Уменьшение эффективной размерности пространства

Алексей Романенко , alexromsput@gmail.com,
Анастасия Зухба, a__l@mail.ru

апрель 2017

Задачи обучения без учителя

Задача кластеризации:
X^ℓ — признаковое описание объектов Y — отсутствует
$f_j\colon X \to D_j,\ \ j = 1, \ldots, n$ — признаки объектов (features).

Типы признаков:

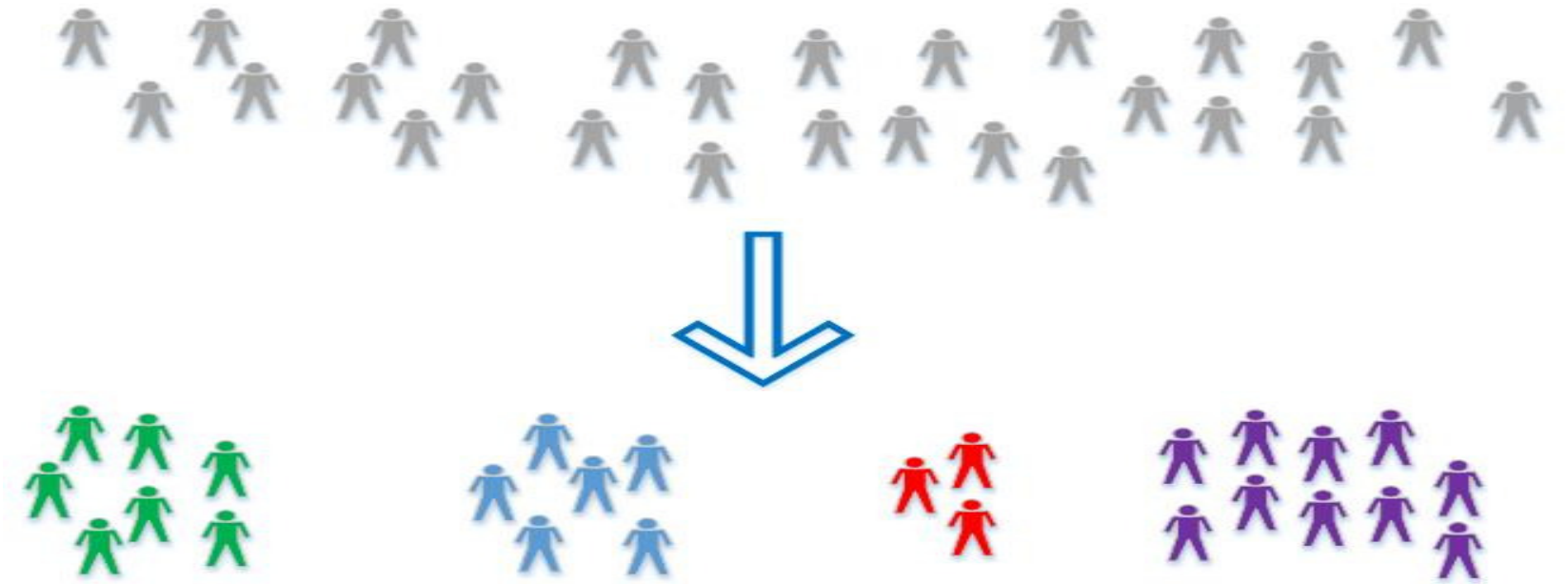
- $D_j = \{0, 1\}$ — *бинарный* признак f_j;
- $|D_j| < \infty$ — *номинальный* признак f_j;
- $|D_j| < \infty$, D_j упорядочено — *порядковый* признак f_j;
- $D_j = \mathbb{R}$ — *количественный* признак f_j.

Вектор $\big(f_1(x), \ldots, f_n(x)\big)$ — *признаковое описание* объекта x.

Матрица «*объекты–признаки*» (features data)

$$F = \big\|f_j(x_i)\big\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}$$

