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)»

Видеолекции: http://shad.yandex.ru/lectures

ШАД Яндекс • 14 марта 2017

Содержание

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;
  $y_i = y(x_i)$, $y : X \to Y$ — неизвестная зависимость;
- $a(x) = f(x, \alpha)$ — модель зависимости,
  $\alpha \in \mathbb{R}^p$ — вектор параметров модели.
- Метод наименьших квадратов (МНК):

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

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:

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

Метод максимума правдоподобия (ММП):

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

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

Теорема

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

Многомерная линейная регрессия

$f_1(x), \ldots, f_n(x)$ — числовые признаки;
Модель многомерной линейной регрессии:

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

Матричные обозначения:

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

Функционал квадрата ошибки:

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

Нормальная система уравнений

Необходимое условие минимума в матричном виде:

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

откуда следует *нормальная система* задачи МНК:

$$F^{\mathsf{T}}F\alpha = F^{\mathsf{T}}y,$$

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

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

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

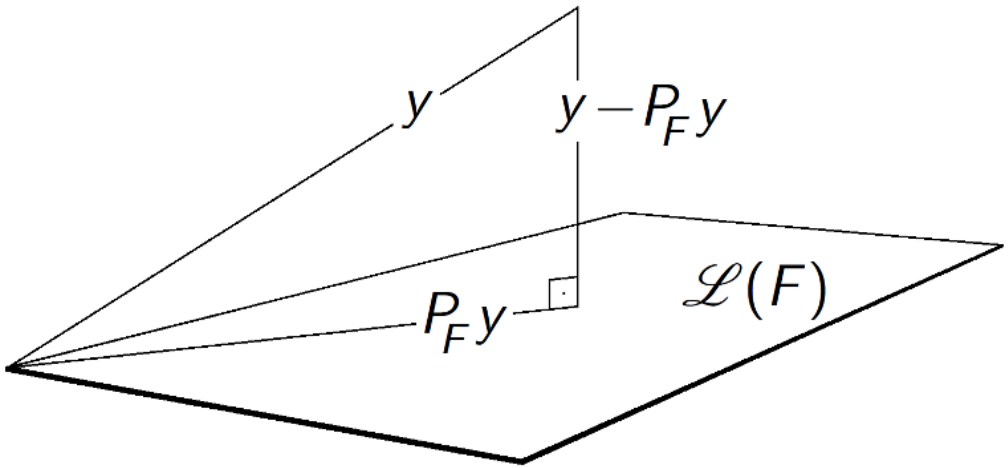
где $P_F = FF^{+} = F(F^{\mathsf{T}}F)^{-1}F^{\mathsf{T}}$ — *проекционная матрица*.

Геометрическая интерпретация МНК

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

$$\mathscr{L}(F) = \left\{ \sum_{j=1}^{n} \alpha_j f_j \;\middle|\; \alpha \in \mathbb{R}^n \right\}$$

$P_F = F(F^\mathsf{T}F)^{-1}F^\mathsf{T}$ — проекционная матрица
$P_F y$ — проекция вектора $y \in \mathbb{R}^\ell$ на подпространство $\mathscr{L}(F)$
$(I_\ell - P_F)y$ — проекция y на его ортогональное дополнение



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

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, $D = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}\right)$, $\lambda_j \geqslant 0$ — собственные значения матриц $F^{\mathsf{T}}F$ и FF^{T}.

Решение МНК через сингулярное разложение

Псевдообратная F^+, вектор МНК-решения α^*,
МНК-аппроксимация целевого вектора $F\alpha^*$:

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

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

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

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

Проблема мультиколлинеарности

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

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

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

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

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

Проблема мультиколлинеарности и переобучения

Если матрица $S = F^{\mathsf{T}}F$ плохо обусловлена, то:

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

Стратегии устранения мультиколлинеарности и переобучения:

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

Регуляризация (гребневая регрессия)

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

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

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

Вероятностная интерпретация: априорное распределение вектора α — гауссовское с ковариационной матрицей σI_n.

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

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

Преимущество сингулярного разложения:
можно подбирать параметр τ, вычислив SVD только один раз.

