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 анализах
- 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

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

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

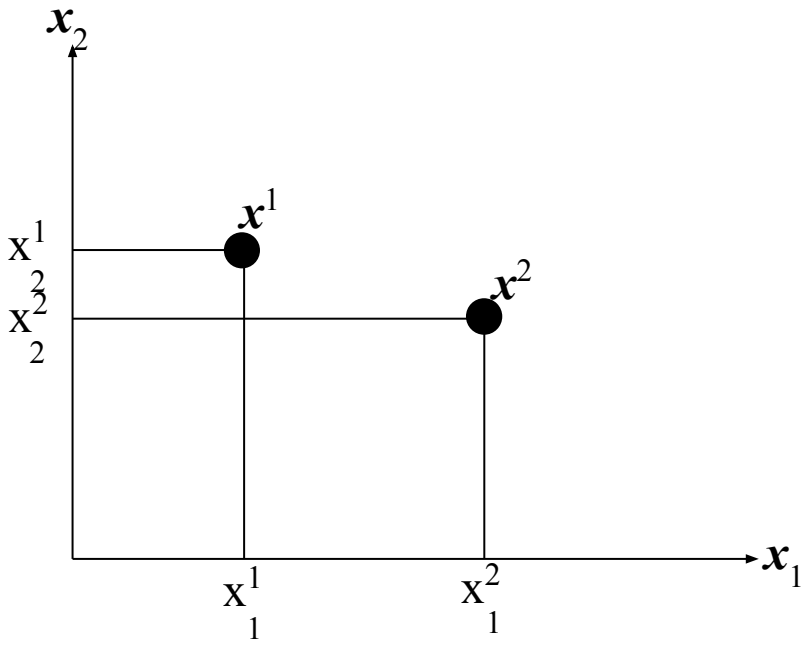
Процедуры многомерного анализа данных могут работать с матрицами дескрипторов X, содержащими большое количество столбцов, многие из которых линейно-зависимы

Центрирование данных для многомерного анализа

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

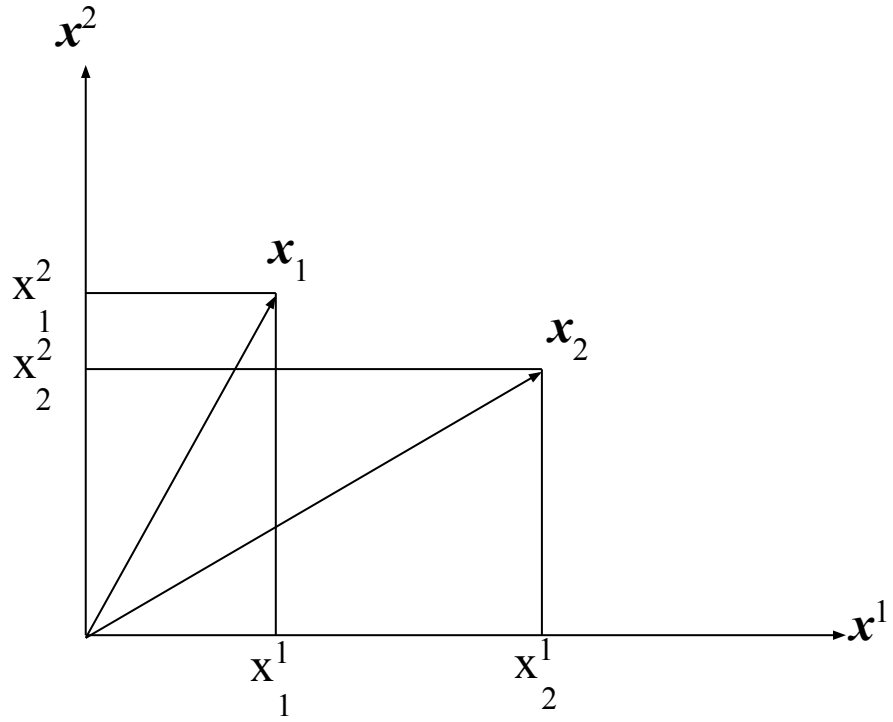
$$\underline{Y} = \left\| \underline{y}_i^j \right\| \quad \text{, где} \quad \underline{y}_i^j = y_i^j - \bar{y}_i \qquad \bar{y}_i = \frac{1}{N}\sum_{k=1}^{N} y_i^k \qquad \forall i = \overline{1, P}$$