Задачи обучения без учителя

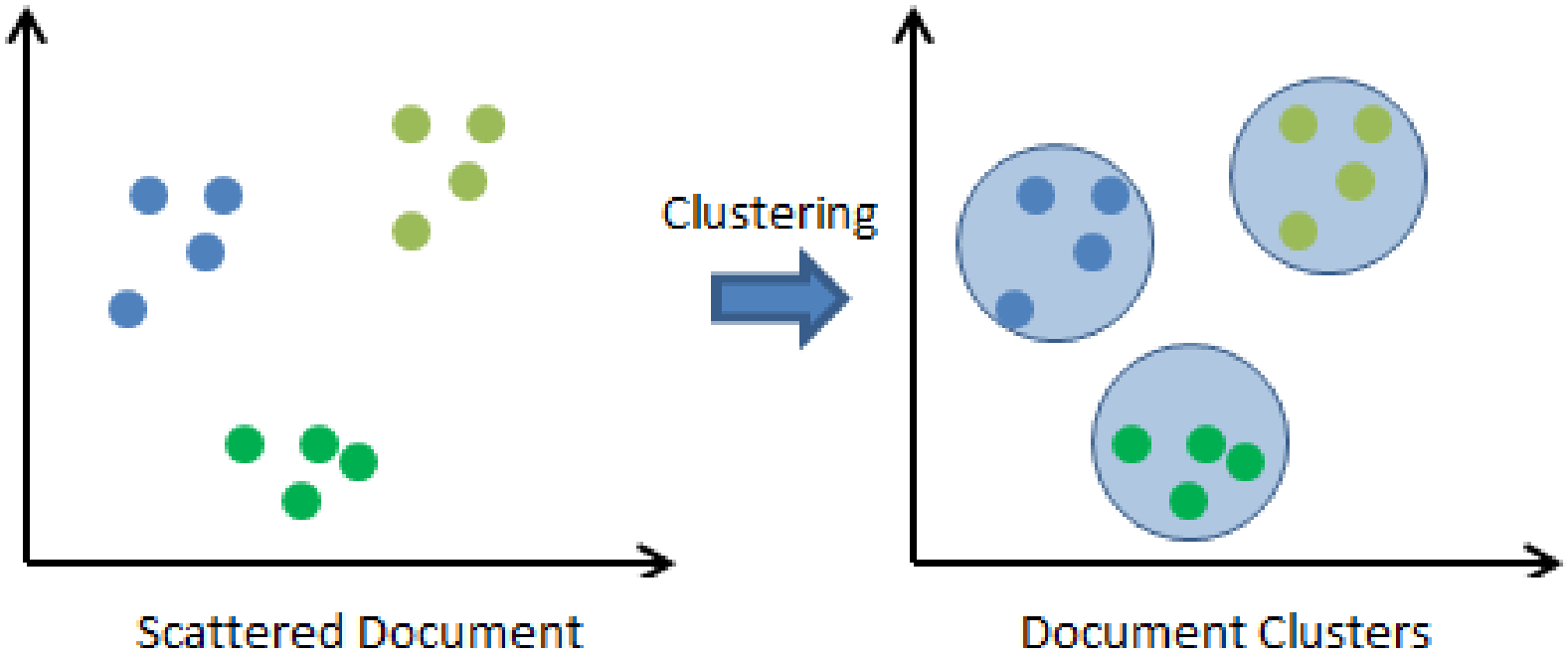


Типы задач без учителя:

- Кластеризация
- Поиск правил ассоциации
- Сокращение размерности
- Визуализация данных

