

# Многомерная линейная регрессия. Метод главных компонент

Воронцов Константин Вячеславович  
vokov@forecsys.ru

<http://www.MachineLearning.ru/wiki?title=User:Vokov>

Этот курс доступен на странице вики-ресурса  
<http://www.MachineLearning.ru/wiki>  
«Машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ • 9 октября 2021

## 1 Многомерная линейная регрессия

- Метод наименьших квадратов
- Многомерная линейная регрессия
- Сингулярное разложение

## 2 Регуляризация

- $L_2$ -регуляризация: гребневая регрессия
- $L_1$ -регуляризация: лассо Тибширани
- Негладкие регуляризаторы

## 3 Метод главных компонент

- Постановка задачи и основная теорема
- Метод главных компонент для линейной регрессии
- Обобщения

## Метод наименьших квадратов

- $X$  — объекты (часто  $\mathbb{R}^n$ );  $Y$  — ответы (часто  $\mathbb{R}$ , реже  $\mathbb{R}^m$ );  
 $X^\ell = (x_i, y_i)_{i=1}^\ell$  — обучающая выборка;  
 $y_i = y(x_i)$ ,  $y: X \rightarrow Y$  — неизвестная зависимость;
- $a(x) = f(x, \alpha)$  — модель зависимости,  
 $\alpha \in \mathbb{R}^p$  — вектор параметров модели.
- Метод наименьших квадратов (МНК):

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

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

$Q(\alpha^*, X^\ell)$  — остаточная сумма квадратов  
 (residual sum of squares, RSS).

## Метод максимума правдоподобия

Модель данных с некоррелированным гауссовским шумом:

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

Метод максимума правдоподобия (ММП):

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

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

### Теорема

*Постановки МНК и ММП, совпадают, причём веса объектов обратно пропорциональны дисперсии шума,  $w_i = \sigma_i^{-2}$ .*

## Многомерная линейная регрессия

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

Модель многомерной линейной регрессии:

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

Матричные обозначения:

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

Функционал квадрата ошибки:

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

## Нормальная система уравнений

Необходимое условие минимума в матричном виде:

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

откуда следует *нормальная система* задачи МНК:

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

где  $F^T F$  — матрица размера  $n \times n$ .

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

Значение функционала:  $Q(\alpha^*) = \|P_F y - y\|^2$ ,

где  $P_F = FF^+ = F(F^T F)^{-1} F^T$  — *проекционная матрица*.

## Геометрическая интерпретация МНК

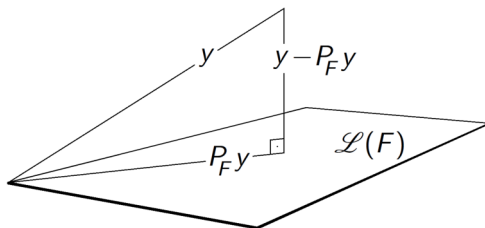
Линейная оболочка столбцов матрицы  $F = (f_1, \dots, f_n)$ ,  $f_j \in \mathbb{R}^\ell$ :

$$\mathcal{L}(F) = \left\{ \sum_{j=1}^n \alpha_j f_j \mid \alpha \in \mathbb{R}^n \right\}$$

$P_F = F(F^T F)^{-1} F^T$  — проекционная матрица

$P_F y$  — проекция вектора  $y \in \mathbb{R}^\ell$  на подпространство  $\mathcal{L}(F)$

$(I_\ell - P_F)y$  — проекция  $y$  на его ортогональное дополнение



МНК — это опускание перпендикуляра в  $\mathbb{R}^\ell$  из  $y$  на  $\mathcal{L}(F)$

## 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

Псевдообратная  $F^+ = (F^T F)^{-1} F^T$ , вектор МНК-решения  $\alpha^*$ ,  
 МНК-аппроксимация целевого вектора  $F\alpha^*$ :

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

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

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

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

---

Тождества:  $(AB)^{-1} = B^{-1}A^{-1}$ ,  $(AB)^T = B^T A^T$ ,  $\|\alpha\|^2 = \alpha^T \alpha$

