Машинное обучение в науках о Земле

Михаил Криницкий

к.т.н.
Зав. лабораторией машинного обучения в науках о Земле МФТИ
с.н.с. Института океанологии РАН им. П.П. Ширшова

Сокращение размерности

Михаил Криницкий

к.т.н.
Зав. лабораторией машинного обучения в науках о Земле МФТИ
с.н.с. Института океанологии РАН им. П.П. Ширшова

ПЛАН ЛЕКЦИИ

- Сокращение размерности
  - PCA (Principal Components Analysis, Метод главных компонент)
  - t-SNE (t-distributed Stochastic Neighbor Embedding)*
  - UMAP (Uniform Manifold Approximation and Projection)**

*Hinton, Geoffrey E., and Sam Roweis. "Stochastic neighbor embedding." Advances in neural information processing systems 15 (2002).

**McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction // arXiv:1802.03426 [cs, stat]. 2020.

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение с учителем»
  - восстановление регрессии
  - классификация
- «Обучение без учителя»
  - Снижение размерности
  - Кластеризация

Снижение размерности

что у меня есть?

признаковое описание объектов

что я хочу?

снизить размерность признакового описания объектов

Снижение размерности

что у меня есть?

признаковое описание объектов

Так а чо, какие проблемы-то?

что я хочу?

снизить размерность признакового описания объектов

Снижение размерности

что у меня есть?

- признаковое описание объектов

что я хочу?

- снизить размерность признакового описания объектов

какие я предъявляю пожелания?

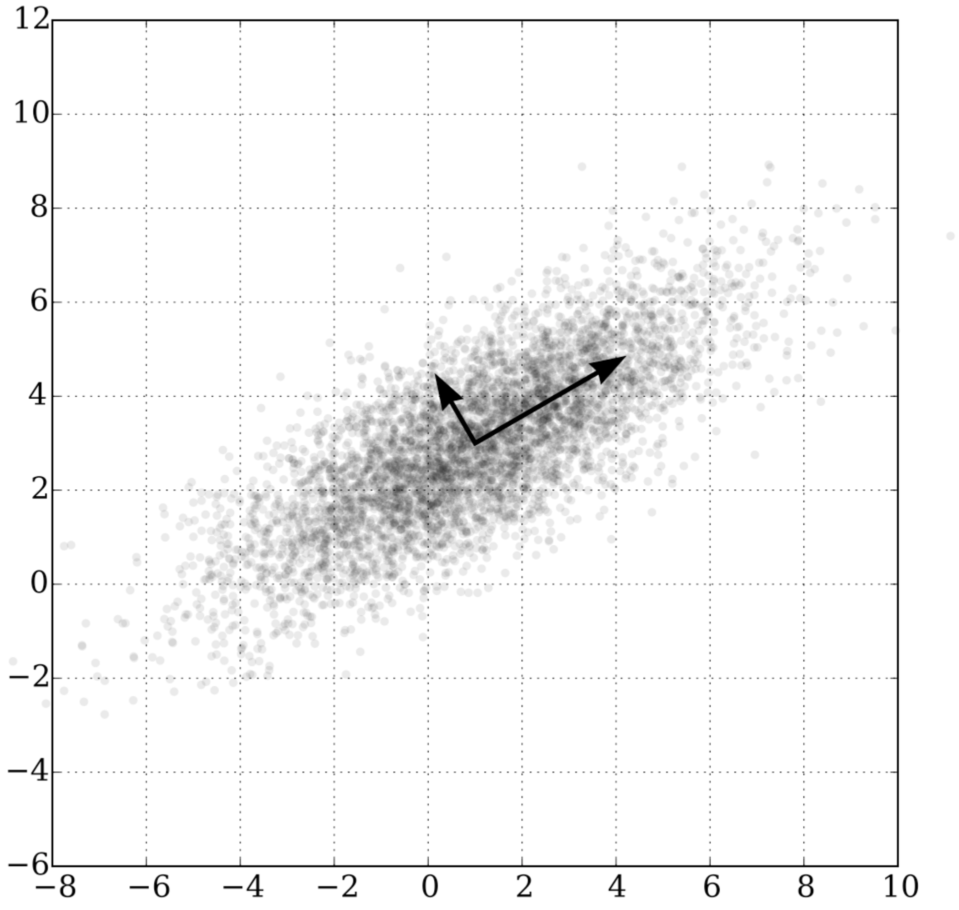
- сохранение отношений близости
- сохранение порядка
- (?) визуализация структуры в данных
- (?) сохранение статистических свойств данных

