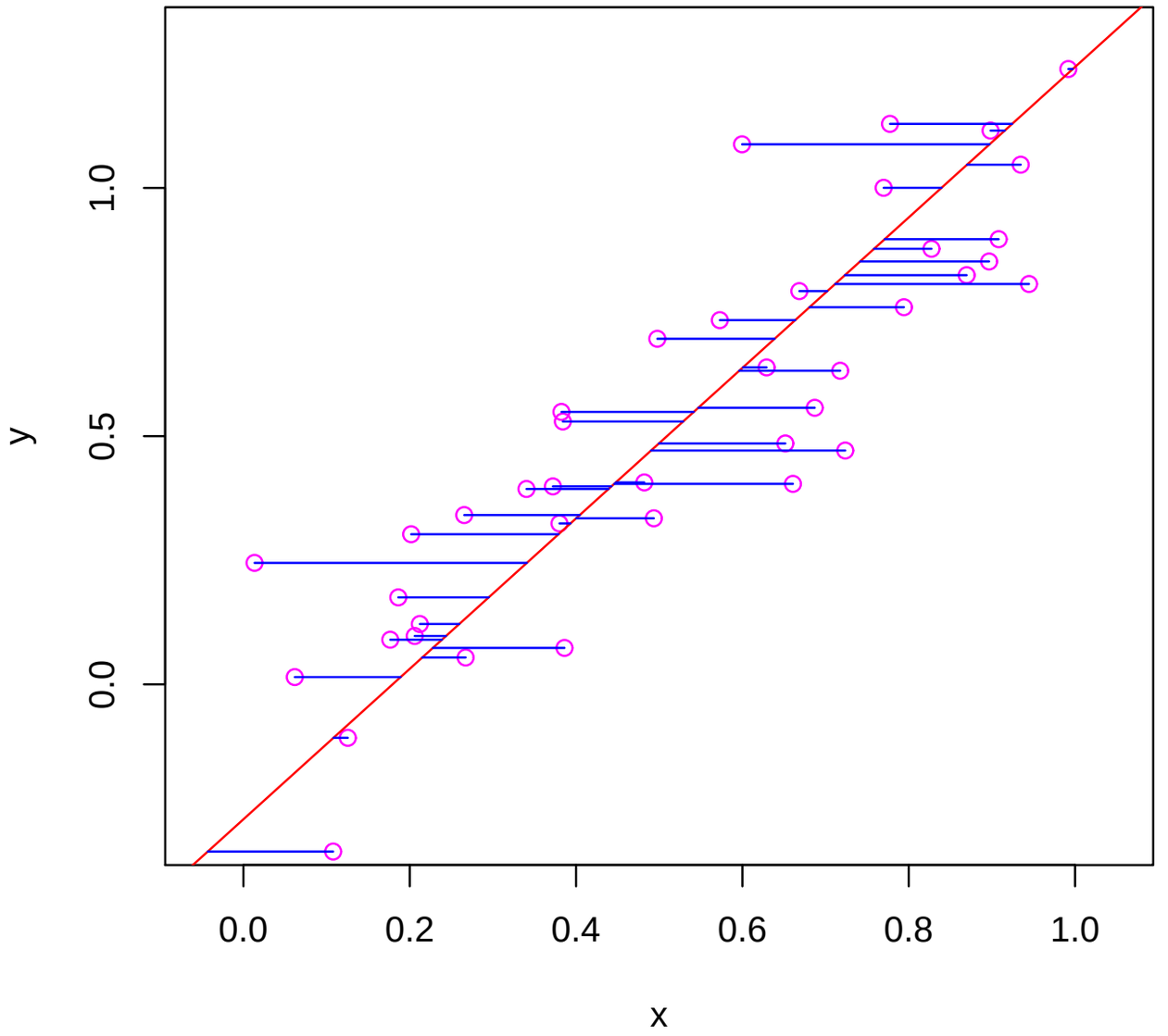
PCA и метод наименьших квадратов — это разные вещи!



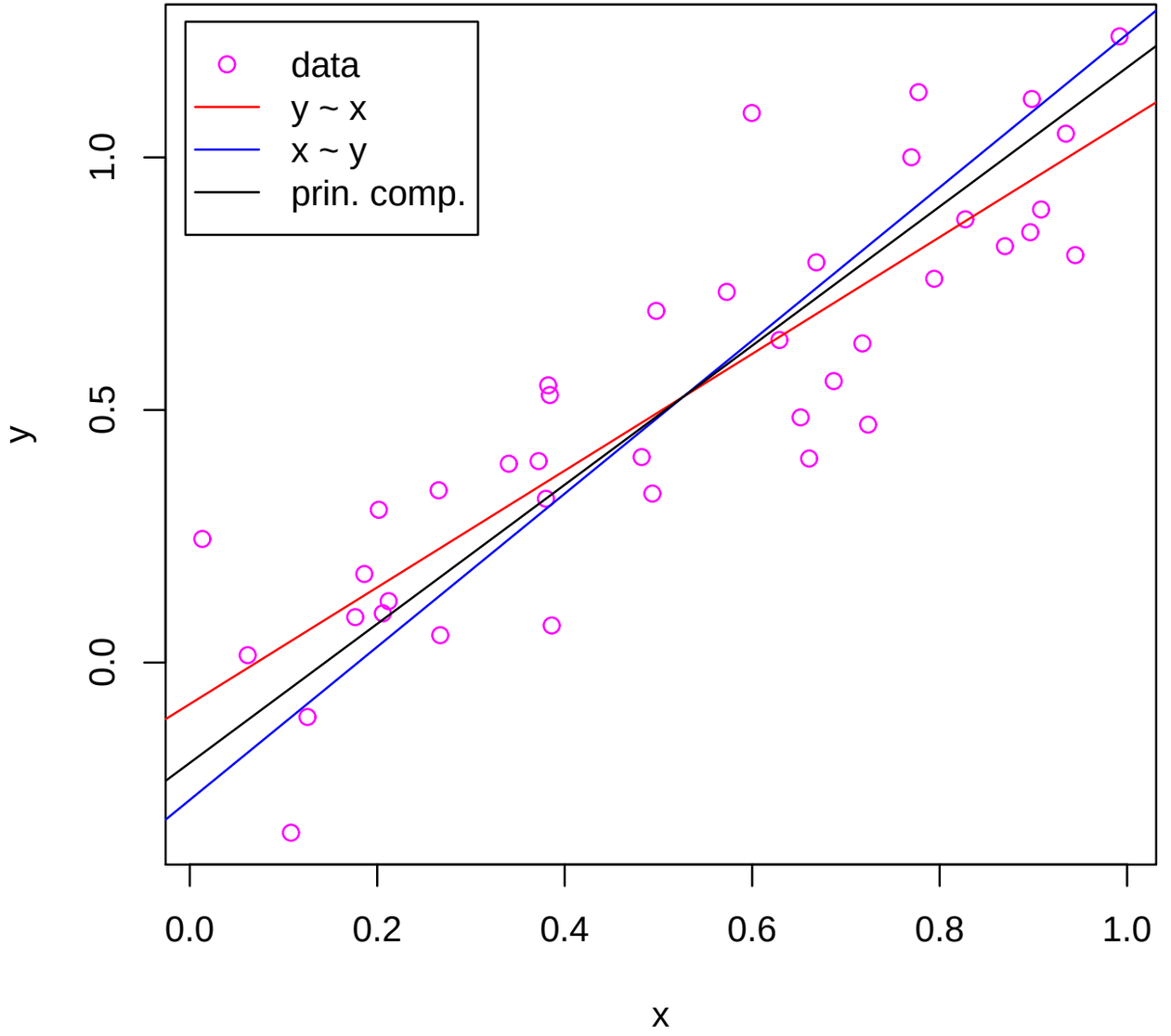
PCA и метод наименьших квадратов — это разные вещи!