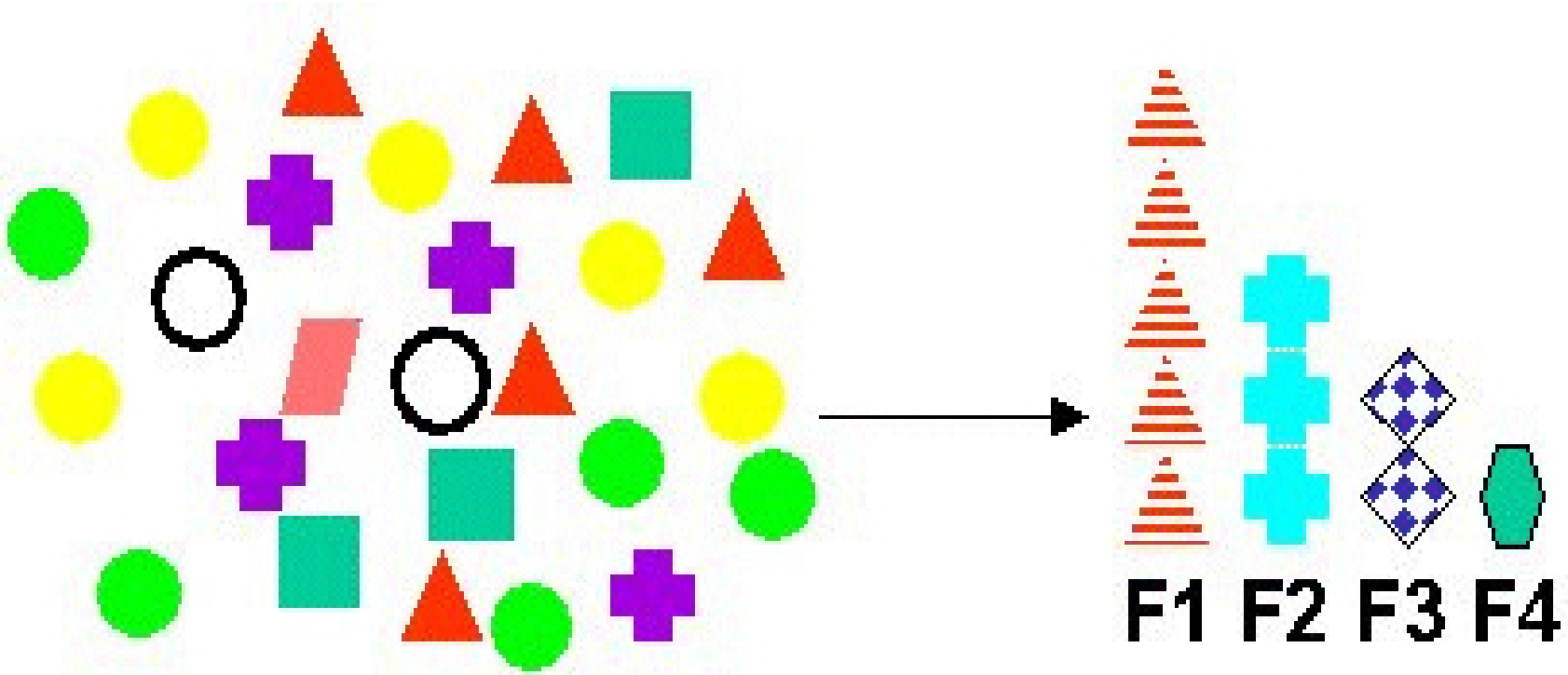
Примеры задач

Кластеризация



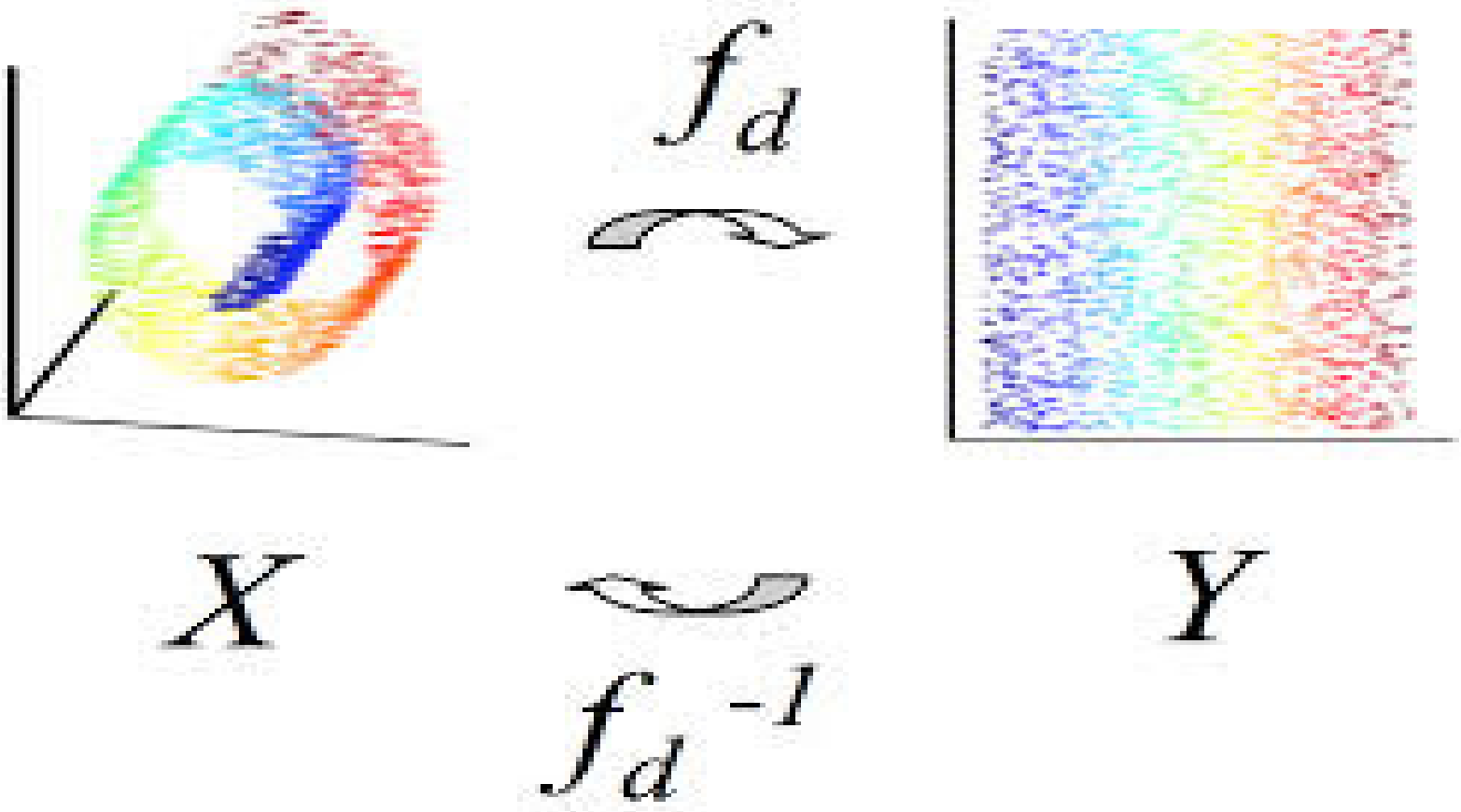
Примеры задач

Feature Extraction



Примеры задач

Снижение размерности



Проблемы больших размерностей

Почему большая размерность признакового пространства - это плохо?

Проблемы больших размерностей

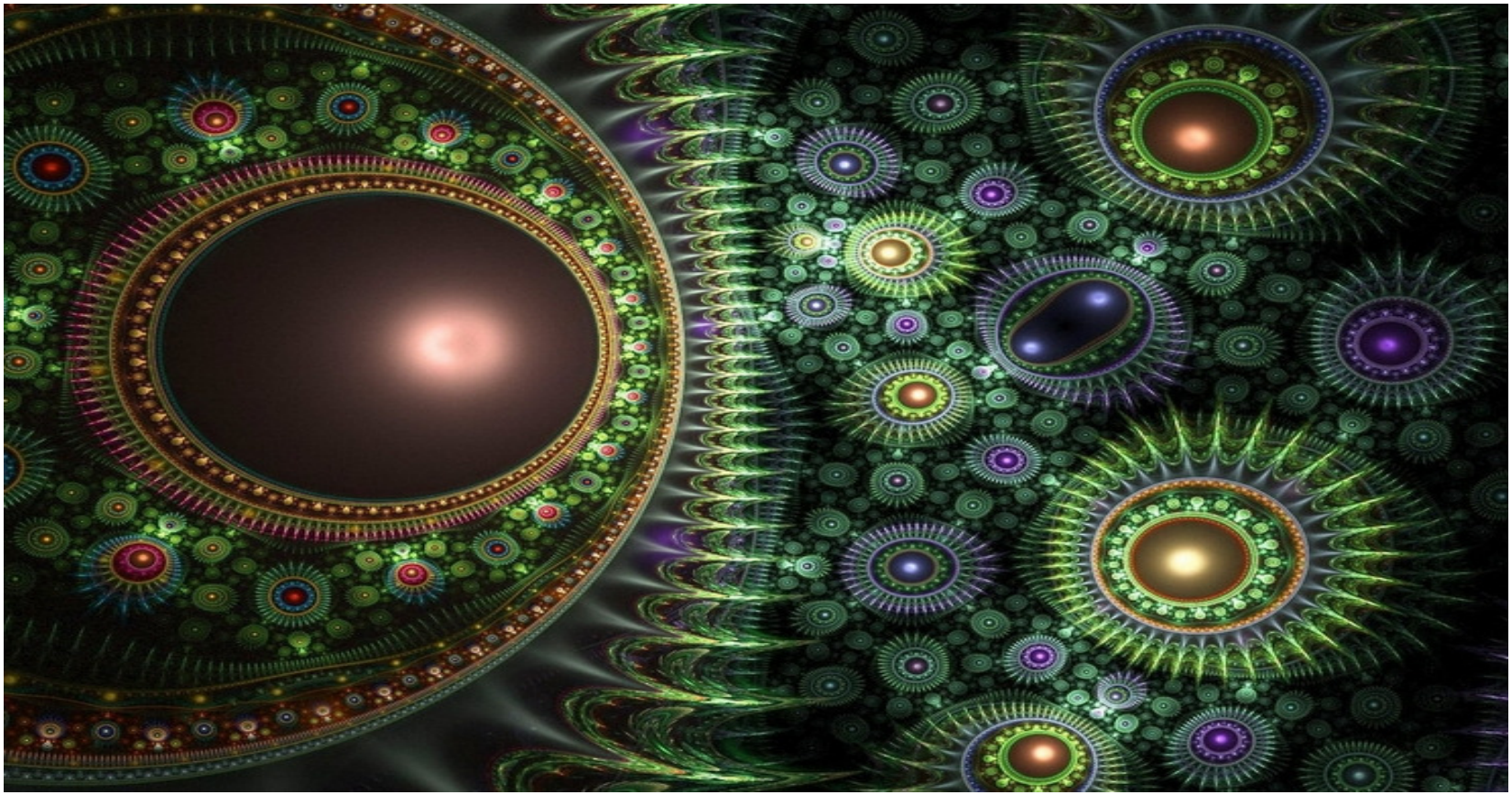
Почему большая размерность признакового пространства - это плохо?

