

Статистические методы в QSAR

Часть 2

Многомерный анализ данных

Затронутые темы

- Многомерный анализ данных
- Понятие о дескрипторном пространстве, химическом расстоянии
- Понятие о дискриминантном и кластерном анализе
- Метрика дескрипторного пространства.
Коллинеарные и ортогональные дескрипторы
- Латентные дескрипторы, оценки (scores) и нагрузки (loading)
- Понятие о факторном анализе и методе главных компонент (PCA)
- Метод частичных наименьших квадратов (PLS)

Многомерный анализ данных

$$X = \begin{pmatrix} x_1^1 & x_2^1 & \boxtimes & x_M^1 \\ x_1^2 & x_2^2 & \boxtimes & x_M^2 \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ x_1^N & x_2^N & \boxtimes & x_M^N \end{pmatrix} \quad Y = \begin{pmatrix} y_1^1 & y_2^1 & \boxtimes & y_P^1 \\ y_1^2 & y_2^2 & \boxtimes & y_P^2 \\ \boxtimes & \boxtimes & \boxtimes & \boxtimes \\ y_1^N & y_2^N & \boxtimes & y_P^N \end{pmatrix}$$

Традиционные регрессионные процедуры - число столбцов в матрицах дескрипторов X относительно невелико, и между ними отсутствуют линейные зависимости

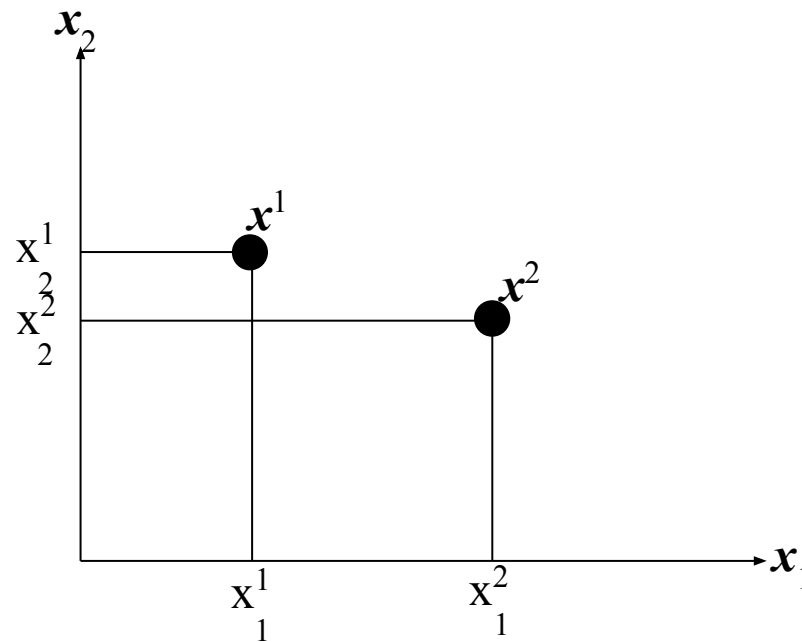
Процедуры многомерного анализа данных могут работать с матрицами дескрипторов X , содержащими большое количество столбцов, многие из которых линейно-зависимы

Центрирование данных для многомерного анализа

$$\underline{X} = \left\| \underline{x}_i^j \right\|, \text{ где } \underline{x}_i^j = x_i^j - \bar{x}_i \quad \bar{x}_i = \frac{1}{N} \sum_{k=1}^N x_i^k \quad \forall i = \overline{1, M}$$

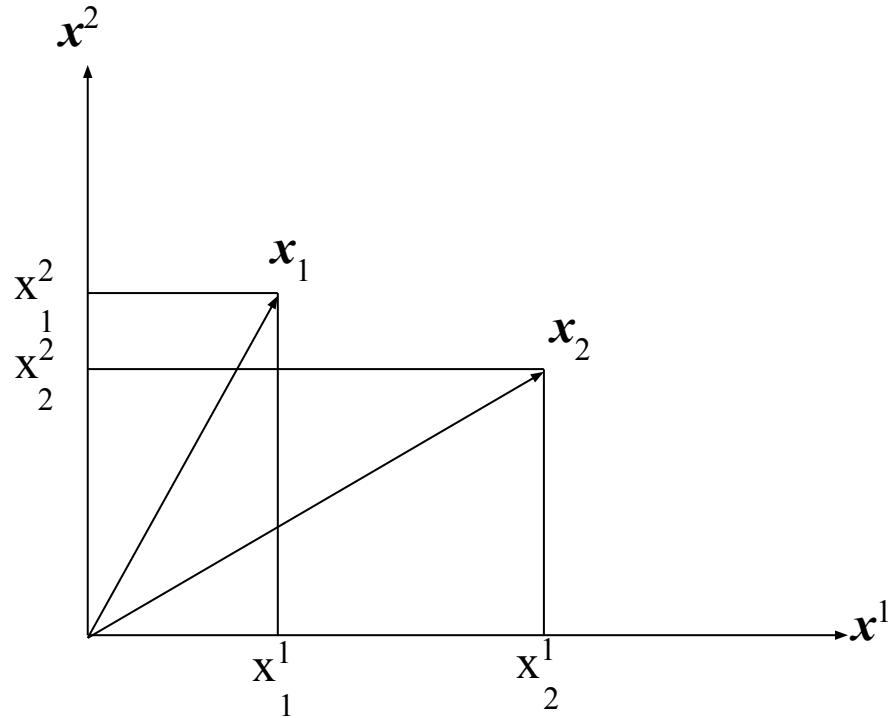
$$\underline{Y} = \left\| \underline{y}_i^j \right\|, \text{ где } \underline{y}_i^j = y_i^j - \bar{y}_i \quad \bar{y}_i = \frac{1}{N} \sum_{k=1}^N y_i^k \quad \forall i = \overline{1, P}$$