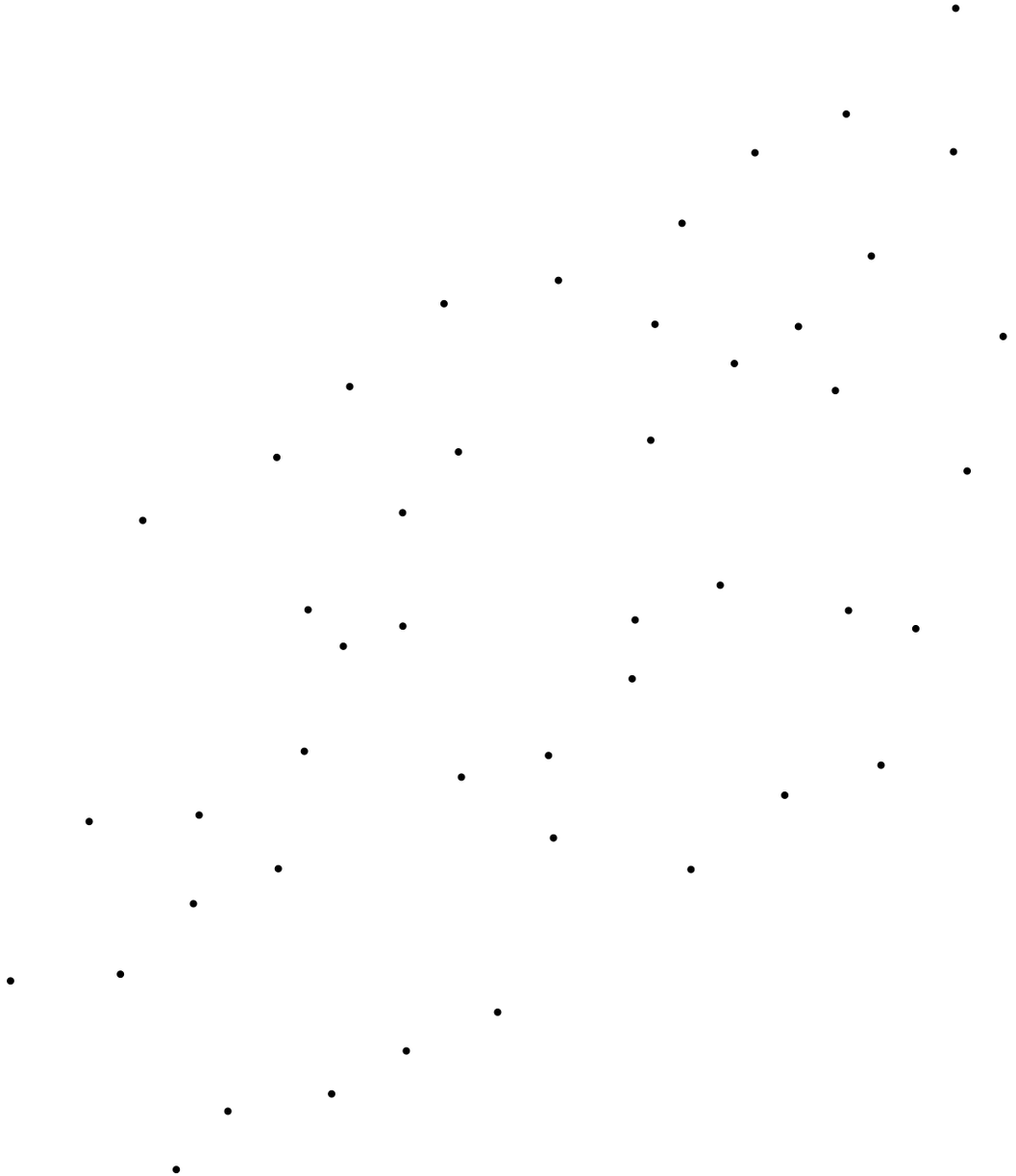


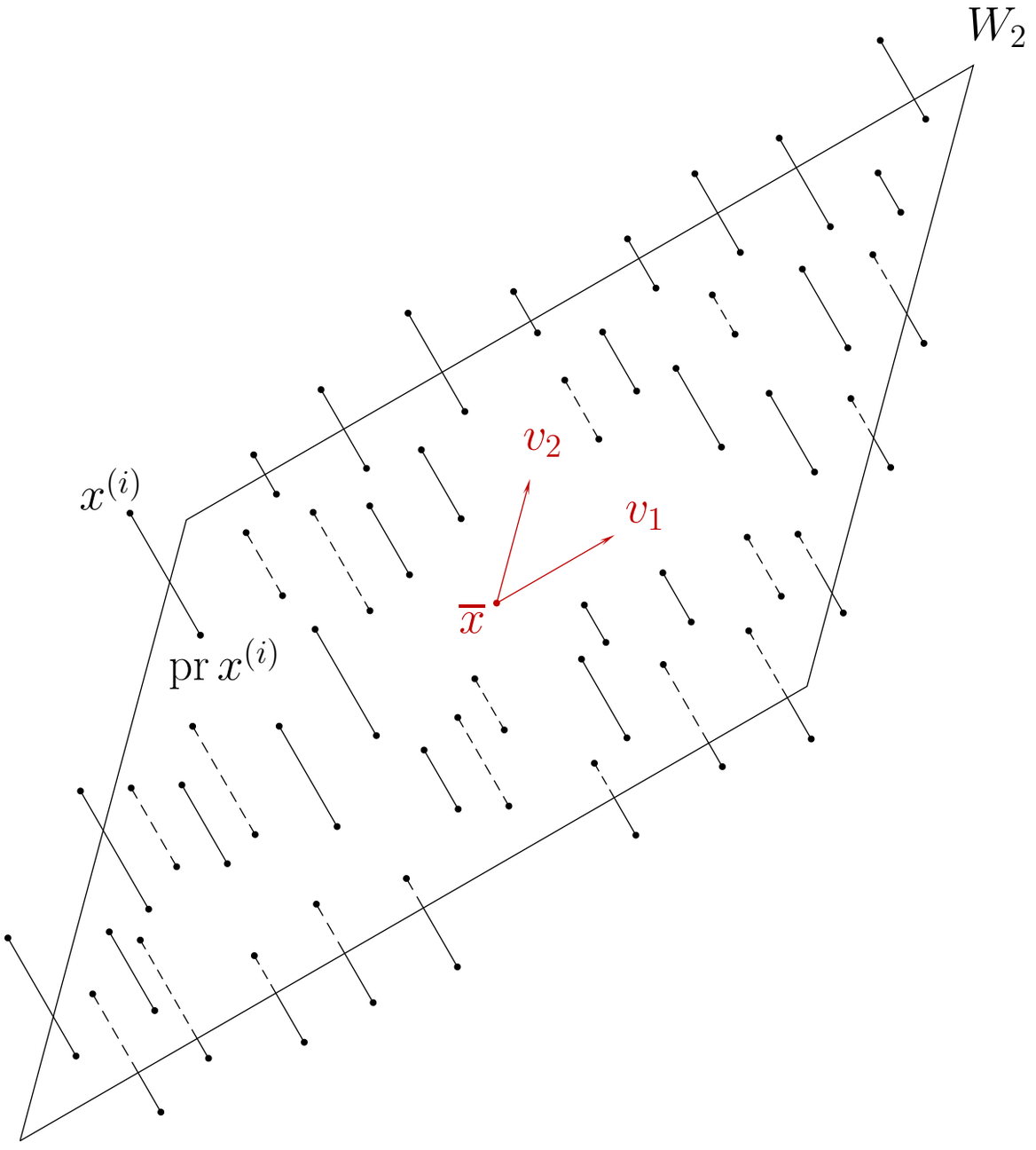
PCA и метод наименьших квадратов — это разные вещи!



PCA и метод наименьших квадратов — это разные вещи!






W_2
v_2
v_1
$x^{(i)}$
$\overline{x}$
$\operatorname{pr} x^{(i)}$

1-я задача. Ищем k-мерное линейное многообразие

$$W_k = v_0 + L(v_1, \ldots, v_k), \quad v_j \in \mathbf{R}^d, \quad \|v_j\| = 1, \quad v_j \perp v_{j'} \quad (j \neq j')$$

такое, что

$$\sum_{i=1}^{N} \operatorname{dist}^2(x^{(i)}, W_k) = \sum_{i=1}^{N} \| \operatorname{ort}_{W_k} x^{(i)} \|^2 = \sum_{i=1}^{N} \left\| x^{(i)} - v_0 - \sum_{j=1}^{k} \left\langle x^{(i)} - v_0,\, v_j \right\rangle v_j \right\|^2 \to \min$$

Оказывается, что $W_0 \subset W_1 \subset \cdots \subset W_k$, $v_0 = \overline{x} = \dfrac{1}{N} \sum\limits_{i=1}^{N} x^{(i)}$
и векторы v_j (*векторы главных компонент*) можно находить итерационно!

Упражнение 4.1 Доказать, что можно положить $v_0 = \overline{x}$. Описать множество всех возможных v_0, доставляющих минимум сумме квадратов расстояний до искомого многообразия.

Таким образом, в этой задаче можно считать, что данные центрированы:
$x^{(i)} \leftarrow x^{(i)} - \overline{x}$ $(i = 1, 2, \ldots, N)$.

2-я задача (максимизировать разброс в ортогональной проекции).

Пусть данные уже центрированы: $x^{(i)} \leftarrow x^{(i)} - \overline{x}$ $(i = 1, 2, \ldots, N)$
(очевидно, что дисперсия в ортогональной проекции от этого не изменится)

Ищем подпространство

$$W_k = L(v_1, v_2, \ldots, v_k), \qquad v_j \in \mathbf{R}^d, \quad \|v_j\| = 1, \quad v_j \perp v_{j'} \quad (j \neq j'),$$

такое, что

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\operatorname{pr}_{W_k} x^{(i)}\right\|^2 = \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{j=1}^{k}\left\langle x^{(i)},\, v_j\right\rangle v_j\right\|^2 = \sum_{i=1}^{N}\sum_{j=1}^{k}\left\langle x^{(i)},\, v_j\right\rangle^2 \to \max$$

Эта задача эквивалентна предыдущей, так как (по теореме Пифагора)

$$\|\operatorname{pr}_{W_k} x^{(i)}\|^2 = \|x^{(i)}\|^2 - \|\operatorname{ort}_{W_k} x^{(i)}\|^2 \Leftrightarrow \left\|\sum_{j=1}^{k}\left\langle x^{(i)},\, v_j\right\rangle v_j\right\|^2 = \left\|x^{(i)}\right\|^2 - \left\|x^{(i)} - \left\langle x^{(i)},\, v_j\right\rangle v_j\right\|^2$$

и первое слагаемое не зависит от v_j.