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



Оси $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ – дескрипторы, точки $\boldsymbol{x}^1$, $\boldsymbol{x}^2$ – соединения

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



Оси $\boldsymbol{x}^1$, $\boldsymbol{x}^2$ – соединения, вектора $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ – 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

$$D_{ij}^{Manhat\,\tan} = \sum_{k=1}^{M}\left|x_k^i - x_k^j\right|$$

Метрика Минковского

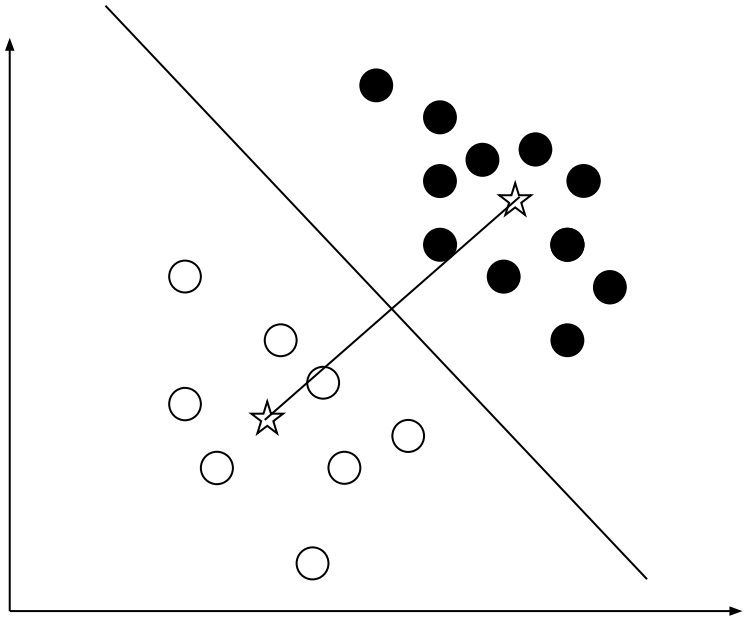
$$D_{ij}^{Minkowski} = \sqrt[r]{\sum_{k=1}^{M}\left|x_k^i - x_k^j\right|^r}$$