Дескрипторное пространство (пространство признаков, feature space, M-пространство)



Оси x_1 , x_2 – дескрипторы, точки x^1 , x^2 – соединения

Пространство соединений (пространство объектов, object space, N-пространство)



Оси x^1, x^2 – соединения, вектора x_1, x_2 – дескрипторы

Метрика дескрипторного пространства (химическое расстояние)

Эвклидово расстояние

$$D_{ij}^{Euclid} = \sqrt{\sum_{k=1}^M (x_k^i - x_k^j)^2}$$

Манхэттоновское расстояние

$$D_{ij}^{Manhattan} = \sum_{k=1}^M |x_k^i - x_k^j|$$

Метрика Минковского

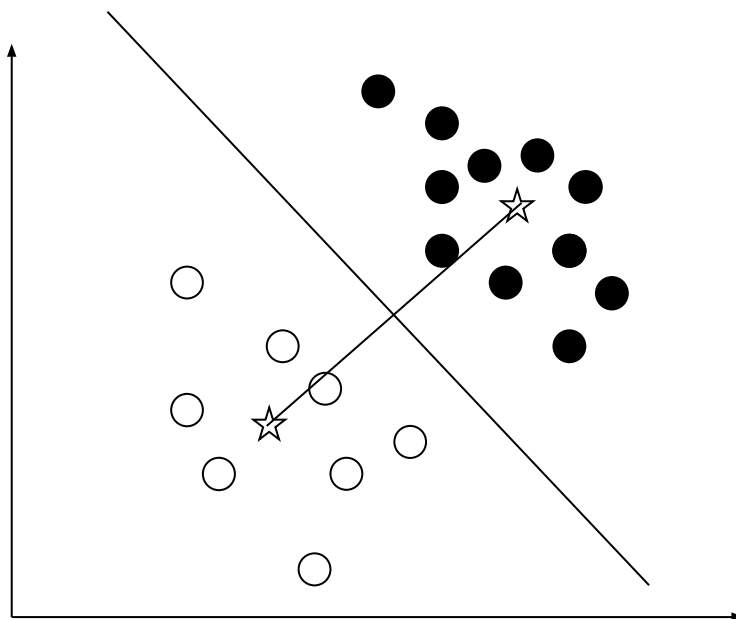
$$D_{ij}^{Minkowski} = \sqrt[r]{\sum_{k=1}^M |x_k^i - x_k^j|^r}$$

Принцип сходства (Similarity Principle)

Постулируется принцип: структурно близкие химические соединения обладают сходными свойствами

Предполагается, что всегда можно найти такой набор дескрипторов и такую метрику дескрипторного пространства, чтобы этот принцип выполнялся

