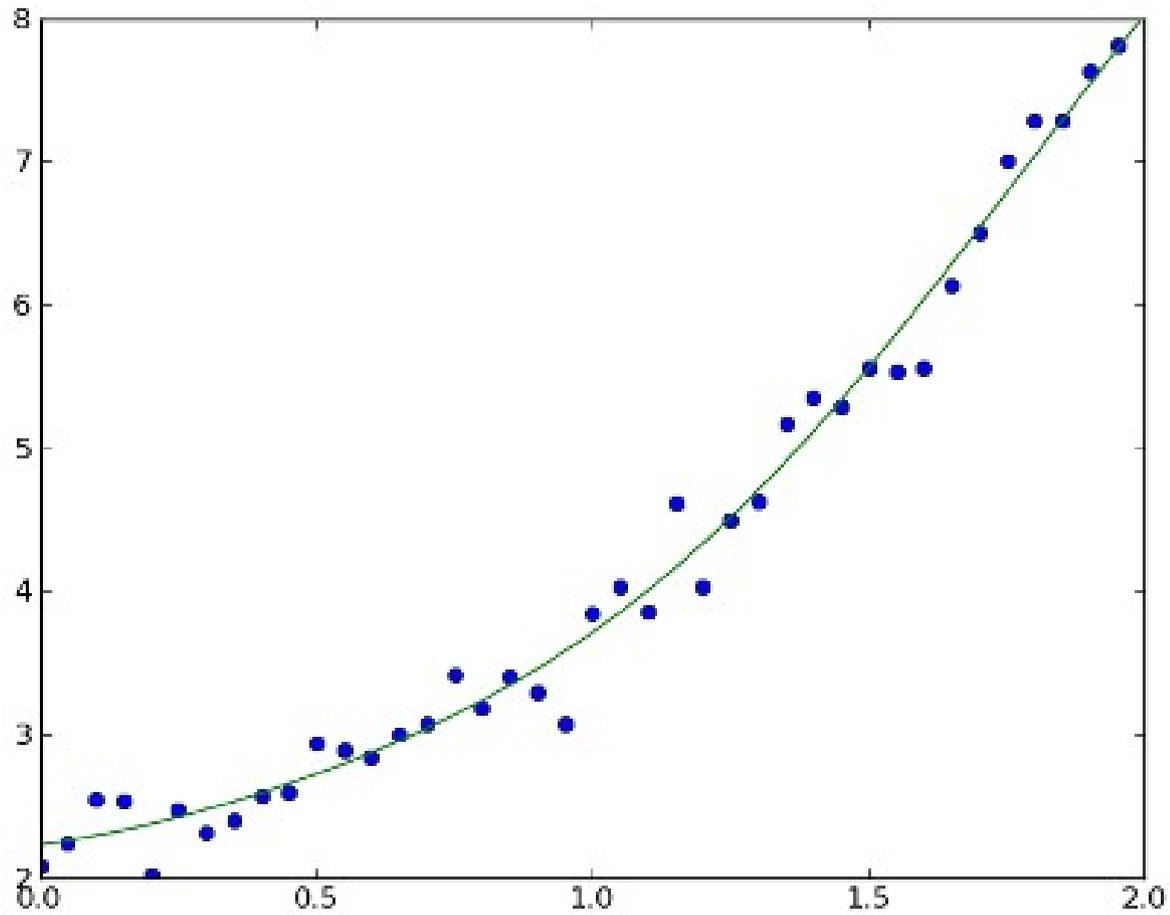


# Машинное обучение

## Методы восстановления регрессии



# Содержание лекции

- Метод наименьших квадратов
- Геометрический смысл
- Регуляризация
- Сингулярное разложение
- Непараметрическая регрессия

# Метод наименьших квадратов

- $X = \mathbb{R}^n, Y = \mathbb{R}$
- Модель:  $a(x) = f(x, \alpha)$
- Метод наименьших квадратов (МНК):

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} w_i (f(x_i, \alpha) - y_i)^2 \rightarrow \min_{\alpha}$$

- $w_i$  — вес, степень важности  $i$ -го объекта

# Многомерная линейная регрессия

- $f_1(x), \dots, f_n(x)$  — числовые признаки;
- Модель:

$$f(x, \alpha) = \sum_{j=1}^n \alpha_j f_j(x), \quad \alpha \in \mathbb{R}^n$$

- Матричная форма:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad y_{\ell \times 1} = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \quad \alpha_{n \times 1} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix}$$

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} (f(x_i, \alpha) - y_i)^2 = \|F\alpha - y\|^2 \rightarrow \min_{\alpha}$$

# Нормальная система уравнений

- Необходимое условие минимума

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^T(F\alpha - y) = 0$$

$$F^T F \alpha = F^T y$$

- где  $F^T F$  — ковариационная матрица  $n \times n$  набора признаков  $f_1, \dots, f_n$
- Решение системы:  $\alpha^* = (F^T F)^{-1} F^T y = F^+ y$
- Значение функционала:  $Q(\alpha^*) = \|P_F y - y\|^2$   
где  $P_F$  - проекционная матрица

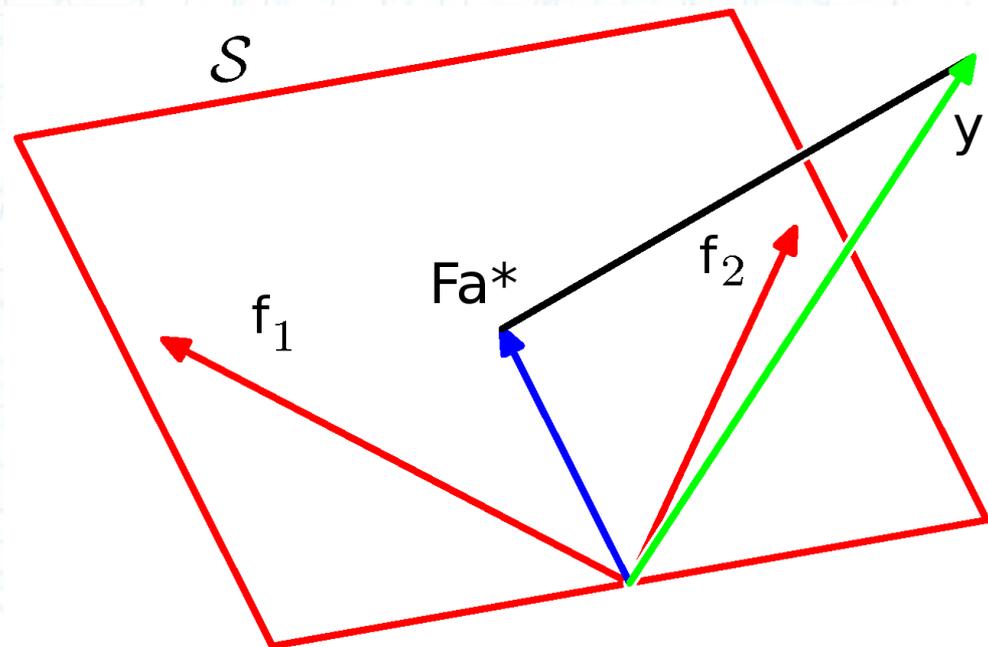
$$P_F = FF^+ = F(F^T F)^{-1} F^T$$

# Геометрический смысл

- Любой вектор вида  $y = F\alpha$  – линейная комбинация признаков

$$\|F\alpha - y\|^2 \rightarrow \min_{\alpha}$$

- $F\alpha^*$  – аппроксимация вектора  $y$  с наименьшим квадратом тогда и только тогда, когда  $F\alpha^*$  – проекция  $y$  на подпространство признаков



# Вероятностный подход

- Модель данных с некоррелированным гауссовским шумом:

$$y(x_i) = f(x_i, \alpha) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad i = 1, \dots, \ell.$$

- Принцип максимума правдоподобия:

