

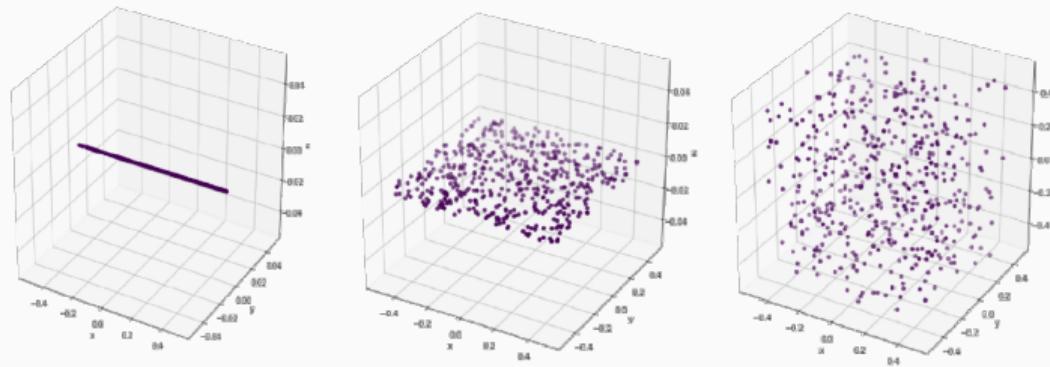
Методы понижения размерности данных

Поглазов Никита
2024

Введение

Введение

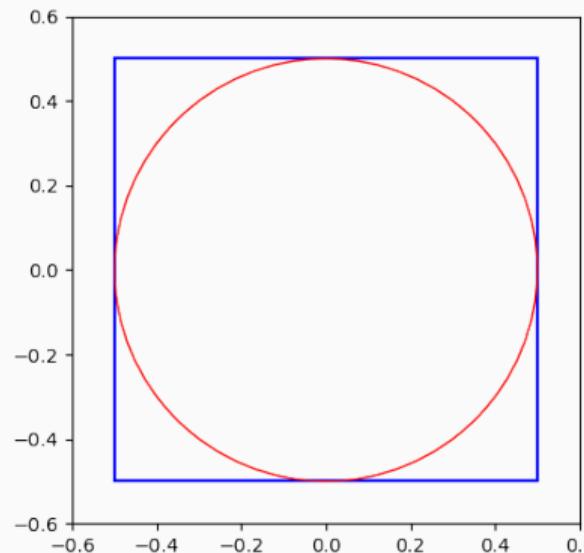
Проклятие размерности: данные высокой размерности сложны для анализа, требуют много вычислительных ресурсов и часто содержат шум.



Мотивация

Что такое "проклятие размерности"?

$$S_{square} = 1 \quad S_{circle} = \pi * (0.5)^2 = \frac{\pi}{4} \approx 0.79$$

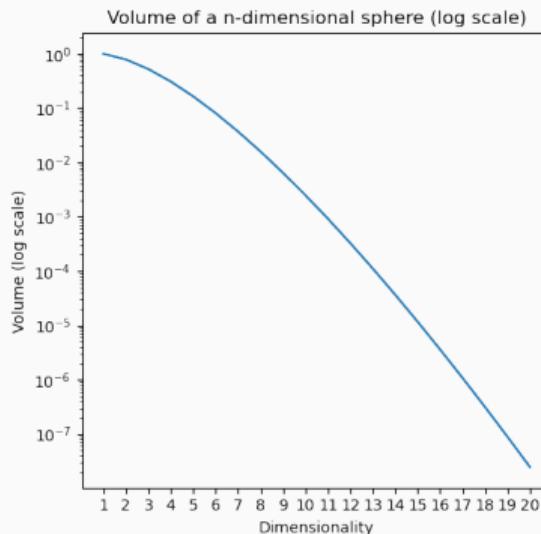
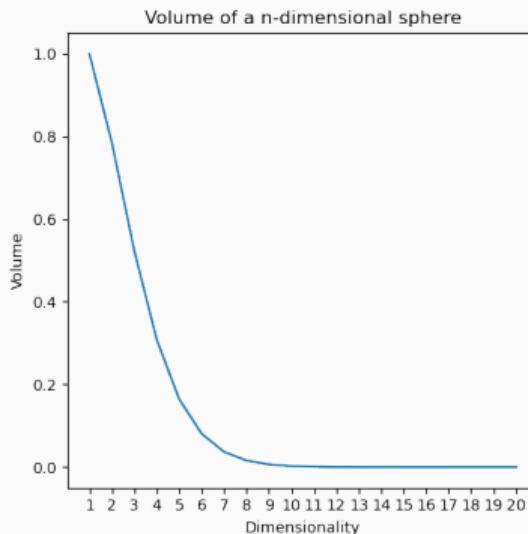


Гиперсфера и гиперкуб

- Объем гиперсферы стремится к нулю при росте размерности:

$$V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} R^n$$

- Диагональ гиперкуба увеличивается как \sqrt{n} .



