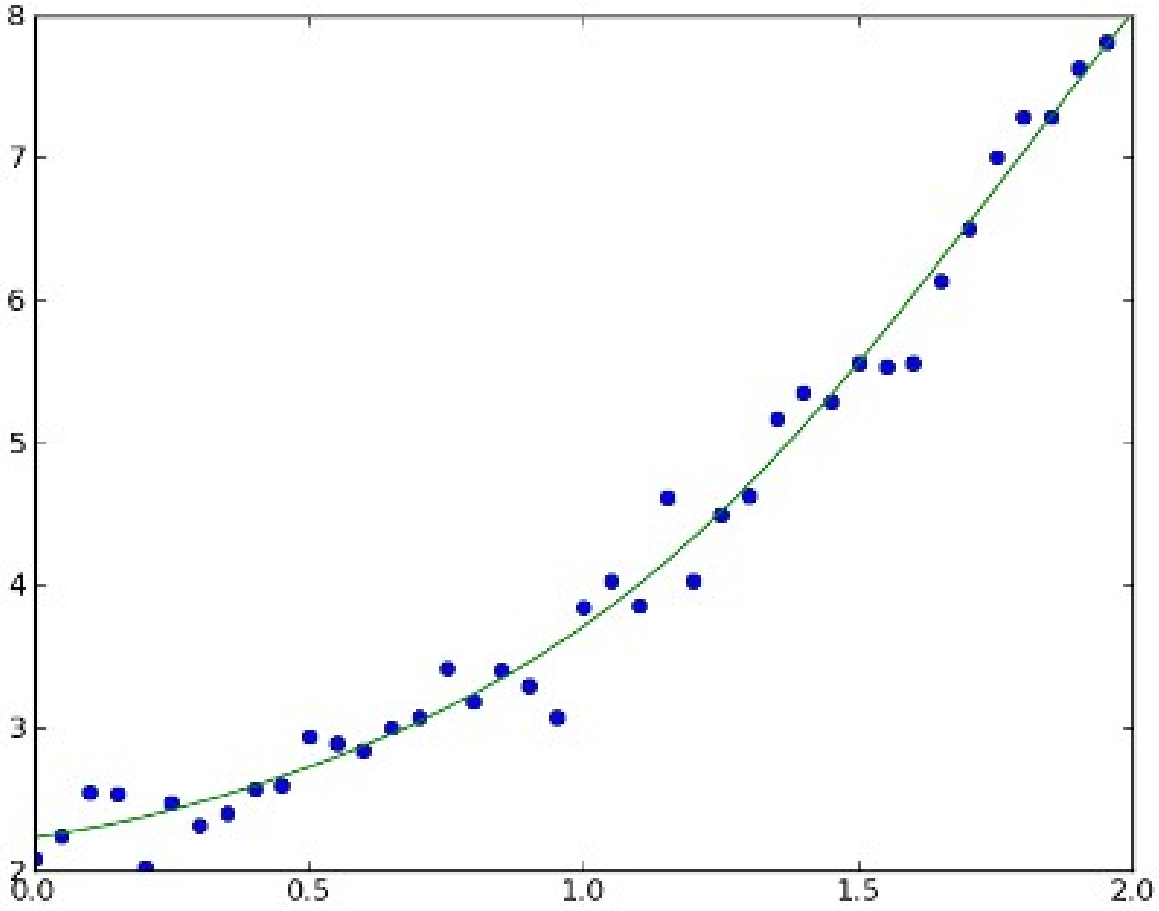
Машинное обучение
Методы восстановления регрессии



Содержание лекции

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

Метод наименьших квадратов

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

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

- w_i — вес, степень важности i-го объекта

Многомерная линейная регрессия

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

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

- Матричная форма:

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}, \quad \underset{\ell\times 1}{y} = \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}, \quad \underset{n\times 1}{\alpha} = \begin{pmatrix} \alpha_1 \\ \ldots \\ \alpha_n \end{pmatrix}$$

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

Нормальная система уравнений

- Необходимое условие минимума

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

$$F^{\intercal}F\alpha = F^{\intercal}y$$

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

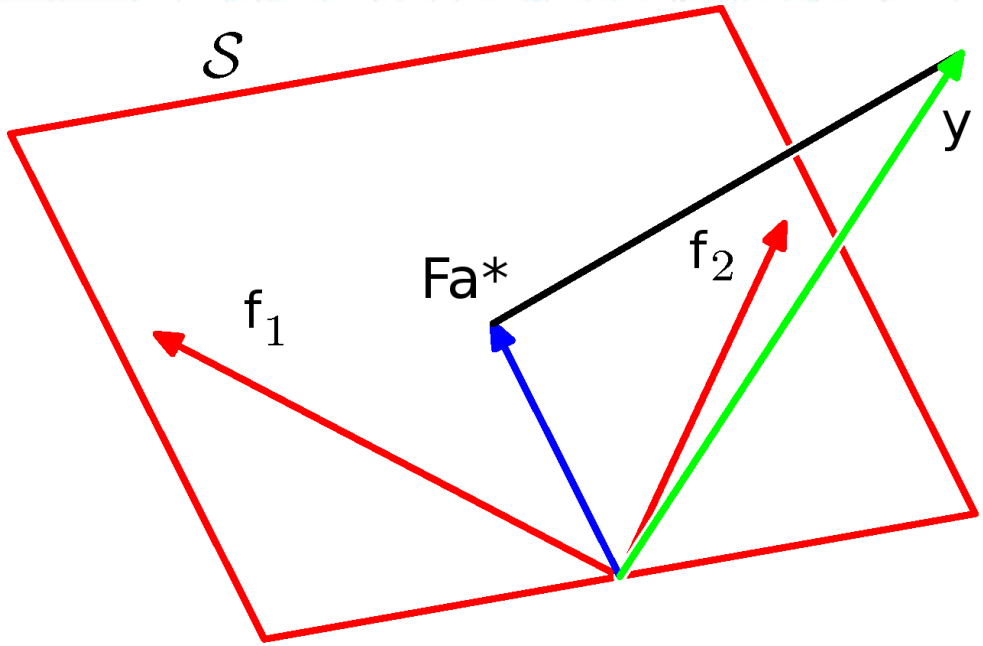
$$P_F = FF^{+} = F(F^{\intercal}F)^{-1}F^{\intercal}$$

Геометрический смысл

- Любой вектор вида y = Fα – линейная комбинация признаков

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

- Fα* – аппроксимация вектора y с наименьшим квадратом тогда и только тогда, когда Fα* - проекция y на подпространство признаков



Вероятностный подход

- Модель данных с некоррелированным гауссовским шумом:

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

- Принцип максимума правдоподобия:

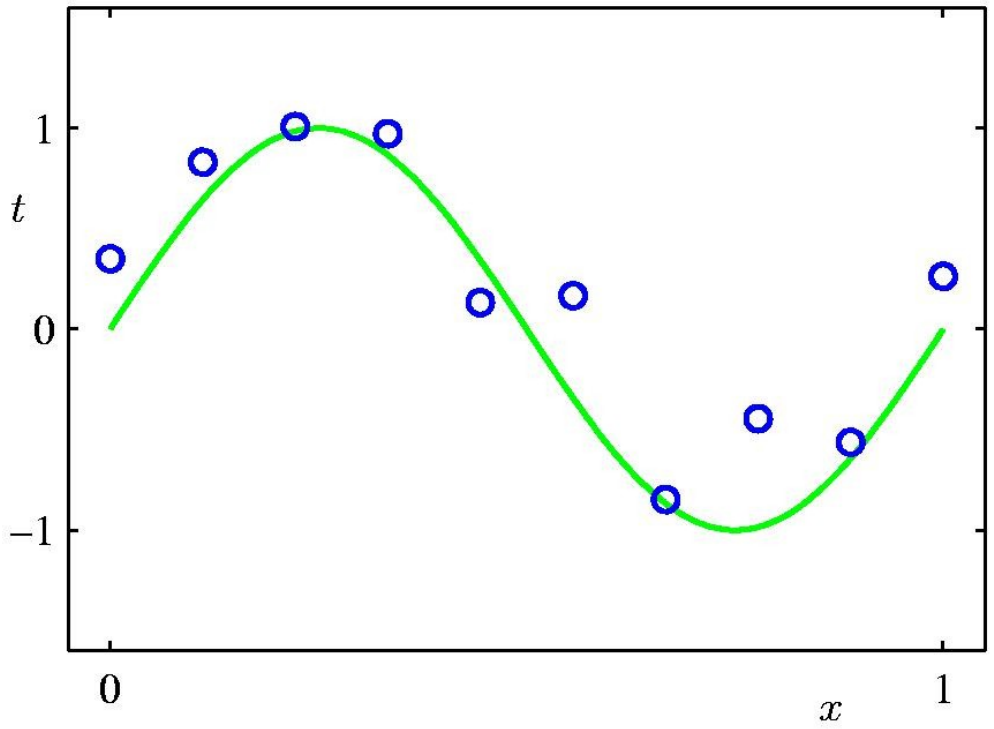
$$L(\varepsilon_1, \ldots, \varepsilon_\ell | \alpha) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\varepsilon_i^2\right) \to \max_\alpha$$