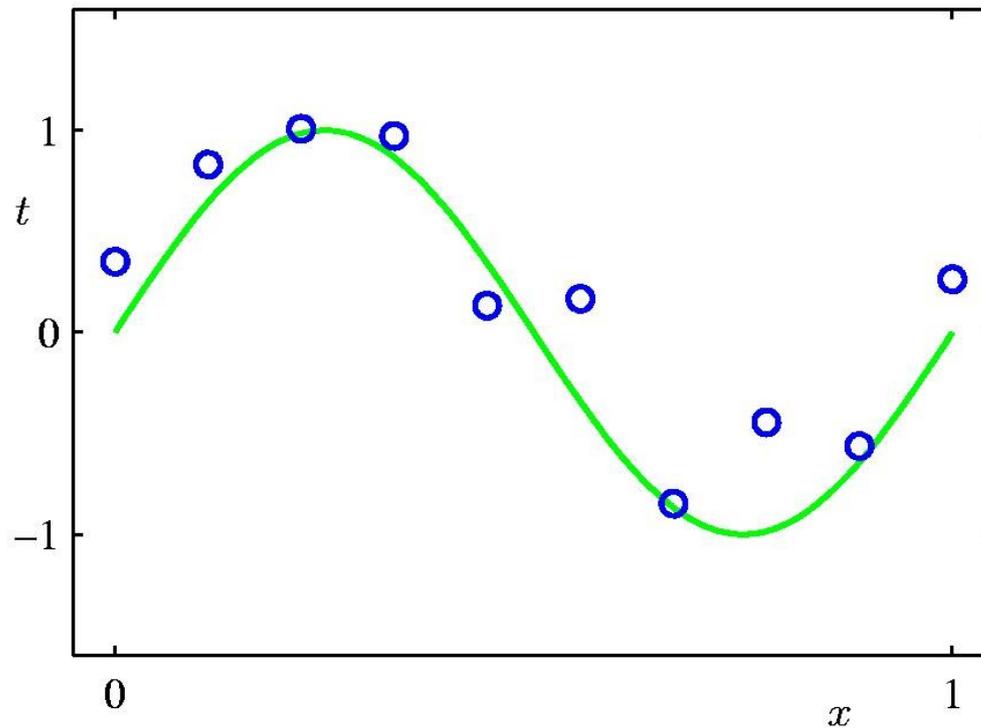
$$L(\varepsilon_1, \dots, \varepsilon_\ell | \alpha) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2} \varepsilon_i^2\right) \rightarrow \max_{\alpha}$$

$$-\ln L(\varepsilon_1, \dots, \varepsilon_\ell | \alpha) = \text{const}(\alpha) + \frac{1}{2} \sum_{i=1}^{\ell} \frac{1}{\sigma_i^2} (f(x_i, \alpha) - y_i)^2 \rightarrow \min_{\alpha}$$