Классификация методов

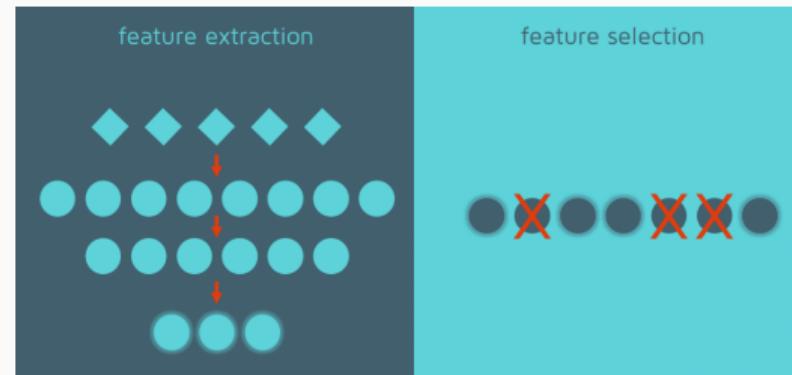
Два подхода к понижению размерности

Отбор признаков:

- Выбор подмножества исходных признаков.
- Сохранение информации без преобразования данных.

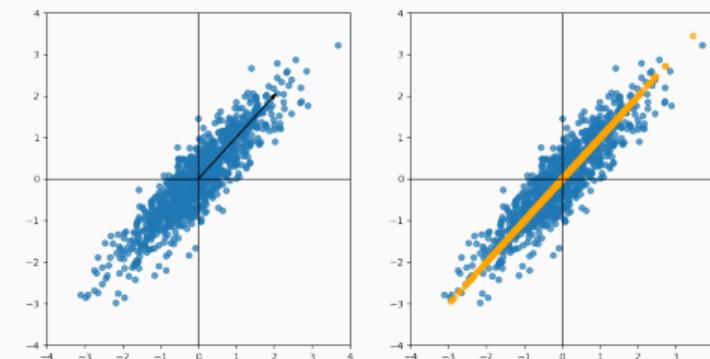
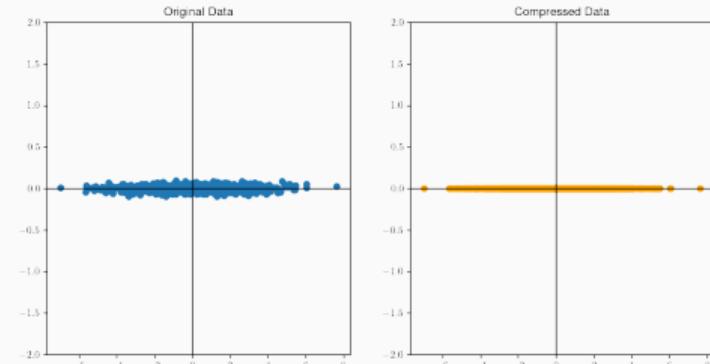
Преобразование признаков:

- Трансформация данных в новое пространство меньшей размерности.
- Сохраняет наиболее значимые свойства данных.