Мультиколлинеарность

Проблемы больших размерностей

Почему большая размерность признакового пространства - это плохо?

Проклятие размерности



Проблемы больших размерностей

Почему большая размерность признакового пространства - это плохо?

Неинтерпретируемость

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 $\underset{n\times n}{F^{\intercal}F}$ — ковариационная матрица 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 $P_F = FF^{+} = F(F^{\intercal}F)^{-1}F^{\intercal}$ — *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^{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 имеются $\lambda_j \to 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо: $\|\alpha\| \to \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»: $Q(\alpha^*) = \|F\alpha^* - y\|^2 \to 0$;
- мультиколлинеарность влечёт *переобучение*.

Проблема мультиколлинеарности

Если имеются $\lambda_j \to 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо: $\|\alpha\| \to \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»: $Q(\alpha^*) = \|F\alpha^* - y\|^2 \to 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

Проблема мультиколлинеарности

Если имеются $\lambda_j \to 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо: $\|\alpha\| \to \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»: $Q(\alpha^*) = \|F\alpha^* - y\|^2 \to 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \to \min$;

Проблема мультиколлинеарности

Если имеются $\lambda_j \to 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо: $\|\alpha\| \to \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»: $Q(\alpha^*) = \|F\alpha^* - y\|^2 \to 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \to \min$;
- Отбор признаков: $f_1, \ldots, f_n \to f_{j_1}, \ldots, f_{j_m}$, $m \ll n$.

Проблема мультиколлинеарности

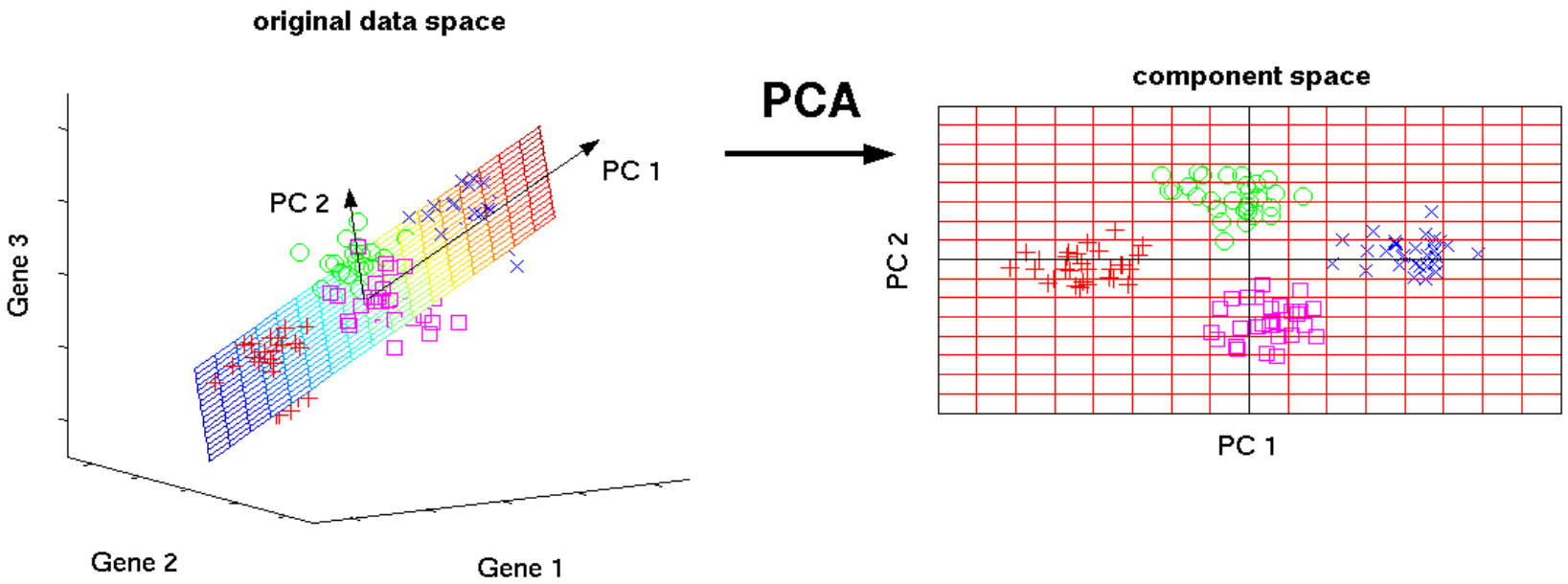
Если имеются $\lambda_j \to 0$, то

- МНК-решение α^* неустойчиво и неинтерпретируемо: $\|\alpha\| \to \infty$;
- ответы на новых объектах $y' = F'\alpha^*$ неустойчивы;
- в то время как на обучении, казалось бы, «всё хорошо»: $Q(\alpha^*) = \|F\alpha^* - y\|^2 \to 0$;
- мультиколлинеарность влечёт *переобучение*.

Три стратегии устранения мультиколлинеарности:

- Регуляризация: $\|\alpha\| \to \min$;
- Отбор признаков: $f_1, \ldots, f_n \to f_{j_1}, \ldots, f_{j_m}, \quad m \ll n$.
- Преобразование признаков: $f_1, \ldots, f_n \to g_1, \ldots, g_m, \quad m \ll n$;

