Сокращение размерности

1. **Размерности** — переменные или характеристики данных (столбцы в матрице X). X ($N \times P$): N — число наблюдений (строк), P — число признаков (столбцов). Чем больше признаков, тем сложнее данные. Алгоритм лучше всего работает именно с непрерывными данными.
2. **Сокращение размерности**: мы берём данные с высокой размерностью ($p > 4$) и «упрощаем» их (до $p = 2$). То есть, мы представляем полное пространство данных высокой размерности в подпространстве меньшей размерности.

Например, МГК (PCA) — метод, уменьшающий размерность и максимизирующий общую дисперсию. Первая ГК определяется через направление наибольшей вариации данных, последующие компоненты ортогональны предыдущим, что отражает уникальные данные. Определение числа компонент зависит от исследователя, но количество компонент должно быть $< p$, так как из любого пространства X можно извлечь до p компонент.

UMAP — изучение геометрического пространства и геометрических паттернов, которые лежат в пространстве высокой размерности, и, основываясь на расстояниях между наблюдениями, уменьшает размерность исходных данных. Свои основания UMAP (Uniform Manifold Approximation and Projection) берёт в методе t-SNE (t-distributed Stochastic Neighbor Embedding).

Про массив данных: термометр чувств от 1 до 100 по поводу того, какие чувства респонденты испытывают к тому или иному человеку/феномену. Ценность в том, что они, скорее всего, коллинеарны. **Гипотеза**: структура ответов на эти термометры будет формироваться по партийной линии.

LLE имеет схожую с PCA интуицию: основан на идее «многообразного» обучения через изучение локальной структуры (из соседних значений) и дальнейшей её интерпретации на глобальную структуру. Метод подходит, если данные не выражаются линейно. Он сохраняет локальные связи между точками, что позволяет работать с более сложными структурами данных. Схожим образом работают t-SNE и UMAP, а также нейронные модели: SOM и autoencoders.

Чем PCA отличается от факторного анализа? Факторный анализ предполагает ряд предварительных допущений о гауссовском распределении и существовании условной независимости между переменными. Фокус на латентных переменных, каузальная модель, которая пытается выявить скрытые факторы.

Очистка данных. Два подхода для очистки данных в R: базовый синтаксис и tidy-подход. Tidy-подход — использование операторов `%>%` для написания нескольких операций в одной функции. Вместо удаления используется импутация, то есть вычисленные правдоподобные значения. Также рекомендуется задуматься и о причинах пропусков значений. В случае случайных пропусков используется множественная импутация с помощью kNN: выбираются ближайшие непустые значения и берётся их среднее.