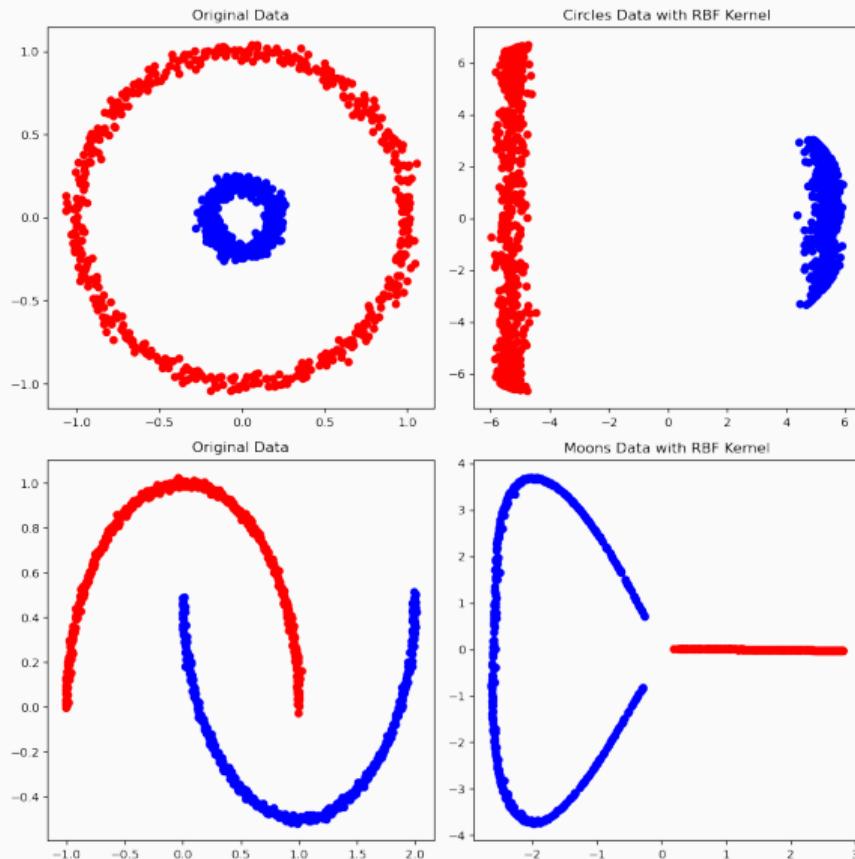
Линейные методы

- Предполагается, что данные имеют **линейные зависимости**.
- Методы находят наилучшую (по определенным критериям) проекцию данных на пространство меньшей размерности.



Нелинейные методы

- Данные имеют **сложные взаимосвязи**, которые нельзя описать линейно.
- Методы выявляют **нелинейные структуры** и сворачивают их в более простую форму.

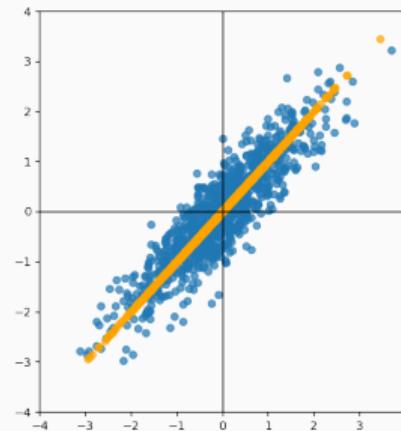
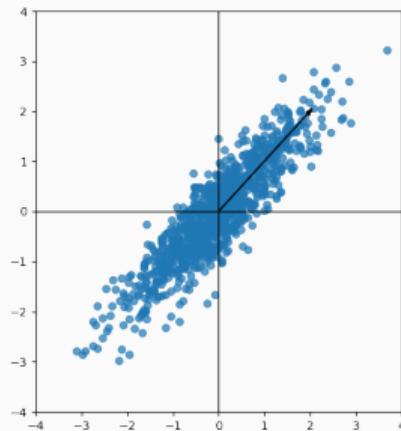


Обзор методов

Principal Component Analysis (Анализ главных компонент, PCA) (1)

Основная идея: Нахождение ортогональных направлений (главных компонент), вдоль которых дисперсия максимальна.

Ключевой момент: Собственные векторы матрицы ковариации данных будут являться главными компонентами.



Principal Component Analysis (Анализ главных компонент, PCA) (2)