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

- В итоге пришли к МНК

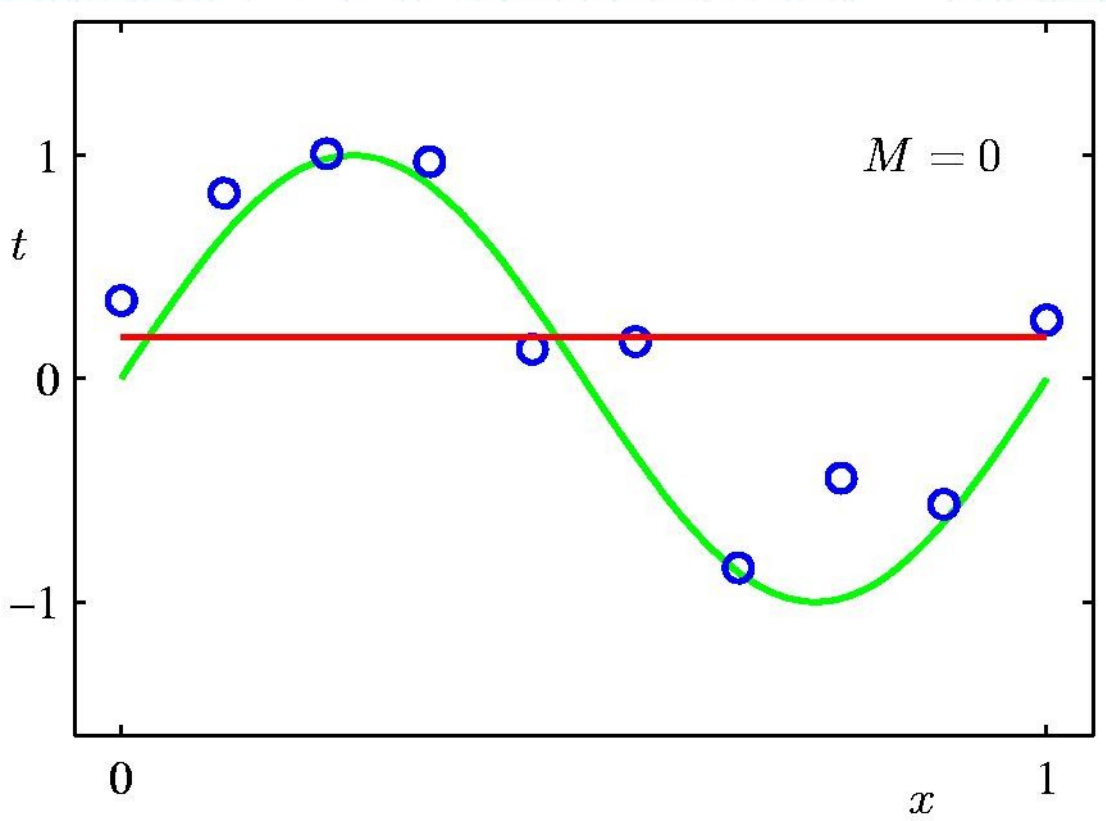
Пример – приближение многочленами

Данные: sin(x) + случайный шум

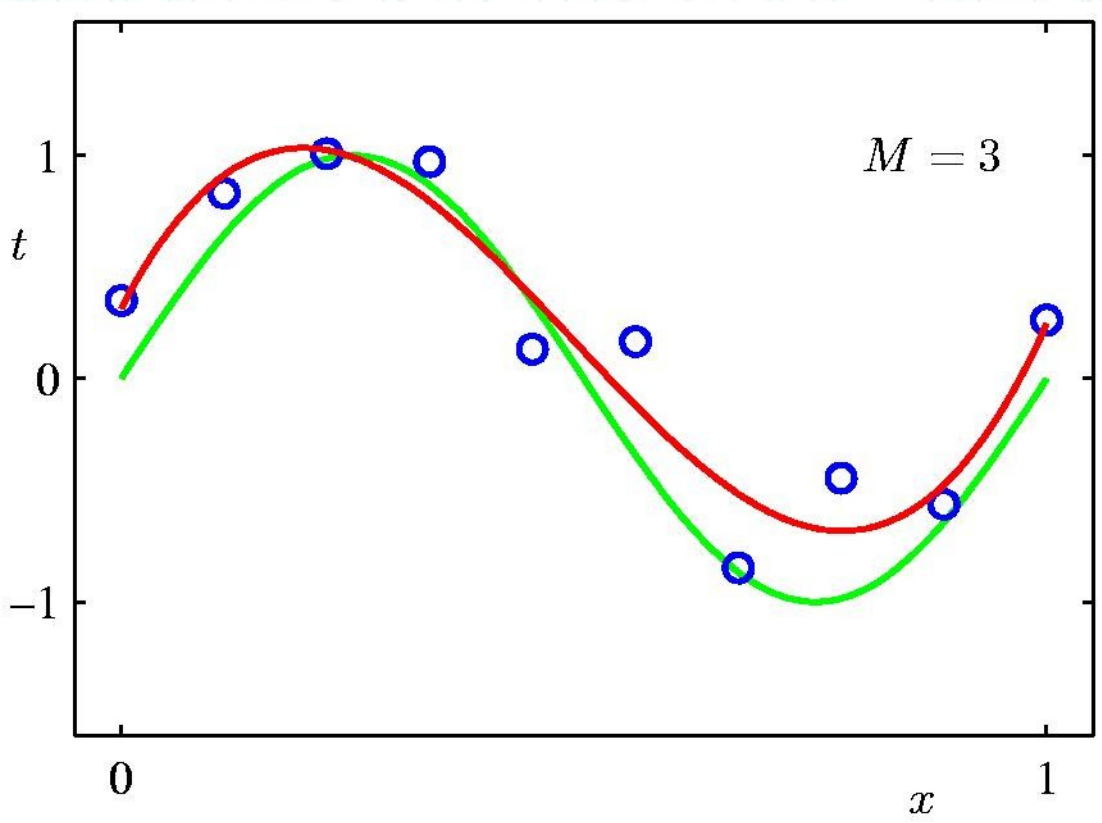


$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

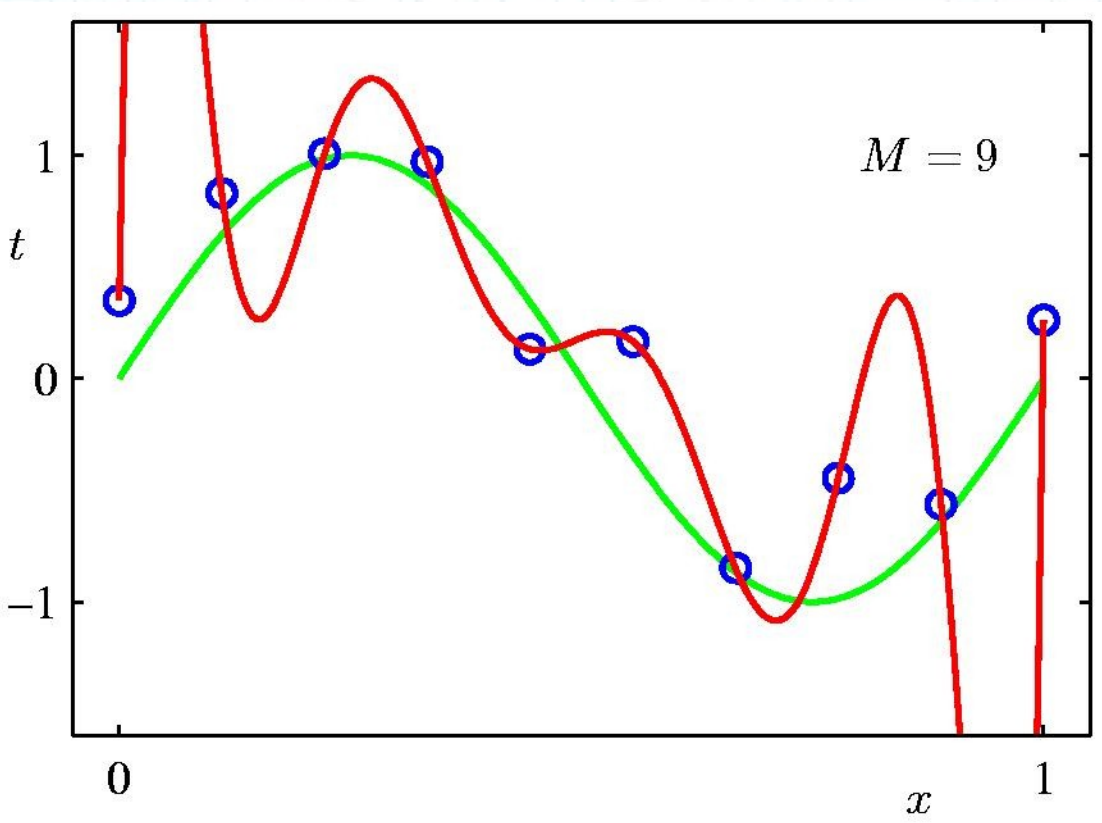
Многочлен степени 0



Многочлен степени 3



Многочлен степени 9



Коэффициенты многочленов

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

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

Гребневая регрессия

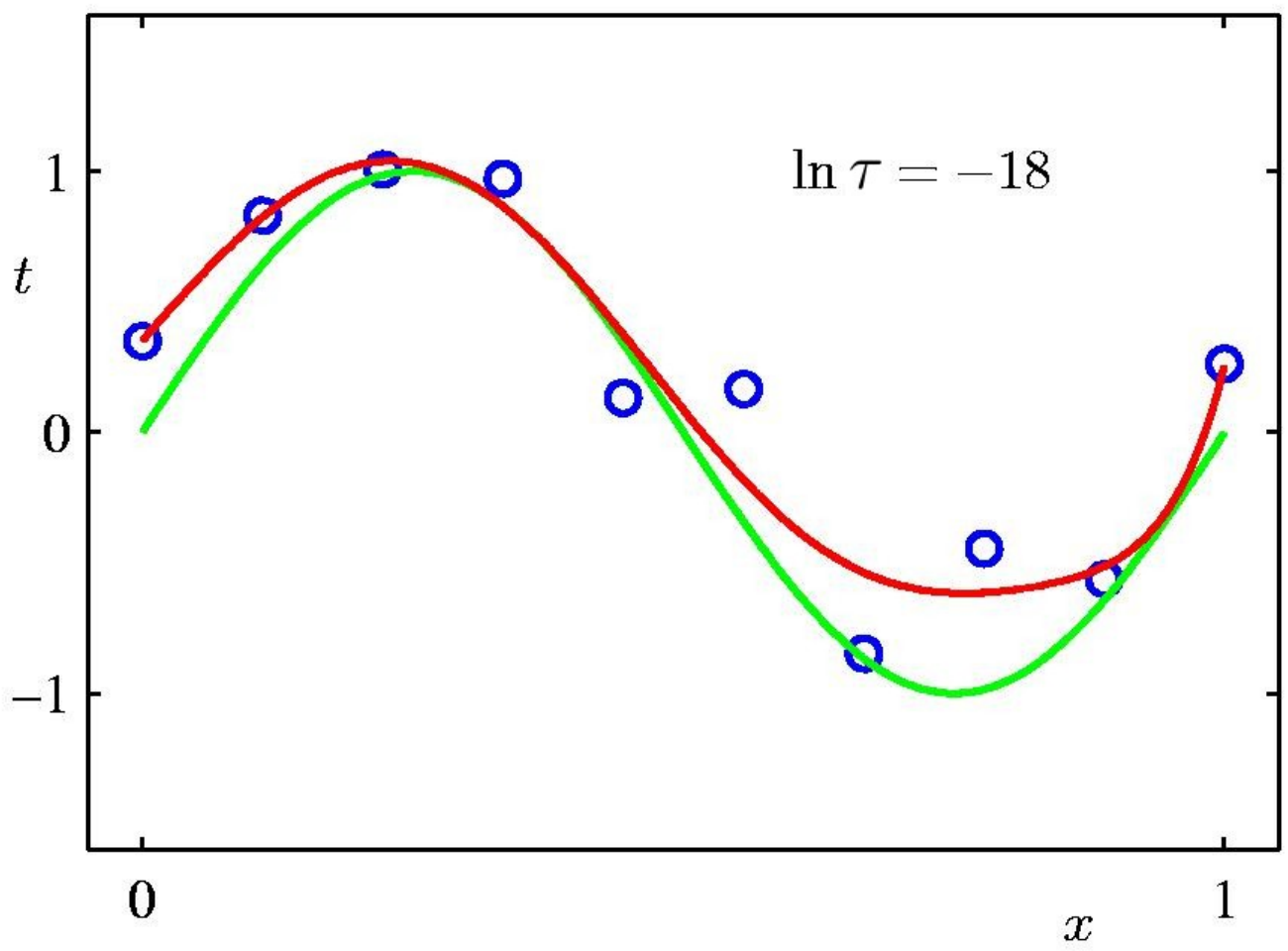
- Штраф за увеличение нормы вектора весов ||α|| :

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

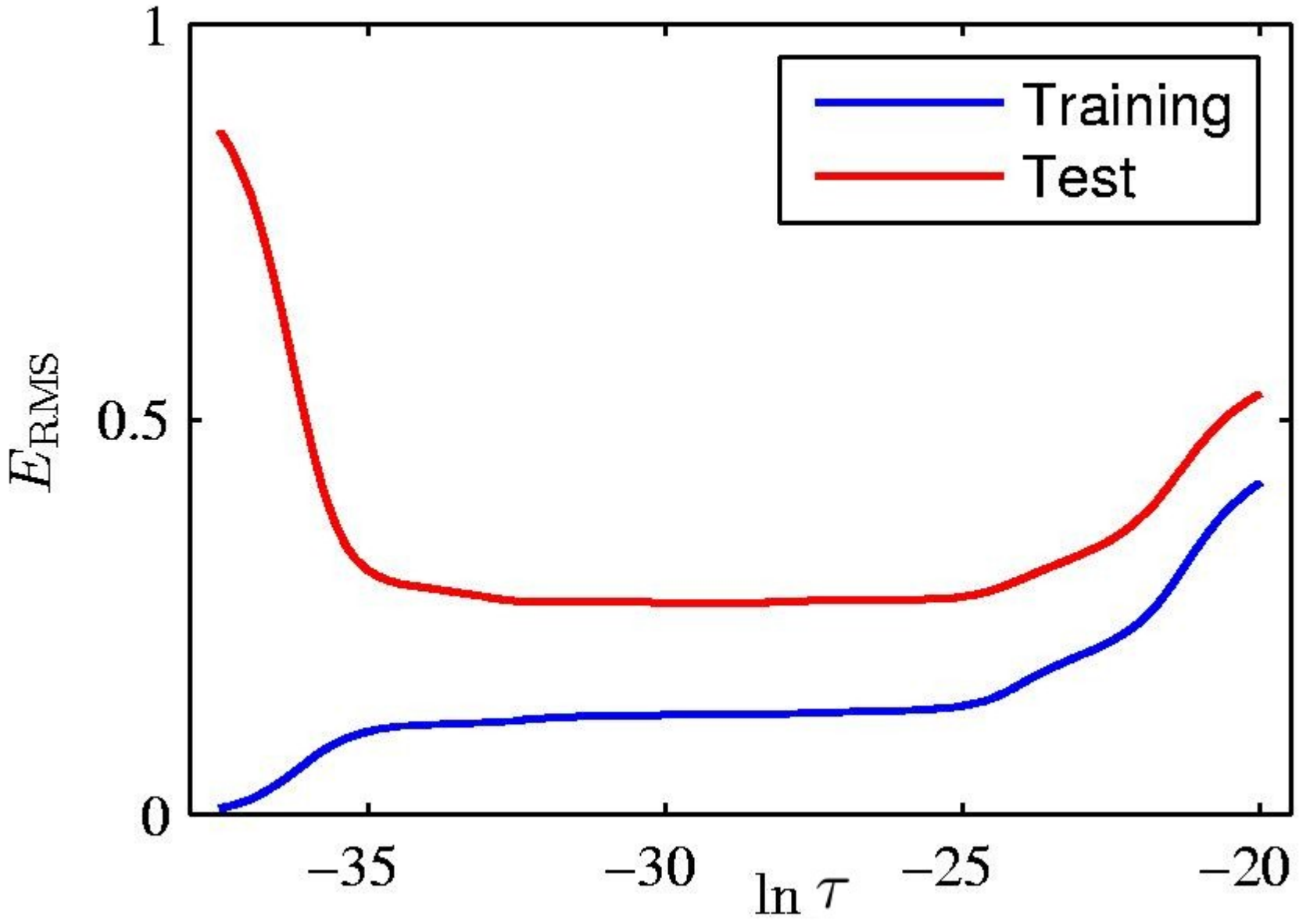
- Модифицированное МНК-решение (τI_n — «гребень»):