- В итоге пришли к МНК

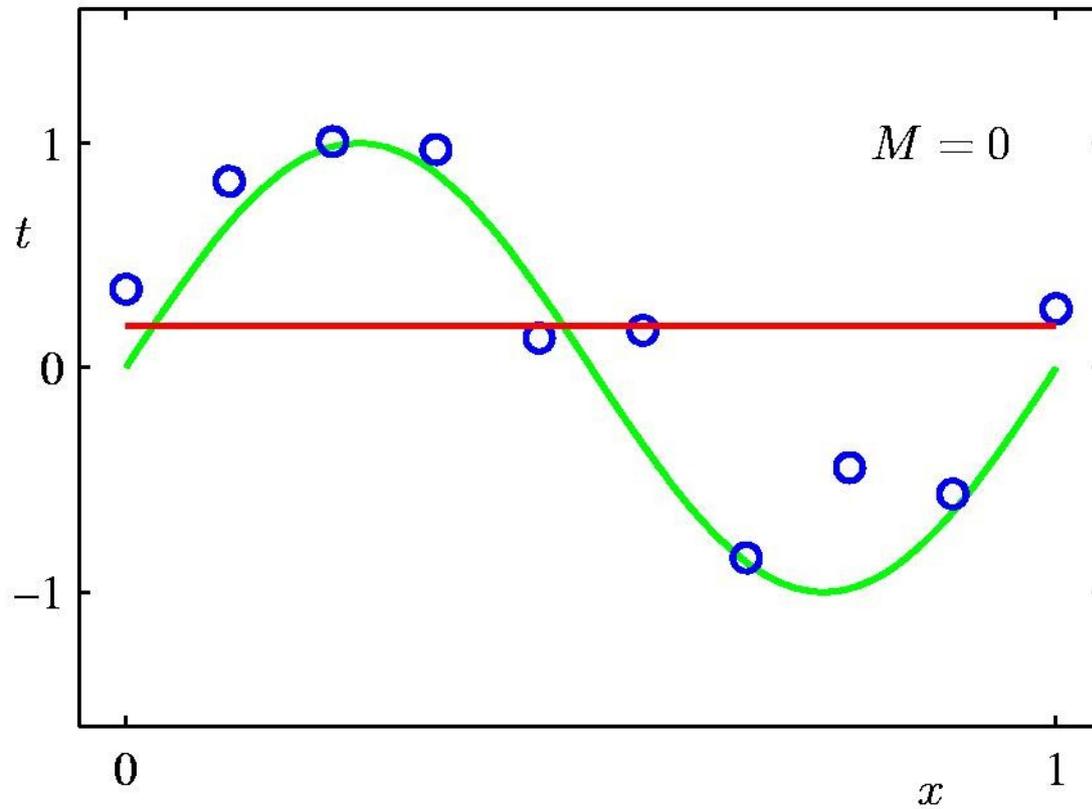
# Пример – приближение многочленами

Данные:  $\sin(x)$  + случайный шум

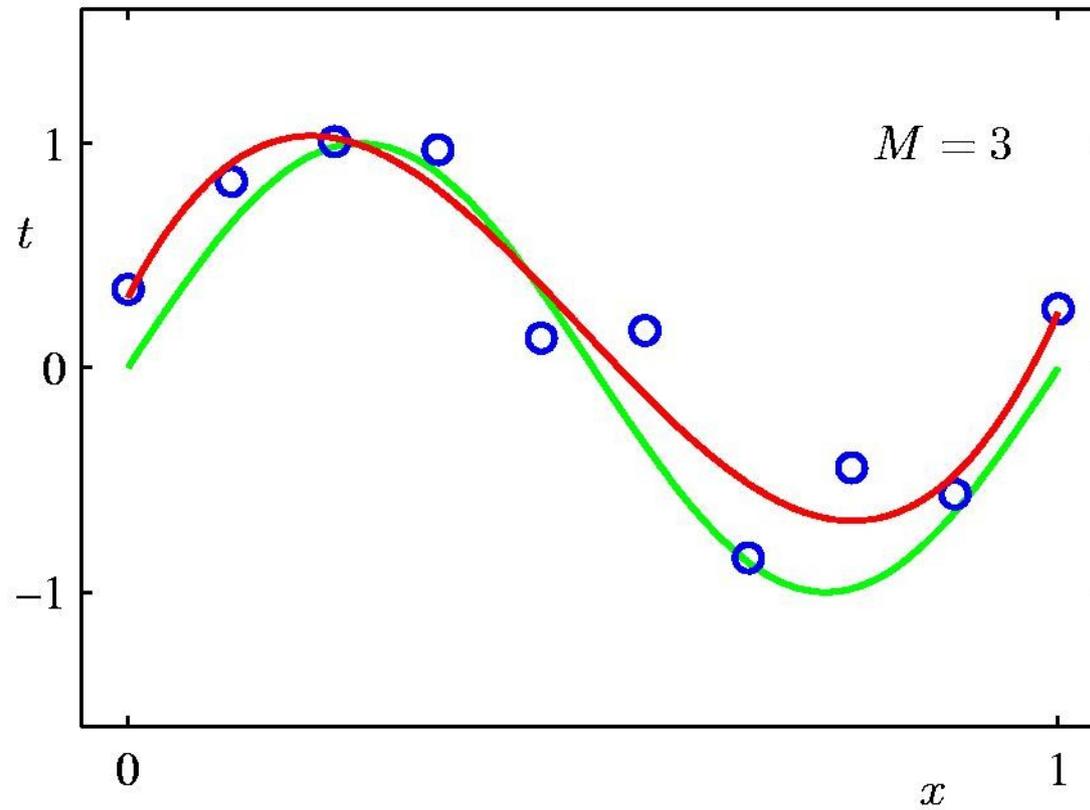


$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

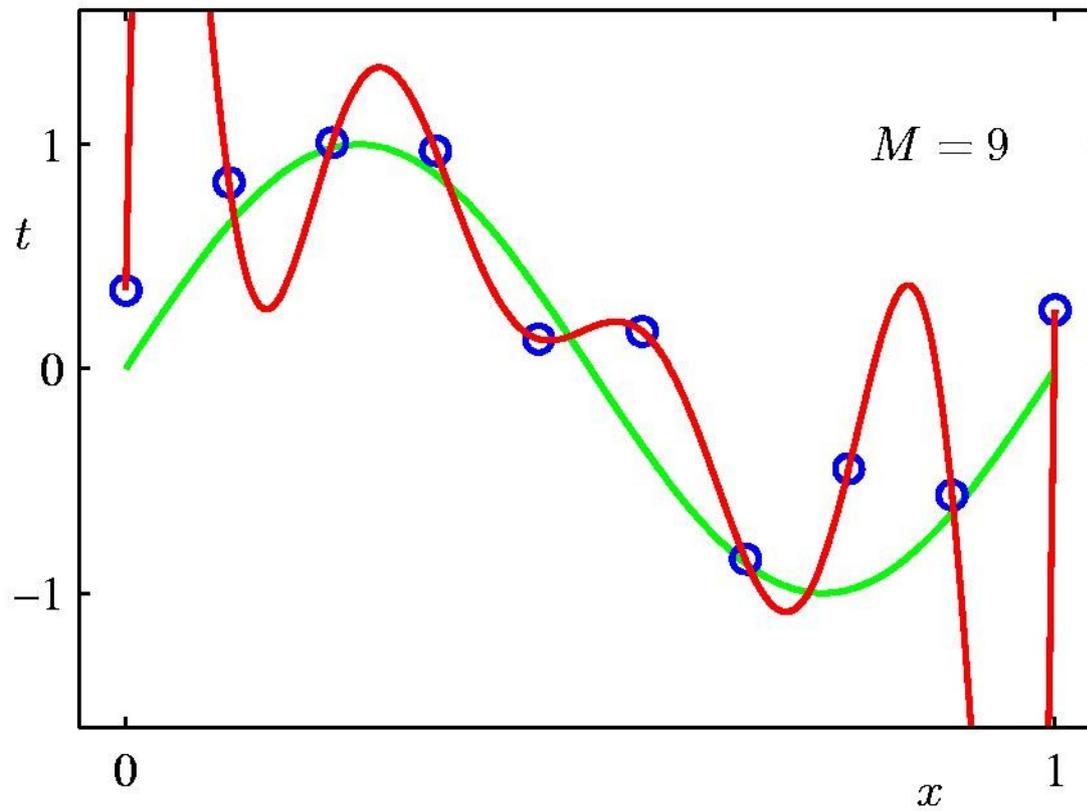
# Многочлен степени 0



# Многочлен степени 3



# Многочлен степени 9



# Коэффициенты многочленов

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
$w_0^*$	0.19	0.82	0.31	0.35
$w_1^*$		-1.27	7.99	232.37
$w_2^*$			-25.43	-5321.83
$w_3^*$			17.37	48568.31
$w_4^*$				-231639.30
$w_5^*$				640042.26
$w_6^*$				-1061800.52
$w_7^*$				1042400.18
$w_8^*$				-557682.99
$w_9^*$				125201.43

В переобученном случае наблюдаются аномально большие коэффициенты многочлена. Выход - регуляризация

# Гребневая регрессия

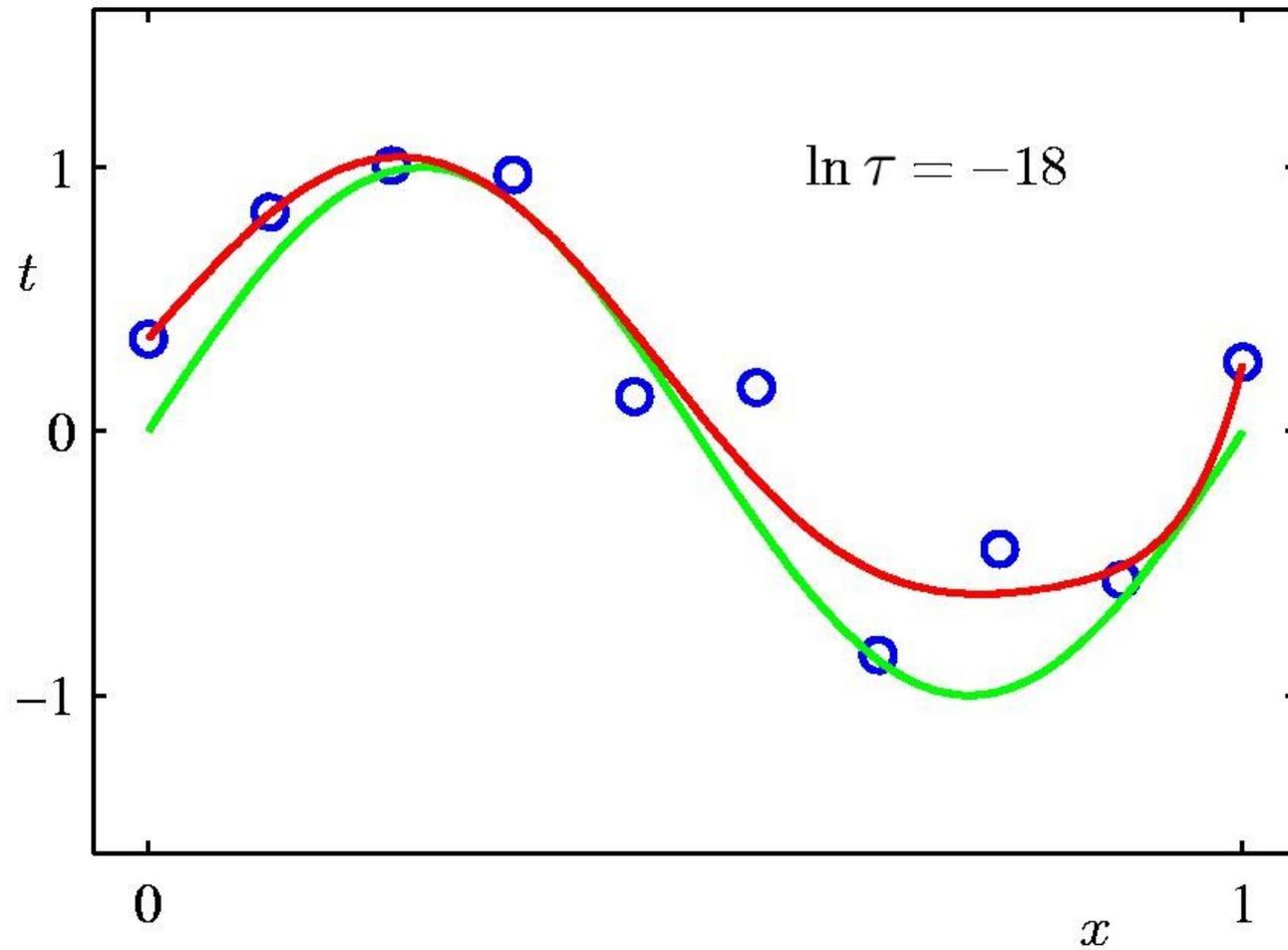
- Штраф за увеличение нормы вектора весов  $\|\alpha\|$  :

$$Q_{\tau}(\alpha) = \|F\alpha - y\|^2 + \frac{1}{\sigma} \|\alpha\|^2 \quad \tau = \frac{1}{\sigma}$$

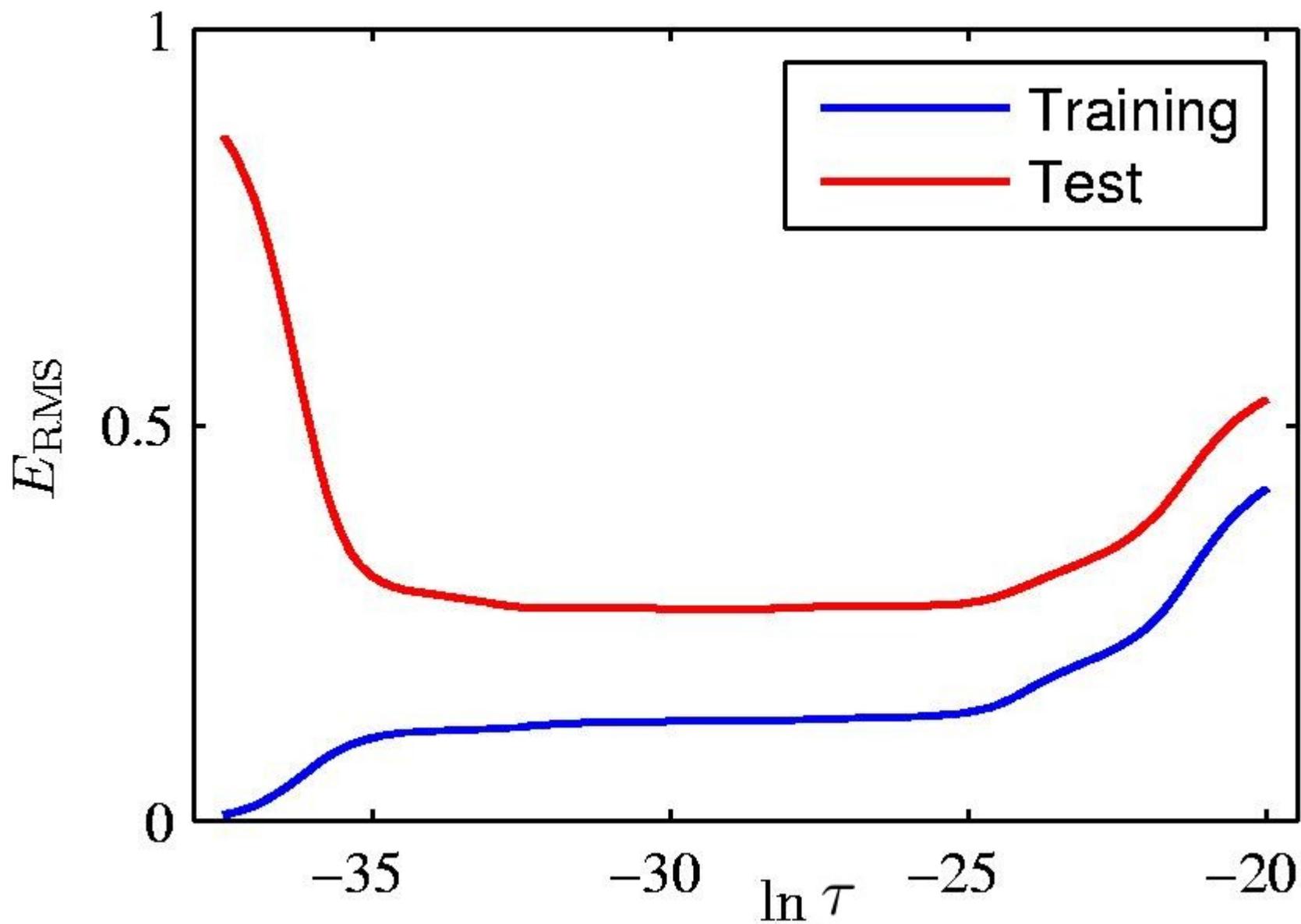
- Модифицированное МНК-решение ( $\tau I_n$  — «гребень»):