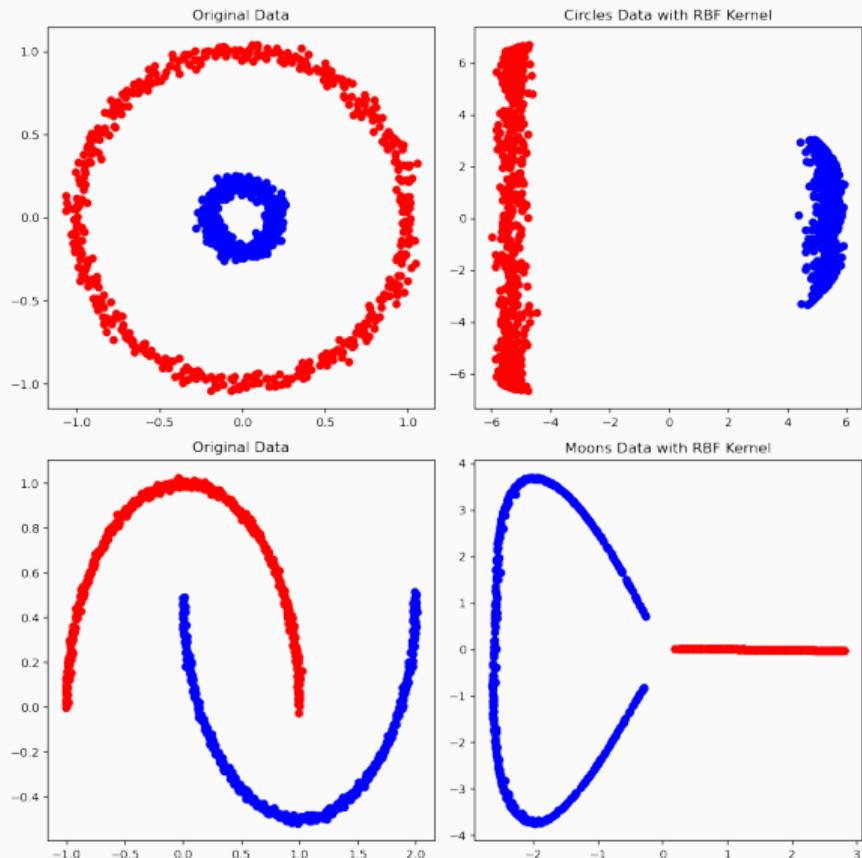


Kernel PCA (Ядерный PCA, КРСА) (1)

Основная идея: Применить PCA в пространстве более высокой размерности.

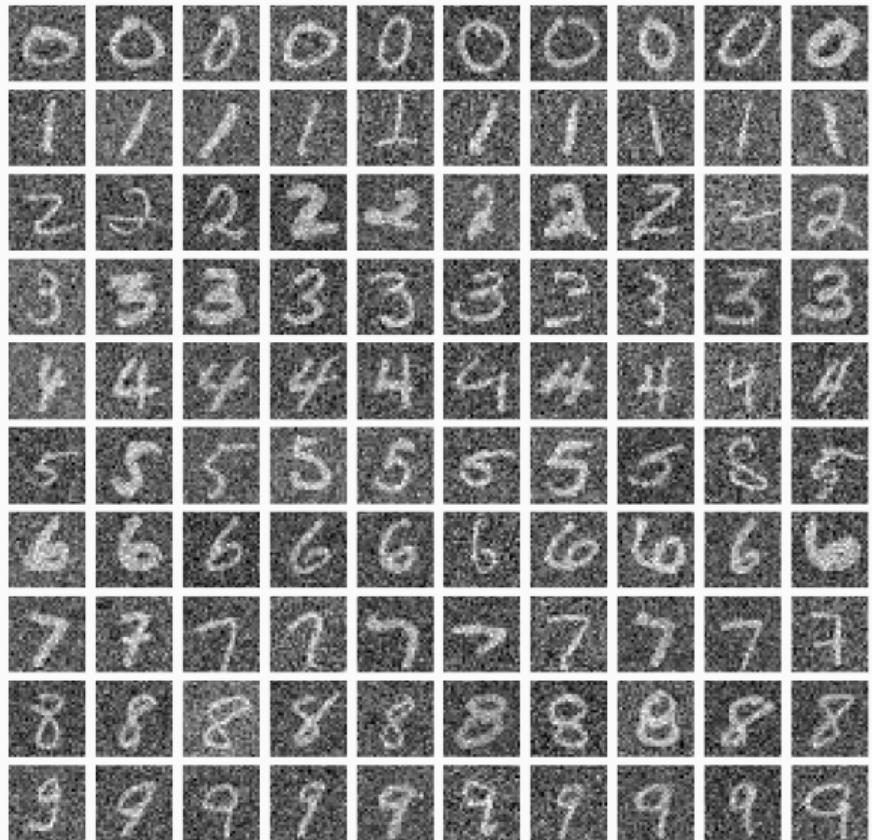
Ключевой момент:

Использование ядерного трюка для избежания прямого преобразования данных в пространство высокой размерности.