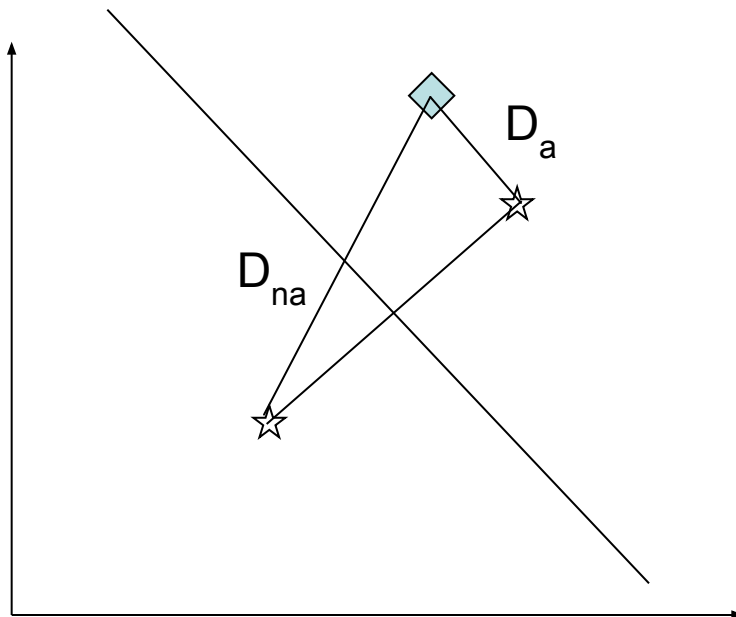
Дискриминантный анализ



● активное соединение

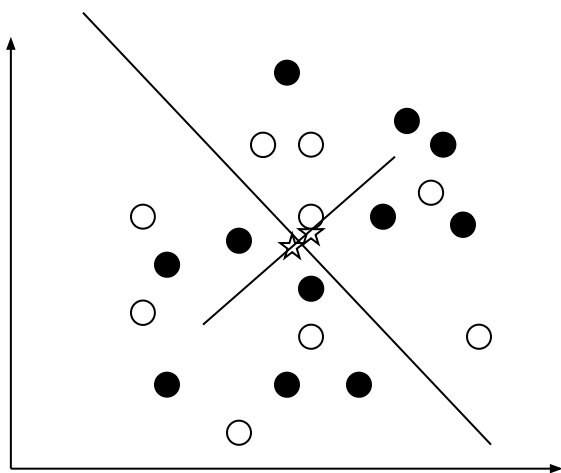
○ неактивное соединение

Дискриминантный анализ

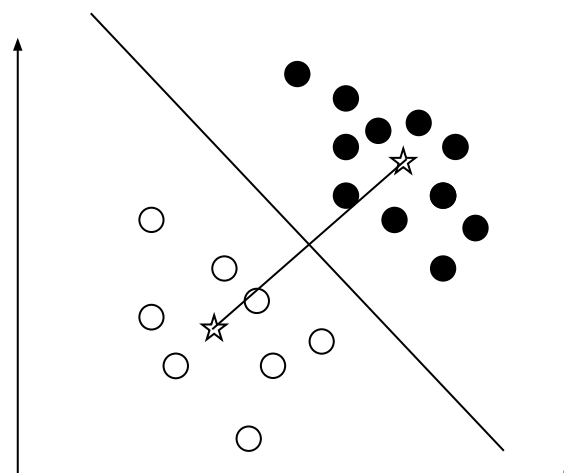


Соединение спрогнозировано как активное, поскольку $D_a < D_{na}$

Дискриминантный анализ (выбор набора дескрипторов)

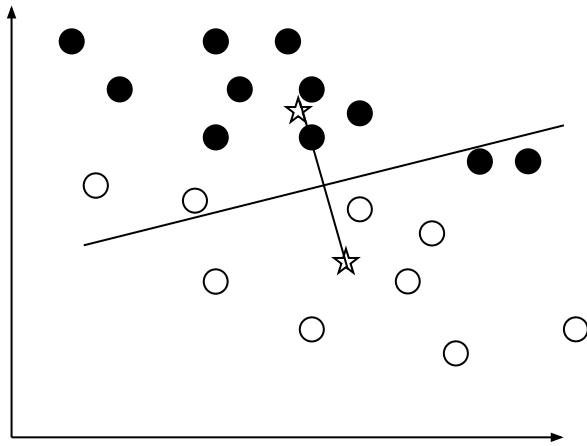


Плохой набор дескрипторов



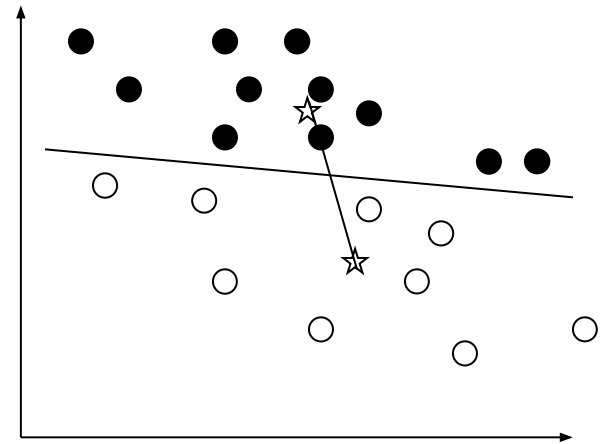
Хороший набор дескрипторов

Дискриминантный анализ (выбор метрики)