## Проблема мультиколлинеарности

Если  $\exists \gamma \in \mathbb{R}^n: F\gamma \approx 0$ , то некоторые  $\lambda_j$  близки к нулю

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

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

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

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

В нашем случае:  $\alpha^* = (F^T F)^{-1} F^T y$ , усиливаются в  $\mu(F^T F)$  раз погрешности измерения признаков  $f_j(x_i)$  и ответов  $y_i$

## Проблема мультиколлинеарности и переобучения

Если матрица  $S = F^T F$  плохо обусловлена, то:

- решение неустойчиво и плохо интерпретируемо,
- $\|\alpha^*\|$  велико;
- возникает переобучение:  
на обучении  $Q(\alpha^*, X^\ell) = \|F\alpha^* - y\|^2$  малó;  
на контроле  $Q(\alpha^*, X^k) = \|F'\alpha^* - y'\|^2$  велико;

Стратегии устранения мультиколлинеарности и переобучения:

- регуляризация:  $\|\alpha\| \rightarrow \min$ ;
- отбор признаков:  $f_1, \dots, f_n \rightarrow f_{j_1}, \dots, f_{j_m}, \quad m \ll n$ .
- преобразование признаков:  $f_1, \dots, f_n \rightarrow g_1, \dots, g_m, \quad m \ll n$ ;

## Гребневая регрессия (ridge regression)

Штраф за увеличение  $L_2$ -нормы вектора весов  $\|\alpha\|$ :

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

где  $\tau = \frac{1}{\sigma}$  — неотрицательный параметр регуляризации.

**Вероятностная интерпретация:** априорное распределение вектора  $\alpha$  — гауссовское с ковариационной матрицей  $\sigma I_n$ .

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

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

**Преимущество** сингулярного разложения:

можно подбирать параметр  $\tau$ , вычислив SVD только один раз.

## Регуляризованный МНК через сингулярное разложение

Вектор регуляризованного МНК-решения  $\alpha_\tau^*$   
 и МНК-аппроксимация целевого вектора  $F\alpha_\tau^*$ :

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

$$F\alpha_\tau^* = V D U^T \alpha_\tau^* = V \operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^T y = \sum_{j=1}^n \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^T y);$$

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

$F\alpha_\tau^* \neq F\alpha^*$ , но зато решение становится гораздо устойчивее.

## Выбор параметра регуляризации $\tau$

Контрольная выборка:  $X^k = (x'_i, y'_i)_{i=1}^k$ ;

$$F'_{k \times n} = \begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix}, \quad y'_{k \times 1} = \begin{pmatrix} y'_1 \\ \dots \\ y'_k \end{pmatrix}.$$

Вычисление функционала  $Q$  на контрольных данных  $T$  раз потребует  $O(kn^2 + knT)$  операций:

$$Q(\alpha_\tau^*, X^k) = \|F' \alpha_\tau^* - y'\|^2 = \left\| \underbrace{F' U}_{k \times n} \operatorname{diag} \left( \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} \right) \underbrace{V^T y}_{n \times 1} - y' \right\|^2.$$

Зависимость  $Q(\tau)$  обычно имеет характерный минимум.

## Регуляризация сокращает «эффективную размерность»

Сжатие (shrinkage) или сокращение весов (weight decay):

$$\|\alpha_{\tau}^*\|^2 = \sum_{j=1}^n \frac{\lambda_j}{(\lambda_j + \tau)^2} (v_j^T y)^2 < \|\alpha^*\|^2 = \sum_{j=1}^n \frac{1}{\lambda_j} (v_j^T y)^2.$$

Почему говорят о сокращении эффективной размерности?