3-я задача (максимизировать в ортогональной проекции среднеквадратичное расстояние между каждой парой точек).

Пусть данные уже центрированы: $x^{(i)} \leftarrow x^{(i)} - \overline{x}$ $(i = 1, 2, \ldots, N)$.

Ищем подпространство

$$W_k = L(v_1, v_2, \ldots, v_k), \qquad v_j \in \mathbf{R}^d, \quad \|v_j\| = 1, \quad v_j \perp v_{j'} \quad (j \neq j'),$$

такое, что

$$\frac{1}{N(N-1)} \sum_{i \neq i'} \left\| \mathrm{pr}_{W_j} x^{(i)} - \mathrm{pr}_{W_j} x^{(i')} \right\|^2 \to \max$$

Эта задача эквивалентна 2-й задаче (а, следовательно, и 1-й), так как для любых векторов $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ в $\mathbf{R}^d$

$$\frac{1}{N(N-1)} \sum_{i \neq i'} \left\| x^{(i)} - x^{(i')} \right\|^2 = \frac{2N^2}{N(N-1)} \left(\frac{1}{N} \sum_{i=1}^{N} \|x^{(i)}\|^2 - \left\| \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \right\|^2 \right)$$

(справа в больших скобках стоит выборочная дисперсия).

$k = 1$:

$$\sum_{i=1}^{N} \left\langle x^{(i)},\, v_1 \right\rangle^2 = \|\mathbf{X}v_1\|^2 = v_1^\top \mathbf{X}^\top \mathbf{X} v_1 \to \max_{\|v_1\|=1}$$

Функция Лагранжа:

$$\mathscr{L} = v_1^\top \mathbf{X}^\top \mathbf{X} v_1 - \lambda_1(\|v_1\|^2 - 1)$$

$$\frac{\partial \mathscr{L}}{\partial v_1} = 2\mathbf{X}^\top \mathbf{X} v_1 - 2\lambda_1 v_1 = 0 \qquad \Leftrightarrow \qquad \mathbf{X}^\top \mathbf{X} v_1 = \lambda_1 v_1,$$

т. е. v_1 — собственный вектор матрицы $\mathbf{X}^\top \mathbf{X}$, относящийся к максимальному с. ч. λ_1.

$k = 2$:

$$\sum_{i=1}^{N} \left\langle x^{(i)},\, v_2 \right\rangle^2 = \|\mathbf{X}v_2\|^2 = v_2^\top \mathbf{X}^\top \mathbf{X} v_2 \to \max_{\substack{\|v_2\|=1 \\ v_1 \perp v_2}}$$

Функция Лагранжа:

$$\mathscr{L} = v_2^\top \mathbf{X}^\top \mathbf{X} v_2 - \lambda_2(\|v_2\|^2 - 1) - \mu_{12} v_1^\top v_2 \quad \Rightarrow \quad \frac{\partial \mathscr{L}}{\partial v_2} = 2\mathbf{X}^\top \mathbf{X} v_2 - 2\lambda_2 v_2 - \mu_{12} v_1 = 0,$$

умножая на $v_1^\top$ слева:

$$2v_1^\top \mathbf{X}^\top \mathbf{X} v_2 - 2\lambda_2 v_1^\top v_2 - \mu_{12} v_1^\top v_1 = 0 \quad \Rightarrow \quad 2\lambda_1 v_1^\top v_2 - 2\lambda_2 v_1^\top v_2 - \mu_{12} v_1^\top v_1 = 0 \quad \Rightarrow \quad \mu_{12} = 0,$$

откуда $\mathbf{X}^\top \mathbf{X} v_2 = \lambda_2 v_2$ т. е. v_2 — с. в. матрицы $\mathbf{X}^\top \mathbf{X}$ для следующего с. ч. λ_2.

$$v_1 = \underset{\|v\|=1}{\operatorname{argmax}} \|\mathbf{X}v\|^2 \quad (1) \qquad v_2 = \underset{\substack{\|v\|=1 \\ v \perp v_1}}{\operatorname{argmax}} \|\mathbf{X}v\|^2 \quad (2) \quad \dots \quad v_k = \underset{\substack{\|v\|=1 \\ v \perp v_1, \dots, v \perp v_{k-1}}}{\operatorname{argmax}} \|\mathbf{X}v\|^2 \quad (k)$$

Докажем, что жадный алгоритм работает, т. е.

$$\|\mathbf{X}v_1\|^2 + \dots + \|\mathbf{X}v_k\|^2 = \max_{\substack{\|w_i\|=1 \\ w_i \perp w_j}} \left(\|\mathbf{X}w_1\|^2 + \dots + \|\mathbf{X}w_k\|^2\right)$$

$k = 1$: из (1)

$k = 2$:

Пусть

$$w_2 \perp v_1, \qquad W_2 = L(w_1, w_2) \tag{$*$}$$

$$\begin{matrix} \|\mathbf{X}v_1\|^2 \geq \|\mathbf{X}w_1\|^2 \text{ из } (1) \\ \|\mathbf{X}v_2\|^2 \geq \|\mathbf{X}w_2\|^2 \text{ из } (2) \text{ и } (*) \end{matrix} \quad \Rightarrow \quad \|\mathbf{X}v_1\|^2 + \|\mathbf{X}v_2\|^2 \geq \|\mathbf{X}w_1\|^2 + \|\mathbf{X}w_2\|^2$$

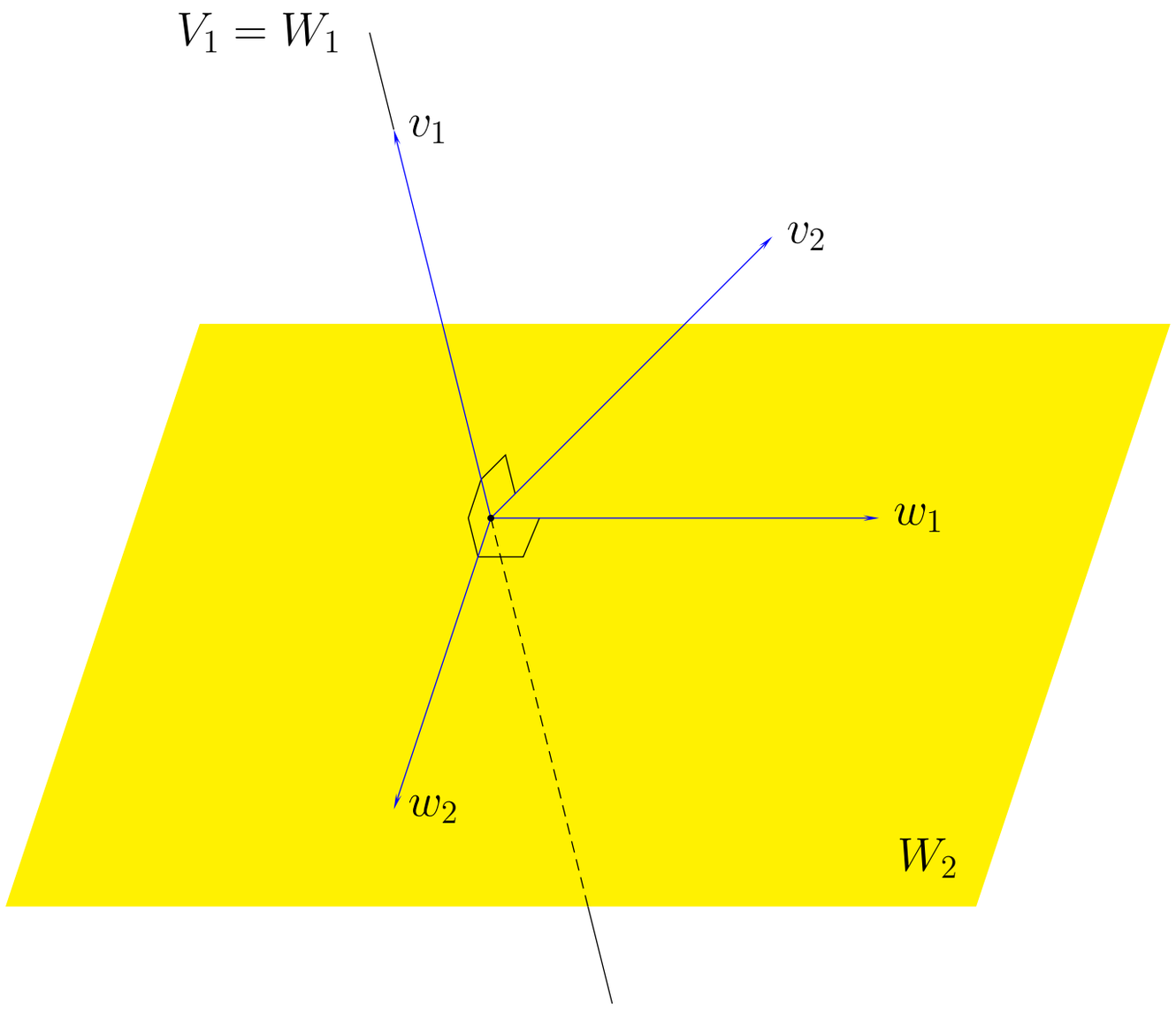
$k - 1 \to k$:

