# 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:

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

откуда следует *нормальная система* задачи МНК:

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

где $F^{\intercal}F$ — $n \times n$-матрица.

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

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

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

# 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 $FF^{\mathsf{T}}$;
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 $FF^{\mathsf{T}}$, $\sqrt{\lambda_j}$ — сингулярные числа матрицы $F$.

# Решение МНК через сингулярное разложение

Псевдообратная $F^+$, вектор МНК-решения $\alpha^*$,
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:** мультиколлинеарность при $\lambda_j \to 0$.

# Проблема мультиколлинеарности и переобучения

Если имеются сингулярные числа, близкие к нулю, то:

- матрица $\Sigma = F^{\mathsf{T}}F$ плохо обусловлена;
- решение становится неустойчивым и неинтерпретируемым, слишком большие коэффициенты $\|\alpha_j^*\|$ разных знаков;
- 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:

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

# Резюме

- Задача многомерной линейной регрессии может быть решена через сингулярное разложение
- Мультиколлинеарность приводит к плохой обусловленности, неустойчивости и переобучению
- Методы устранения мультиколлинеарности (гребневая регрессия, метод главных компонент) также связаны с сингулярным разложением (об этом в следующих лекциях)

# Регуляризация (гребневая регрессия)

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

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

где $\tau$ — неотрицательный *параметр регуляризации*.

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

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

**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 $F\alpha_\tau^*$:

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

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

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

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

# Выбор параметра регуляризации $\tau$

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

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

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

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

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

# Регуляризация сокращает «эффективную размерность»

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

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

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

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

При использовании регуляризации:

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

# LASSO приводит к отбору признаков в линейных моделях

LASSO — Least Absolute Shrinkage and Selection Operator, два эквивалентных варианта постановки задачи:

$$Q(\alpha) = \|F\alpha - y\|^2 \to \min_{\alpha} \quad \text{при} \quad \sum_{j=1}^{n} |\alpha_j| \leqslant \varkappa;$$

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

После замены переменных

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

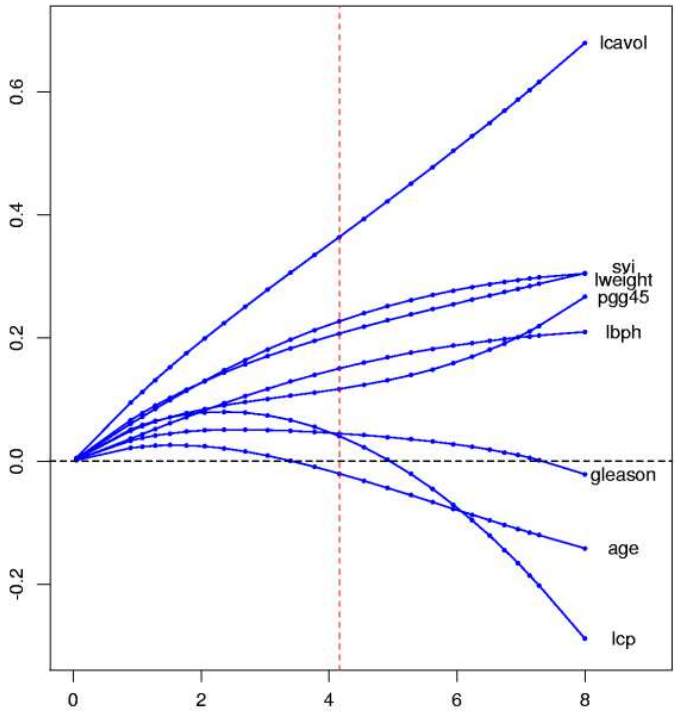
ограничения принимают канонический вид:

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

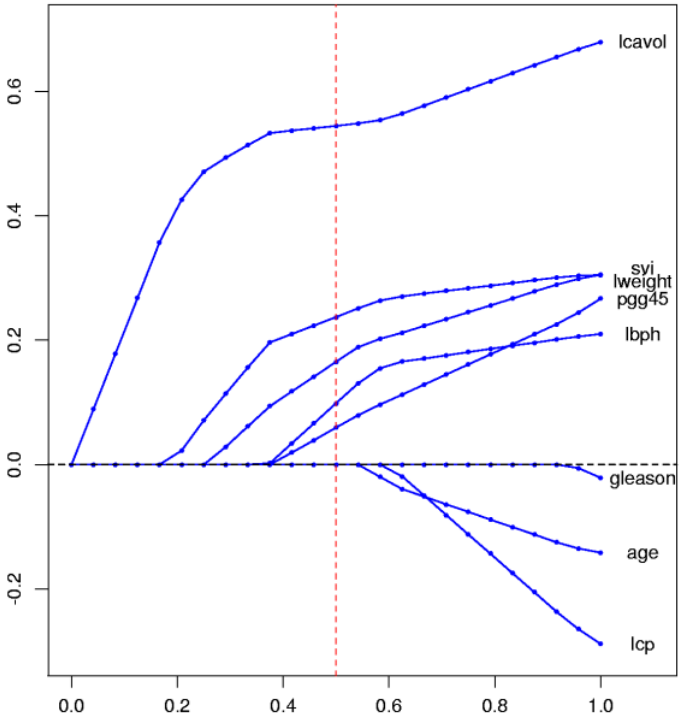
Чем меньше $\varkappa$, тем больше $j$ таких, что $\alpha_j^+ = \alpha_j^- = 0$.