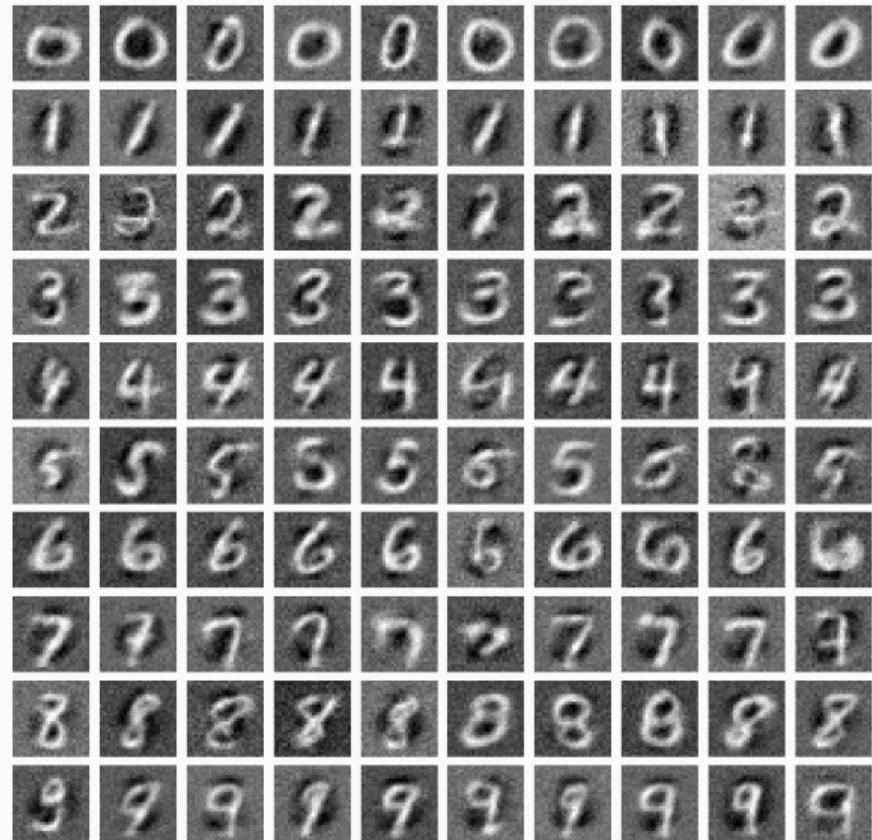


Kernel PCA (Ядерный PCA, KPCA) (2)

Noisy data



RBF KPCA denoise



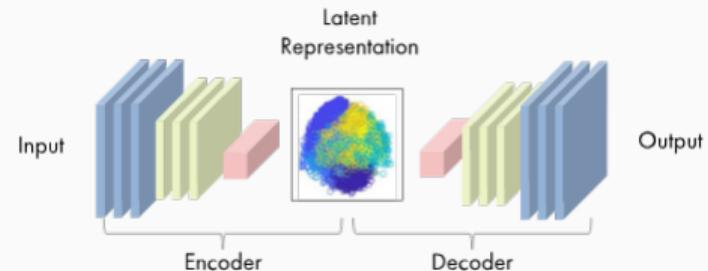
AutoEncoders (AEs) (1)

Основная идея: Нахождение компактных нелинейных представлений данных.

Ключевой момент: Использование нейронных сетей, обучающихся восстанавливать входные данные, для построения нелинейных преобразований.

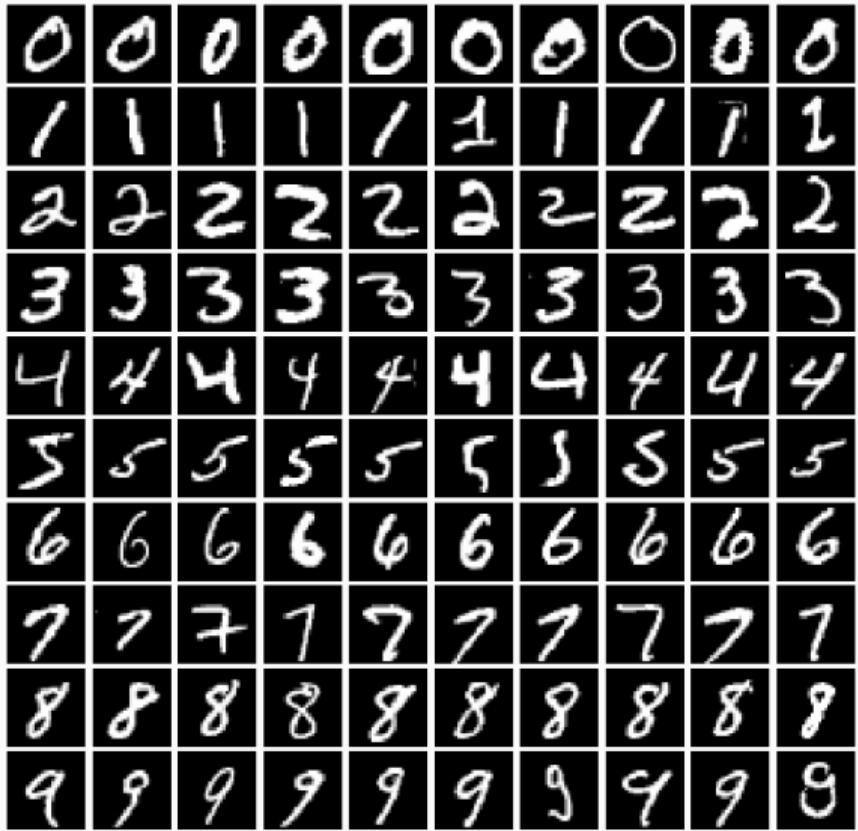
Архитектура:

- Кодировщик (encoder): преобразует входные данные в компактное представление.
- Декодировщик (decoder): восстанавливает данные из сжатого представления.