Метод главных компонент: постановка задачи



PCA visalisation:
http://setosa.io/ev/principal-component-analysis/

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

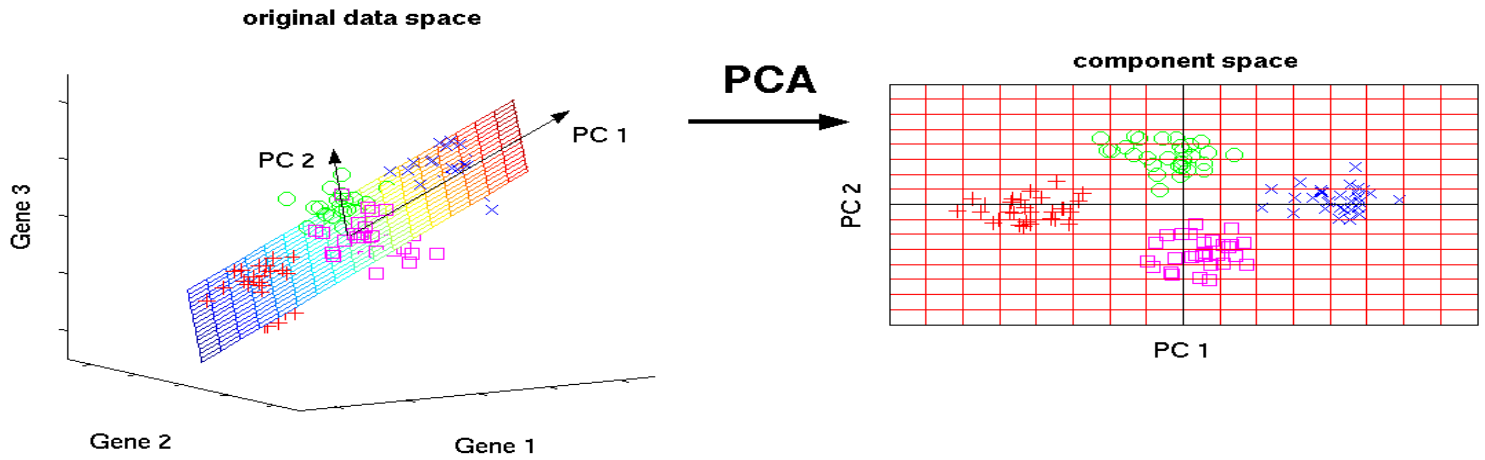
Теорема

Если $m \leqslant \operatorname{rk} F$, *то минимум* $\left\|GU^{\mathrm{T}} - F\right\|^2$ *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:

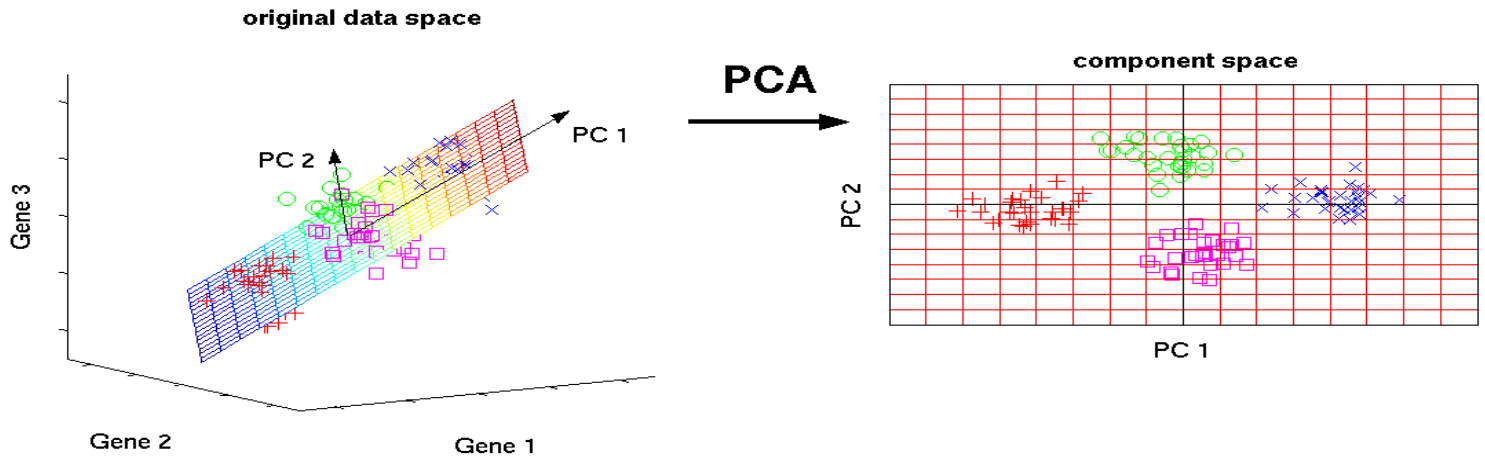
1. *матрица* U *ортонормирована:* $U^{\mathrm{T}}U = I_m$;
2. *матрица* G *ортогональна:* $G^{\mathrm{T}}G = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$;
3. $U\Lambda = F^{\mathrm{T}}FU$; $G\Lambda = FF^{\mathrm{T}}G$;
4. $\left\|GU^{\mathrm{T}} - F\right\|^2 = \|F\|^2 - \operatorname{tr}\Lambda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

Три интерпретации метода PCA



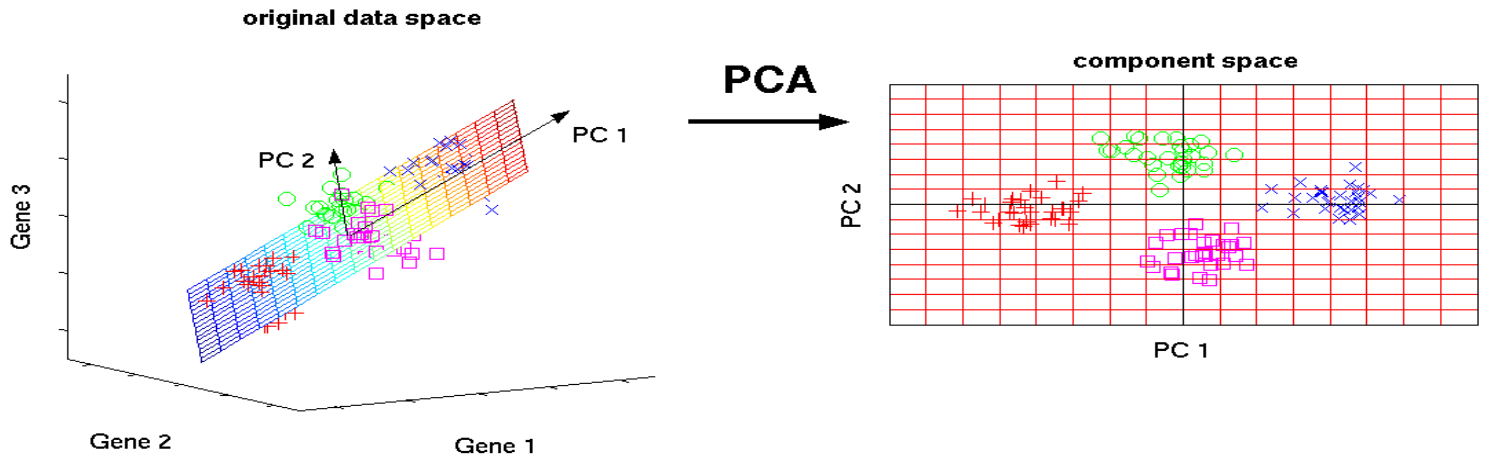
1. найти подпространства меньшей размерности, в ортогональной проекции на которые разброс данных (то есть среднеквадратичное отклонение от среднего значения) максимален;