$$\alpha_{\tau}^* = (F^T F + \tau I_n)^{-1} F^T y$$

# Многочлен степени 9 с регуляризацией



# Гребневая регрессия



# Сингулярное разложение

Произвольная  $\ell \times n$ -матрица представима в виде *сингулярного разложения* (singular value decomposition, SVD):

$$F = VDU^T.$$

**Основные свойства сингулярного разложения:**

- 1  $\ell \times n$ -матрица  $V = (v_1, \dots, v_n)$  ортогональна,  $V^T V = I_n$ , столбцы  $v_j$  — собственные векторы матрицы  $FF^T$ ;
- 2  $n \times n$ -матрица  $U = (u_1, \dots, u_n)$  ортогональна,  $U^T U = I_n$ , столбцы  $u_j$  — собственные векторы матрицы  $F^T F$ ;
- 3  $n \times n$ -матрица  $D$  диагональна,  $D = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$ ,  $\lambda_j \geq 0$  — собственные значения матриц  $F^T F$  и  $FF^T$ .

# Решение МНК через сингулярное разложение

$$\alpha^* = (F^T F)^{-1} F^T y = F^+ y$$

$$F^+ = (UDV^T VDU^T)^{-1} UDV^T = UD^{-1} V^T = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j v_j^T$$

$$\alpha^* = F^+ y = UD^{-1} V^T y = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y)$$

$$F\alpha^* = P_F y = (VDU^T) UD^{-1} V^T y = VV^T y = \sum_{j=1}^n v_j (v_j^T y)$$

$$\|\alpha^*\|^2 = \|D^{-1} V^T y\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j} (v_j^T y)^2$$

# Проблема мультиколлинеарности

- Число обусловленности  $n \times n$ -матрицы  $\Sigma$ :

$$\mu(\Sigma) = \|\Sigma\| \|\Sigma^{-1}\| = \frac{\max_{u: \|u\|=1} \|\Sigma u\|}{\min_{u: \|u\|=1} \|\Sigma u\|} = \frac{\lambda_{\max}}{\lambda_{\min}}$$

- При умножении обратной матрицы на вектор,  $z = \Sigma^{-1} u$ , относительная погрешность усиливается в  $\mu(\Sigma)$  раз:

$$\frac{\|\delta z\|}{\|z\|} \leq \mu(\Sigma) \frac{\|\delta u\|}{\|u\|}$$

# Проблема мультиколлинеарности

- Если матрица  $\Sigma = F^T F$  плохо обусловлена, то:
  - решение становится неустойчивым и неинтерпретируемым,  $\|\alpha^*\|$  велико;
  - возникает переобучение:  
на обучении  $Q(\alpha^*, X^\ell) = \|F\alpha^* - y\|^2$  мало  
на контроле  $Q(\alpha^*, X^k) = \|F'\alpha^* - y'\|^2$  велико
- Стратегии устранения мультиколлинеарности и переобучения:
  - регуляризация
  - отбор признаков
  - преобразование признаков

# Регуляризация с точки зрения SVD-разложения

$$\alpha_{\tau}^* = (F^T F + \tau I_n)^{-1} F^T y$$

$$\alpha_{\tau}^* = U(D^2 + \tau I_n)^{-1} D V^T y = \sum_{j=1}^n \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^T y)$$

Без регуляризации:

$$\alpha^* = F^+ y = U D^{-1} V^T y = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y)$$

# Отбор признаков

- LASSO — Least Absolute Shrinkage and Selection Operator

$$\begin{cases} Q(\alpha) = \|F\alpha - y\|^2 \rightarrow \min_{\alpha}; \\ \sum_{j=1}^n |\alpha_j| \leq \kappa; \end{cases}$$

- Чем меньше  $\kappa$ , тем больше нулевых  $\alpha_j$

# Метод главных компонент (РСА)

- $f_1(x), \dots, f_n(x)$  — исходные числовые признаки;
- $g_1(x), \dots, g_m(x)$  — новые числовые признаки,  $m < n$ ;
- Требование: старые признаки должны линейно восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^m g_s(x) u_{js}, \quad j = 1, \dots, n, \quad \forall x \in X$$

как можно точнее на обучающей выборке:

$$\sum_{i=1}^{\ell} \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 \rightarrow \min_{\{g_s(x_i)\}, \{u_{js}\}}$$

# Постановка задачи РСА в матричной форме

$$\hat{F} = GU^T \stackrel{\text{ХОТИМ}}{\approx} F$$

Найти: и новые признаки  $G$ , и преобразование  $U$ :

$$\sum_{i=1}^{\ell} \sum_{j=1}^n (\hat{f}_j(x_i) - f_j(x_i))^2 = \|GU^T - F\|^2 \rightarrow \min_{G,U}$$

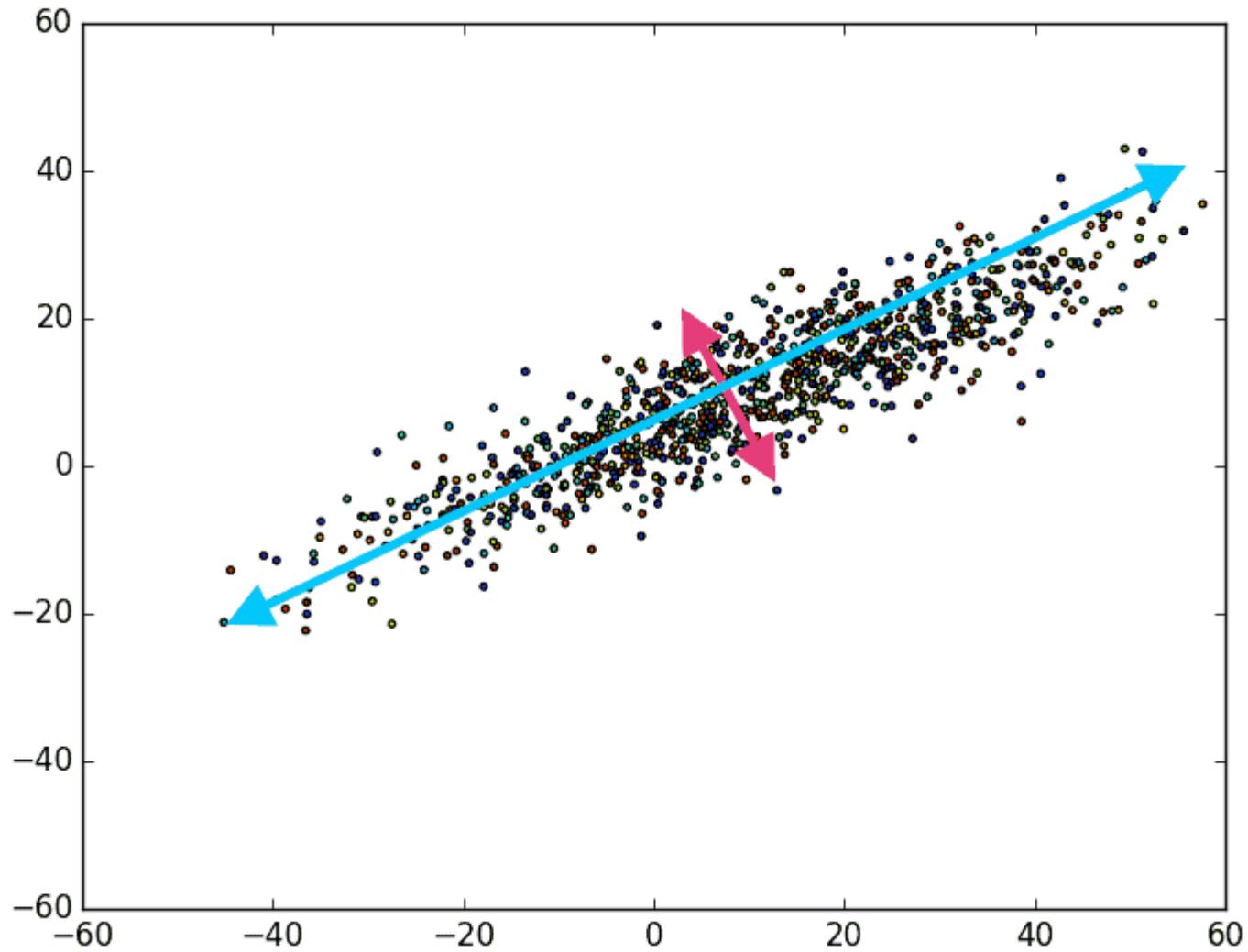
# Теорема

Если  $m \leq \text{rk } F$ , то минимум  $\|GU^T - F\|^2$  достигается, когда столбцы  $U$  — это с.в. матрицы  $F^T F$ , соответствующие  $m$  максимальным с.з.  $\lambda_1, \dots, \lambda_m$ , а матрица  $G = FU$ .