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

Многочлен степени 9 с регуляризацией



Гребневая регрессия



Сингулярное разложение

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

$$F = VDU^{\mathsf{T}}.$$

Основные свойства сингулярного разложения:

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

Решение МНК через сингулярное разложение

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

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

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

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

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

Проблема мультиколлинеарности

- Число обусловленности n×n-матрицы Σ:

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

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

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

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

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

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

Без регуляризации:

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

Отбор признаков

- LASSO — Least Absolute Shrinkage and Selection Operator

$$\begin{cases} Q(\alpha) = \|F\alpha - y\|^2 \to \min\limits_{\alpha}; \\ \sum\limits_{j=1}^{n} |\alpha_j| \leqslant \varkappa; \end{cases}$$

- Чем меньше $\varkappa$, тем больше нулевых α_j

Метод главных компонент (PCA)

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

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

как можно точнее на обучающей выборке:

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

Постановка задачи PCA в матричной форме

$$\hat{F} = GU^{\mathsf{T}} \overset{\text{хотим}}{\approx} F$$

Найти: и новые признаки G , и преобразование U:

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

Теорема

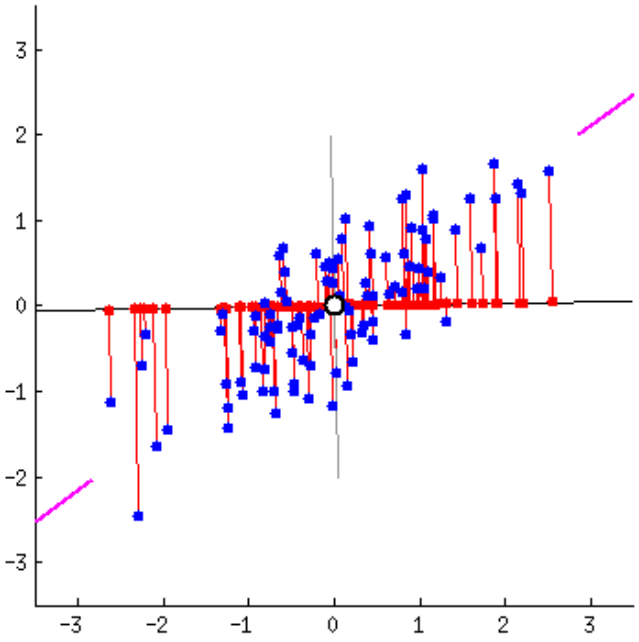
Если $m \leqslant \operatorname{rk} F$, *то минимум* $\left\| GU^{\mathrm{T}} - F \right\|^2$ *достигается, когда столбцы* U *— это с.в. матрицы* $F^{\mathrm{T}}F$, *соответствующие* m *максимальным с.з.* $\lambda_1, \ldots, \lambda_m$, *а матрица* $G = FU$.

При этом:

1. *матрица* U *ортонормирована:* $U^{\mathrm{T}}U = I_m$;
2. *матрица* G *ортогональна:* $G^{\mathrm{T}}G = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$;
3. $U\Lambda = F^{\mathrm{T}}FU$; $G\Lambda = FF^{\mathrm{T}}G$;
4. $\left\| GU^{\mathrm{T}} - F \right\|^2 = \|F\|^2 - \operatorname{tr}\Lambda = \sum\limits_{j=m+1}^{n} \lambda_j$.

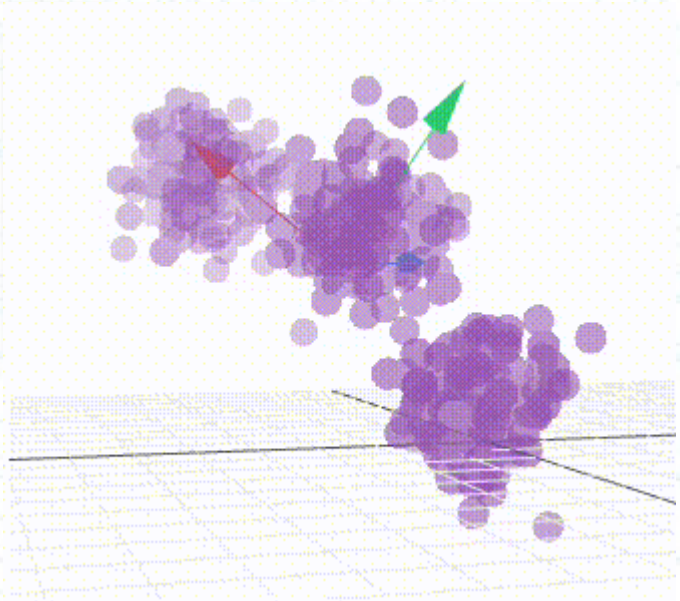
Главная компонента датасета - направление наибольшей вариации точек

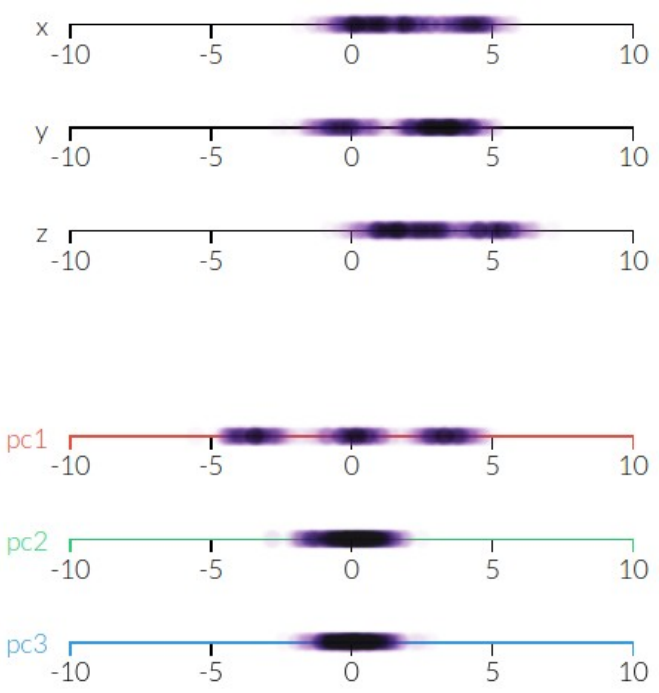
$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w} \right)^2 \right\}$$



Альтернативное определение:
главная компонента задает прямую, среднее расстояние точек датасета до которой минимально

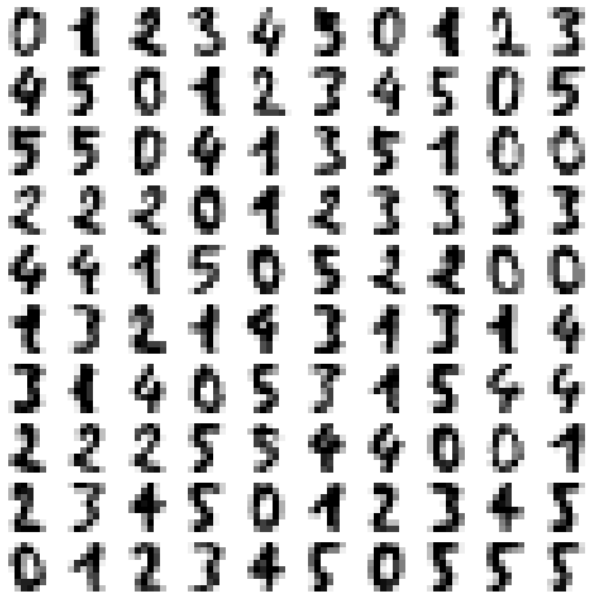
Следующие компоненты - такие же главные в ортогональном дополнении к предыдущим