Три интерпретации метода PCA



1. ...
2. найти подпространства меньшей размерности, в ортогональной проекции на которые среднеквадратичное расстояние между точками максимально;

Три интерпретации метода PCA



1. ...
2. ...
3. построить такое ортогональное преобразование координат, в результате которого корреляции между отдельными координатами обратятся в нуль.

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).

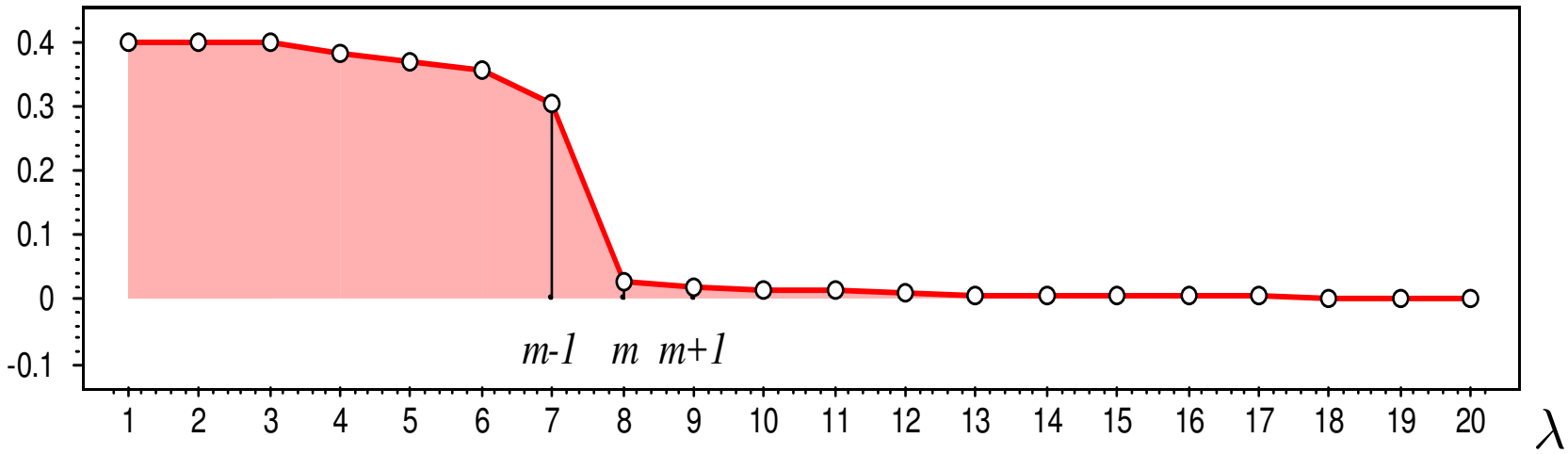
Эффективная размерность выборки

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 F на её приближение $GU^\intercal$:

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

Связь нового и старого вектора коэффициентов:

$$\alpha = U\beta; \qquad \beta = U^\intercal\alpha.$$

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

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

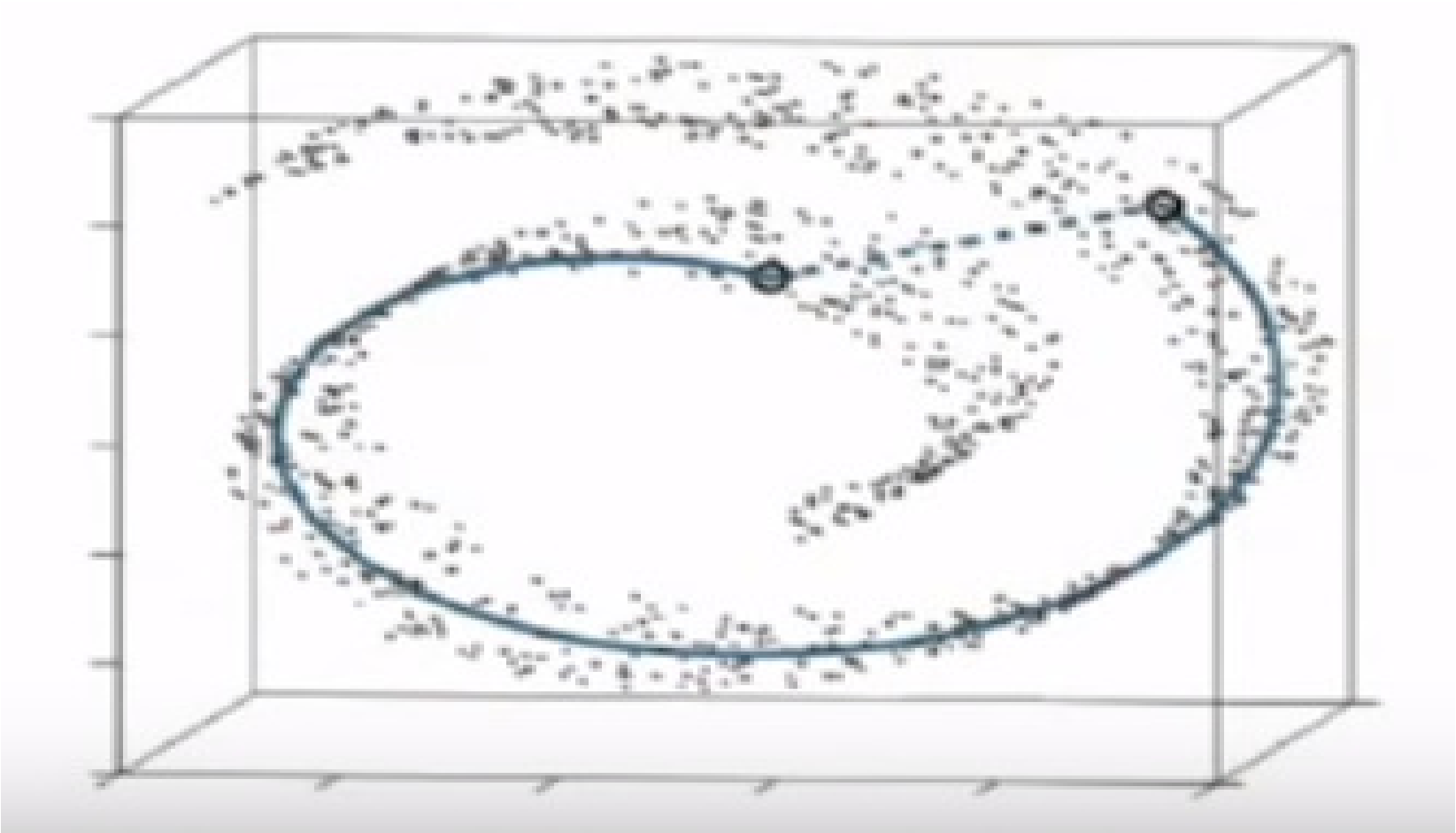
$$G\beta^* = VV^\intercal y = \sum_{j=1}^{m} v_j(v_j^\intercal y);$$