Роль размерности играет след проекционной матрицы:

$$\text{tr } F(F^T F)^{-1} F^T = \text{tr } (F^T F)^{-1} F^T F = \text{tr } I_n = n.$$

При использовании регуляризации:

$$\text{tr } F(F^T F + \tau I_n)^{-1} F^T = \text{tr } \text{diag} \left( \frac{\lambda_j}{\lambda_j + \tau} \right) = \sum_{j=1}^n \frac{\lambda_j}{\lambda_j + \tau} < n.$$

---

Тождество:  $\text{tr}(AB) = \text{tr}(BA)$

## Регуляризация по $L_1$ -норме для отбора признаков

LASSO — Least Absolute Shrinkage and Selection Operator

$$\|F\alpha - y\|^2 + \mu \sum_{j=1}^n |\alpha_j| \rightarrow \min_{\alpha} \iff \begin{cases} \|F\alpha - y\|^2 \rightarrow \min_{\alpha} \\ \sum_{j=1}^n |\alpha_j| \leq \varkappa; \end{cases}$$

После замены переменных

$$\begin{cases} \alpha_j = \alpha_j^+ - \alpha_j^-; \\ |\alpha_j| = \alpha_j^+ + \alpha_j^-; \end{cases} \quad \alpha_j^+ \geq 0; \quad \alpha_j^- \geq 0.$$

ограничения принимают канонический вид:

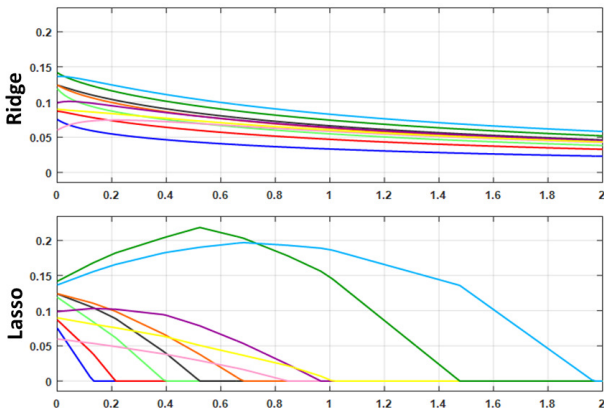
$$\sum_{j=1}^n \alpha_j^+ + \alpha_j^- \leq \varkappa; \quad \alpha_j^+ \geq 0; \quad \alpha_j^- \geq 0.$$

Чем меньше  $\varkappa$ , тем больше  $j$  таких, что  $\alpha_j^+ = \alpha_j^- = 0$ .



## Сравнение $L_2$ (Ridge) и $L_1$ (LASSO) регуляризации

Типичный вид зависимости весов  $\alpha_j$  от селективности  $\mu$



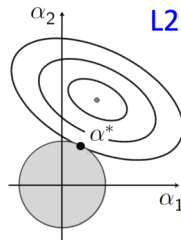
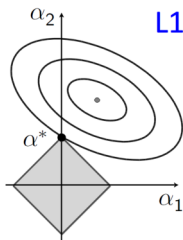
В LASSO с увеличением  $\mu$  усиливается отбор признаков

## Геометрическая интерпретация отбора признаков

Эквивалентная постановка задачи  $\|F\alpha - y\|^2 \rightarrow \min_{\alpha}$   
 с регуляризатором в виде ограничения-неравенства:

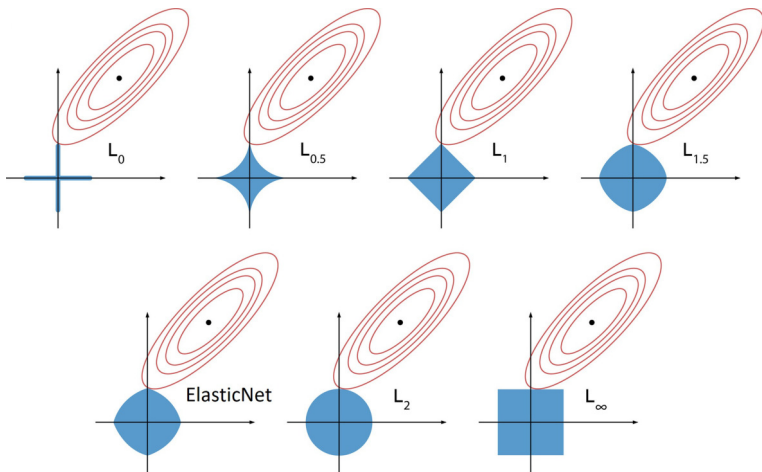
Lasso (L1):  $\sum_{j=1}^n |\alpha_j| \leq \kappa$