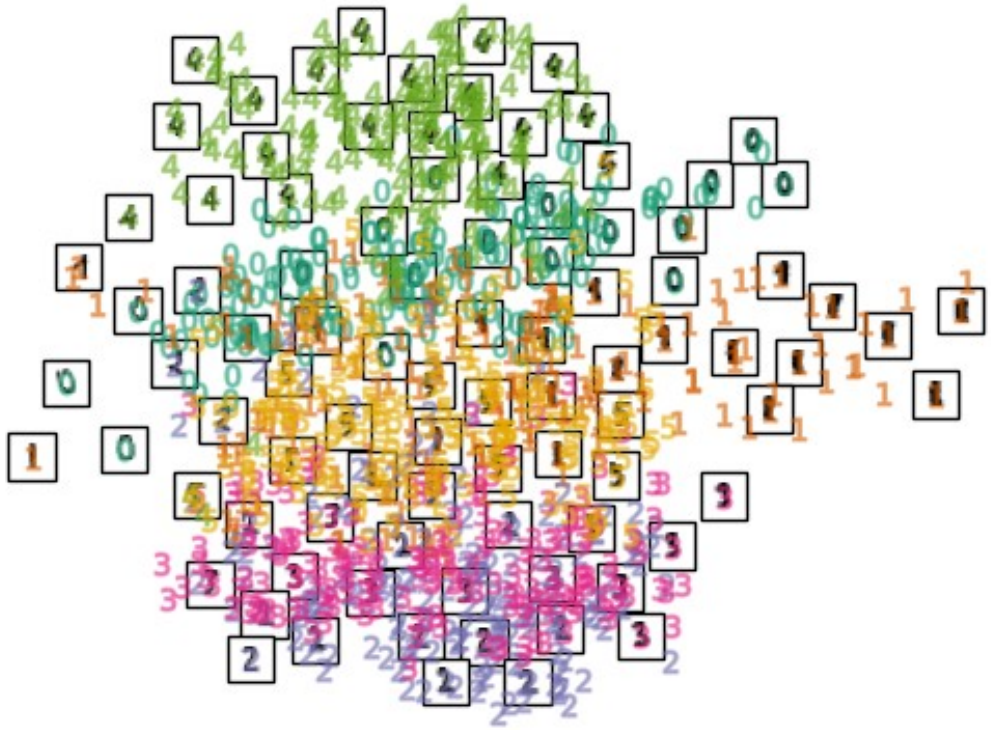




Самое распространенное приложение: понижение размерности датасета

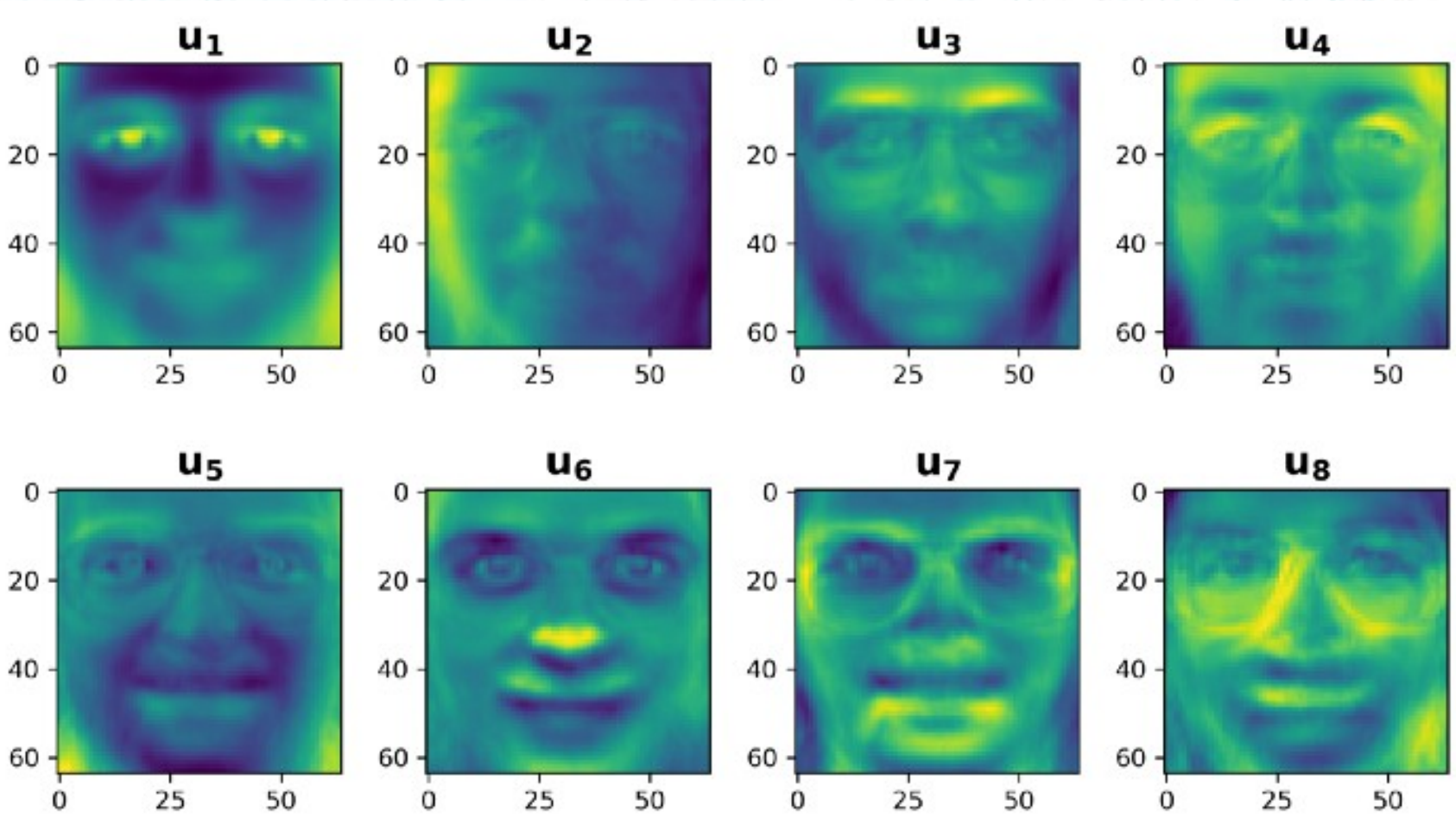
A selection from the 64-dimensional digits dataset



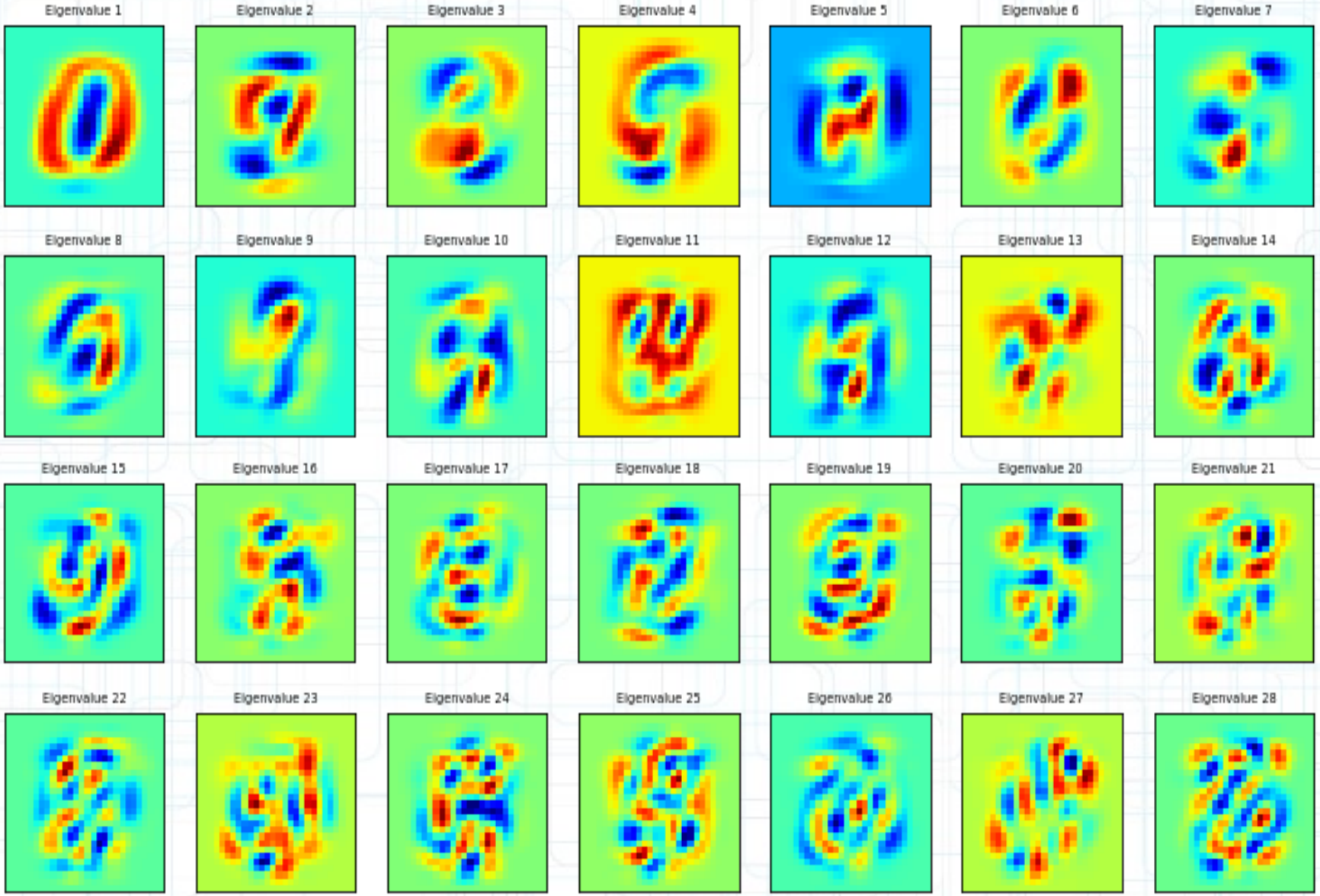


Главные компоненты датасета Olivetti faces

Показывают ортогональные направления, вдоль которых лица датасета меняются сильнее всего