При этом:

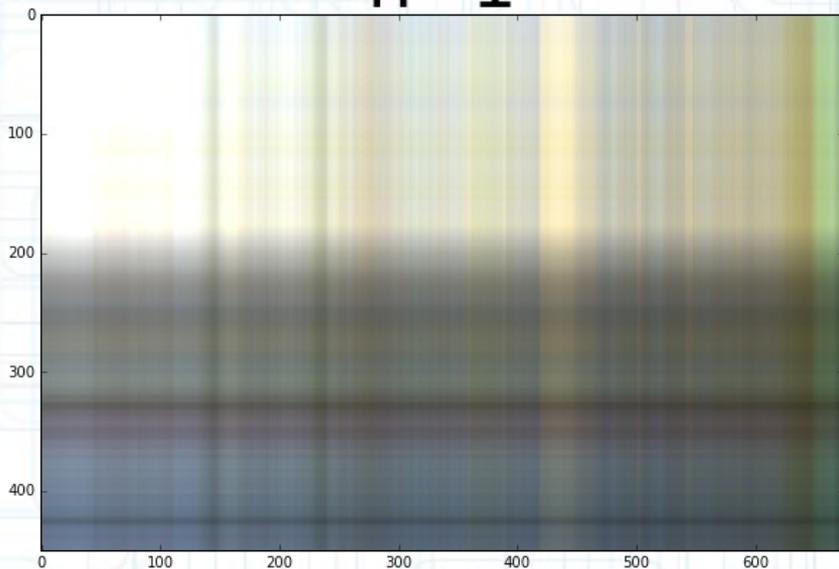
- 1 матрица  $U$  ортонормирована:  $U^T U = I_m$ ;
- 2 матрица  $G$  ортогональна:  $G^T G = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ ;
- 3  $U\Lambda = F^T F U$ ;  $G\Lambda = FF^T G$ ;
- 4  $\|GU^T - F\|^2 = \|F\|^2 - \text{tr } \Lambda = \sum_{j=m+1}^n \lambda_j$ .

# РСА для двумерной матрицы $F$



# Применение RSA к сжатию изображений

$n=1$



$n=10$



$n=30$

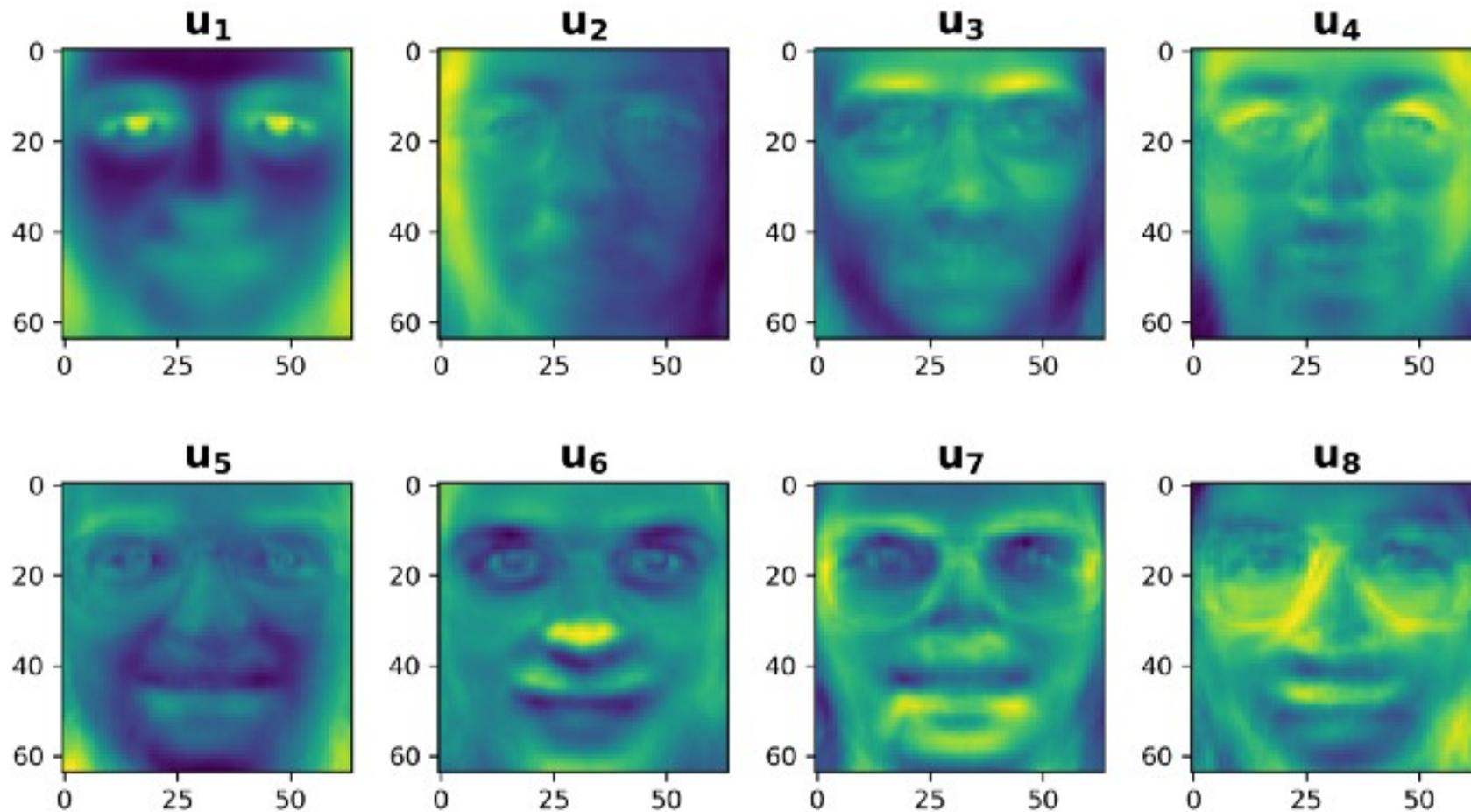


$n=100$

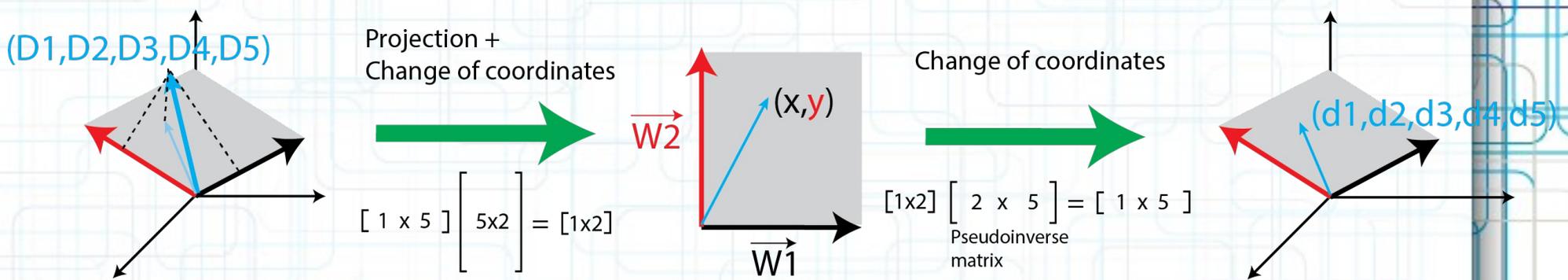
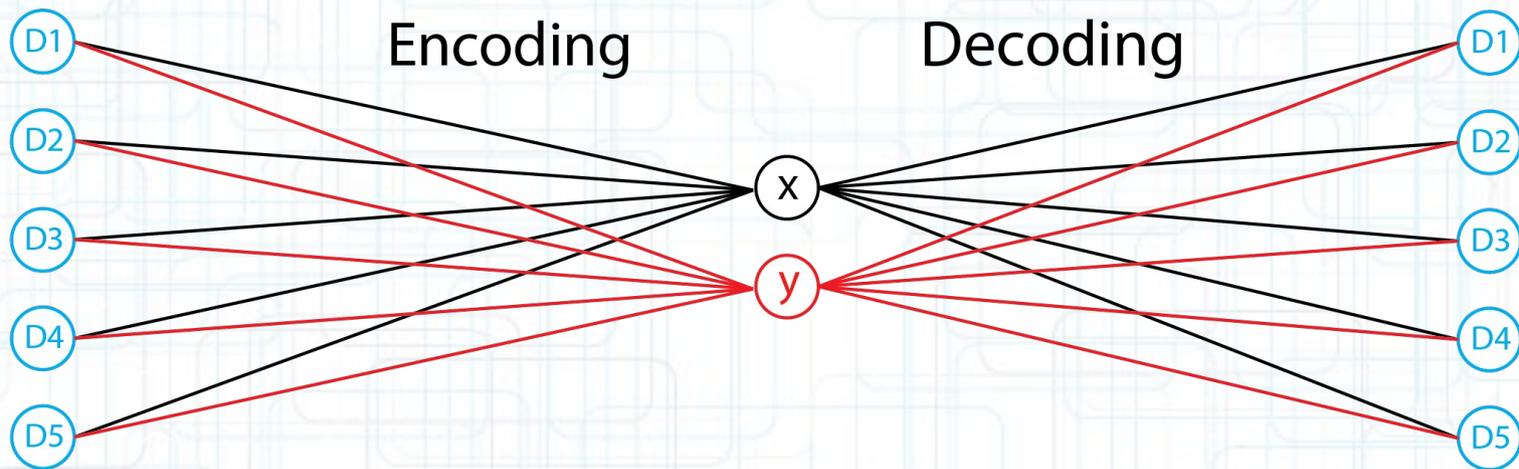