Ridge (L2):  $\sum_{j=1}^n \alpha_j^2 \leq \kappa$



## Геометрическая интерпретация отбора признаков

Сравнение регуляризаторов по различным  $L_p$ -нормам:



## Другие негладкие регуляризаторы для отбора признаков

Общий вид регуляризаторов ( $\mu$  — параметр селективности):

$$\frac{1}{2} \|F\alpha - y\|^2 + \tau \sum_{j=1}^n R_{\mu}(\alpha_j) \rightarrow \min_{\alpha}$$

Регуляризаторы с эффектом группировки зависимых признаков:

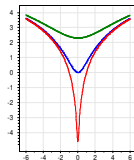
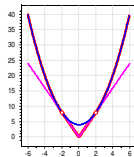
**Elastic Net:**  $R_{\mu}(\alpha) = \mu|\alpha| + \alpha^2$

**Support Features Machine (SFM):**

$$R_{\mu}(\alpha) = \begin{cases} 2\mu|\alpha|, & |\alpha| \leq \mu; \\ \mu^2 + \alpha^2, & |\alpha| \geq \mu; \end{cases}$$

**Relevance Features Machine (RFM):**

$$R_{\mu}(\alpha) = \ln(\mu\alpha^2 + 1)$$



## Метод главных компонент (Principal Component Analysis, PCA)

$f_1(x), \dots, f_n(x)$  — исходные числовые признаки;

$g_1(x), \dots, g_m(x)$  — новые числовые признаки,  $m \leq n$ ;

**Требование:** старые признаки  $f_j(x)$  должны линейно восстанавливаться по новым признакам  $g_s(x)$ :

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

как можно точнее на обучающей выборке  $x_1, \dots, x_\ell$ :

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

Задача преобразования признаков (feature transformation)

— это задача обучения без учителя, тут нет ответов  $y_i$

## Матричные обозначения

Матрицы «объекты–признаки», старая и новая:

$$F_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}; \quad G_{\ell \times m} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_\ell) & \dots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$U_{n \times m} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \quad \hat{F} = GU^T \overset{\text{ХОТИМ}}{\approx} F.$$

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

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

## Основная теорема метода главных компонент

### 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  $\|GU^T - F\|^2 = \|F\|^2 - \text{tr } \Lambda = \sum_{j=m+1}^n \lambda_j$ .

Тождество:  $\|F\|^2 = \text{tr}(F^T F) = \sum_{j=1}^n \lambda_j$

## Связь с сингулярным разложением

Если взять  $m = n$ , то:

①  $\|GU^T - F\|^2 = 0$ ;

② представление  $\hat{F} = GU^T = F$  точное и совпадает с сингулярным разложением при  $G = V\sqrt{\Lambda}$ :

$$F = GU^T = V\sqrt{\Lambda}U^T; \quad U^TU = I_m; \quad V^TV = I_m.$$

③ линейное преобразование  $U$  работает в обе стороны:

$$F = GU^T; \quad G = FU.$$

Поскольку новые признаки некоррелированы ( $G^TG = \Lambda$ ), преобразование  $U$  называется *декоррелирующим* (или преобразованием Карунена–Лоэва).



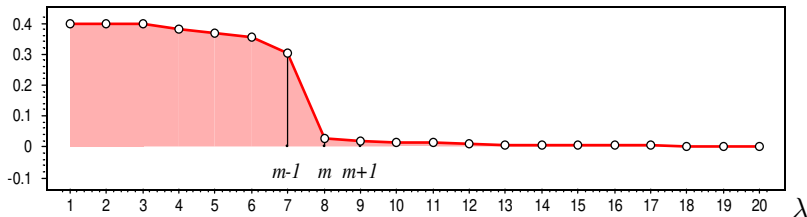
## Эффективная размерность выборки

Упорядочим с.з.  $F^T F$  по убыванию:  $\lambda_1 \geq \dots \geq \lambda_n \geq 0$ .