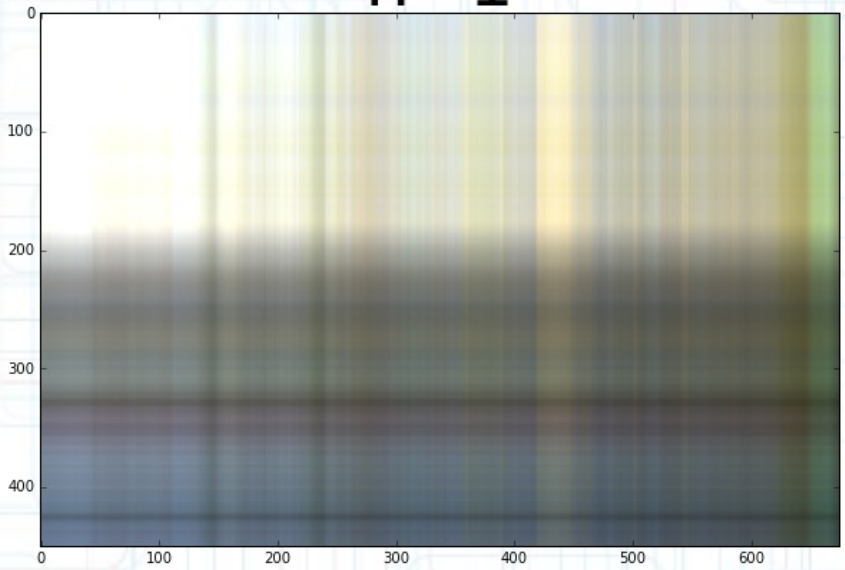


Главные компоненты датасета рукописных цифр



Применение PCA к сжатию изображений

n=1

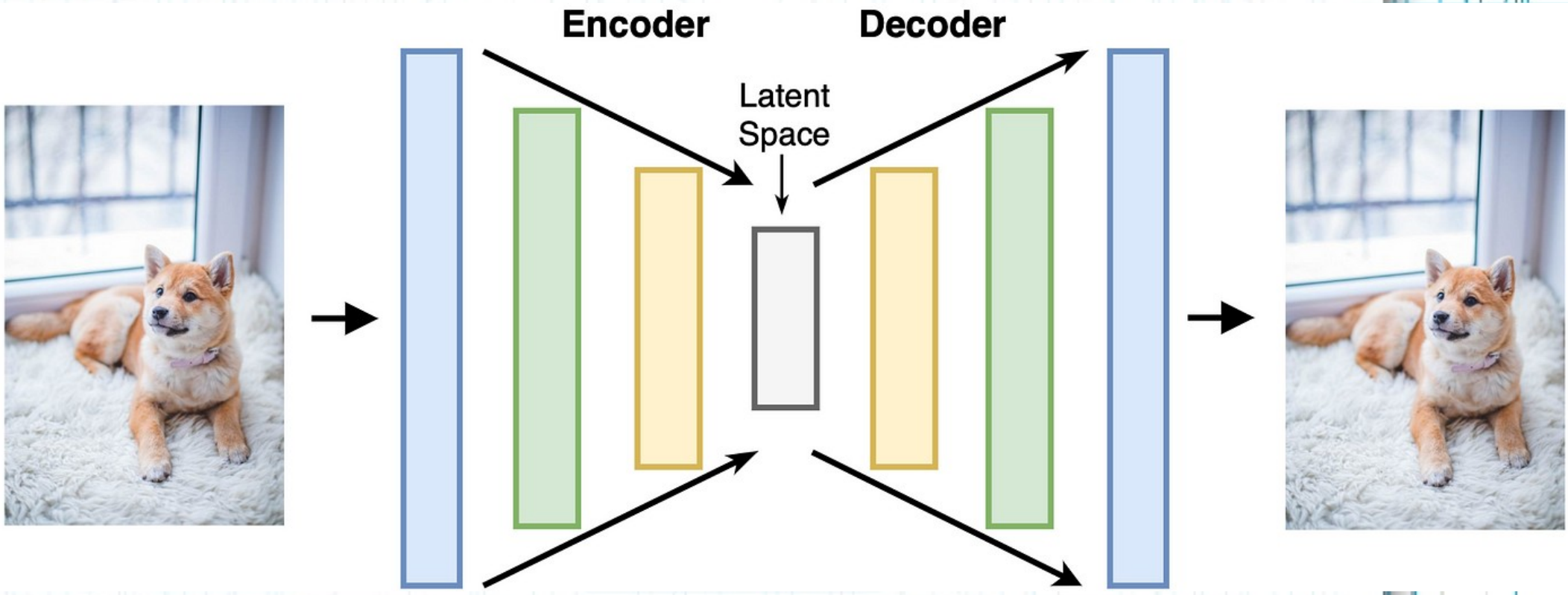


n=10

n=30

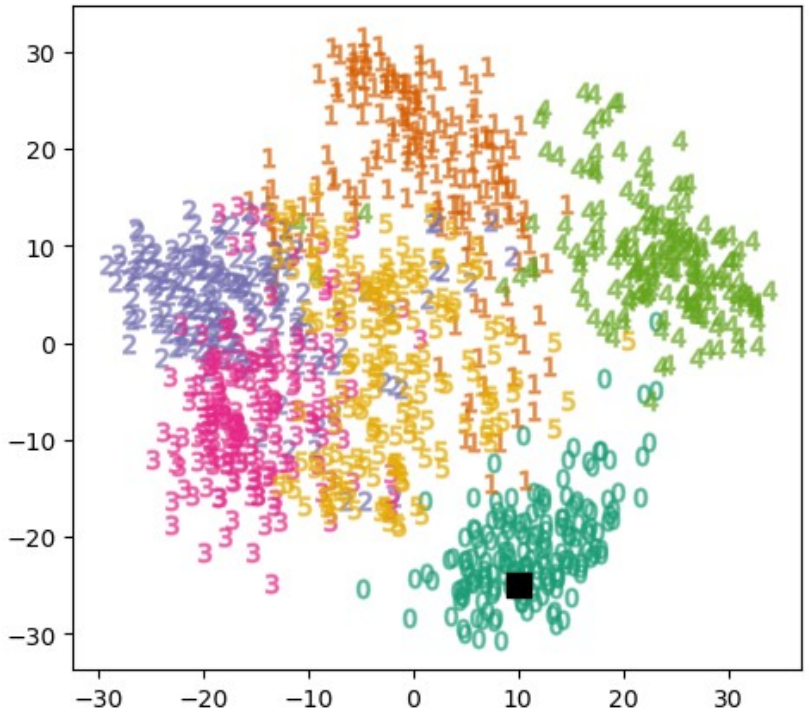
n=100

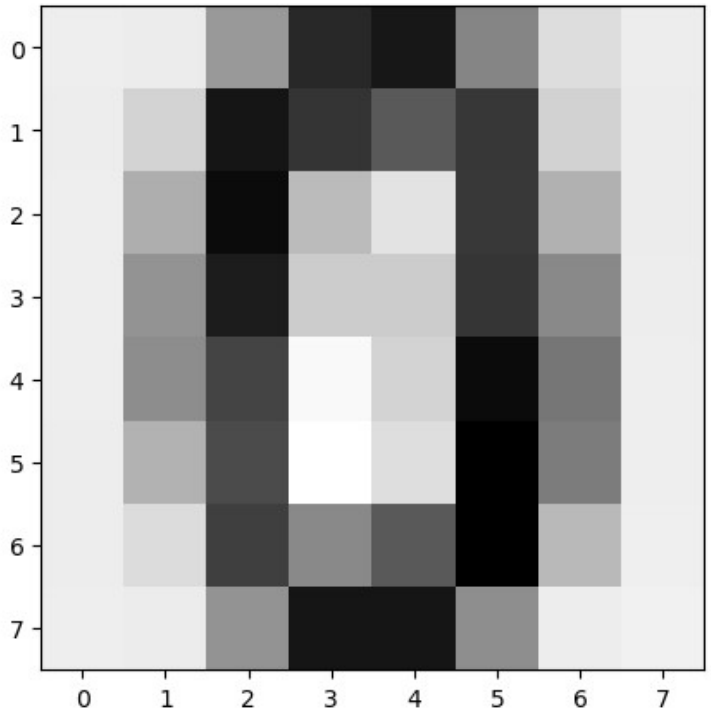
PCA – простейший пример трансформера



PCA - простейший линейный трансформер

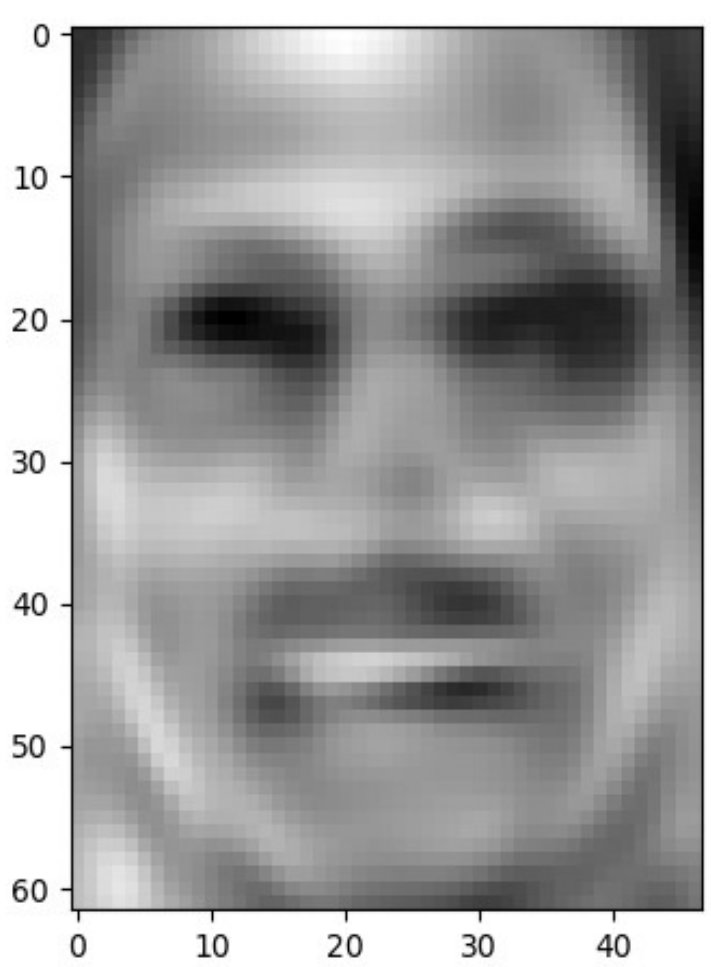
- Латентное пространство PCA - линейная оболочка, натянутая на первые несколько главных компонент
- Выбирая точки в латентном пространстве и применяя декодер, можно генерировать “искусственные” объекты, отсутствующие в датасете