# Главные компоненты датасета Olivetti faces

Показывают ортогональные направления, вдоль которых лица датасета меняются сильнее всего

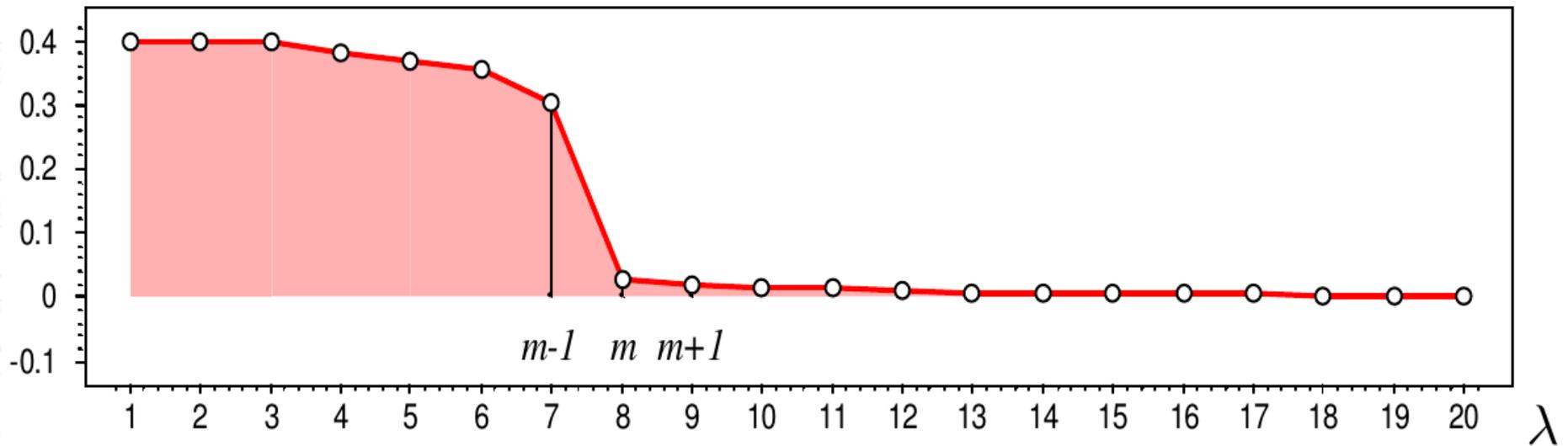


# РСА – простейший пример трансформера



# Сколько главных компонент брать?

- Критерий “крутого склона”:



# Непараметрическая регрессия

- Обычная задача МНК:

$$Q(\alpha, X^\ell) = \sum_{i=1}^{\ell} w_i (f(x_i, \alpha) - y_i)^2 \rightarrow \min_{\alpha}$$

- Приближение константой  $f(x, \alpha) = \alpha$  в окрестности  $x \in X$

$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x) (\alpha - y_i)^2 \rightarrow \min_{\alpha \in \mathbb{R}}$$

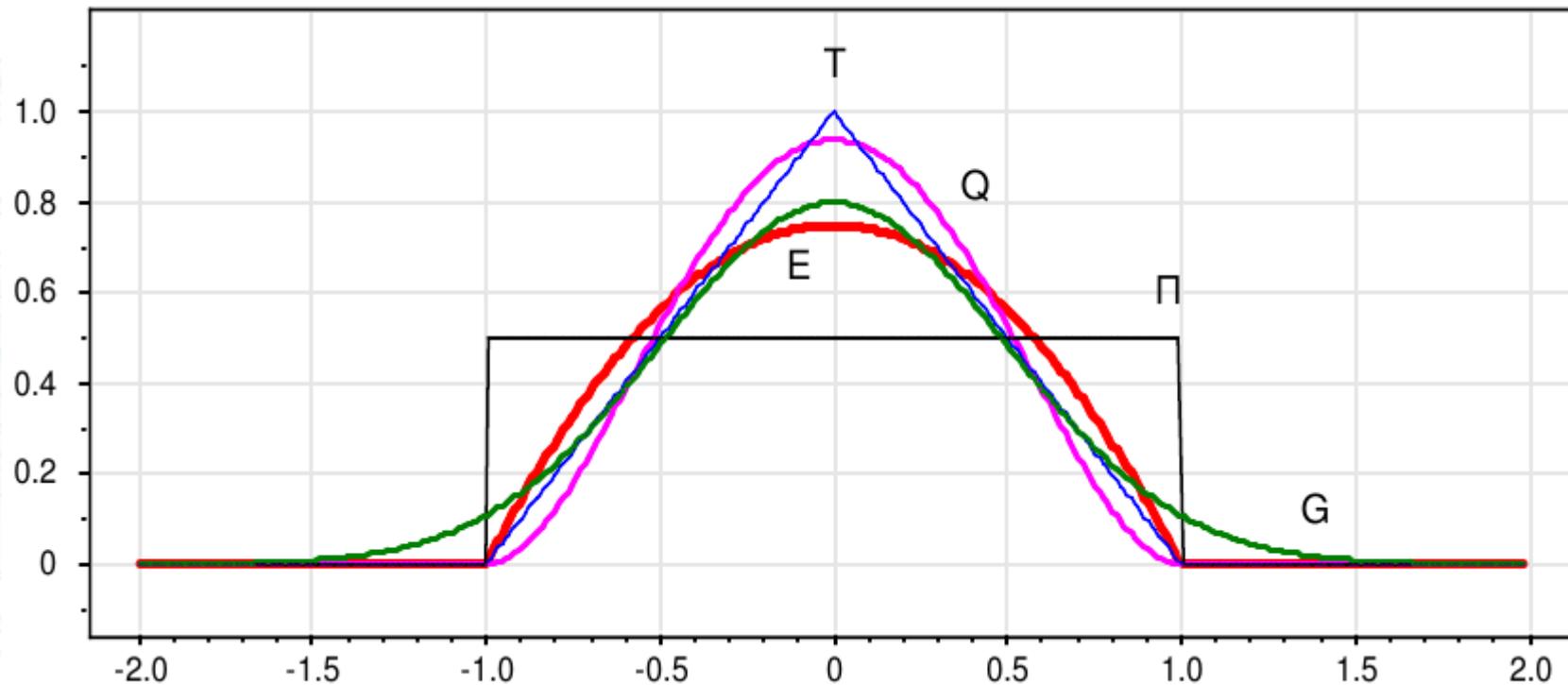
$$w_i(x) = K \left( \frac{\rho(x, x_i)}{h} \right) - \text{веса объектов } x_i$$

относительно  $x$ ;

# Формула ядерного сглаживания Надарая–Ватсона

$$a_h(x; X^\ell) = \frac{\sum_{i=1}^{\ell} y_i w_i(x)}{\sum_{i=1}^{\ell} w_i(x)} = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)}$$