ProCons по PCA, SVD

- PCA помогает избавиться от "лишних"размерностей
- PCA строит оптимальное линейное преобразование
- PCA применим для больших размерностей
- существуют ядерные вариант PCA (KernelPCA)

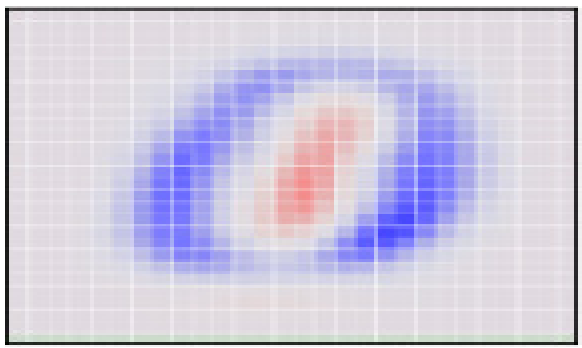
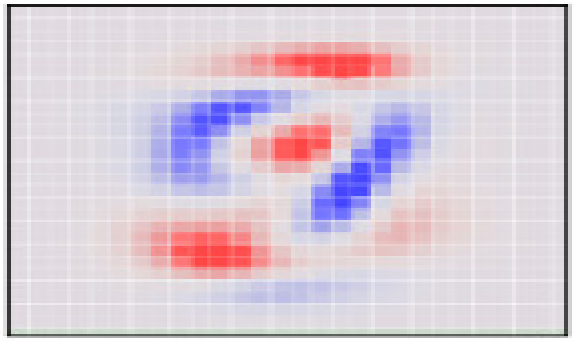
- если требуется сложное нелинейное преобразование, PCA не работает
- не подходит для визуализации на двухмерной плоскости при большой размерности исходных пространств

ProCons по PCA, SVD

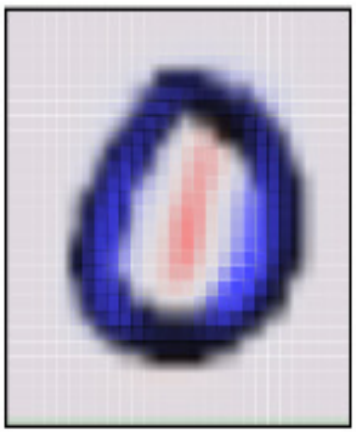


ProCons по PCA, SVD

MNIST with PCA

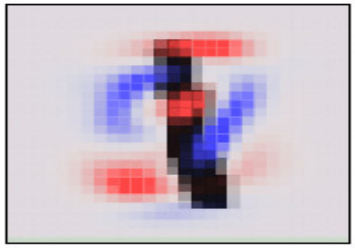



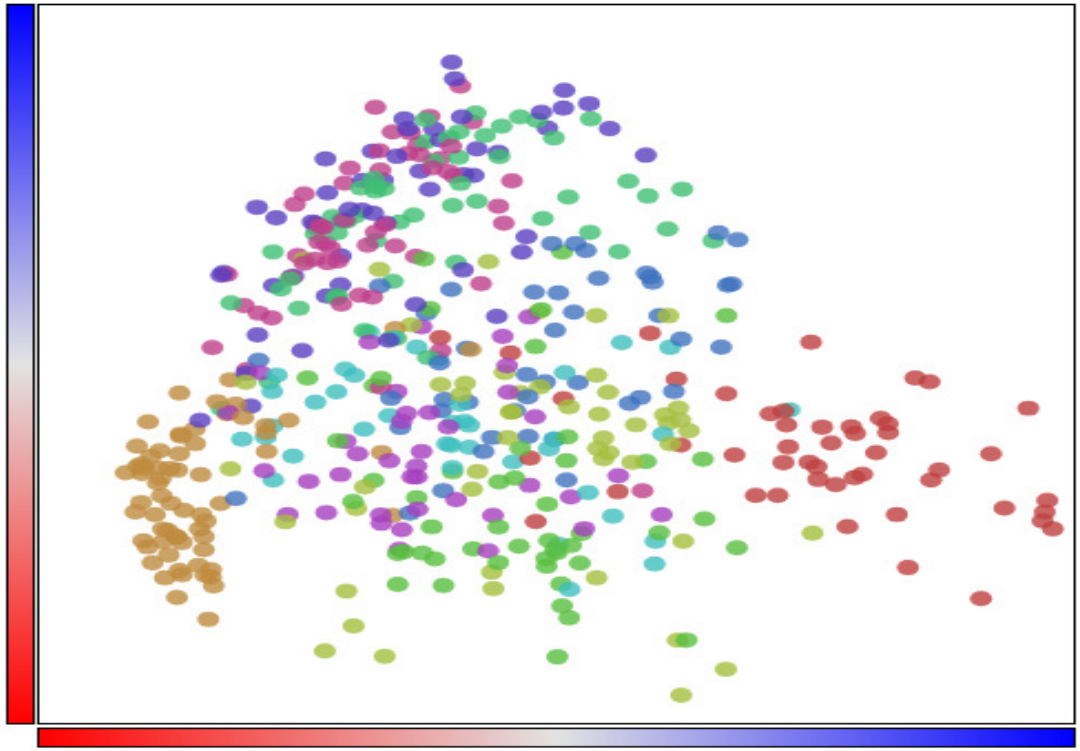





ProCons по PCA, SVD

MNIST with PCA





t-SNE. Шаг первый.