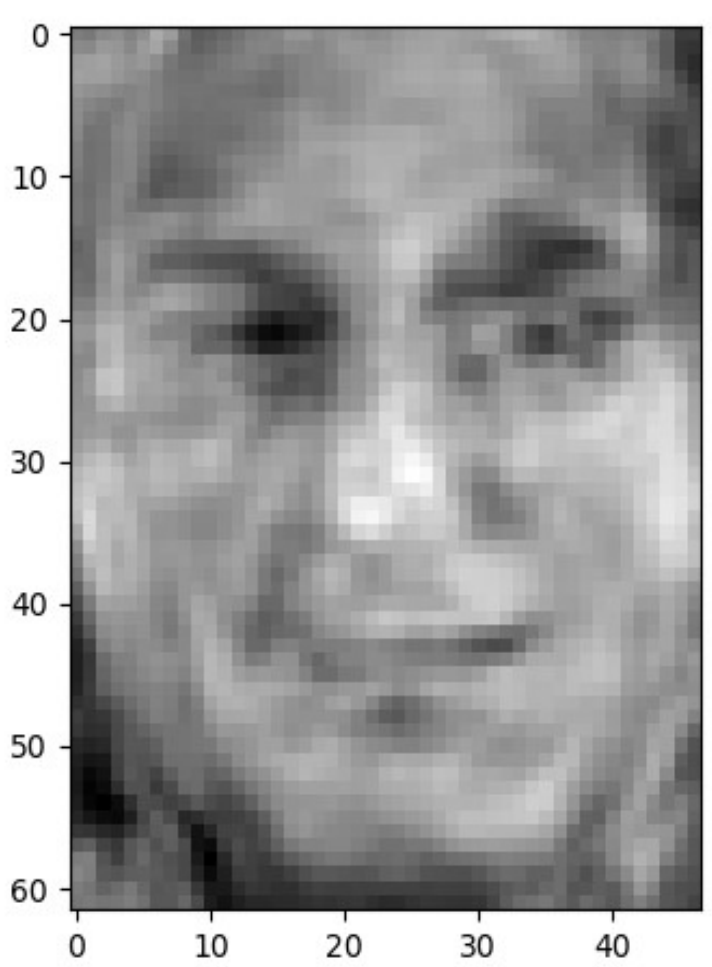


Генерация фото с помощью PCA

150 компонент

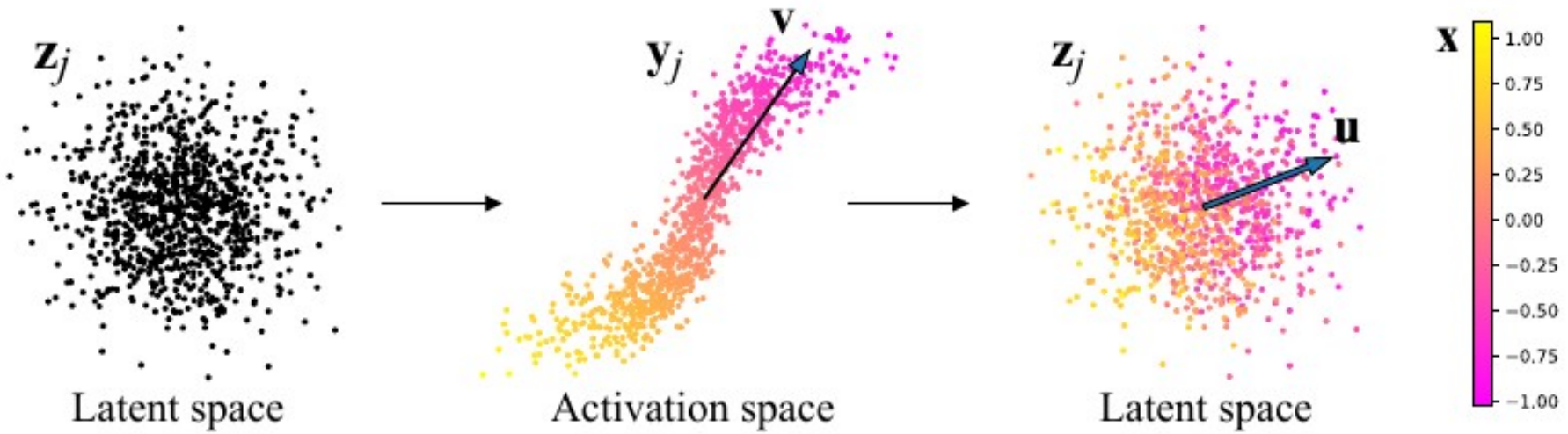


1500 компонент



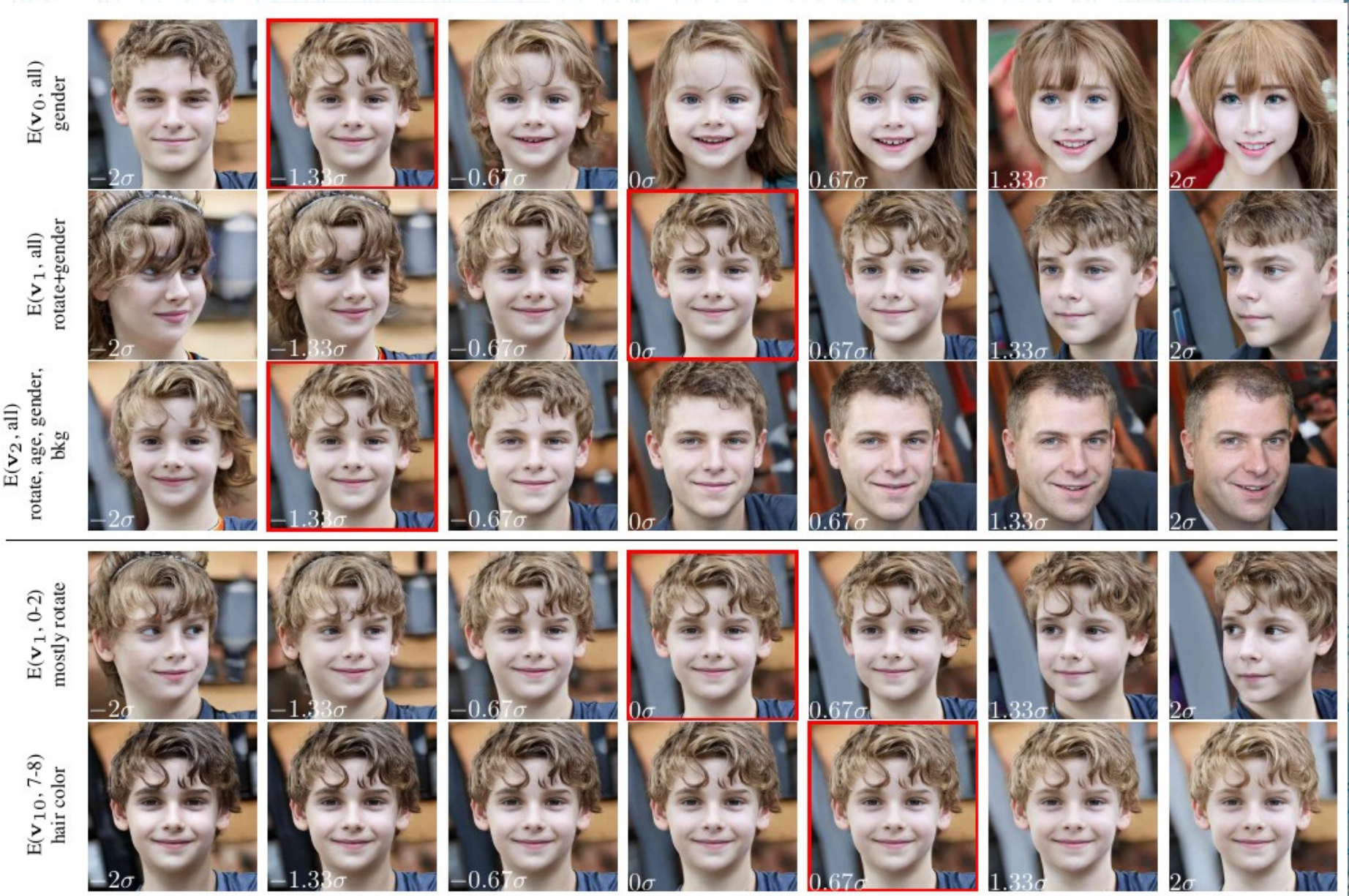
Управление генерацией в GAN

- Для генерации изображений с помощью GAN берут случайные точки в латентном пространстве
- Оказывается, что после трансформации нормально распределенной выборки на одном скрытом слое, главные компоненты в следующих слоях поддаются интерпретации



E.Härkönen, A.Hertzmann, J.Lehtinen, S.Paris GANSpace: Discovering Interpretable GAN Controls. arXiv:2004.02546

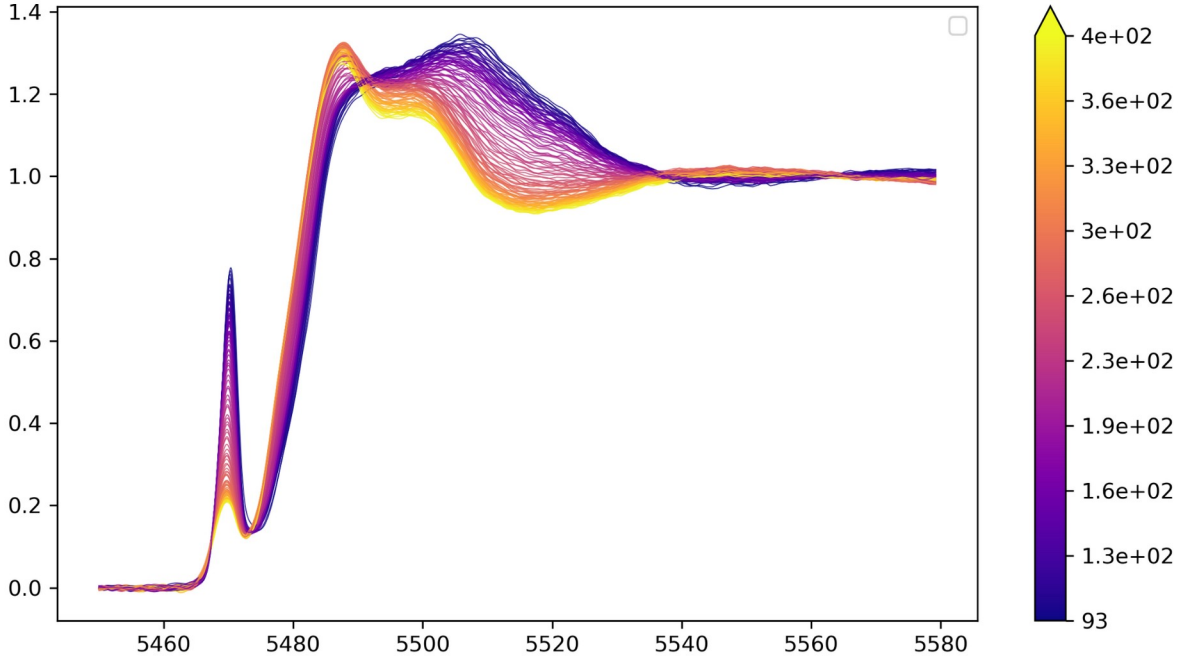
Управление генерацией в GAN



Управление генерацией в GAN

Поиск компонент смесей

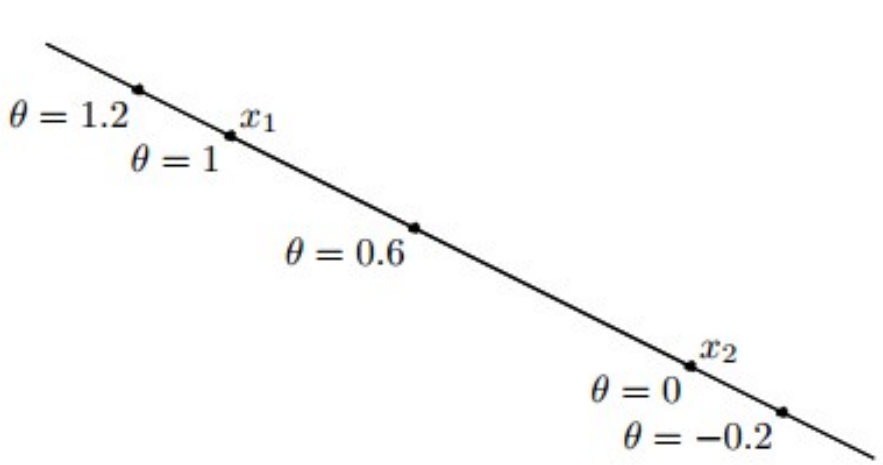
- Дано: серия спектров для смеси веществ в процессе проведения некоторого эксперимента (изменяется температура/ давление/… и в результате меняются концентрации компонент)
- Найти: спектры чистых компонент и зависимость концентрации от меняющегося параметра эксперимента

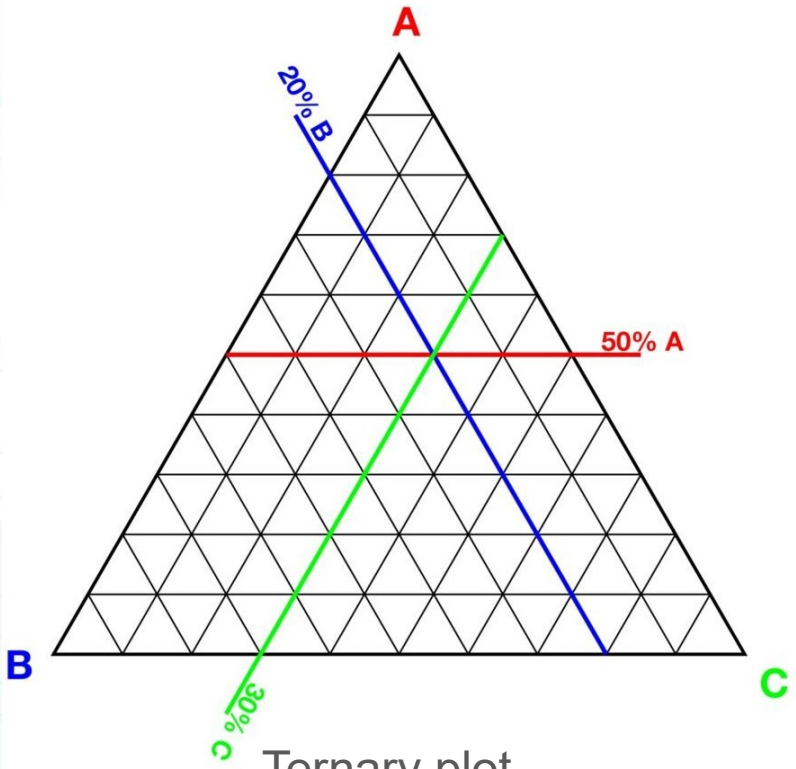


Подход на основе PCA

- Благодаря линейности преобразования в латентном пространстве смеси тоже являются линейными комбинациями точек и имеют вид (для 2-х компонент):
  $c1(t)\cdot Spectr1+c2(t)\cdot Spectr2$
- Что за фигура получается в результате изменения температуры t ?
- А если компонент три:
  $c1(t)\cdot Spectr1+c2(t)\cdot Spectr2+c3(t)\cdot Spectr3$?

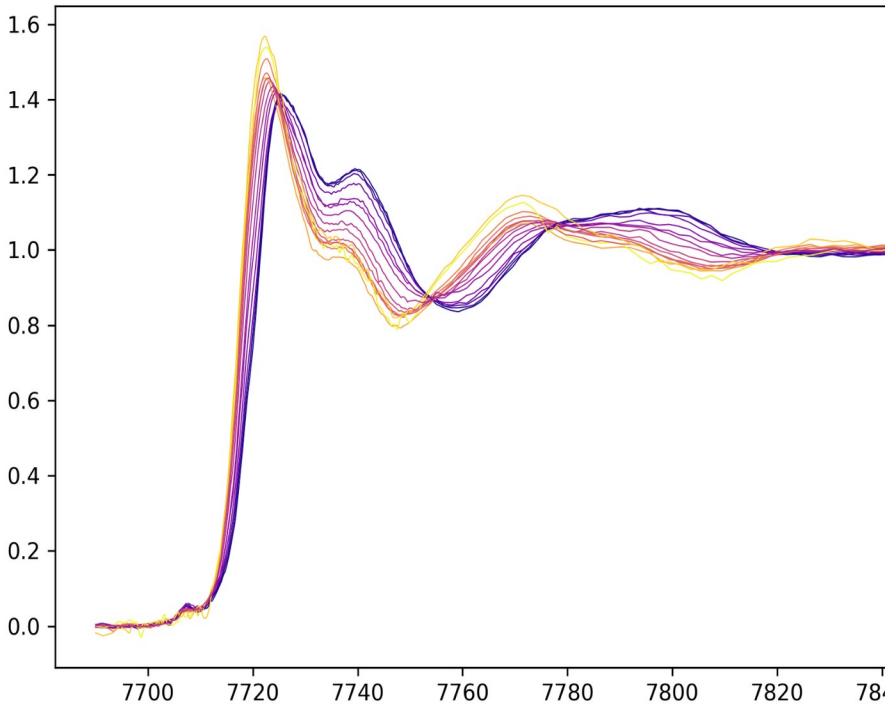
Смеси двух и трёх компонент в латентном пространстве