Пусть

$$w_k \perp v_1, v_2, \dots, v_{k-1}, \qquad W_k = L(w_1, \dots, w_k) \tag{$**$}$$

$$\begin{matrix} \|\mathbf{X}v_1\|^2 + \dots + \|\mathbf{X}v_{k-1}\|^2 \geq \|\mathbf{X}w_1\|^2 + \dots + \|\mathbf{X}w_{k-1}\|^2 \text{ по инд. предположению} \\ \|\mathbf{X}v_k\|^2 \geq \|\mathbf{X}w_k\|^2 \qquad\qquad \text{из } (k) \text{ и } (**) \end{matrix} \quad \Rightarrow$$

$$\|\mathbf{X}v_1\|^2 + \dots + \|\mathbf{X}v_k\|^2 \geq \|\mathbf{X}w_1\|^2 + \dots + \|\mathbf{X}w_k\|^2$$


$V_1 = W_1$
v_1
v_2
w_1
w_2
W_2

4-я задача. Пусть X — d-мерная случайная величина.

Требуется найти такой ортонормированный базис $v_1, v_2, \ldots, v_d$, в котором ковариация между различными координатами равна нулю.

После преобразования к этому базису

$$\mathsf{Cov}(X_j, X_{j'}) = 0 \qquad (j \neq j'),$$

где $\mathsf{Cov}(X_j, X_{j'}) = \mathsf{E}\left((X_j - \mathsf{E}\,X_j)(X_{j'} - \mathsf{E}\,X_{j'})\right)$ — коэффициент ковариации.

Т. е. матрица $\mathsf{Cov}\,X$ должна стать диагональной.

Искомый базис — это базис из собственных векторов матрицы $\mathsf{Cov}\,X$.

Первые три задачи являются аппроксимацией к 4-й задаче.

Дисперсия по направлению w, $\|w\| = 1$, равна $w^\top \mathsf{Cov}\,X w$

Ее максимизаци тоже приводит к той же задаче.

Ковариационная матрица (X — столбец):

$$\mathsf{Cov}\, X = \mathsf{E}\left((X - \mathsf{E}\, X) \cdot (X - \mathsf{E}\, X)^\top\right).$$

Эмпирическая (или выборочная) ковариационная матрица ($x^{(i)}$ — столбец):

$$\widehat{\mathsf{Cov}}\, X = \frac{1}{N-1}\sum_{i=1}^{N}(x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^\top.$$