Метрика Эвклида

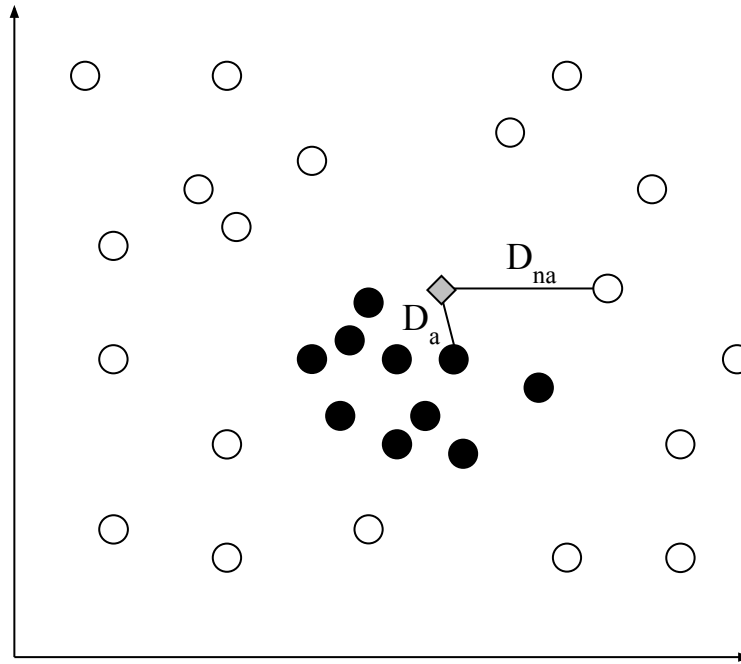
$$D^2 = (\underline{a} - \underline{b})^T (\underline{a} - \underline{b})$$



Метрика Махаланобиса

$$D^2 = (\underline{a} - \underline{b})^T C^{-1} (\underline{a} - \underline{b})$$

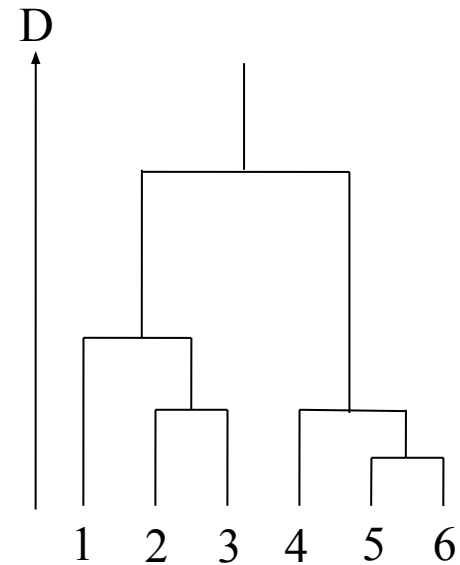
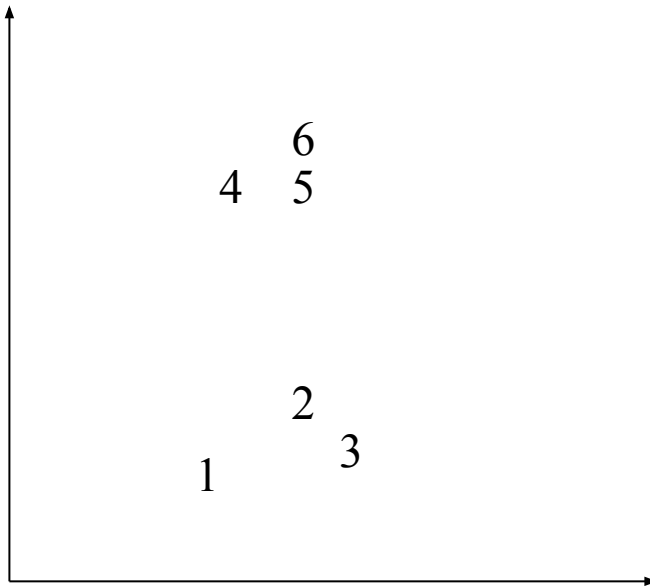
Метод ближайших соседей (kNN – k Nearest Neighbours)



Соединение прогнозируется как активное, поскольку $D_a < D_{na}$

Кластерный анализ

Задача кластерного анализа – изучение внутренней структуры и выявление группировки данных



Дендограмма

Свойства пространства соединений

$$\underline{\underline{x}}_i \cdot \underline{\underline{x}}_j = \sum_{k=1}^N \underline{x}_i^k \underline{x}_j^k \quad - \text{ скалярное произведение векторов}$$

$$\cos \alpha = \frac{\sum_{k=1}^N \underline{x}_i^k \underline{x}_j^k}{\sum_{k=1}^N (\underline{x}_i^k)^2 \sum_{k=1}^N (\underline{x}_j^k)^2} \quad - \text{ косинус угла между векторами}$$

$$c_{ij} = \frac{1}{N-1} \sum_{k=1}^N \underline{x}_i^k \underline{x}_j^k \quad - \text{ ковариация}$$

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}} \sqrt{c_{jj}}} \quad - \text{ коэффициент корреляции}$$

Свойства пространства соединений

$$r_{ij} = \cos \alpha$$

Коллинеарные вектора – дескрипторы статистически эквивалентны

Перпендикулярные вектора – дескрипторы линейно независимы

Латентные переменные

Одной из главных задач многомерного анализа данных является выявление таких комбинаций исходных переменных (дескрипторов), которые бы позволили эффективно решать актуальные задачи:

1. Описать данные наименьшим числом переменных (факторный анализ)
2. Добиться максимального разделения классов (факторный дискриминантных анализ)
3. Построить регрессионную модель с наилучшей прогнозирующей способностью (метод частичный наименьших квадратов)
4. и т.д.