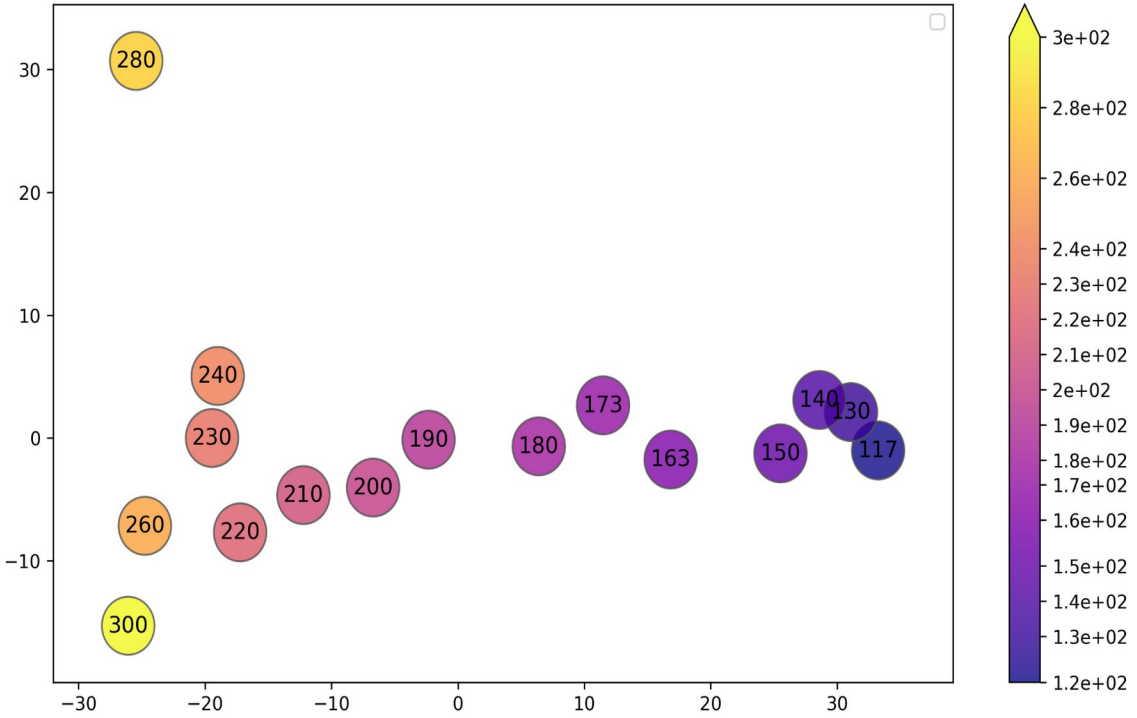




Ternary plot

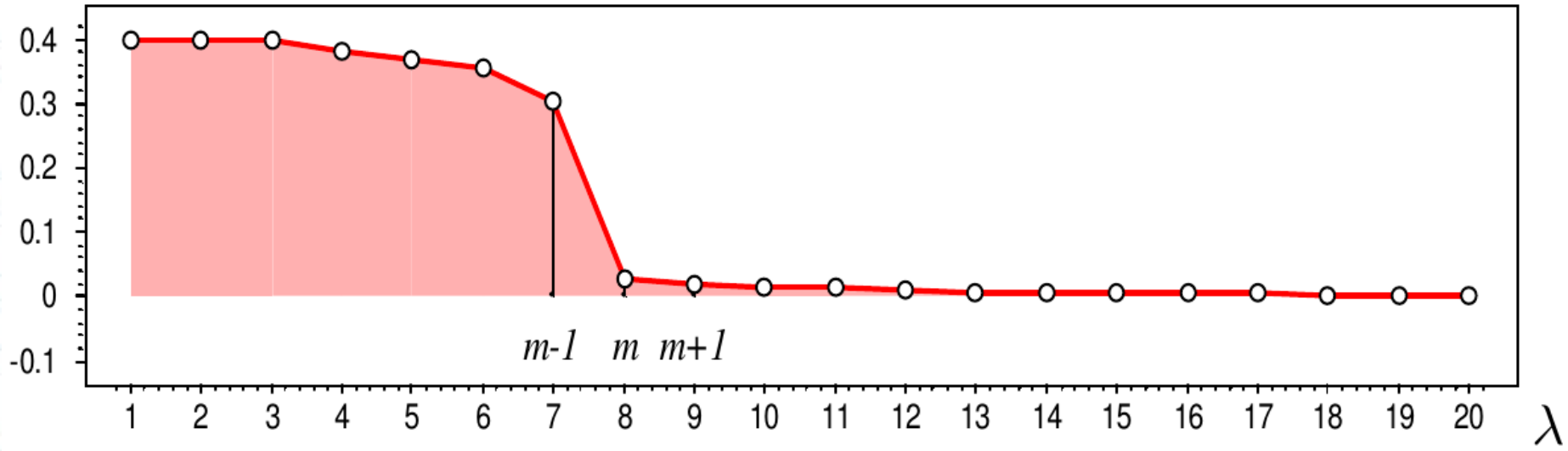
Пример Co TEMPO





Сколько главных компонент брать?

- Критерий “крутого склона”:



Непараметрическая регрессия

- Обычная задача МНК:

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

- Приближение константой f (x, α) = α в окрестности x ∈ X

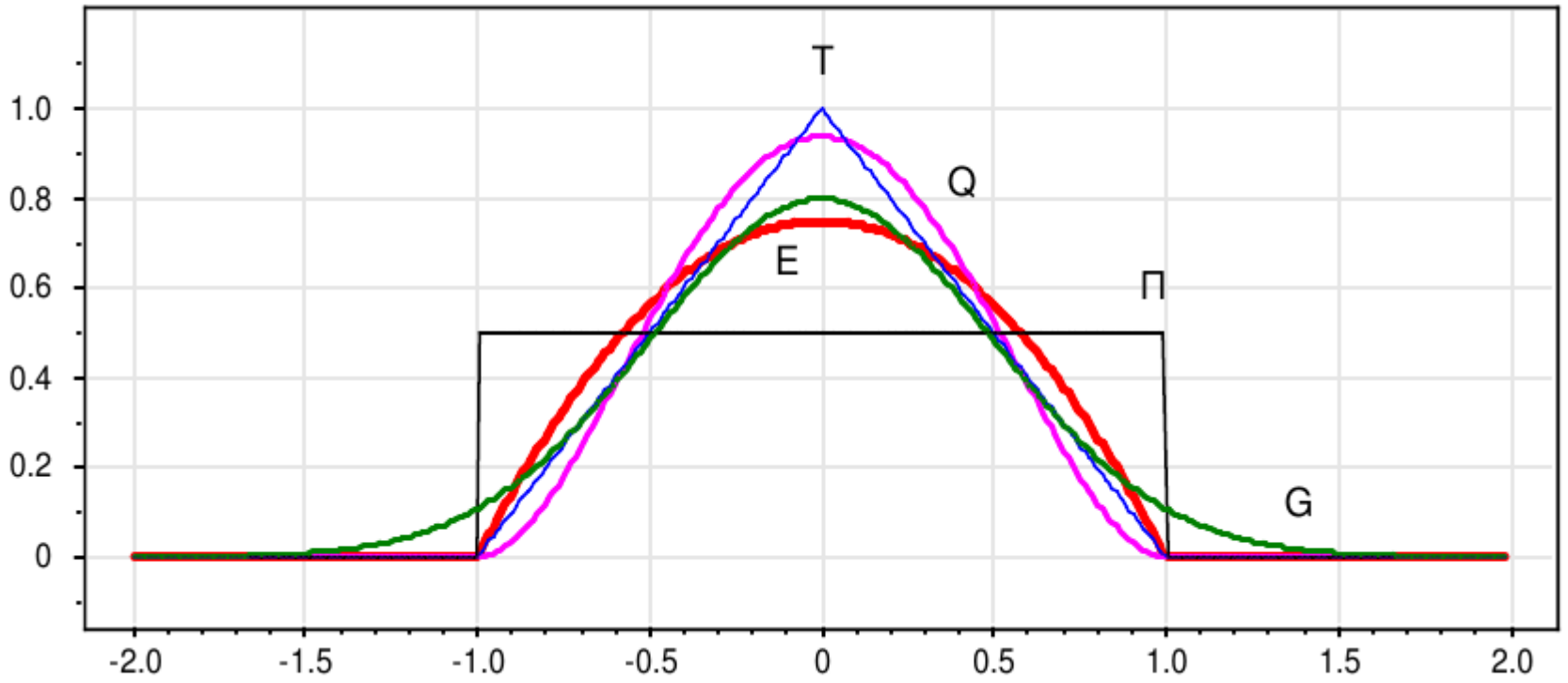
$$Q(\alpha; X^\ell) = \sum_{i=1}^{\ell} w_i(x) \big(\alpha - y_i\big)^2 \to \min_{\alpha \in \mathbb{R}}$$

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



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

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

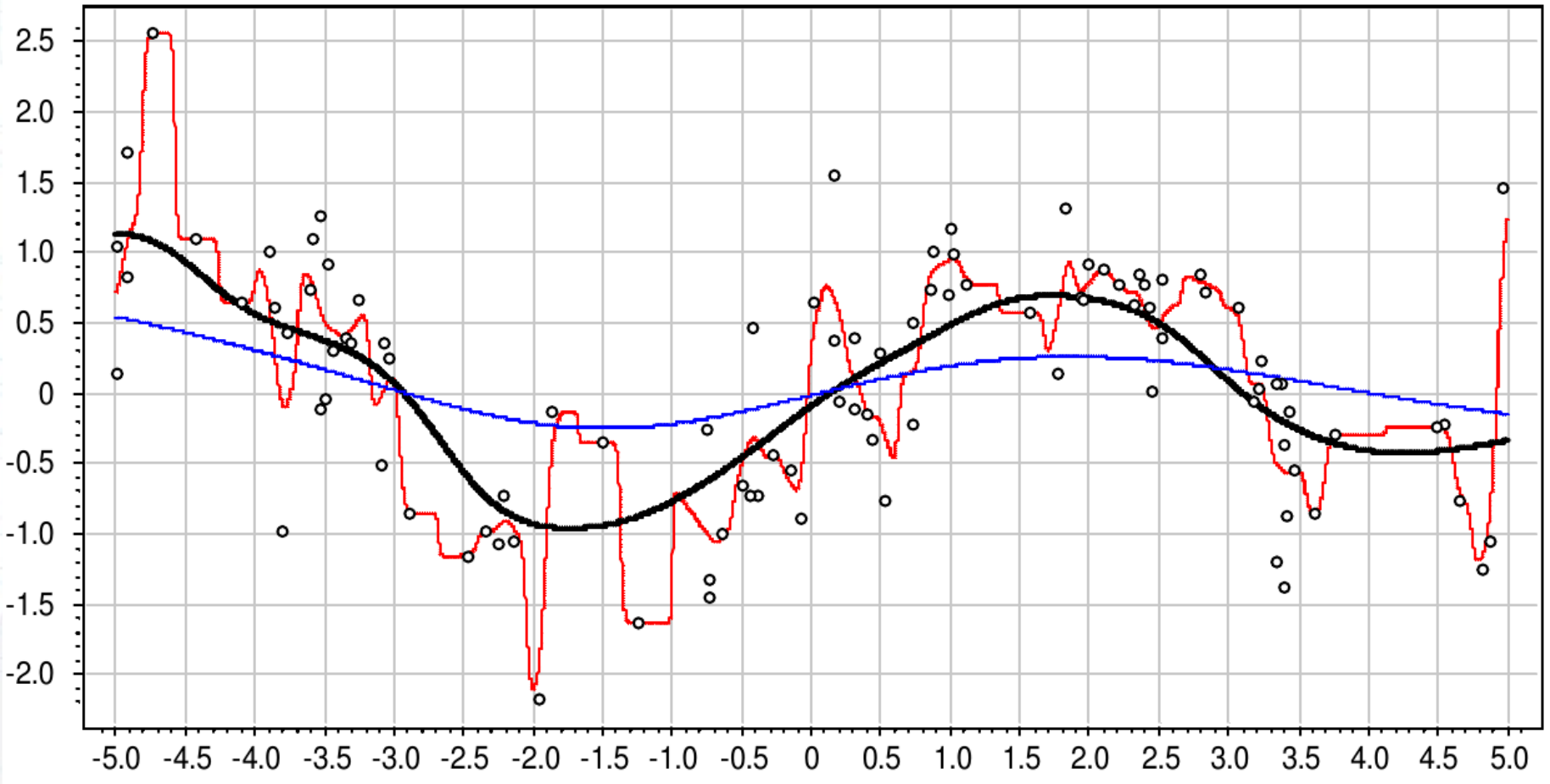
$E(r) = \big(1 - r^2\big)\big[|r| \leqslant 1\big]$ — квадратичное (Епанечникова)