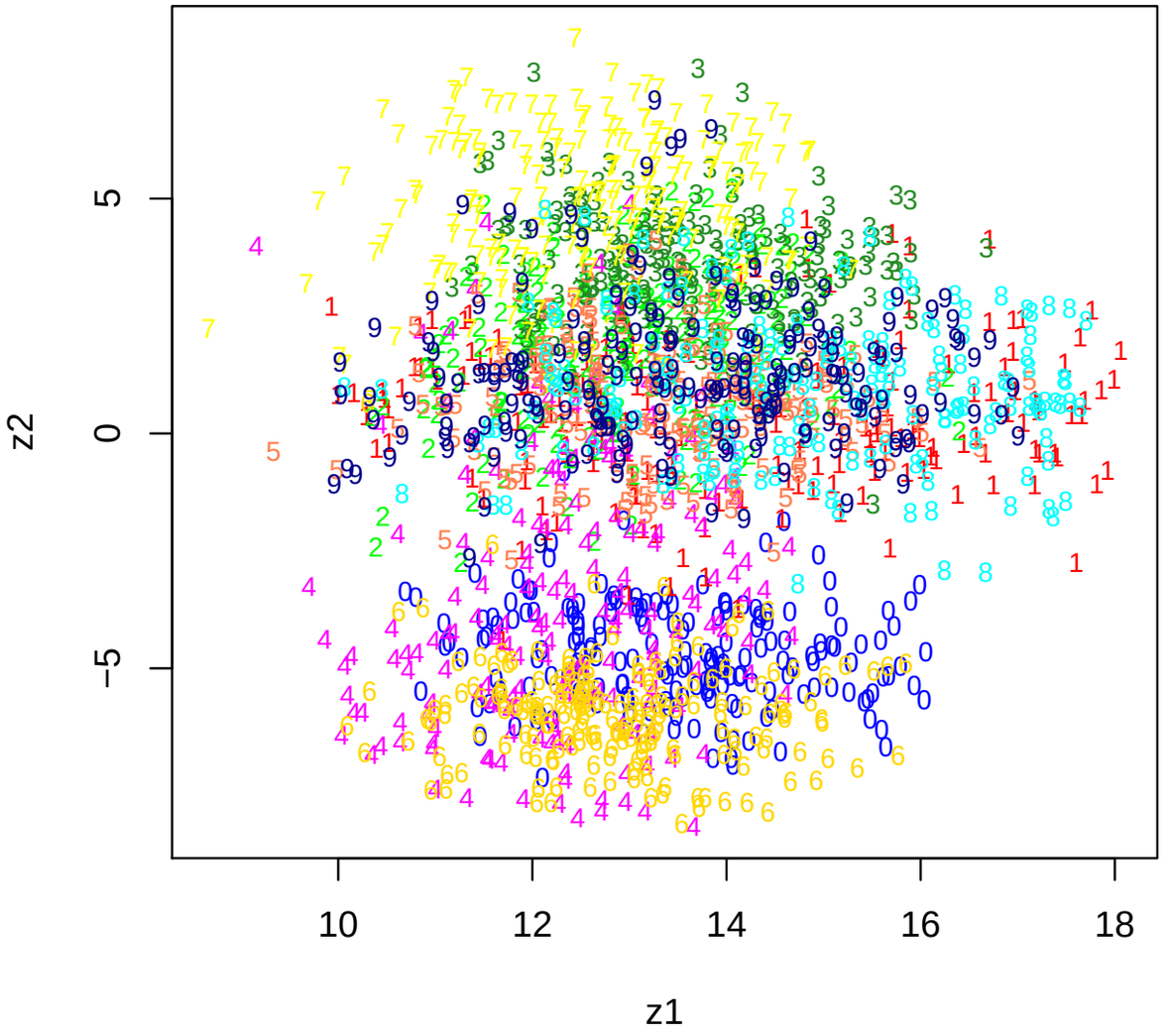
Если данные центрированы, то

$$\widehat{\mathsf{Cov}}\, X = \frac{1}{N-1}\mathbf{X}^\top\mathbf{X}.$$

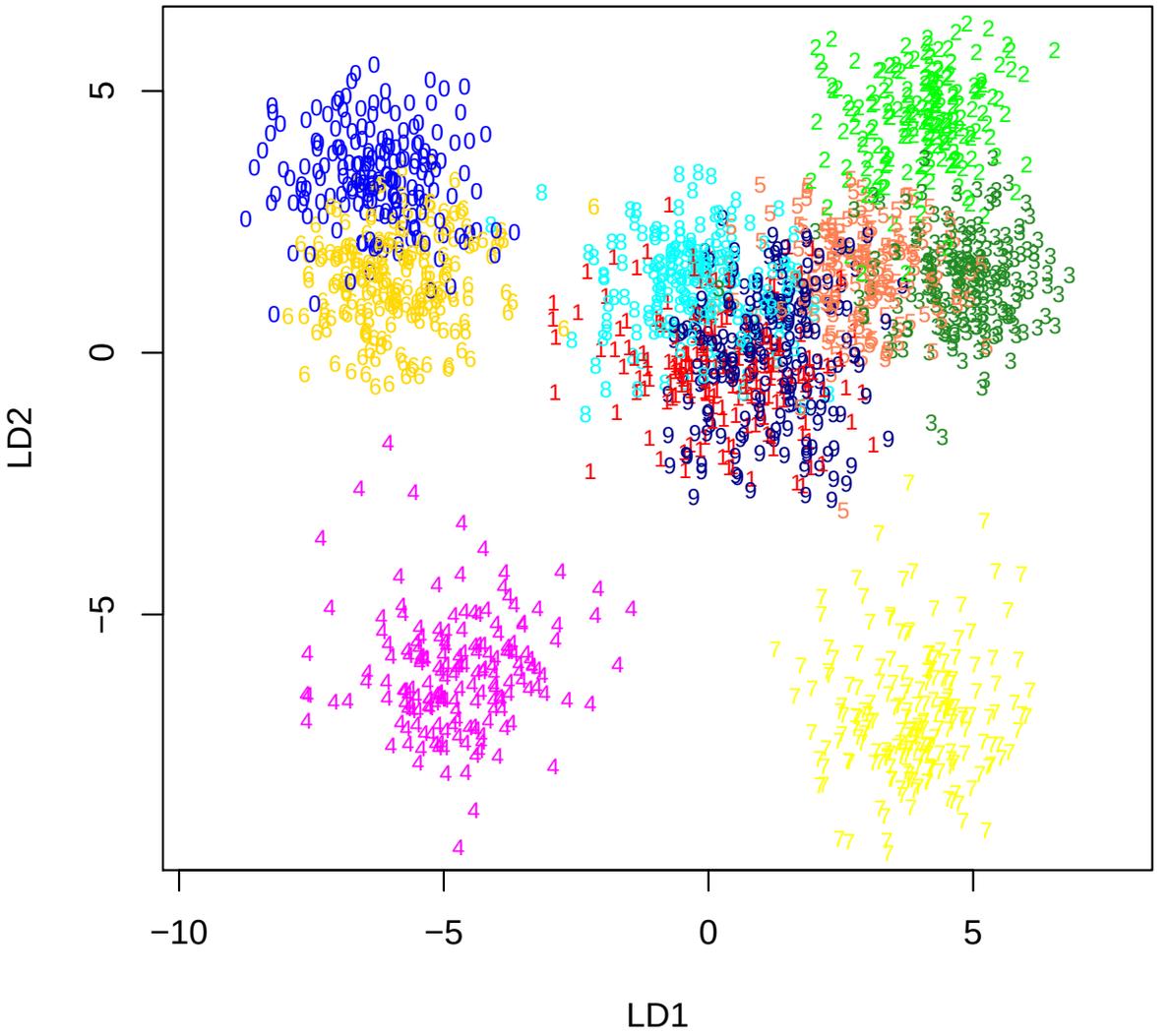
Векторы главных компонент $v_1, v_2, \ldots, v_d$ — это ортонормированный набор собственных векторов матрицы $\widehat{\mathsf{Cov}}\, X$, расположенных в порядке убывания собственных значений

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

Рукописные цифры. $d = 1024$, $k = 2$



Рукописные цифры. LDA

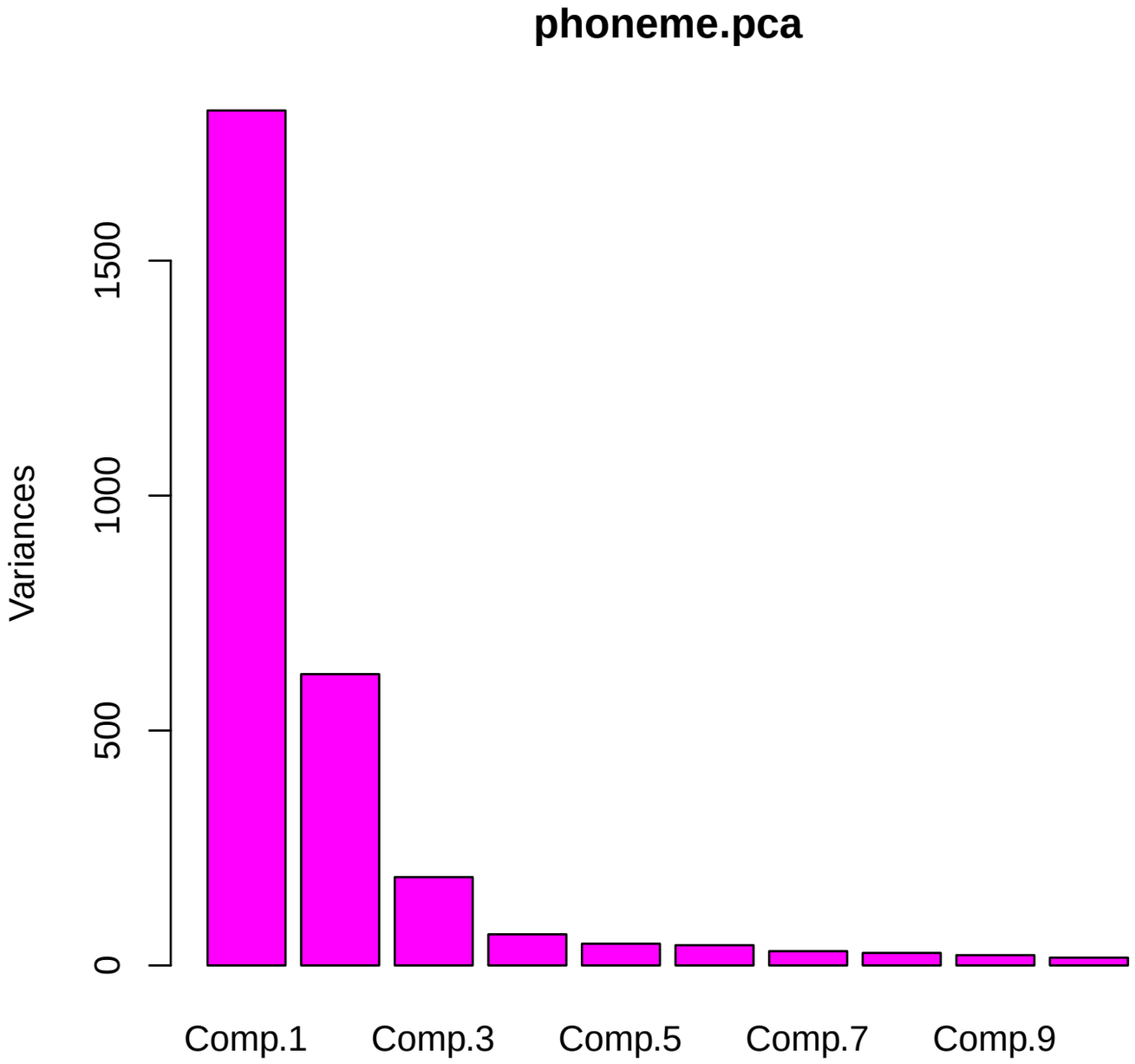


Phoneme — распознавание голоса.

(Andreas Buja, Werner Stuetzle and Martin Maechler)

4509 log-периодограмм длины $d = 256$ (все признаки количественные).

[a:]	695
[C]	1022
[i:]	1163
[d]	757
[s]	873


phoneme.pca
Variances
0
500
1000
1500
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9

Какое количество главных компонент выбрать?

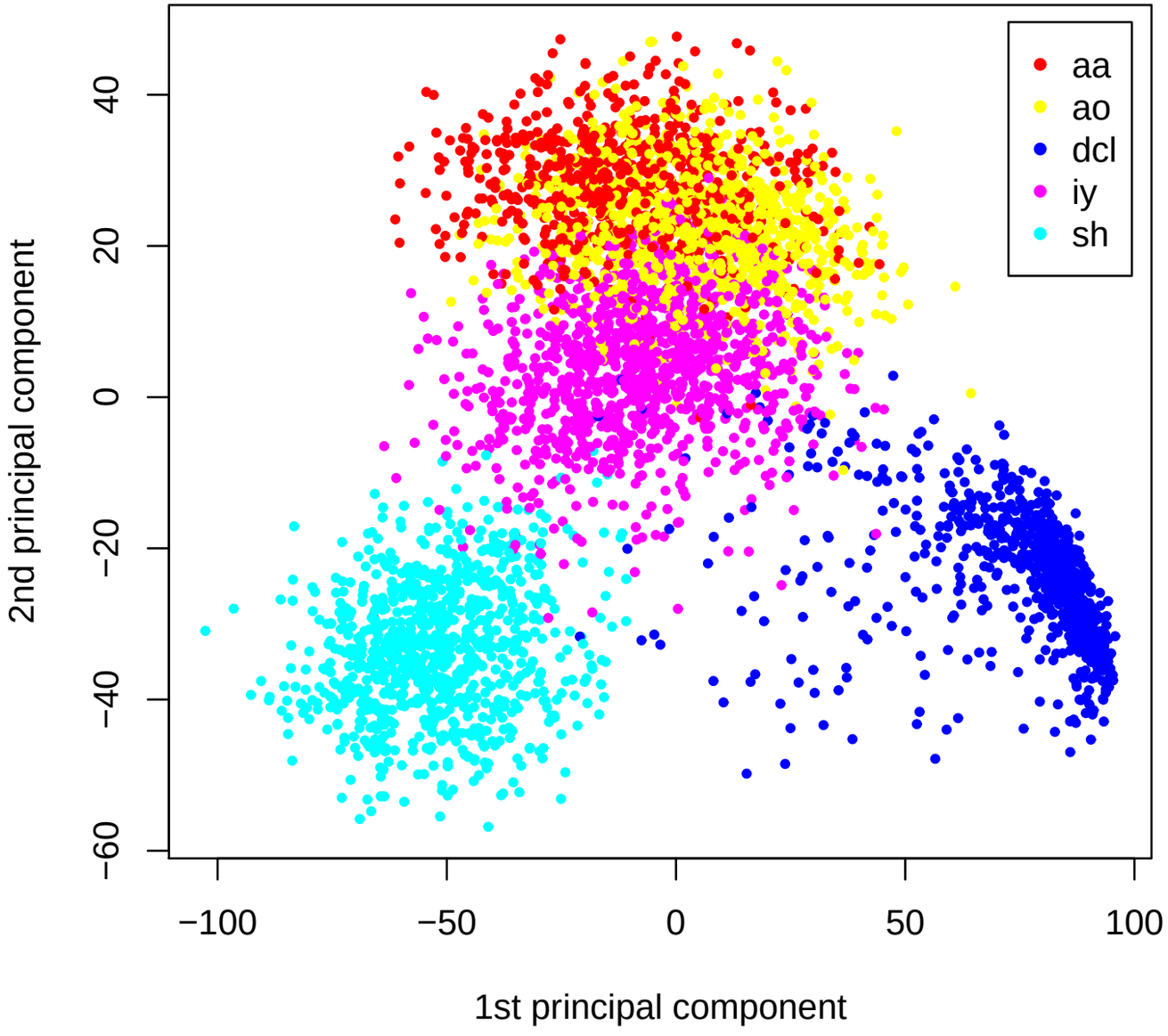
Существует много (эвристических) подходов.