Снижение размерности: метод главных компонент (PCA)

Идея:

- Предположим, что признаки в наборе данных скоррелированы;
- Перейдем в новые координаты, где признаки полностью раскоррелированы;
- Будем делать это «жадно»: выбирая новые признаки один за другим таким образом, чтобы суммарная объясненная дисперсия с каждой итерацией прирастала максимально.

- Для идеального восстановления исходных признаков необходима исходная размерность новых признаков;
- Для восстановления с потерями достаточно меньшего количества признаков; будем сохранять не менее 9x% дисперсии.

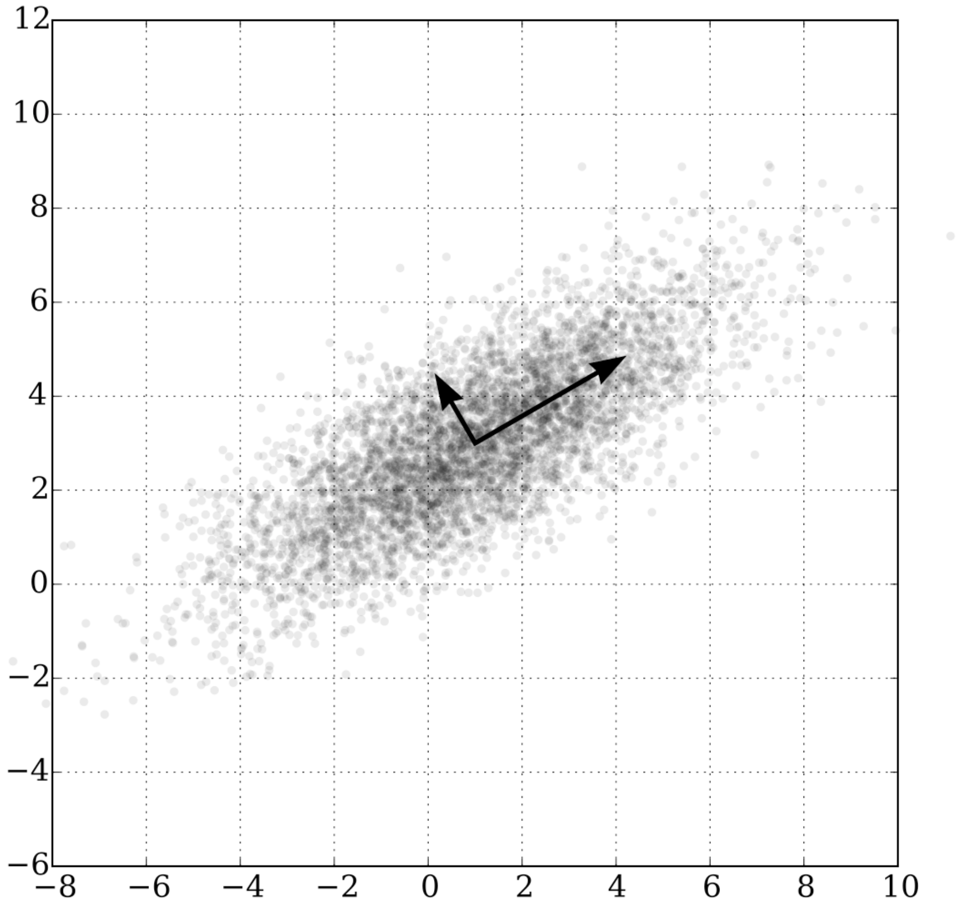


Снижение размерности: метод главных компонент (PCA)