*Эффективная размерность выборки* — это наименьшее целое  $m$ , при котором

$$E_m = \frac{\|GU^T - F\|^2}{\|F\|^2} = \frac{\lambda_{m+1} + \dots + \lambda_n}{\lambda_1 + \dots + \lambda_n} \leq \varepsilon.$$

*Критерий «крутого склона»*: находим  $m$ :  $E_{m-1} \gg E_m$ :



## Решение задачи НК в новых признаках

Заменим  $F_{\ell \cdot n}$  на её приближение  $G_{\ell \cdot m} \cdot U^T_{m \cdot n}$ , предполагая  $m \leq n$ :

$$\|G \underbrace{U^T \alpha}_{\beta} - y\|^2 = \|G\beta - y\|^2 \rightarrow \min_{\beta}.$$

Связь нового и старого вектора коэффициентов:

$$\beta = U^T \alpha; \quad \alpha = U\beta.$$

Решение задачи наименьших квадратов относительно  $\beta$   
 (единственное отличие —  $m$  слагаемых вместо  $n$ ):

$$\beta^* = D^{-1} V^T y; \quad \alpha^* = U D^{-1} V^T y = \sum_{j=1}^m \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y);$$

$$G\beta^* = V V^T y = \sum_{j=1}^m v_j (v_j^T y);$$

## Спектральный метод наименьших квадратов

1. Построить SVD-разложение, упорядочить  $\lambda_1 \geq \dots \geq \lambda_n$
2. Отделить  $n - m$  наименьших с.з. от нуля:  $\lambda'_j := \lambda_j + \delta_j$

Частные случаи:

- $\lambda'_j := \lambda_j + \tau$  — гребневая регрессия
- $\lambda'_j := \lambda_j + [j > m] \cdot \infty$  — метод главных компонент

3. Применить формулы SVD для модификации МНК-решения:

$$\alpha^* = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} u_j (v_j^T y) \quad \longrightarrow \quad \alpha^* = \sum_{j=1}^n \frac{\sqrt{\lambda_j}}{\lambda'_j} u_j (v_j^T y)$$

$$F\alpha^* = \sum_{j=1}^n v_j (v_j^T y) \quad \longrightarrow \quad F\alpha^* = \sum_{j=1}^n \frac{\lambda_j}{\lambda'_j} v_j (v_j^T y)$$

## Задачи низкорангового матричного разложения

- Понижение размерности в задачах регрессии/классификации
- Генерация новых признаков
- Формирование сжатого представления данных

**Дано:** матрица  $Z = \|z_{ij}\|_{n \times m}$ ,  $(i, j) \in \Omega \subseteq \{1..n\} \times \{1..m\}$

**Найти:** матрицы  $X = \|x_{it}\|_{n \times k}$  и  $Y = \|y_{tj}\|_{k \times m}$  такие, что

$$\|Z - XY\|^2 = \sum_{(i,j) \in \Omega} \left( z_{ij} - \sum_t x_{it} y_{tj} \right)^2 \rightarrow \min_{X, Y}$$

Дополнительные ограничения, вынуждающие отказаться от SVD:

- неквадратичная функция потерь
- неотрицательное матричное разложение:  $x_{it} \geq 0$ ,  $y_{tj} \geq 0$
- разреженные данные:  $|\Omega| \ll nm$

- Метод наименьших квадратов (МНК)
  - нормальный некоррелированный шум
- Многомерная линейная регрессия
  - через сингулярное разложение
- Три приёма против мультиколлинеарности и переобучения
  - регуляризация, отбор и преобразование признаков
- $L_2$ -регуляризация, она же гребневая регрессия
  - тоже через сингулярное разложение
- $L_1$ -регуляризация (LASSO) и др. негладкие регуляризаторы
  - регулируемый отбор признаков
- Метод главных компонент — задача матричного разложения
  - снова через сингулярное разложение
- Другие методы матричных разложений и их приложения
  - в следующем семестре