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

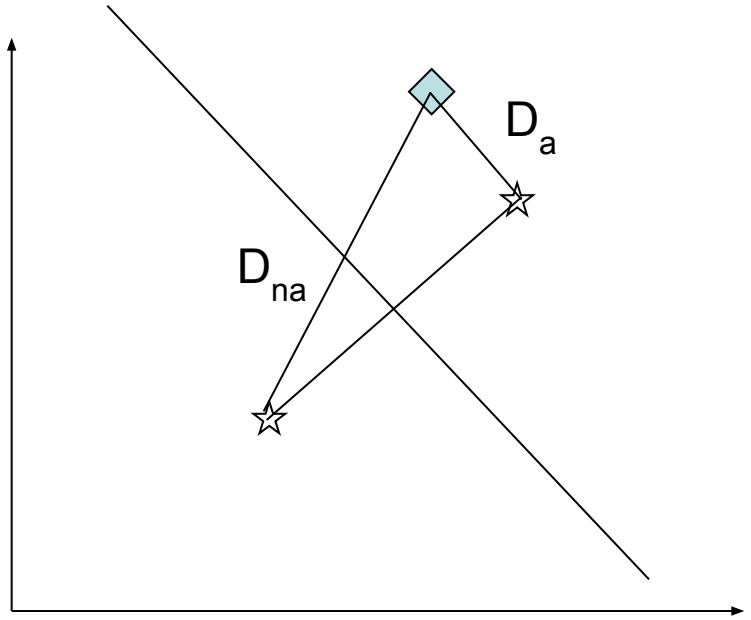
Дискриминантный анализ



● активное соединение

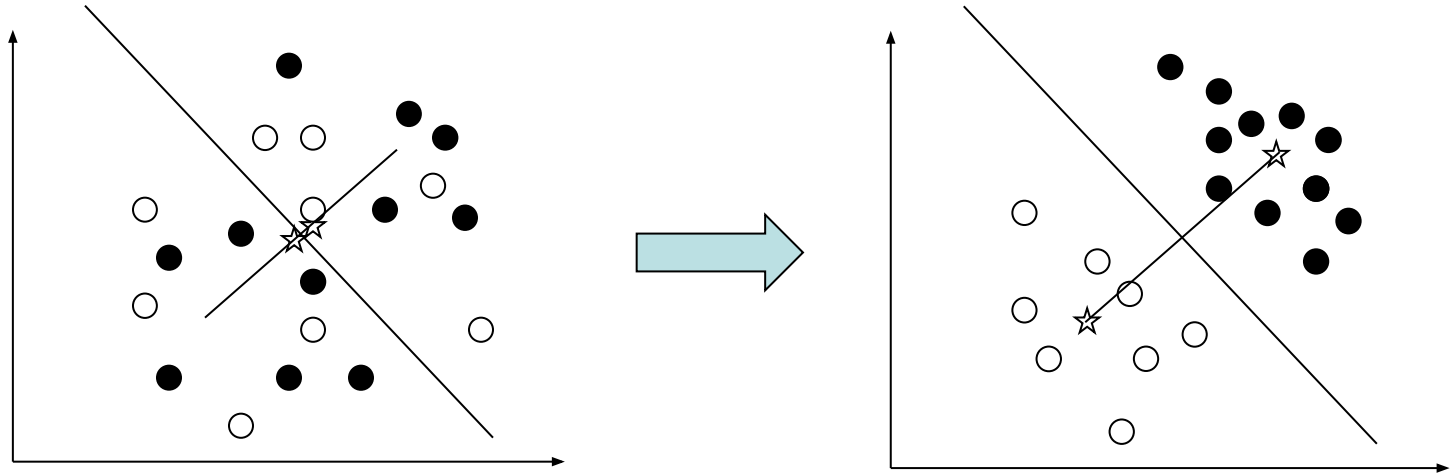
○ неактивное соединение

Дискриминантный анализ



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

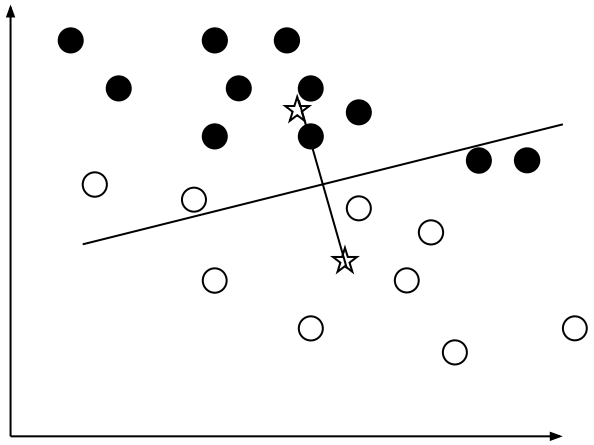
Дискриминантный анализ
(выбор набора дескрипторов)



