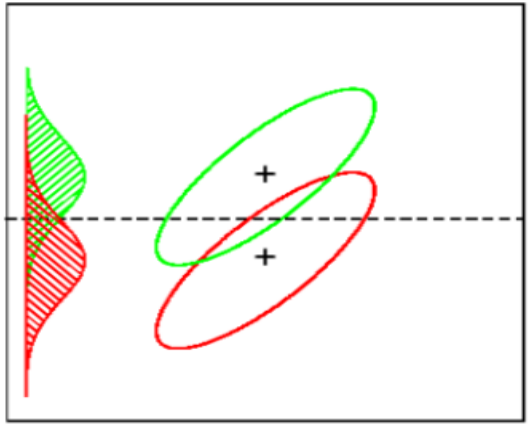
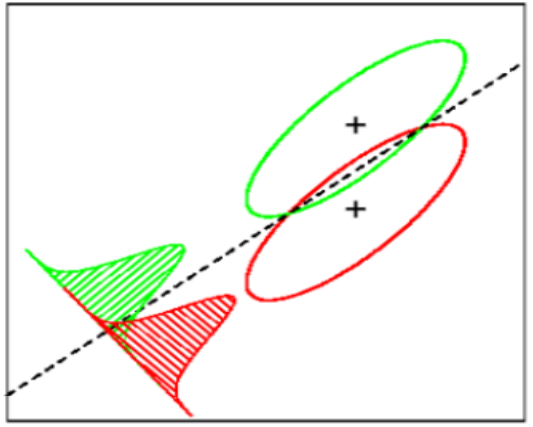
Отбор переменных. Снижение размерности данных

Подготовка фичей

- Масштабирование фичей
  - Как правило вычитаем среднее, делим на стандартное отклонение
- Отбор фичей(переменных).

Отбор переменных(фичей) для линейных моделей

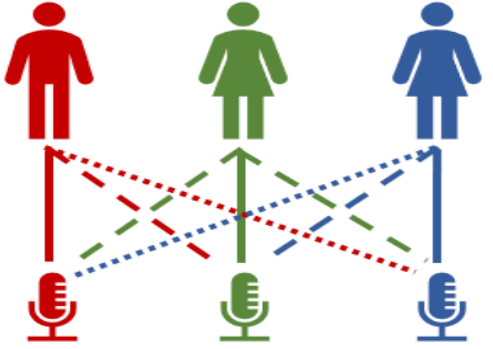




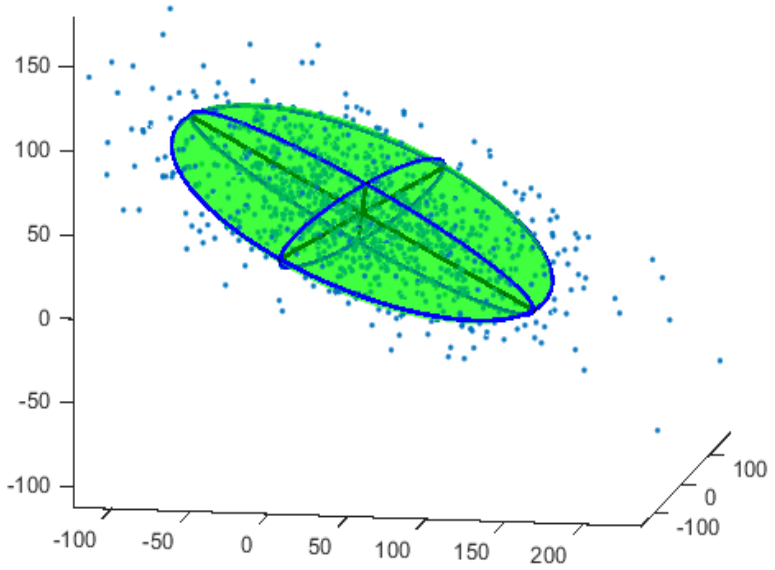
- Переменные должны коррелировать с моделируемой переменой:
  - Кореляция (для регрессии)
  - AU-ROC(для классификации)
  - Взаимная информация (для классификации)
- Переменные должны быть максимально независимы:
  - Минимальная корреляция(Pearson)
  - Мнимальная мультиколлинеарность(VIF)
- Оптимальное число переменных пропорционально количеству объектов(в случае некоррелированных переменных) или корню квадратному из количества объектов

Снижение размерности данных

- Может использоваться в рамках задачи сокращения количества фичей (переменных)
- Может использоваться для самостоятельных задач:
  - Topic modelind
  - Coctail party problem