Подобные комбинации исходных переменных называются латентными переменными (скрытыми факторами, оценками)

Линейные латентные переменные

$$\underline{s}_i = l_{1i} \underline{x}_1 + l_{2i} \underline{x}_2 + \dots + l_{Mi} \underline{x}_M$$

$$S = \underline{XL}$$

$$S = \begin{pmatrix} s_{11} & \dots & s_{1R} \\ \dots & \dots & \dots \\ s_{N1} & \dots & s_{NR} \end{pmatrix} \quad L = \begin{pmatrix} l_{11} & \dots & l_{1R} \\ \dots & \dots & \dots \\ l_{M1} & \dots & l_{MR} \end{pmatrix}$$

Матрица оценок (scores)

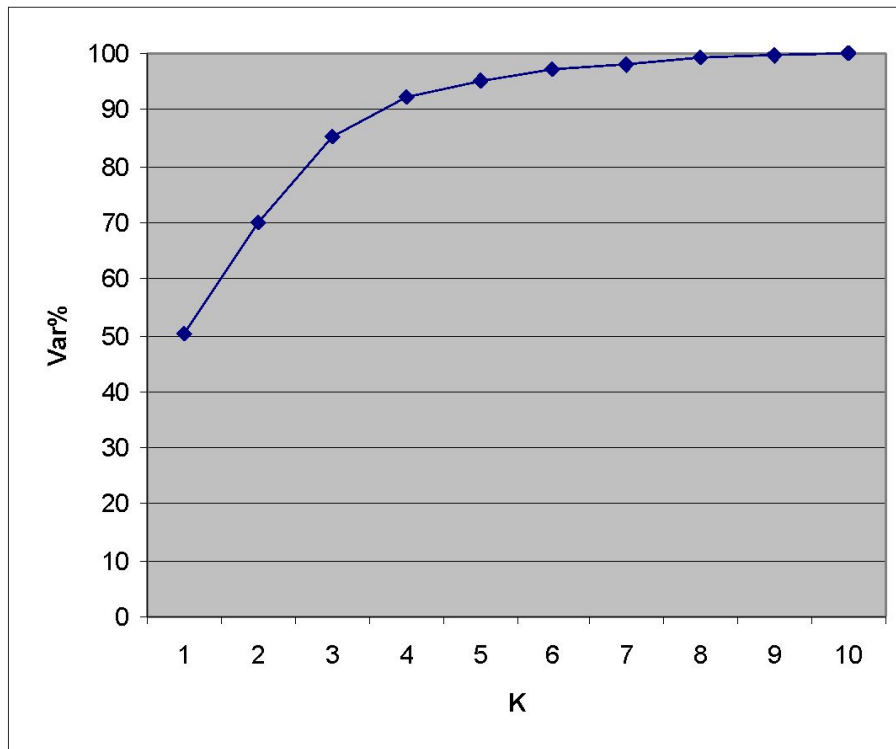
Матрица нагрузок (loading)

Вектора \mathbf{s} обычно берутся ортогональными,
т.е. латентные переменные линейно независимы