$X \in \mathbb{R}^D$

$$cov(X_i, X_j) = \mathbb{E}\left((X_i - \mu_i)(X_j - \mu_j)\right)$$

$$cov(X_i, X_j) = \mathbb{E}(X_i X_j) - \mu_i \mu_j$$



Снижение размерности: метод главных компонент (PCA)

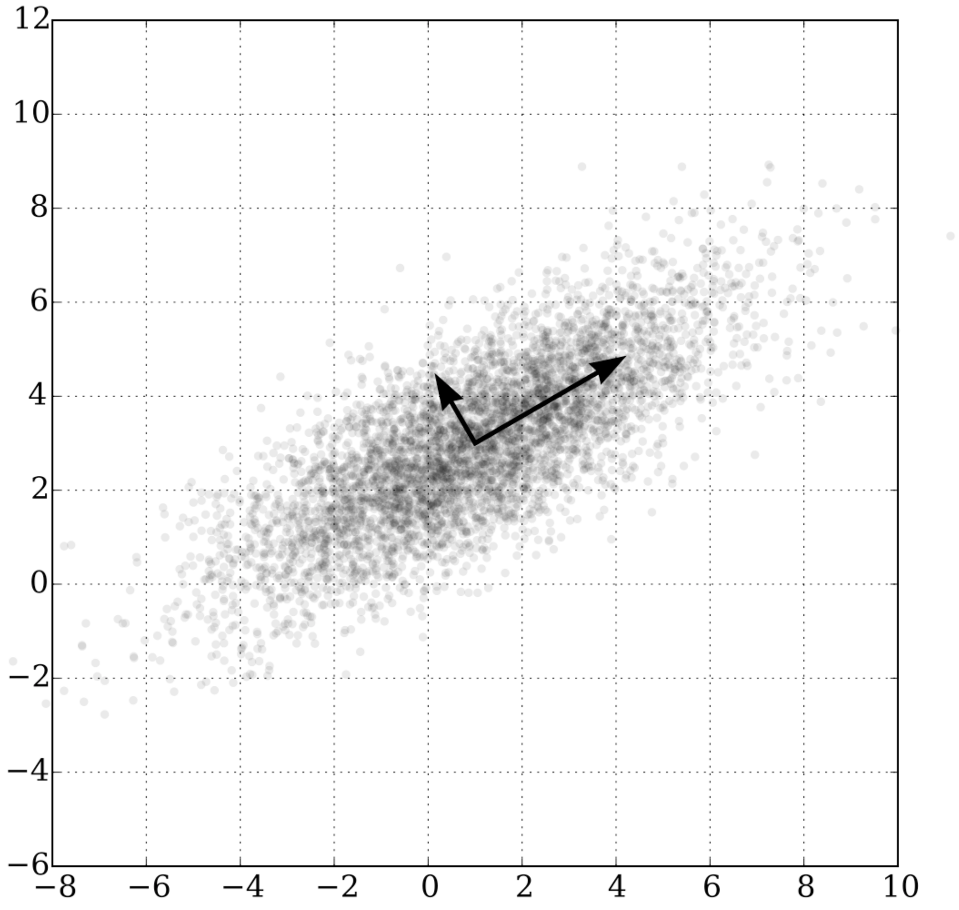
$X \in \mathbb{R}^D$

$$cov(X_i, X_j) = \mathbb{E}\Big((X_i - \mu_i)(X_j - \mu_j)\Big)$$

$$cov(X_i, X_j) = \mathbb{E}(X_i X_j) - \mu_i \mu_j$$

$$cov(X_i, X_j) = cov(X_j, X_i)$$

$$cov(X_i, X_i) = Var(X_i)$$



Снижение размерности: метод главных компонент (PCA)