Регуляризованный МНК через сингулярное разложение

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

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

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

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

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

Выбор параметра регуляризации τ

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

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

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

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

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

Регуляризация сокращает «эффективную размерность»

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

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

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

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

При использовании регуляризации:

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

Регуляризация для отбора признаков

LASSO — Least Absolute Shrinkage and Selection Operator

$$\|F\alpha - y\|^2 + \mu \sum_{j=1}^{n} |\alpha_j| \to \min_{\alpha} \quad \Longleftrightarrow \quad \begin{cases} \|F\alpha - y\|^2 \to \min\limits_{\alpha}; \\ \sum\limits_{j=1}^{n} |\alpha_j| \leqslant \varkappa; \end{cases}$$

После замены переменных

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

ограничения принимают канонический вид:

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

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

Негладкие регуляризаторы

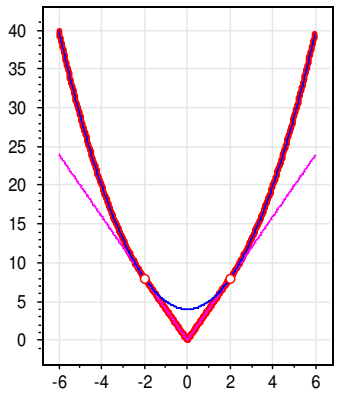
Elastic Net:

$$\frac{1}{2}\|F\alpha - y\|^2 + \mu\sum_{j=1}^{n}|\alpha_j| + \frac{\tau}{2}\sum_{j=1}^{n}\alpha_j^2 \;\to\; \min_{\alpha}.$$

Support Features Machine:

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

$$R_\mu(\alpha_j) = \begin{cases} 2\mu|\alpha_j|, & |\alpha_j| \leqslant \mu; \\ \mu^2 + \alpha_j^2, & |\alpha_j| \geqslant \mu; \end{cases}$$



Применение этих методов требует выбора траектории регуляризации (regularization path) в пространстве (μ, τ)

Метод главных компонент: постановка задачи

$f_1(x), \ldots, f_n(x)$ — исходные числовые признаки;
$g_1(x), \ldots, g_m(x)$ — новые числовые признаки, $m \leqslant n$;

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 $x_1, \ldots, x_\ell$:

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

Матричные обозначения

Матрицы «объекты–признаки», старая и новая:

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}; \quad \underset{\ell\times m}{G} = \begin{pmatrix} g_1(x_1) & \ldots & g_m(x_1) \\ \ldots & \ldots & \ldots \\ g_1(x_\ell) & \ldots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$\underset{n\times m}{U} = \begin{pmatrix} u_{11} & \ldots & u_{1m} \\ \ldots & \ldots & \ldots \\ u_{n1} & \ldots & u_{nm} \end{pmatrix}; \qquad \hat{F} = GU^{\mathsf{T}} \overset{\text{хотим}}{\approx} F.$$

Найти: и новые признаки G, и преобразование U:

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

Основная теорема метода главных компонент

Теорема

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

При этом:

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

Связь с сингулярным разложением

Если взять $m = n$, то:

1. $\left\| GU^{\mathsf{T}} - F \right\|^2 = 0$;
2. представление $\hat{F} = GU^{\mathsf{T}} = F$ точное и совпадает с сингулярным разложением при $G = V\sqrt{\Lambda}$:
$$F = GU^{\mathsf{T}} = V\sqrt{\Lambda}U^{\mathsf{T}}; \quad U^{\mathsf{T}}U = I_m; \quad V^{\mathsf{T}}V = I_m.$$
3. линейное преобразование U работает в обе стороны:
$$F = GU^{\mathsf{T}}; \quad G = FU.$$