Например:

Выбираем нужное количество s главных компонент так, чтобы объясненная дисперсия не меньше некоторого заданного уровня α (например, $\alpha = 0.95$):

$$\frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_s^2}{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_s^2 + \cdots + \sigma_d^2} \geq \alpha$$

Заметим, что $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_s^2 + \cdots + \sigma_d^2 = \operatorname{tr} \widehat{\mathsf{Cov}}\, X$


aa
ao
dcl
iy
sh
2nd principal component
40
20
0
−20
−40
−60
−100
−50
0
50
100
1st principal component

Все данные были случайно поделены на обучающую (4/5) и тестовую (1/5) выборки

Тестовая ошибка:

Метод	*Все признаки*	15 *главных компонент*
SVM	0.076	0.074
Random Forest	0.081	0.083
AdaBoost	0.083	0.088
Bagging	0.099	0.124

4.1.1. Eigen Faces

Sirovich, Kirby (1987) — для задачи реконструкции лица

Matthew Turk and Alex Pentland (1991) — для задачи распознавания

Пример.

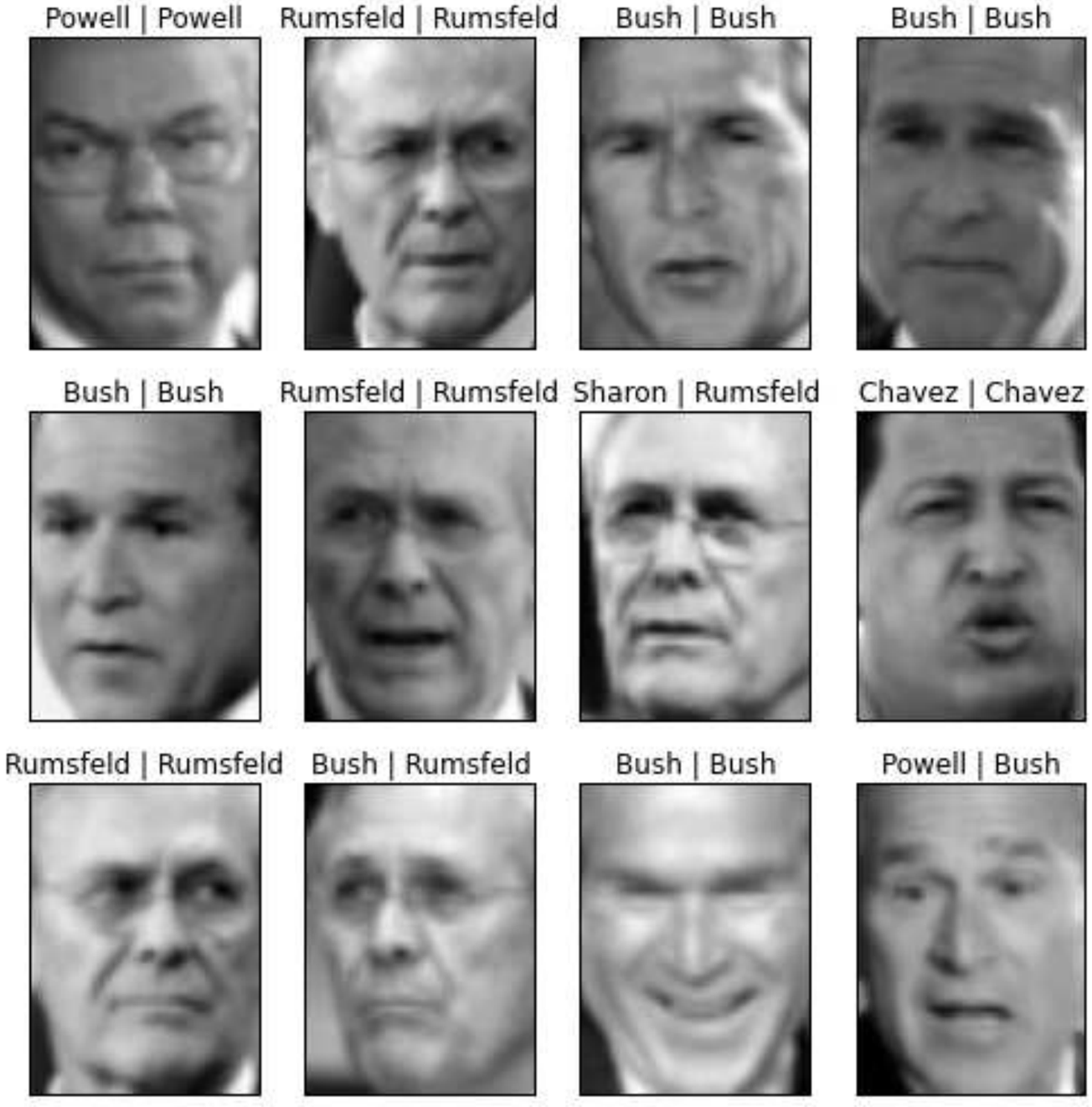
Ariel Sharon	77
Colin Powell	236
Donald Rumsfeld	121
George W. Bush	530
Gerhard Schroeder	109
Hugo Chavez	71
Tony Blair	144

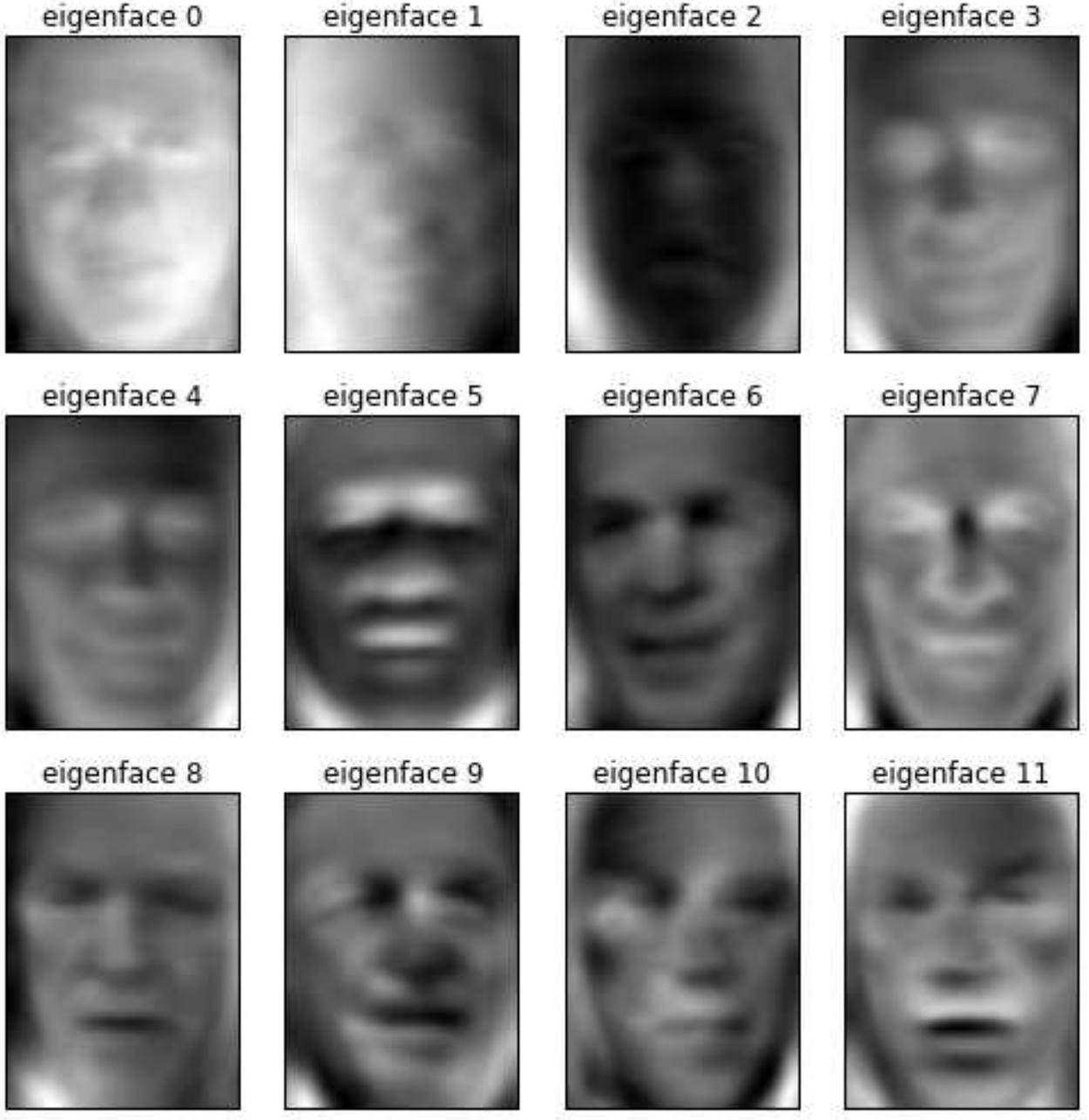
$N = 1288$, $h = 50$, $w = 37$, $d = 50 \times 37 = 1850$