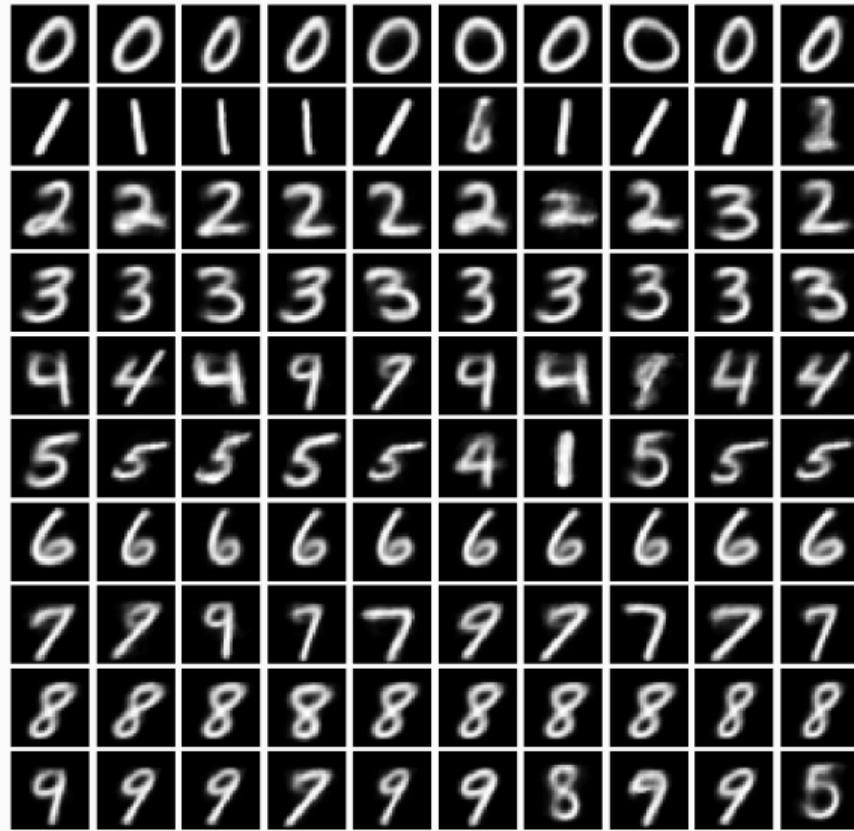


AutoEncoders (AEs) (2)