$Q(r) = \big(1 - r^2\big)^2\big[|r| \leqslant 1\big]$ — квартическое

$G(r) = \exp\big(-2r^2\big)$ — гауссовское

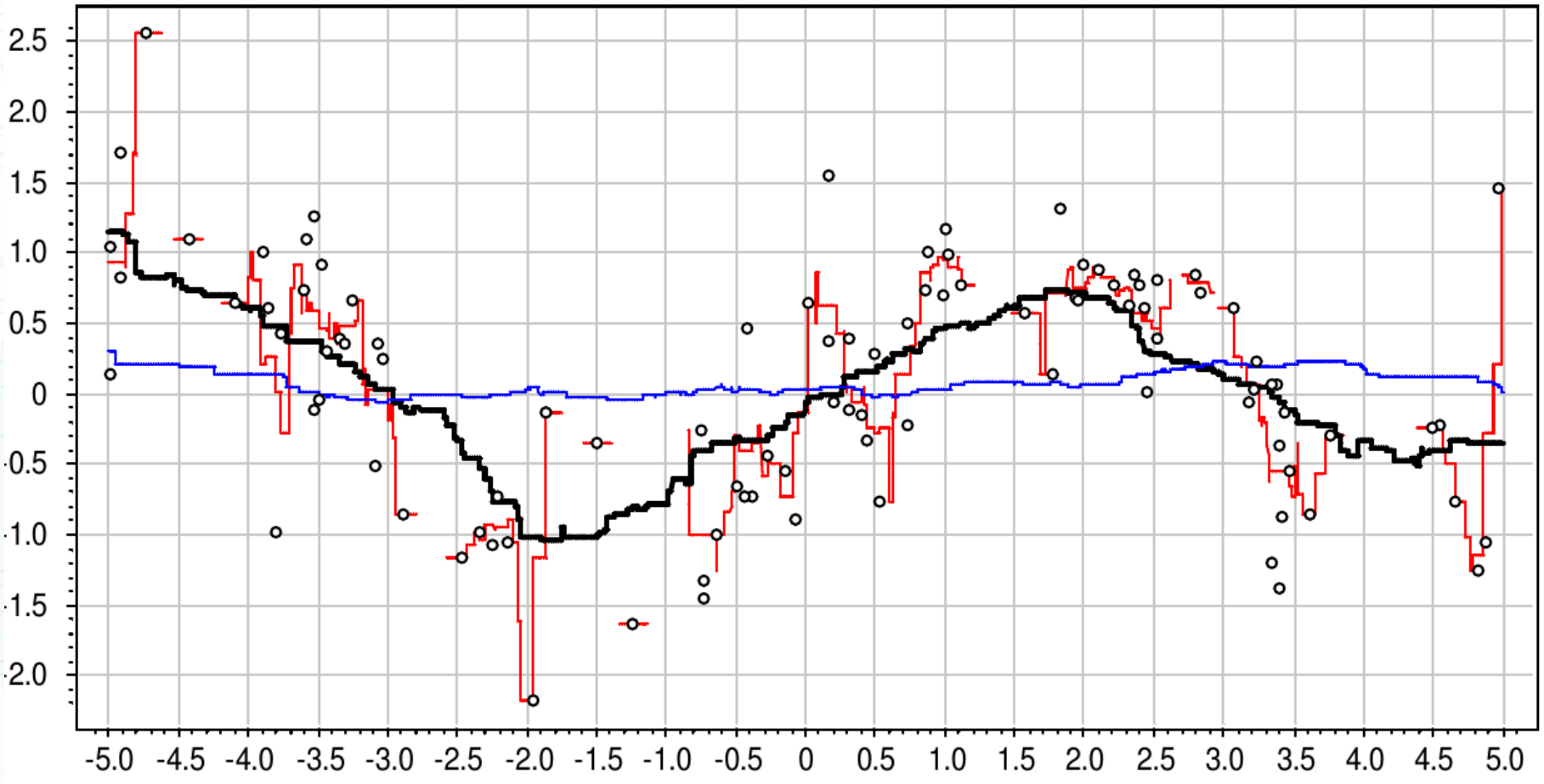
Выбор ширины окна и ядра

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



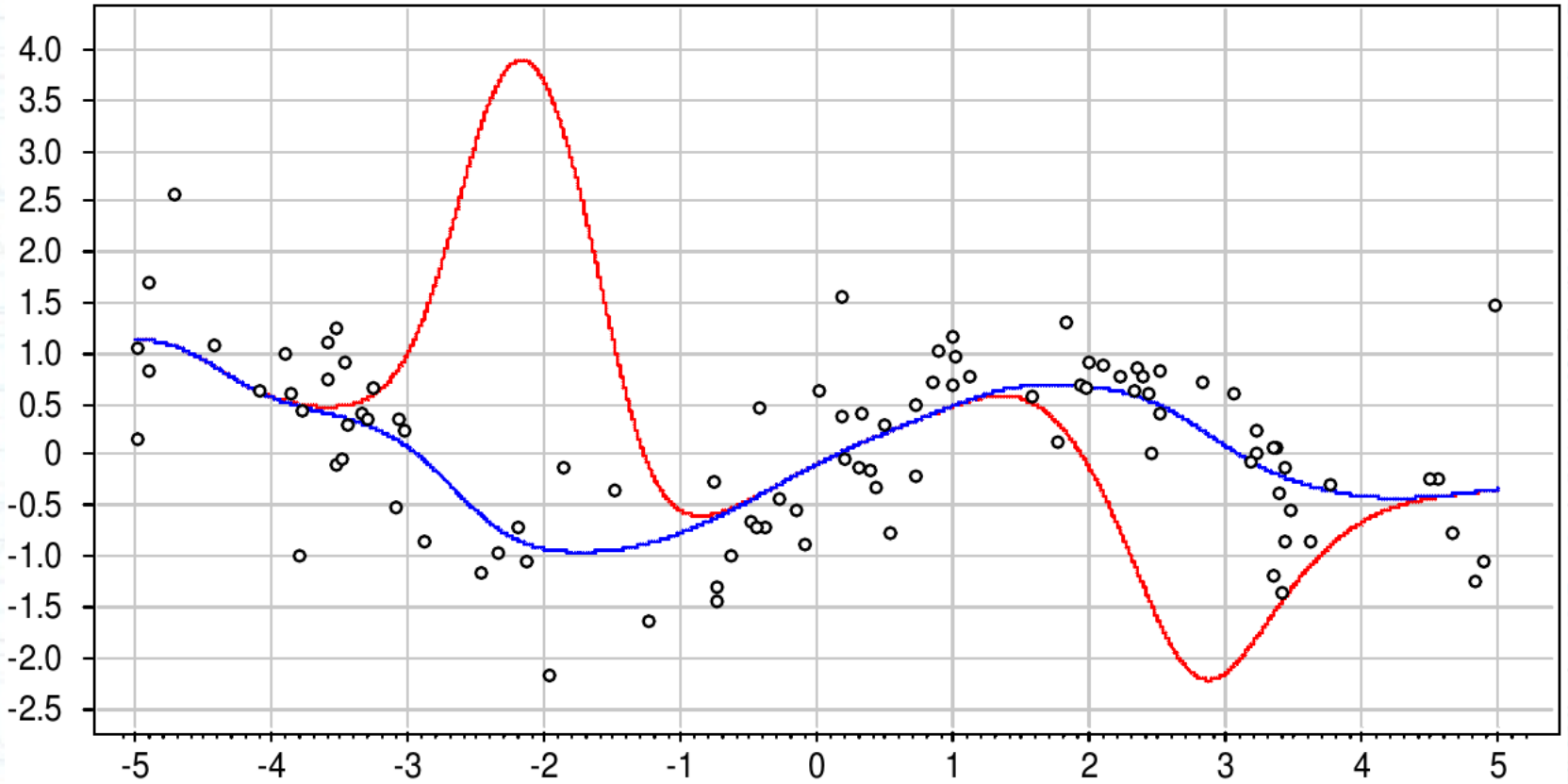
Выбор ширины окна и ядра

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



Проблема выбросов

- **l** = 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 $\varepsilon_i = \left|a_h\left(x_i; X^\ell \setminus \{x_i\}\right) - y_i\right|$ 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: квартическое ядро

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

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

Алгоритм LOWESS (LOcally WEighted Scatter plot Smoothing)

Вход: X^ℓ — обучающая выборка;

Выход: коэффициенты γ_i, $i = 1, \ldots, \ell$;

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

2: **повторять**

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

4: вычислить оценки скользящего контроля:

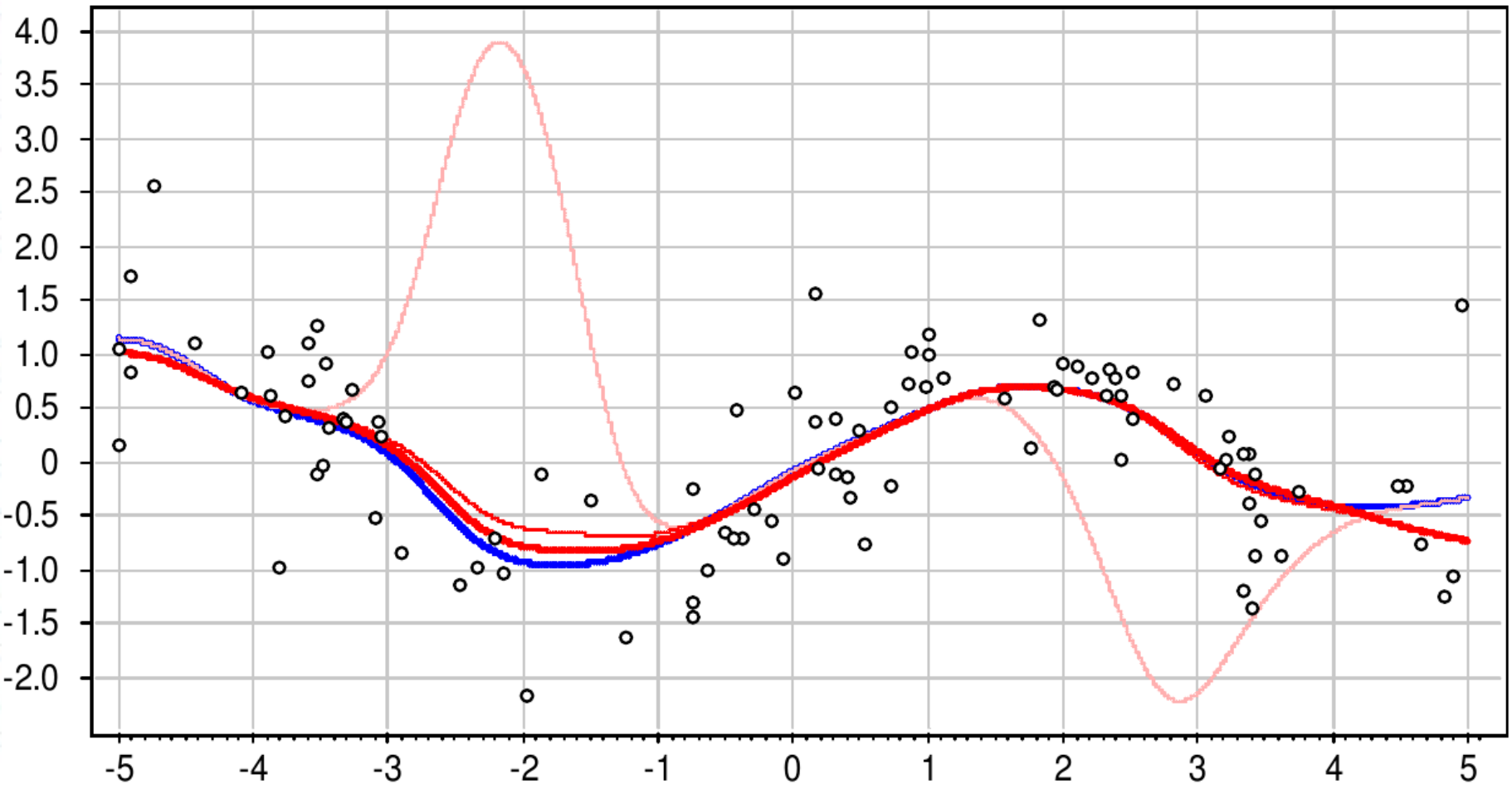
$$a_i := a_h\left(x_i; X^\ell \setminus \{x_i\}\right) = \frac{\sum\limits_{j=1,\, j\neq i}^{\ell} y_j \gamma_j K\left(\frac{\rho(x_i, x_j)}{h(x_i)}\right)}{\sum\limits_{j=1,\, j\neq i}^{\ell} \gamma_j K\left(\frac{\rho(x_i, x_j)}{h(x_i)}\right)};$$

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

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

7: **пока** коэффициенты γ_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