# Часто используемые ядра



$\Pi(r) = [ |r| \leq 1 ]$  — прямоугольное

$T(r) = (1 - |r|) [ |r| \leq 1 ]$  — треугольное

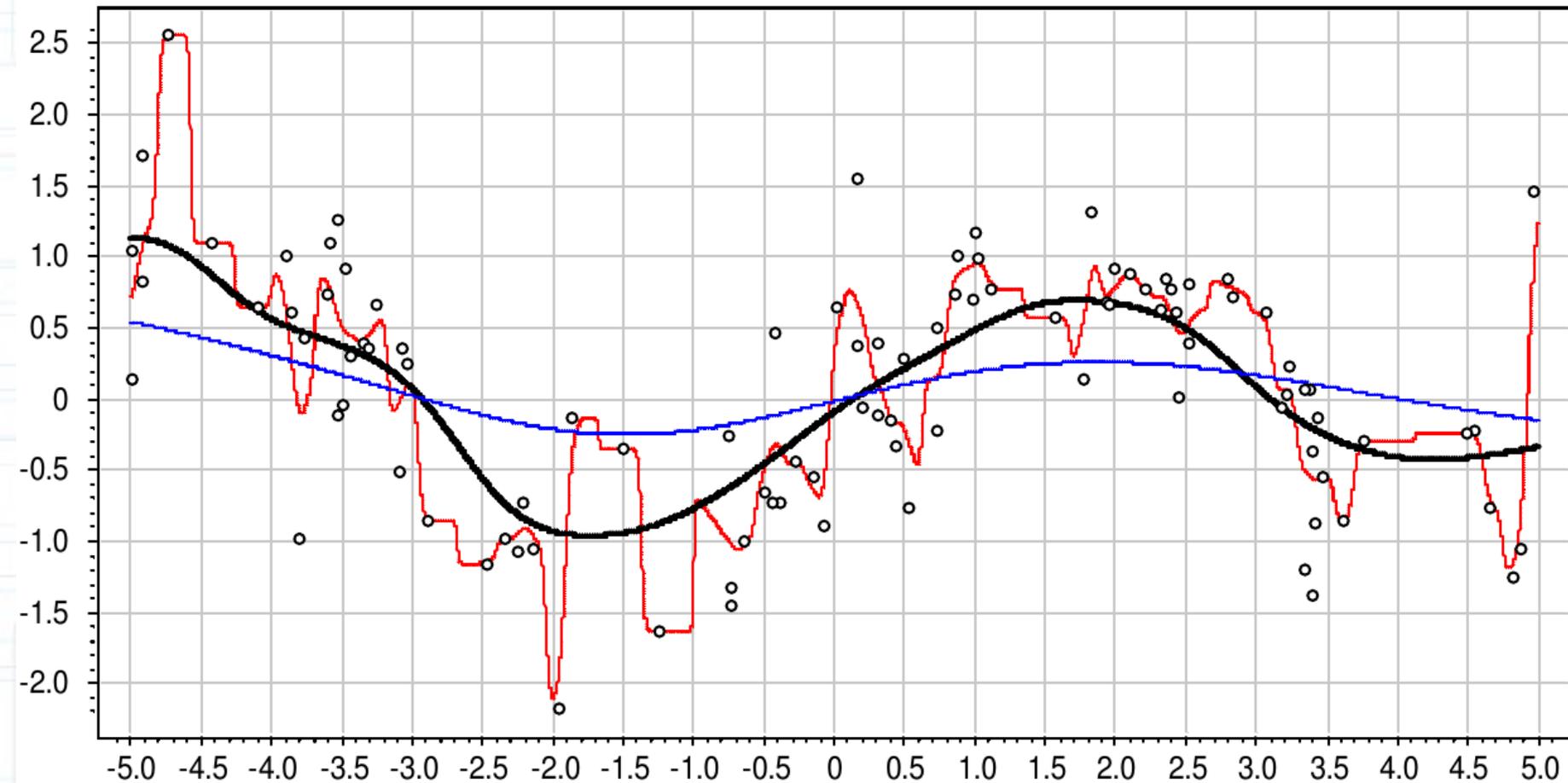
$E(r) = (1 - r^2) [ |r| \leq 1 ]$  — квадратичное (Епанечникова)