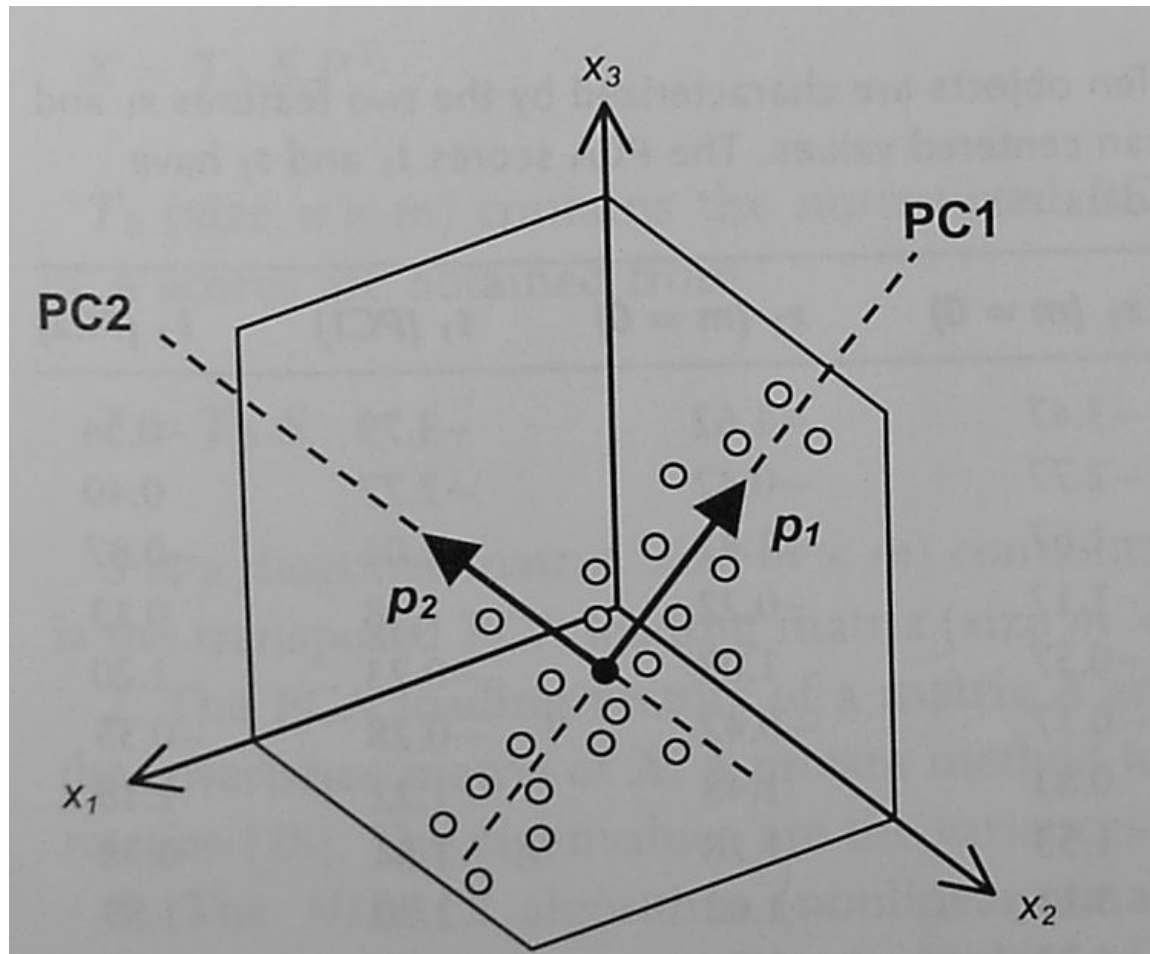
Отбор главных компонент

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_R \geq 0$$

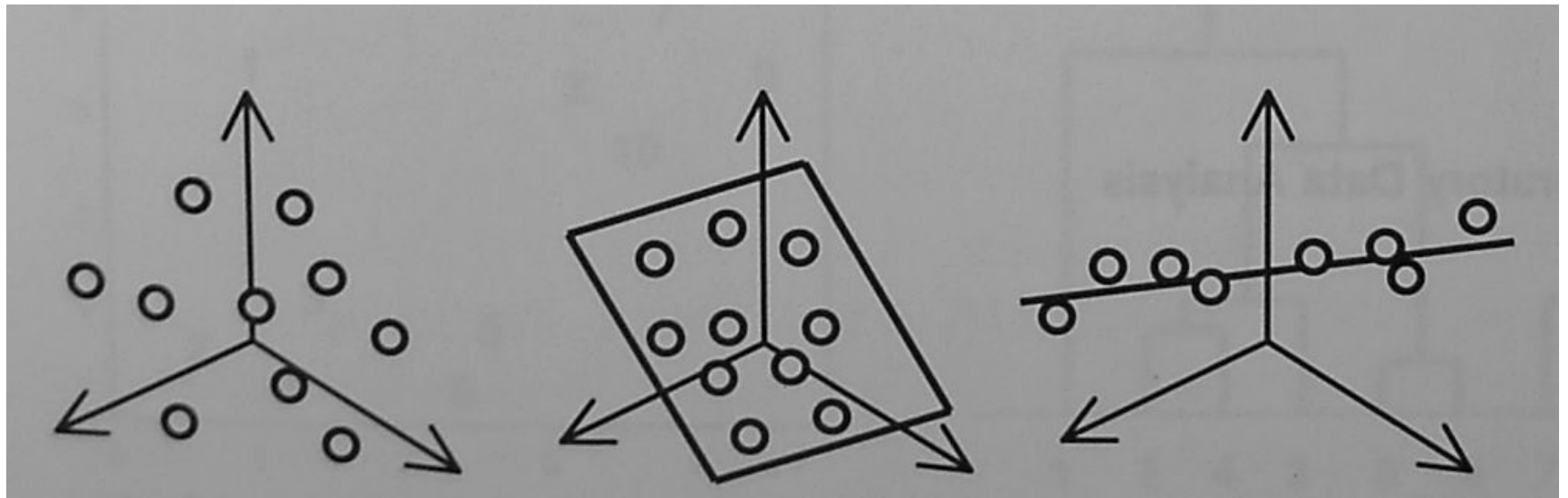
$$Var\% = \frac{100 \cdot \sum_{i=1}^K \lambda_i}{\sum_{i=1}^R \lambda_i}$$