При использовании $s = 100$ главных компонент + машина опорных векторов с ядром RBF получаем ошибку 14%.

При $s > 300$ ошибка только возрастает.

Hugo Chavez


Powell | Powell
Rumsfeld | Rumsfeld
Bush | Bush
Bush | Bush
Bush | Bush
Rumsfeld | Rumsfeld
Sharon | Rumsfeld
Chavez | Chavez
Rumsfeld | Rumsfeld
Bush | Rumsfeld
Bush | Bush
Powell | Bush


eigenface 0
eigenface 1
eigenface 2
eigenface 3
eigenface 4
eigenface 5
eigenface 6
eigenface 7
eigenface 8
eigenface 9
eigenface 10
eigenface 11

4.2. Сингулярное разложение

— это самый «правильный» способ вычислить главные компоненты

(нужно еще для многих других вещей)

Сингулярное разложение матрицы $\mathbf{X}$ — это

$$\underbrace{\mathbf{X}}_{N\times d} = \underbrace{\mathbf{U}}_{N\times N} \cdot \underbrace{\mathbf{\Sigma}}_{N\times d} \cdot \underbrace{\mathbf{V}^{\mathrm{T}}}_{d\times d}$$

Матрицы $\mathbf{U}$ и $\mathbf{V}$ — ортогональные ($\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$, $\mathbf{V}^{-1} = \mathbf{V}^{\mathrm{T}}$)

Матрица $\mathbf{\Sigma}$ — диагональная с неотрицательными диагональными элементами (*сингулярными числами*), записанными в порядке убывания

В матрице $\mathbf{U}$ по столбцам — *левые сингулярные векторы*

В матрице $\mathbf{V}$ по столбцам — *правые сингулярные векторы* (это и есть главные компоненты!)

Пусть $d < N$, тогда лучше использовать *неполное сингулярное разложение*:

$$\underbrace{\mathbf{X}}_{N\times d} = \underbrace{\mathbf{U}}_{N\times d} \cdot \underbrace{\mathbf{\Sigma}}_{d\times d} \cdot \underbrace{\mathbf{V}^{\mathrm{T}}}_{d\times d}$$

Если $d > N$, то наоборот:

$$\underbrace{\mathbf{X}}_{N\times d} = \underbrace{\mathbf{U}}_{N\times N} \cdot \underbrace{\mathbf{\Sigma}}_{N\times N} \cdot \underbrace{\mathbf{V}^{\mathrm{T}}}_{N\times d}$$

Пусть данные в $\mathbf{X}$ центрированы, тогда

$$C = \frac{1}{N-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \frac{1}{N-1}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}) = \frac{1}{N-1}\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^{\mathrm{T}} = \mathbf{V}\left(\frac{1}{N-1}\mathbf{\Sigma}^2\right)\mathbf{V}^{\mathrm{T}}$$

Итак, $\mathbf{V}$ составлена из главных компонент (правых сингулярных векторов).

С точностью до перестановки столбцов $\mathbf{V} = Q$.

$\frac{1}{N-1}\mathbf{\Sigma}^2$ составлена из дисперсий по главным компонентам.

С точностью до перестановки диагональных элементов $\frac{1}{N-1}\mathbf{\Sigma}^2 = D$.

Пусть используется $k \leq d$ главных компонент.

$\mathbf{U}_k$ — первые k столбцов матрицы $\mathbf{U}$

$\mathbf{\Sigma}_k$ — первые k столбцов и k строк матрицы $\mathbf{\Sigma}$

$\mathbf{V}_k$ — первые k столбцов матрицы $\mathbf{V}$

$$\underbrace{\mathbf{X}}_{N\times d} \approx \underbrace{\mathbf{U}_k}_{N\times k} \cdot \underbrace{\mathbf{\Sigma}_k}_{k\times k} \cdot \underbrace{\mathbf{V}_k^\top}_{k\times d}$$

$\mathbf{V}_k$ — векторы главных компонент (по столбцам), или «матрица нагрузок» (loadings)

$\mathbf{X}\mathbf{V}_k = \mathbf{U}_k\mathbf{\Sigma}_k$ — проекции векторов на г.к., или «матрица счетов» (scores)

Упражнение 4.2 Доказать, что $\mathbf{X}\mathbf{V}_k = \mathbf{U}_k\mathbf{\Sigma}_k$

$\mathbf{X}(\mathbf{\Sigma}_k)^{-1}\sqrt{N} = \mathbf{U}_k\sqrt{N}$ — нормировка на единичную дисперсию, или «матрица z-счетов» (z-scores) — whitening

$\mathbf{X} - \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top$ — матрица ошибок

4.3. Напоминание: Проверка статистических гипотез

Пусть функция распределения $P^*(x)$ случайной величины X не известна.

Гипотеза H_0: $P^*(x) = P(x)$.

Альтернативная гипотеза H_1.

Ошибка 1-го рода: Гипотеза H_0 отвергается (и принимается H_1), хотя H_0 верна.

Ошибка 2-го рода: Гипотеза H_0 принимается, хотя она не верна.

Аналогично для параметрических гипотез.

Вначале выбирается α — *вероятностный уровень значимости* критерия, равный вероятности того, что произойдет ошибка 1-го рода (например, 0.1, 0.05, 0.01)

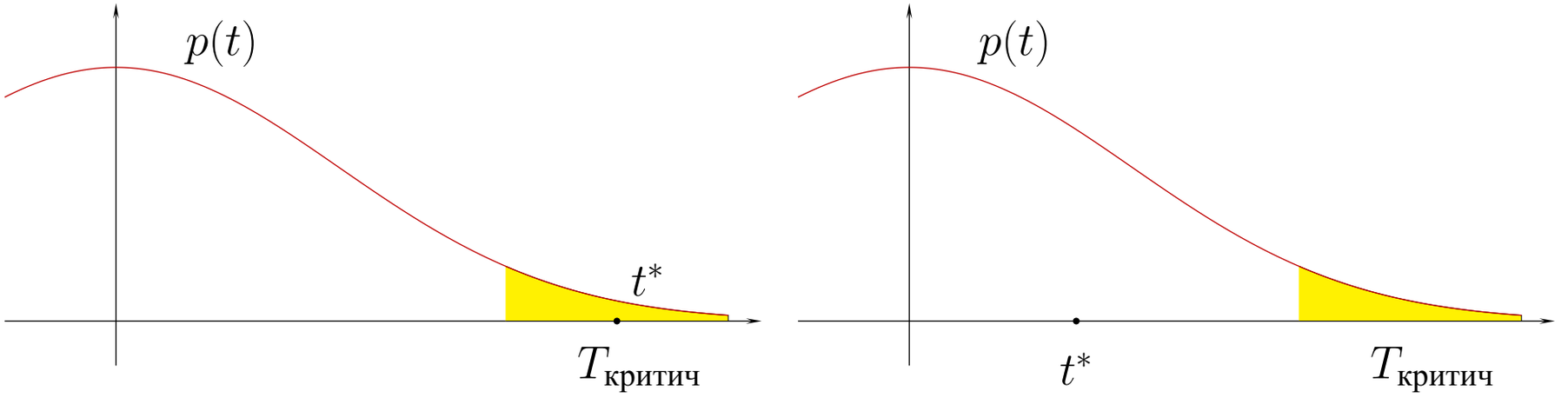
Вычисляется реализация t^* статистики $t(X)$, характеризующей отклонение эмпирических данных от теоретических (соответствующих гипотезе H_0).