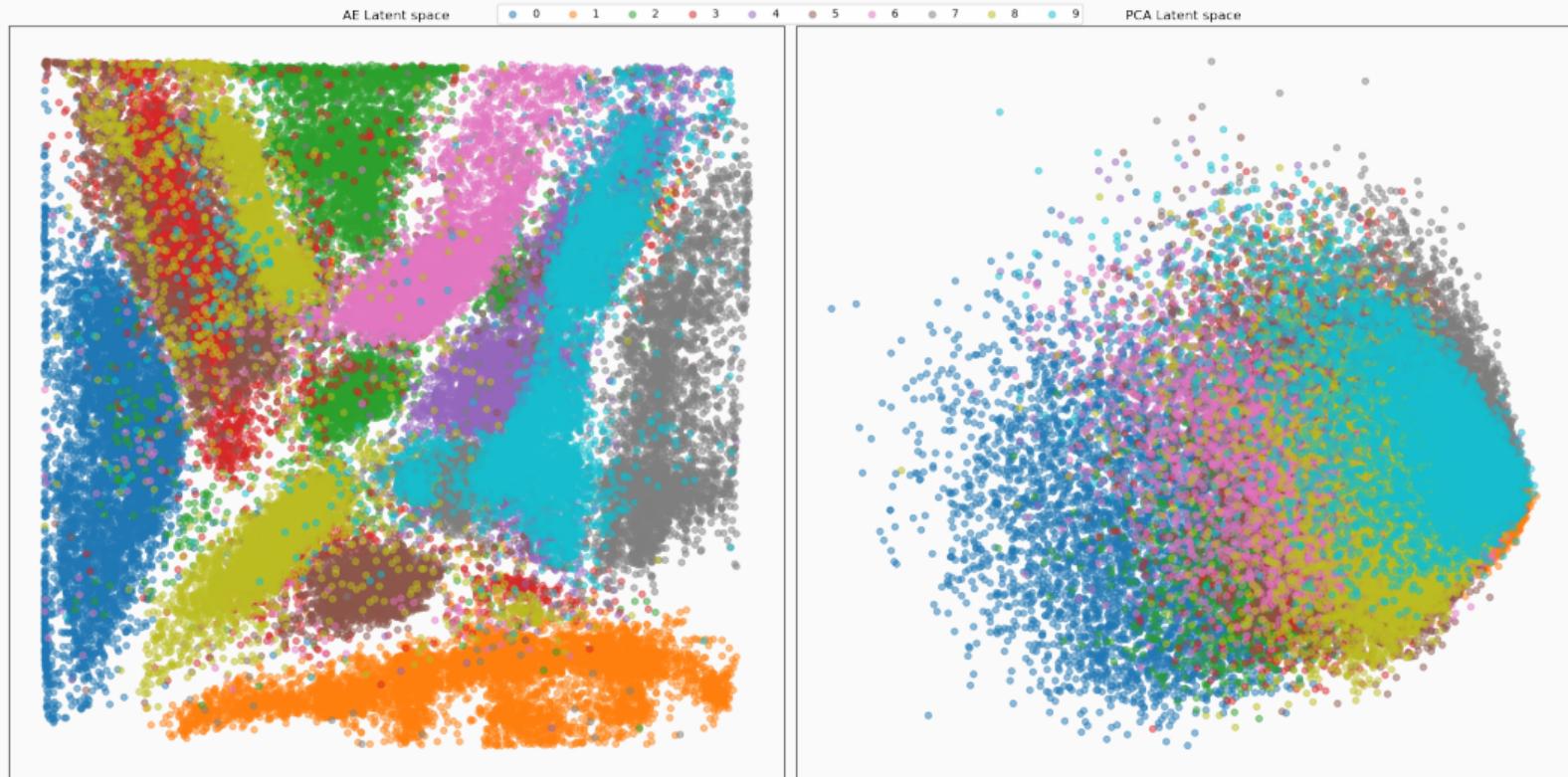
Original digits



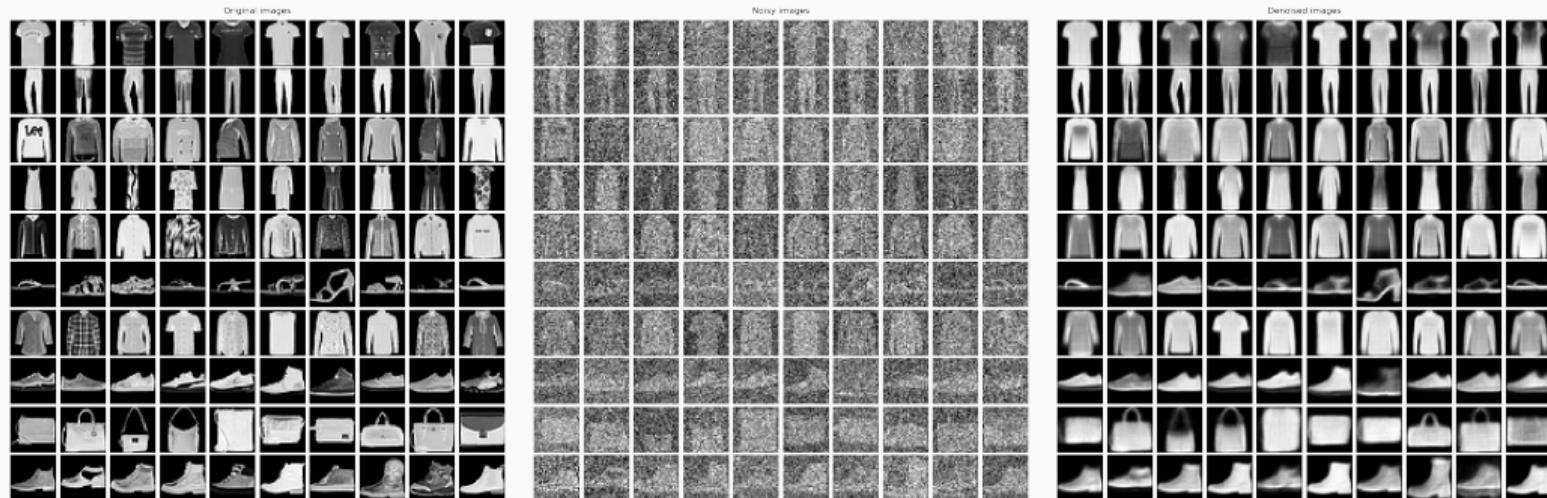
Decoded digits



AutoEncoders (AEs) (3)