$$\boxed{X \in \mathbb{R}^{D \times N}}$$

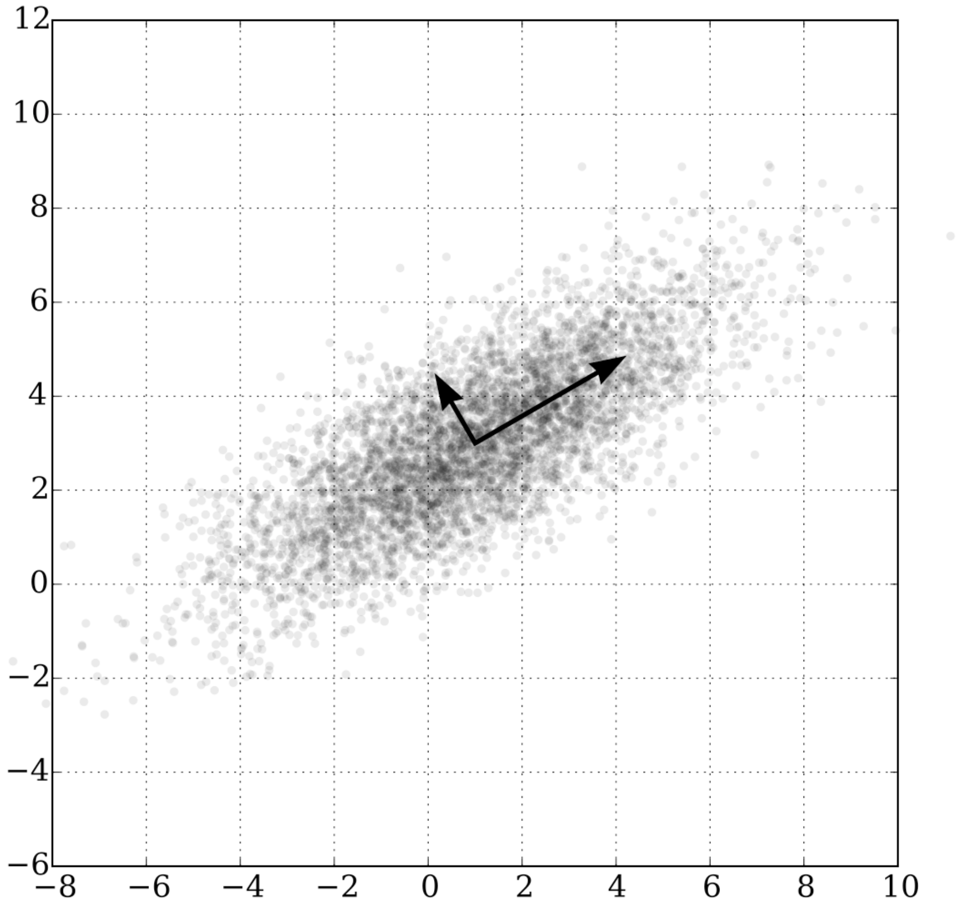
$$cov(X_i, X_j) = \mathbb{E}\Big((X_i - \mu_i)(X_j - \mu_j)\Big)$$

$$cov(X_i, X_j) = \mathbb{E}(X_i X_j) - \mu_i \mu_j$$

$$cov(X_i, X_j) = cov(X_j, X_i)$$

$$cov(X_i, X_i) = Var(X_i)$$

$$\boxed{\begin{gathered}\Sigma = \mathbb{E}[(X - M)(X - M)^T] \\ M = \mathbb{E}(X)\end{gathered}}$$



Снижение размерности: метод главных компонент (PCA)