# Сравнение гребневой регрессии и Лассо

Зависимость $\{\alpha_j\}$ от $\frac{1}{\tau}$




Зависимость $\{\alpha_j\}$ от $\varkappa$




Задача диагностики рака (prostate cancer, UCI)

---

*T.Hastie, R.Tibshirani, J.Friedman.* The Elements of Statistical Learning. Springer, 2001.

# Резюме

- Гребневая регрессия удобно вводится и интерпретируется через сингулярное разложение
- Гребневая регрессия сокращает веса признаков
- LASSO обнуляет веса признаков
- Оба метода имеют параметр регуляризации (селективности), позволяющий определять число признаков (сложность модели) по внешним критериям (по кросс-валидации).

# 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).

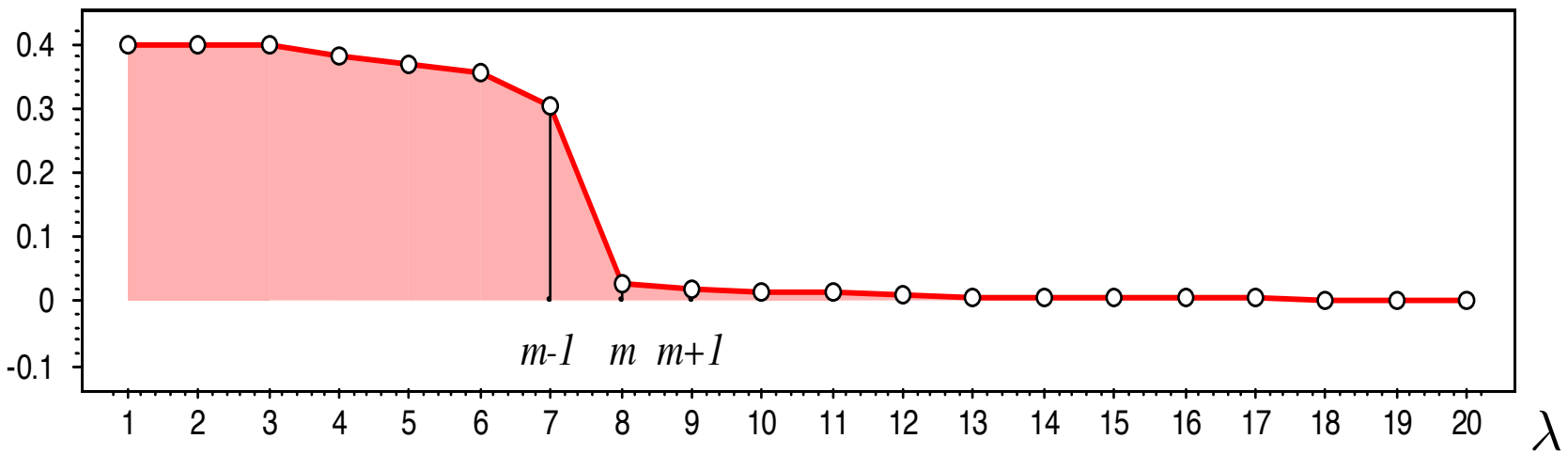
# Эффективная размерность выборки

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 для МЛР в новых признаках

Задача наименьших квадратов для МЛР: $\|F\alpha - y\|^2 \to \min\limits_{\alpha}$.

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 $\beta$ (единственное отличие — $m$ слагаемых вместо $n$):

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

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

# Резюме

- Метод главных компонент позволяет приближать матрицу её низкоранговым разложением.
- Для этого достаточно взять из SVD-разложения первые $m$ сингулярных чисел и векторов матрицы.
- Этот приём широко используется в анализе данных — в задачах регрессии, классификации, сжатия данных, обработки изображений.