$Q(r) = (1 - r^2)^2 [ |r| \leq 1 ]$  — четвертое

$G(r) = \exp(-2r^2)$  — гауссовское

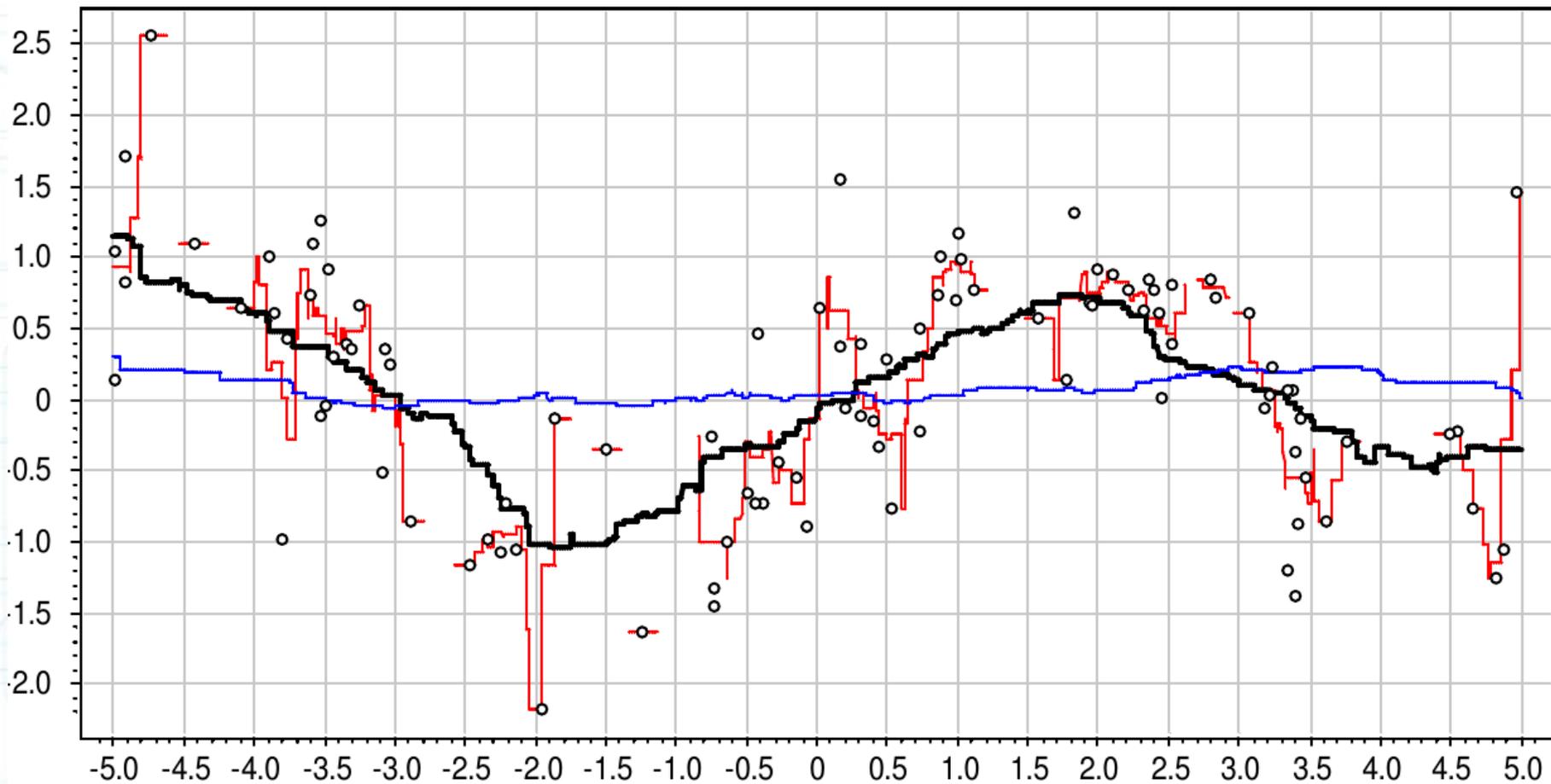
# Выбор ширины окна и ядра

- $h \in \{0.1, 1.0, 3.0\}$ , гауссовское ядро



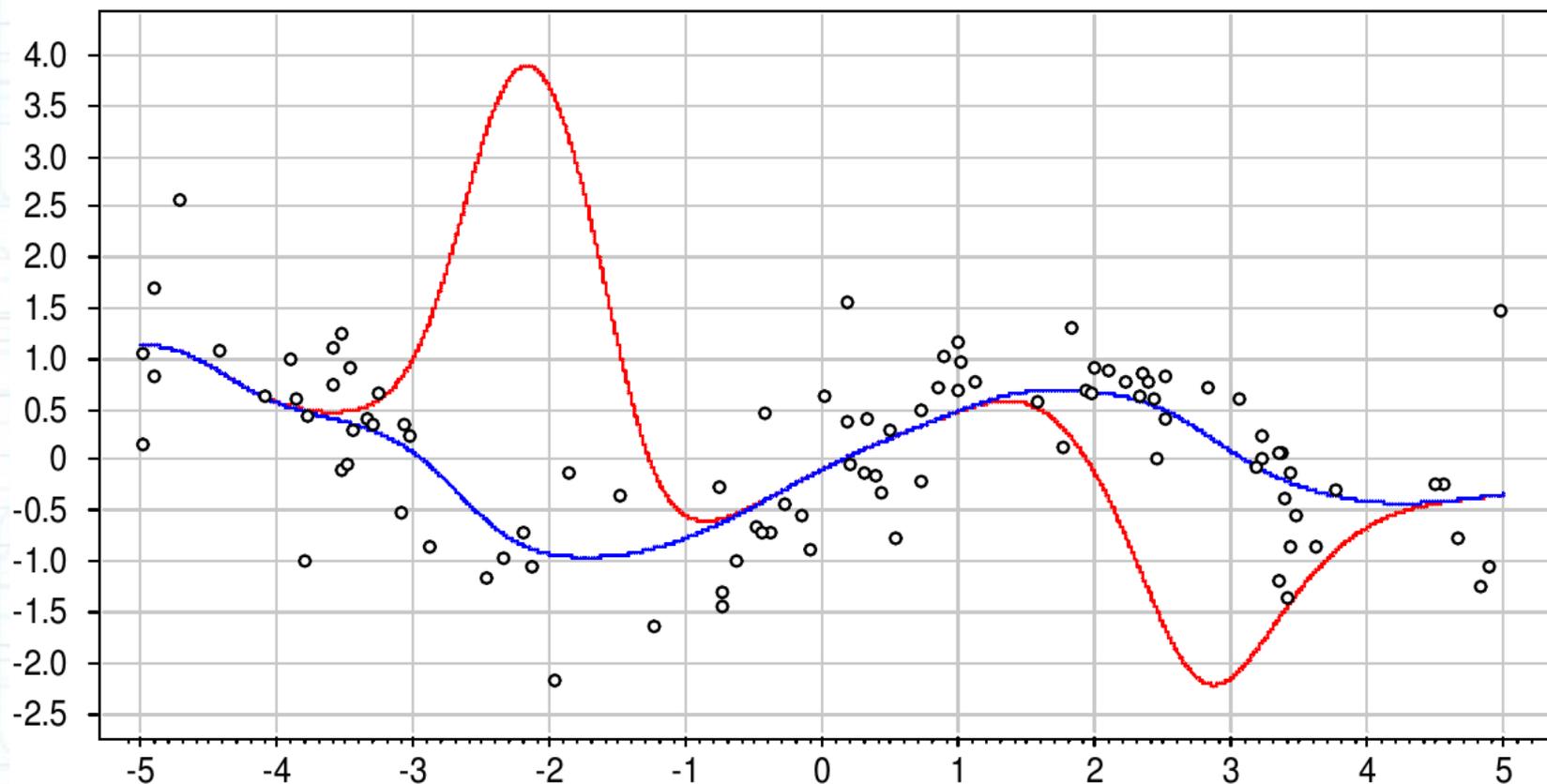
# Выбор ширины окна и ядра

- $h \in \{0.1, 1.0, 3.0\}$ , прямоугольное ядро



# Проблема выбросов

- $n = 100$ ,  $h = 1.0$ , гауссовское ядро  $K(r) = \exp(-2r^2)$
- **Две точки - выбросы с ординатами 40 и -40**
- **Синяя кривая — выбросов нет**



# Локально взвешенное сглаживание

- Основная идея: чем больше величина ошибки  $\varepsilon_i = |a_h(x_i; X^\ell \setminus \{x_i\}) - y_i|$  тем больше прецедент  $(x_i, y_i)$  похож на выброс, тем меньше должен быть его вес  $w_i(x)$ .
- Эвристика: домножить веса  $w_i(x)$  на коэффициенты  $\gamma_i = \tilde{K}(\varepsilon_i)$  где  $\tilde{K}$  — ещё одно ядро
- Рекомендация: кватрическое ядро

$$\tilde{K}(\varepsilon) = K_Q\left(\frac{\varepsilon}{6 \operatorname{med}\{\varepsilon_i\}}\right)$$

где  $\operatorname{med}\{\varepsilon_i\}$  — медиана вариационного ряда ошибок.

# Алгоритм LOWESS (LOcally WEighted Scatter plot Smoothing)

**Вход:**  $X^\ell$  — обучающая выборка;

**Выход:** коэффициенты  $\gamma_i$ ,  $i = 1, \dots, \ell$ ;

---

1: инициализация:  $\gamma_i := 1$ ,  $i = 1, \dots, \ell$ ;

2: **повторять**

3: **для всех** объектов  $i = 1, \dots, \ell$

4: вычислить оценки скользящего контроля:

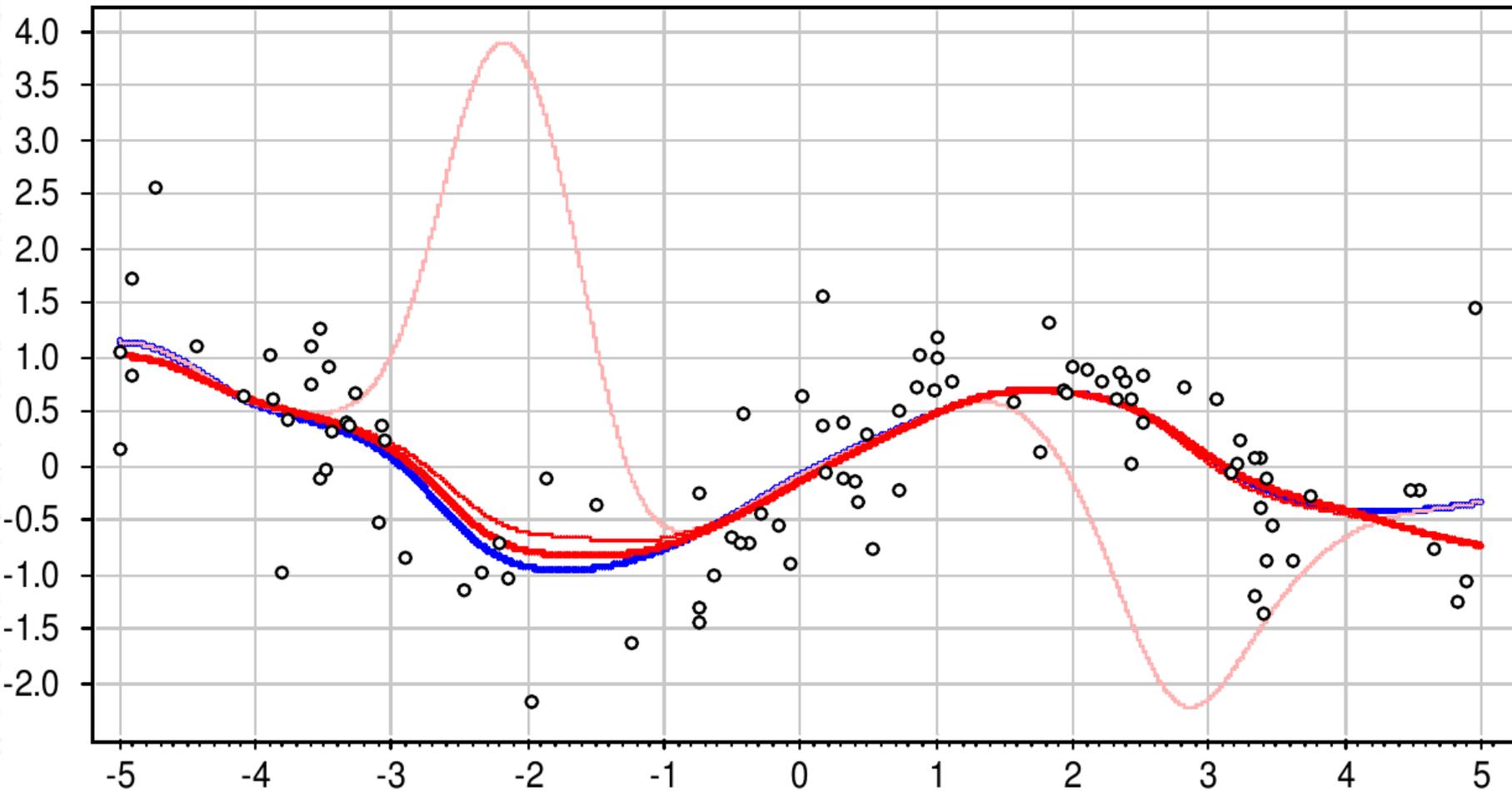
$$a_i := a_h(x_i; X^\ell \setminus \{x_i\}) = \frac{\sum_{j=1, j \neq i}^{\ell} y_j \gamma_j K\left(\frac{\rho(x_i, x_j)}{h(x_i)}\right)}{\sum_{j=1, j \neq i}^{\ell} \gamma_j K\left(\frac{\rho(x_i, x_j)}{h(x_i)}\right)};$$

5: **для всех** объектов  $i = 1, \dots, \ell$

6:  $\gamma_i := \tilde{K}(|a_i - y_i|)$ ;

7: **пока** коэффициенты  $\gamma_i$  не стабилизируются;

# Пример работы LOWESS



# Вопросы для самоконтроля

- Сгенерируйте случайную выборку из 10 объектов с вещественным целевым и одним количественным признаками. Возьмите случайное значение признака  $x$ . Примените, к данной выборке и точке  $x$  формулу ядерного сглаживания Надарая-Ватсона (с некоторым ядром) и предскажите значение целевой функции в точке  $x$

# Вопросы для самоконтроля

- Сгенерируйте случайную выборку из 3 объектов с вещественным целевым и одним количественным признаками. Найдите линейную аппроксимацию неизвестной функции методом наименьших квадратов