$x_1, \dots x_N$ – точки (data points) в исходном пространстве $\mathbb{R}^D$.
$z_1, \dots z_N$ – точки (map points) в пространстве $\mathbb{R}^2$.

Построение плотностей в изначальном пространстве:

$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

Это выражение показывает, насколько точка x_j близка к x_i, при гауссовом распределении вокруг x_i с заданной дисперсией σ_i^2. Дисперсия различна для каждой точки.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

t-SNE. Шаг второй.

Построение приближения в пространстве меньшей размерности:

$$q_{ij} = \frac{(1 + ||z_i - z_j||^2)^{-1}}{\sum_{k \neq m}(1 + ||z_k - z_m||^2)^{-1}}$$

Минимизируется расстояния Кульбака-Лейблера:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

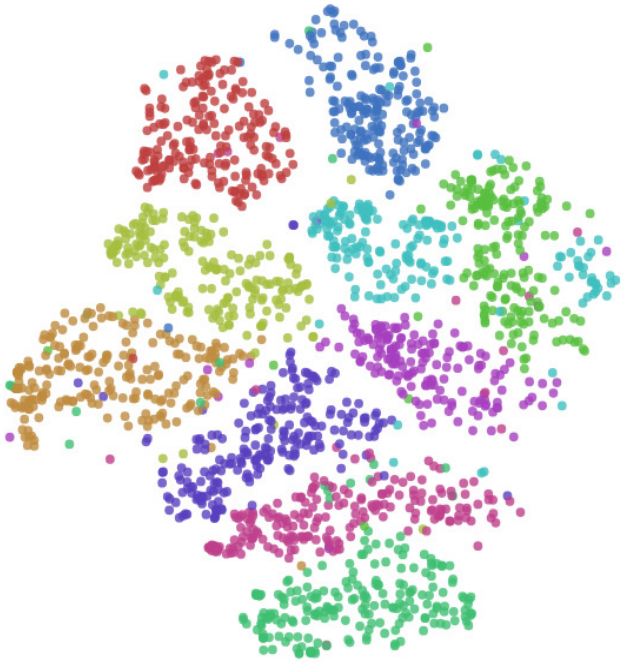
Градиент может быть вычислен аналитически:

$$\frac{\partial KL(P||Q)}{\partial z_i} = 4 \sum_j (p_{ij} - q_{ij}) g(|x_i - x_j|) u_{ij}$$

где $g(z) = \frac{z}{1+z^2}$; u_{ij} – единичный вектор от z_j к z_i.

tSNE for MNIST

MNIST with tSNE



A t-SNE plot of MNIST

Demo see here
http://colah.github.io/posts/2014-10-Visualizing-MNIST/

Multi-dimensional Scaling

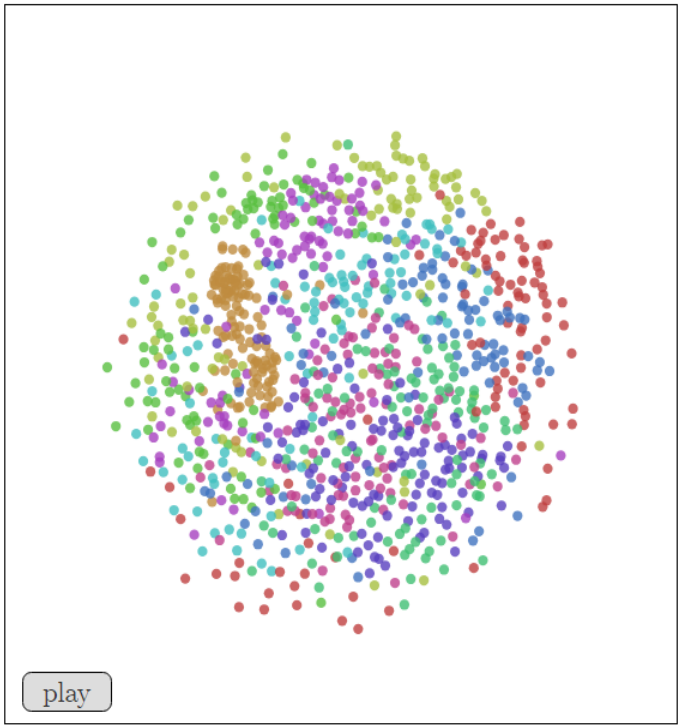
Многомерное масштабирование (Multidimensional scaling, MDS) пытается моделировать сходство данных как расстояния в геометрических пространствах.

MDS ищет для данных представление низкой размерности максимально учитывая расстояния в изначальном многомерном пространстве.

1. **Метрический вариант:** приближает или сохраняет расстояния.
2. **Неметрический вариант:** приближает или сохраняет порядок расстояний.

MDS for MNIST

MNIST with MDS



Visualizing MNIST with MDS

See demo here
http://colah.github.io/posts/2014-10-Visualizing-MNIST/

Заключение по tSNE

PRO tSNE:

- отлично визуализируют (2D, 3D) данные
- испольюзуются для предобработки данных, анализа структуры данных see here http://distill.pub/2016/misread-tsne/
-

Cons tSNE:

- стохастичность (PCA детерминирован)
- интерпретируемость данных
- не применим (сложно применим) для новых данных
- не даёт ответ об оптимальной размерности пространства

Обратная связь

Отзывы о прошедших лекциях и семинарах можно и нужно оставлять здесь:
https://docs.google.com/forms/d/e/1FAIpQLSdefy8neFtoxDlXD3toHi3fW
23APTRj-GuTX8wtAJahQ/viewform?c=0w=1