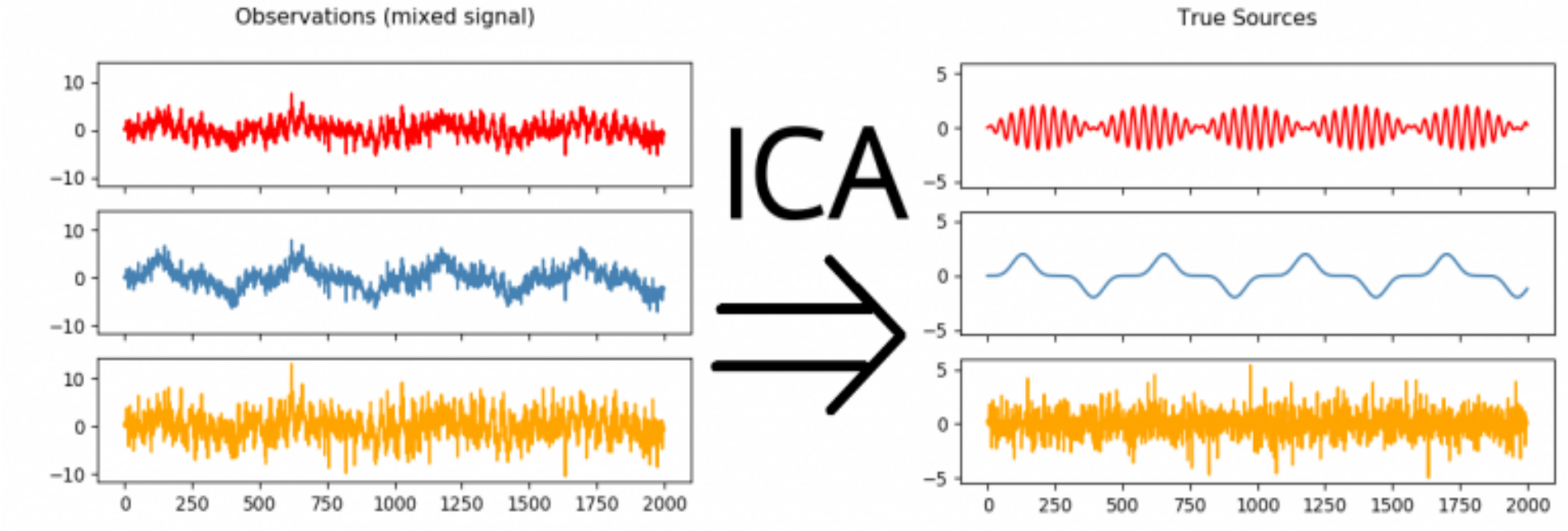
Principal Component Analysis



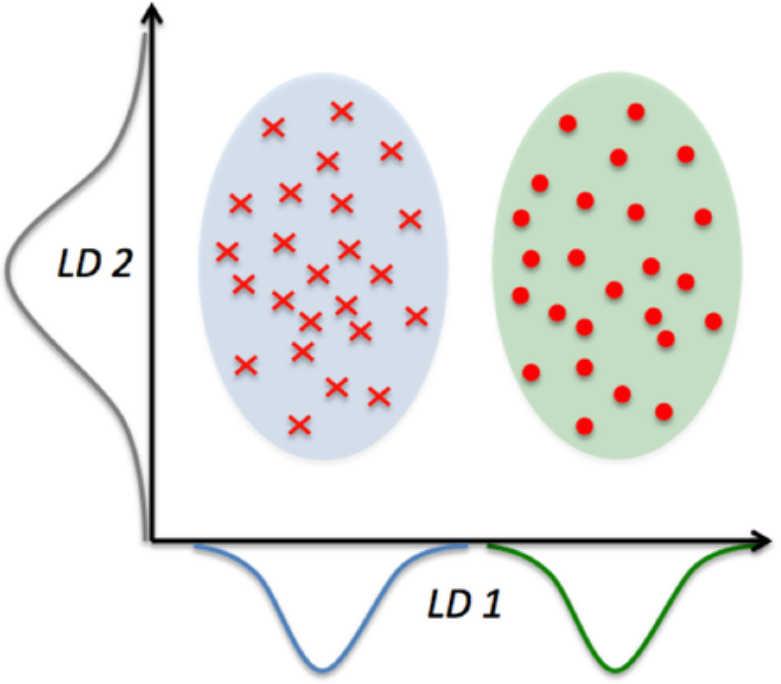
Многомерное шкалирование (MDS)

- Дана матрица расстояний $D=\{d_{ij}\}$ объектов. Необходимо найти координаты объектов так, чтобы минимизировать функционал $stress=\sqrt{\frac{\sum_i (d_i-\hat{d}_i)^2}{\sum_i d_i^2}}$
- Алгоритм:
  1. Вычисляем $A=\{-\frac{1}{2}d_{ij}^2\}$
  2. Вычисляем $B=\{a_{ij}-a_{i.}-a_{j.}+a_{..}\}$, где $a_{i.}$, $a_{.j}$ и $a_{..}$ средние значения по строке i, столбцу j и всей матрице.
  3. Находим собственные значения(λ_i) и соответствующие собственные векторы (L_i). Отбираем векторы соответствующие наибольшим собственным числам. Нормировка векторов: $L_i\bar{L}_i=\lambda_i$, где $\bar{L}_i$ комплексно-сопряженный вектор.
  4. Матрица из собственных векторов $L_{1..p}$ исходная матрица

Independent Component Analysis



Local discriminative analysis(LDA), Corelate components analysis(CCA)



Нелинейные методы снижения размерности

- Isomap:
  - Нелинейная разновидность шкалиировпния натягивает многообразие (mainfold) и проецирует на него объекты таким образом чтобы новая матрица расстояний не отличалась от изначальной.
- Kohonen self-organizing map:
  - Максимально натягивает многообразие меньшей размерности на датасет и проецирует на него данные
- Generative topographic map:
  - Вероятностная модификация SOM с улучшенной сходимостью