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

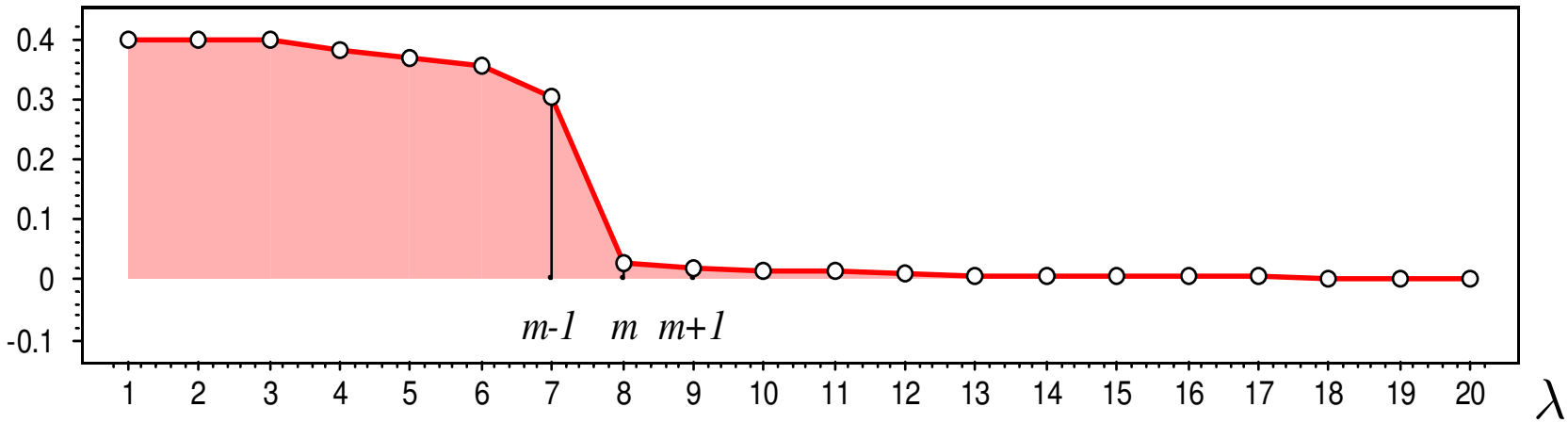
Эффективная размерность выборки

Упорядочим с.з. $F^{\mathsf{T}}F$ по убыванию: $\lambda_1 \geqslant \ldots \geqslant \lambda_n \geqslant 0$.

Эффективная размерность выборки — это наименьшее целое m, при котором

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

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



Решение задачи НК в новых признаках

Заменим $\underset{\ell\cdot n}{F}$ на её приближение $\underset{\ell\cdot m}{G}\cdot\underset{m\cdot n}{U^{\mathsf{T}}}$, предполагая $m \leqslant n$:

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

Связь нового и старого вектора коэффициентов:

$$\beta = U^{\mathsf{T}}\alpha; \qquad \alpha = U\beta.$$

Решение задачи наименьших квадратов относительно β (единственное отличие — m слагаемых вместо n):

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

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

Обобщение. Спектральные методы решения задачи НК

1. Построить SVD-разложение $(\lambda_1 \geqslant \cdots \geqslant \lambda_{m+1} \geqslant \cdots \geqslant \lambda_n)$.
2. Игнорировать m наименьших с. з. или иным способом отделить с. з. от нуля: $\lambda'_j := f(\lambda_j)$.

Частные случаи:

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

3. Применить формулы SVD для модификации МНК-решения:

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

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

Задачи низкорангового матричного разложения

- Понижение размерности в задачах регрессии
- Понижение размерности в задачах классификации
- 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

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

Дополнительные ограничения, вынуждающие отказаться от SVD:

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

Примеры прикладных задач матричного разложения

1. **Выявление интересов в рекомендательных системах (recommender systems, collaborative filtering)**

$$z_{iu} = \sum_t p_{it} q_{tu}$$

дано: z_{iu} — рейтинги товаров i, поставленные пользователем u;
найти: p_{it} — профиль интересов товара i;
q_{tu} — профиль интересов пользователя u.

2. **Разделение смеси химических веществ по данным жидкостной хроматографии**

$$z_{t\lambda} = \sum_i x_{ti} y_{i\lambda}$$

дано: $z_{t\lambda}$ — выход сканирующего УФ-детектора;
найти: x_{ti} — хроматограмма i-го вещества, t — время;
$y_{i\lambda}$ — спектр i-го вещества, λ — длина волны.

Примеры прикладных задач матричного разложения

3. **Латентный семантический анализ коллекций текстов (тематическое моделирование)**

$$z_{wd} = \sum_t \varphi_{wt}\theta_{td}$$

дано: $z_{wd} = p(w|d)$ — частоты слов w в документах d;
найти: $\varphi_{wt} = p(w|t)$ — распределения слов w в темах t,
$\theta_{td} = p(t|d)$ — распределения тем t в документах d.

4. **Оценивание экспрессии генов по данным ДНК-микрочипов с учётом кросс-гибридизации**

$$z_{pk} = \sum_g a_{pg}\, c_{gk}$$

дано: z_{pk} — интенсивность свечения p-й пробы на k-м чипе;
найти: a_{pg} — коэффициент сродства p-й пробы g-му гену,
c_{gk} — концентрация g-го гена на k-м чипе.

Резюме в конце лекции

- Метод наименьших квадратов
  — нормальный некоррелированный шум
- Многомерная линейная регрессия
  — через сингулярное разложение
- Гребневая регрессия
  — тоже через сингулярное разложение
- Метод главных компонент
  — тоже через сингулярное разложение
- Негладкие регуляризаторы: Лассо Тибширани, Elastic Net
  — отбор признаков