Главные компоненты



Определение размерности данных



$K=3$

$K=2$

$K=1$

Графики оценок и нагрузок

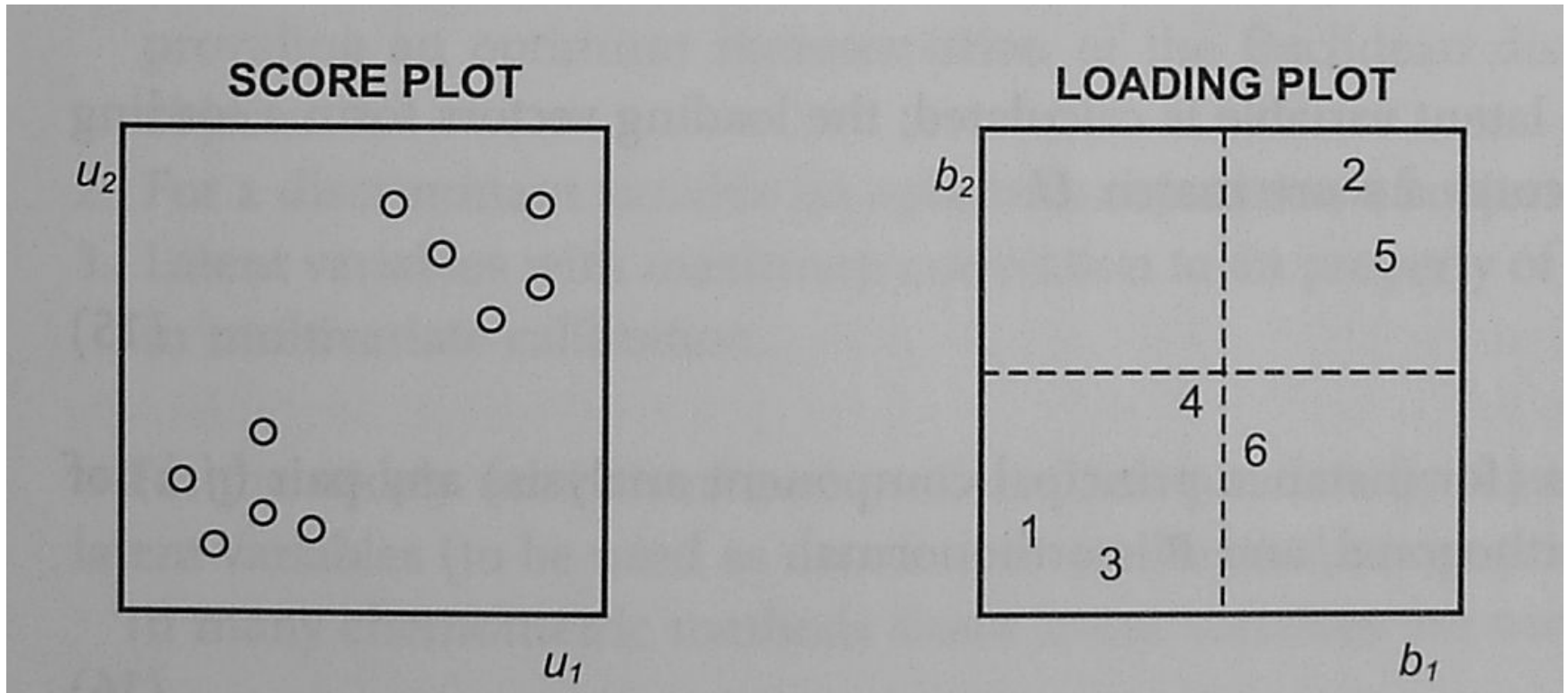


График в координатах
главных оценок

График в координатах
главных нагрузок

Резюме метода главных компонент

- Вычисляется матрица ковариаций
- Находятся ее собственные вектора и собственные значения
- Отбираются латентные переменные, соответствующие двум наибольшим собственным значениям
- Строятся 2-мерные графики оценок и нагрузок

Факторный анализ

- Определяется число латентных переменных, необходимых для воспроизведения данных с заданной точностью
- Путем вращения векторов исходных латентных переменных ищутся легко интерпретируемые варианты