Пусть $T_{\text{критич}} = T_{\text{критич}}(H_0, \alpha)$ — *критическая область* — подмножество значений статистики $t(X)$, такое, что

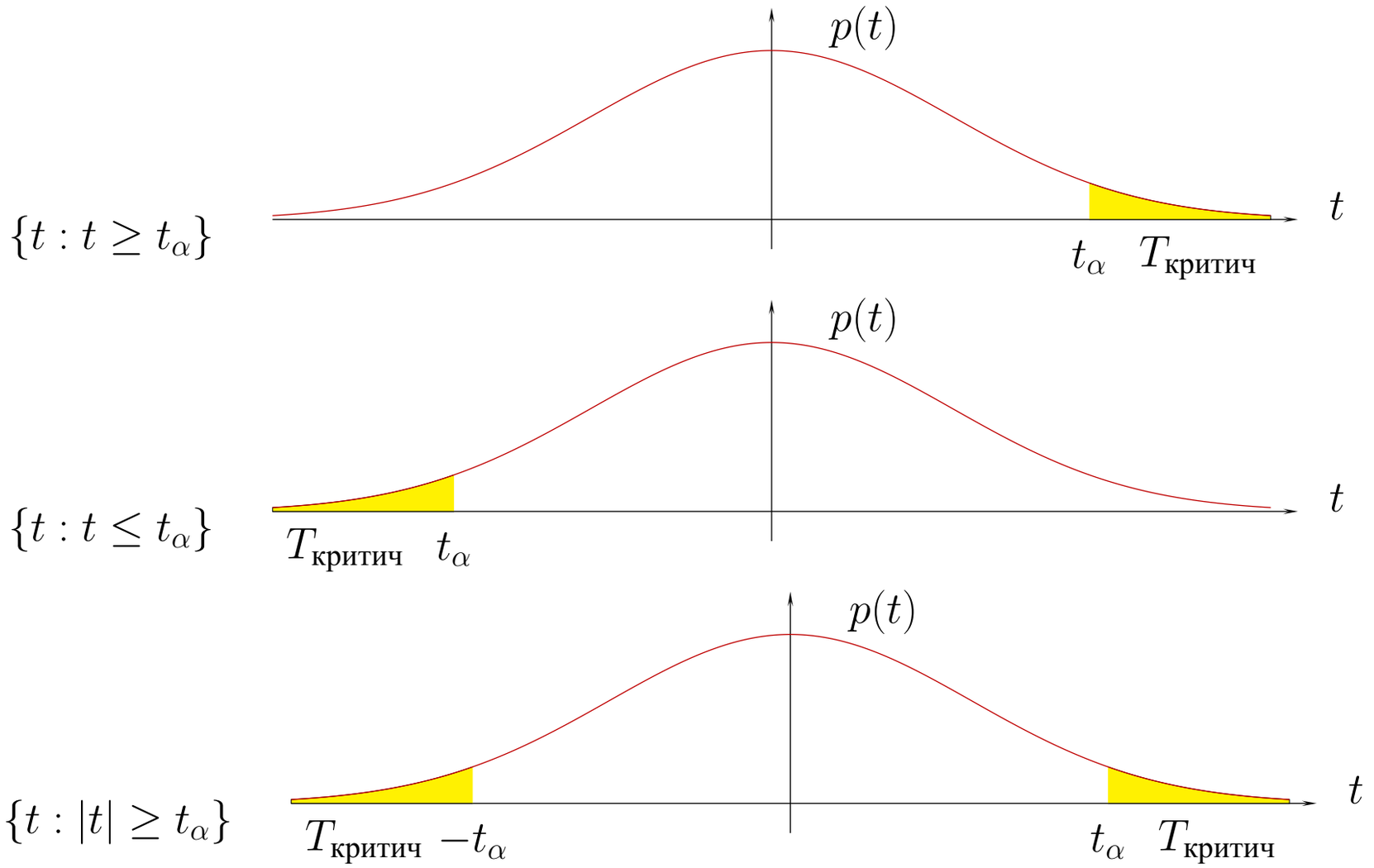
$$\mathsf{Pr}\left(t \in T_{\text{критич}} \,\middle|\, H_0\right) = \alpha.$$

Если $t^* \in T_{\text{критич}}$, то гипотеза H_0 отвергается.

Если $t^* \notin T_{\text{критич}}$, то можно считать, что эмпирические данные согласуются с гипотезой H_0, и гипотеза H_0 принимается.



Часто в качестве $T_{\text{критич}}(H_0, \alpha)$ выбираются области вида



Площадь желтой области на каждом графике равна вероятностному уровню значимости α.

4.3.1. p-value

p-valueпринадлежит интервалу $[0, 1]$ и вычисляется по выборочному значению t^* статистики $t(X)$ (и зависит также от критерия и гипотезы H_0, но не зависит от α).

p-value— это минимальное значение уровня значимости α, при котором данное значение t^* статистики $t(X)$ принадлежит критической области $T_{\text{критич}}(H_0, \alpha)$:

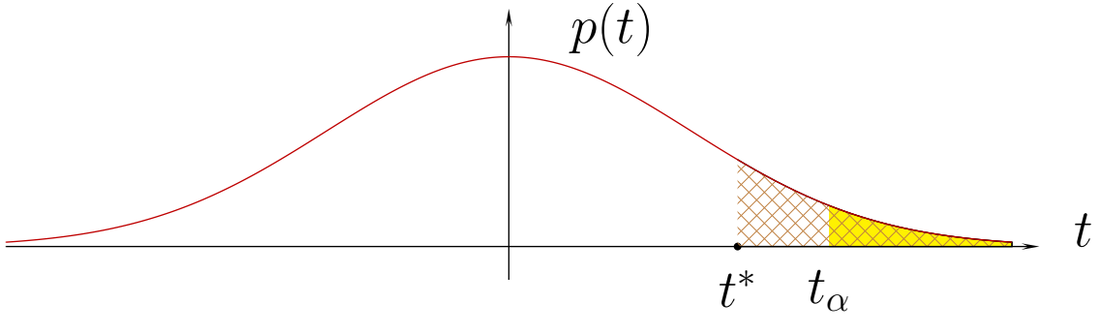
$$\text{p-value}(t^*, H_0) = \inf \{\alpha : \ t^* \in T_{\text{критич}}(H_0, \alpha)\}$$

Из этого определение вытекает следующее основное свойство p-value.

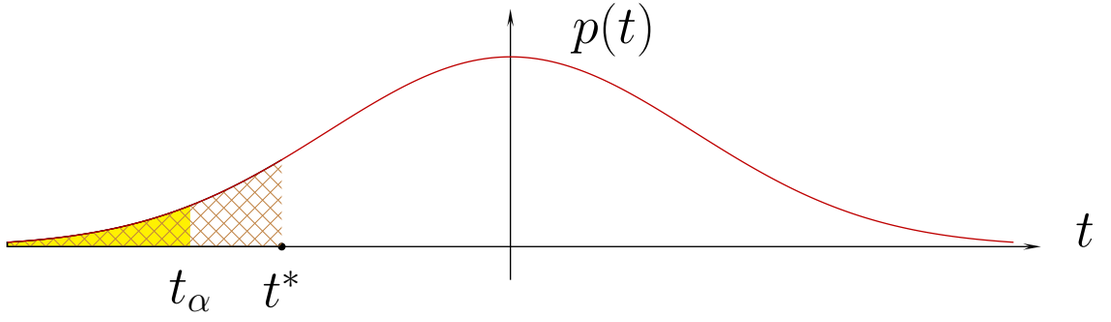
Если p-valueменьше α, то гипотеза H_0 отвергается.

Иначе гипотезу H_0 отвергнуть нельзя.

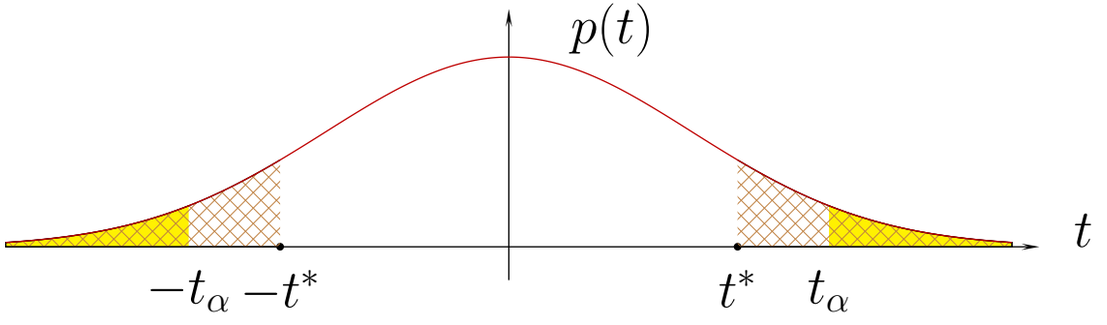
Если $T_{\text{критич}} = \{t : t \geq t_\alpha\}$, то p-value $= \Pr\{t(X) \geq t^*\}$



Если $T_{\text{критич}} = \{t : t \leq t_\alpha\}$, то p-value $= \Pr\{t(X) \leq t^*\}$



Если $T_{\text{критич}} = \{t : |t| \geq t_\alpha\}$, то p-value $= \Pr\{|t(X)| \geq |t^*|\}$



p-value= площади заштрихованной фигуры, α = площади желтой фигуры