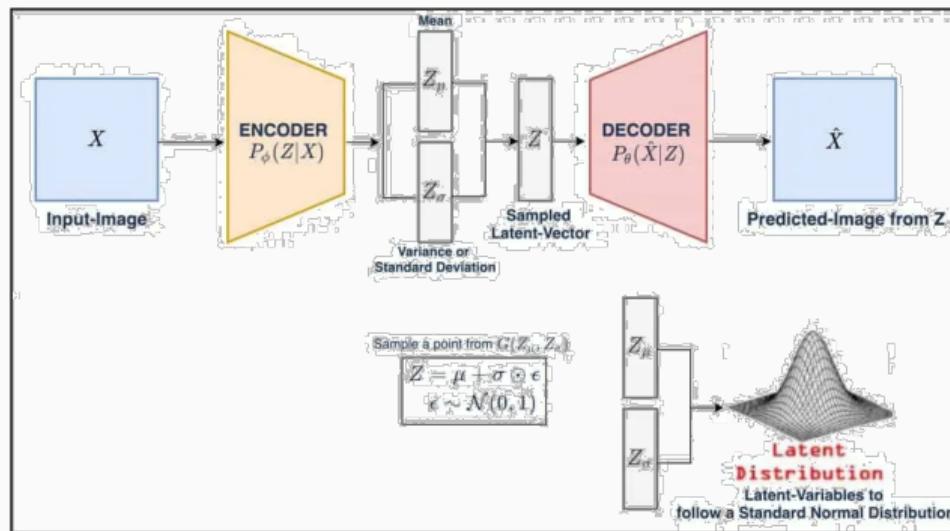


AutoEncoders (AEs) (4)

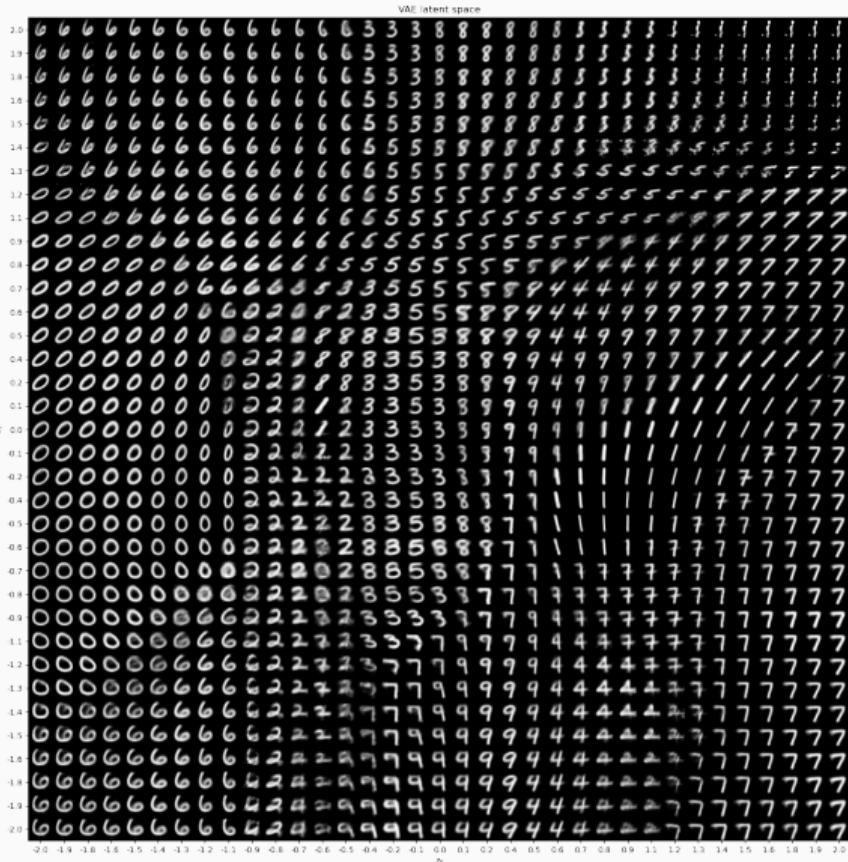


Variational AutoEncoders (VAEs) (1)

Основная идея: Представление скрытого пространства в виде вероятностного распределения, что позволит генерировать новые данные.



Variational AutoEncoders (VAEs) (2)



Заключение

Сравнительный анализ методов

Ключевые отличия:

- **PCA:** базовый линейный метод, идеален для анализа линейных зависимостей.
- **KPCA:** позволяет работать с нелинейной структурой данных, но требует осторожной настройки.
- **AE:** предоставляет большую гибкость благодаря нейронным сетям.
- **VAE:** расширяет автоэнкодеры за счет вероятностной модели, идеально подходит для задач генерации данных и анализа скрытых переменных.

Заключение

Современные вызовы: Работа с высокоразмерными данными требует гибких и мощных инструментов. Методы понижения размерности предоставляют исследователям возможность:

- Повышения информативности данных.
- Улучшения эффективности алгоритмов машинного обучения.
- Удобства визуализации данных.

Вывод: Методы понижения размерности предоставляют исследователям мощный арсенал для анализа данных, повышая информативность, эффективность и удобство визуализации.