$X \in \mathbb{R}^{D\times N}$

$\Sigma = \mathbb{E}\left[X^o X^{oT}\right]$

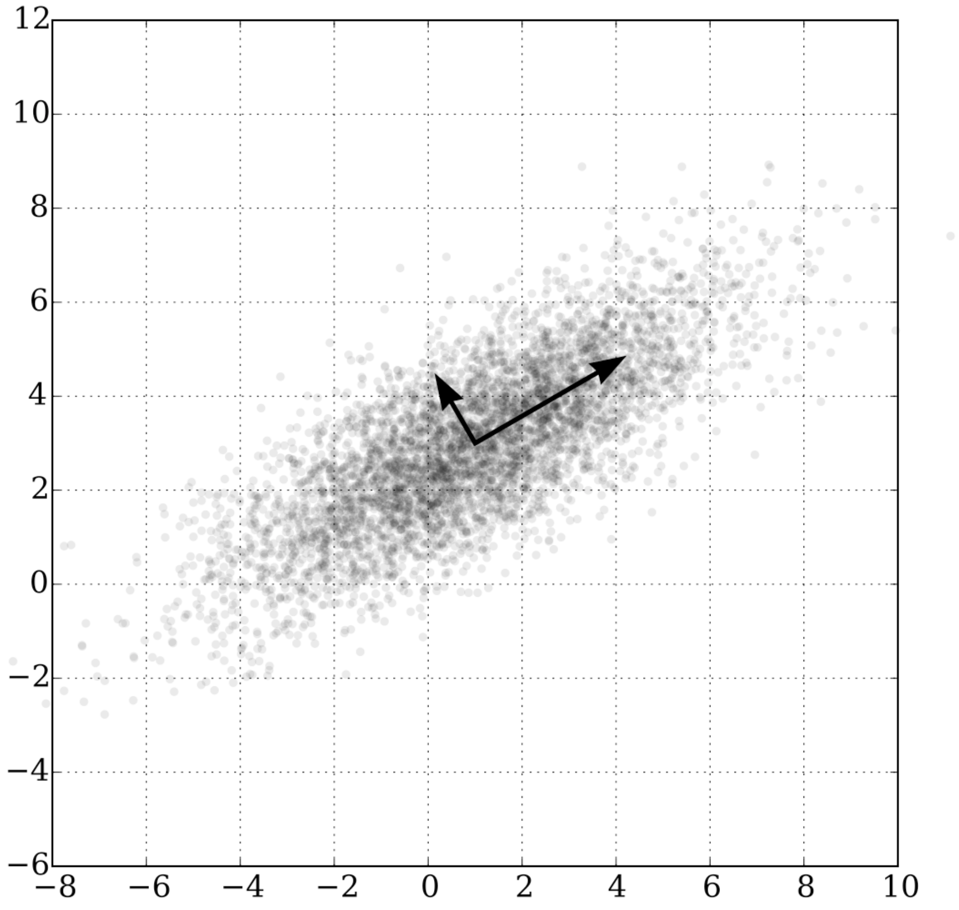
$\Sigma \in \mathbb{R}^{D\times D}$

Σ – квадратная, симметричная, действительная
(=> эрмитова)

Тогда:
Σ имеет D собственных значений λ_d и собственных векторов ξ_d

$\Sigma\xi_d = \lambda_d\xi_d$

(!!!) максимальная дисперсия данных достигается вдоль собственного вектора ξ_d^*, соответствующего максимальному собственному значению λ_d^*.



Снижение размерности: метод главных компонент (PCA)

(!!!) максимальная дисперсия данных достигается вдоль собственного вектора ξ_d^*, соответствующего максимальному собственному значению λ_d^*.

1. Центрируем исходные признаки -> X^o;
2. Найдем Σ, найдем собственные значения $\boldsymbol{\lambda_d}$ и собственные векторы $\boldsymbol{\xi_d}$ матрицы Σ;
3. Отранжируем их по величине собственных значений $\boldsymbol{\lambda_d}$ по убыванию;
4. Возьмем k максимальных собственных значений и соответствующих им собственных векторов;
5. умножим исходную матрицу признаков $X^o \in \mathbb{R}^{N \times D}$ на матрицу k собственных векторов ($V_{(k)} \in \mathbb{R}^{D \times k}$) – получим матрицу $Z \in \mathbb{R}^{N \times k}$ новых признаков-проекций на собственные векторы;