Факторный (канонический) дискриминантный анализ

- Ищутся латентные переменные, позволяющие получить наилучшее разделение классов путем максимизации отношения межгрупповой к общей дисперсии

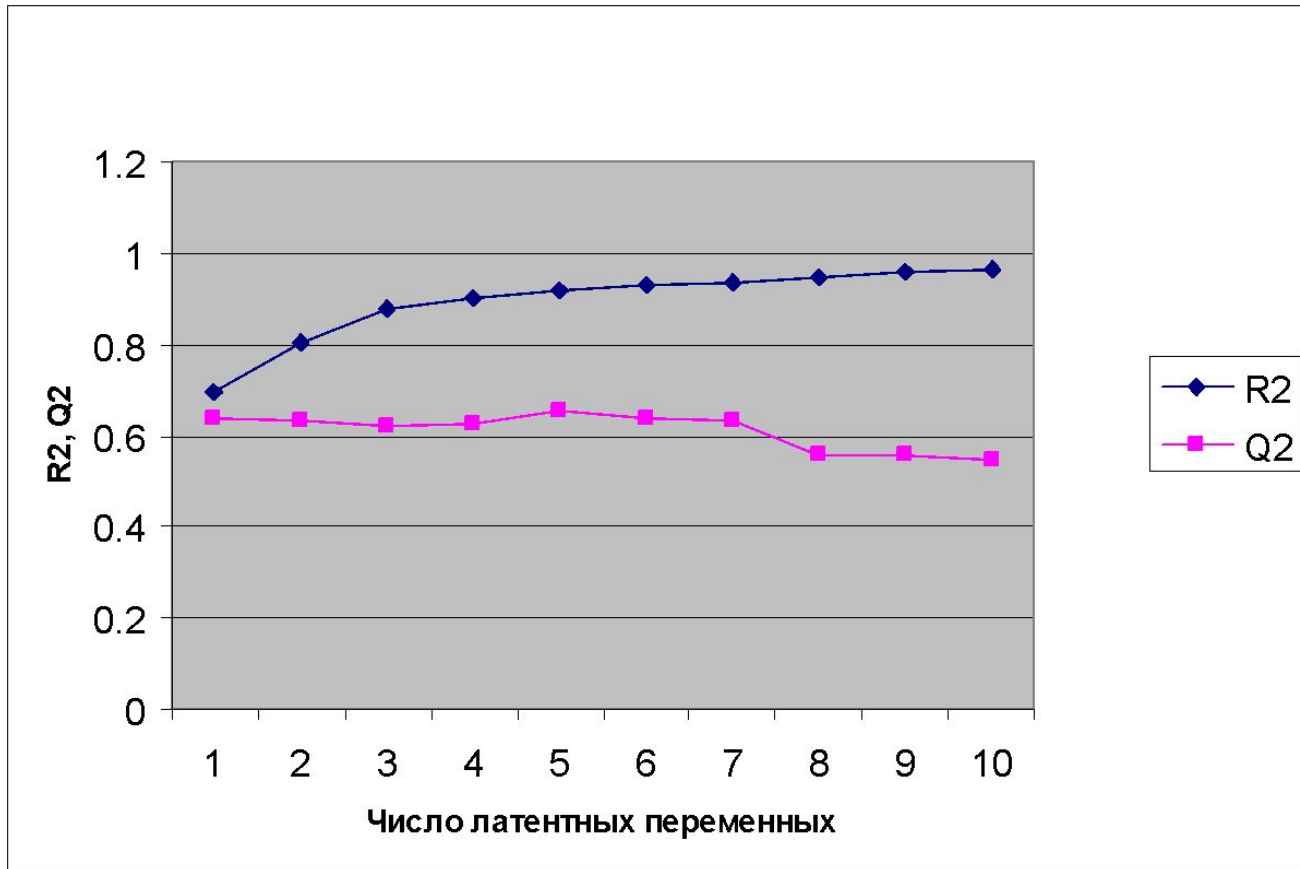
Метод частичных наименьших квадратов (PLS – Partial Least Squares)

В методе частичных наименьших квадратов ищется набор латентных переменных, позволяющий получить регрессионную модель с наилучшей прогнозирующей способностью

$$y^j = \sum_{k=1}^K a_k s_k^j \quad s_k^j = \sum_{i=1}^M l_{ik} x_i^j$$

$$y \propto c_0 + c_1 x_1 + \dots + c_M x_M$$

Определение оптимального числа латентных переменных



Оптимальное число латентных переменных - 5

Резюме метода PLS

- Один за одним отбираются латентные переменные, максимально коллинеарные с векторами свойств или ошибок
- При помощи процедуры скользящего контроля определяется прогнозирующая способность модели
- Выбирается оптимальное число латентных переменных K , максимизирующее критерий Q^2
- Построенная на K латентных переменных регрессионная модель далее используется в для прогноза