Плохой набор дескрипторов

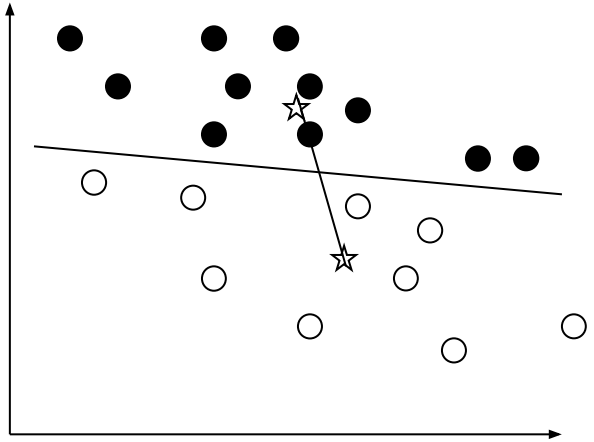
Хороший набор дескрипторов

Дискриминантный анализ
(выбор метрики)



Метрика Эвклида

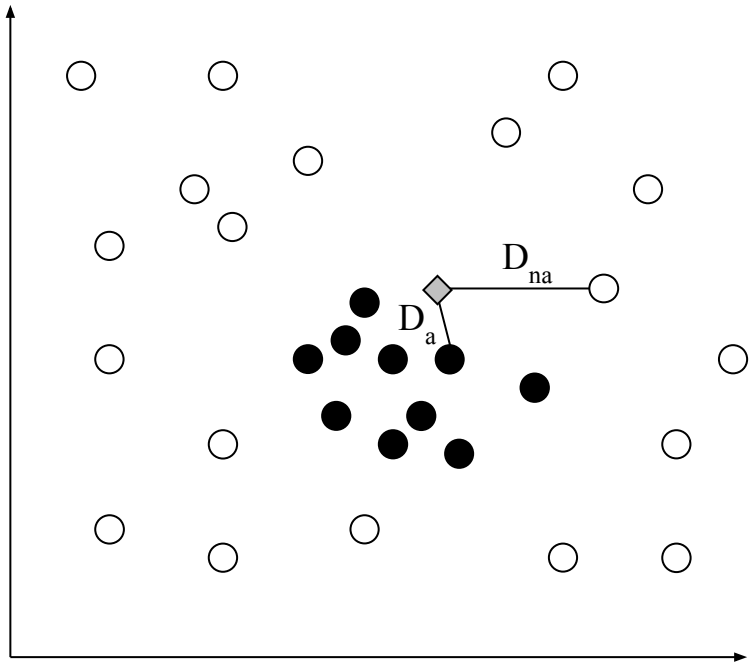
$$D^2 = (\vec{a} - \vec{b})^T (\vec{a} - \vec{b})$$



Метрика Махаланобиса

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

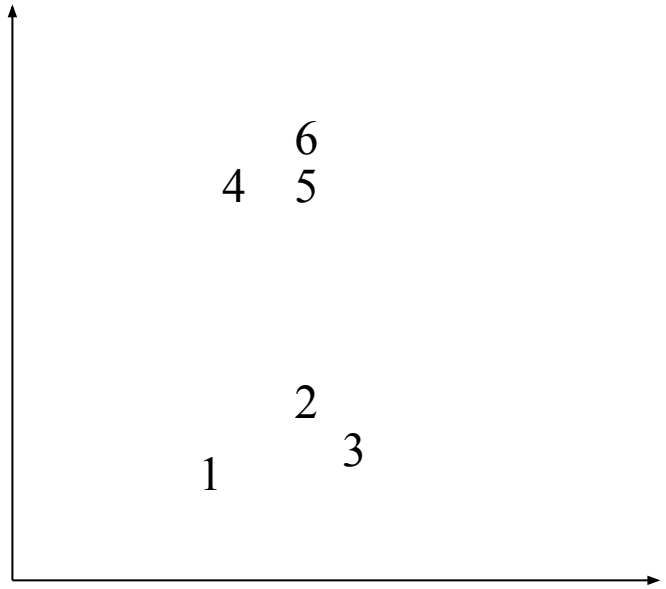
Метод ближайших соседей
(kNN – k Nearest Neighbours)

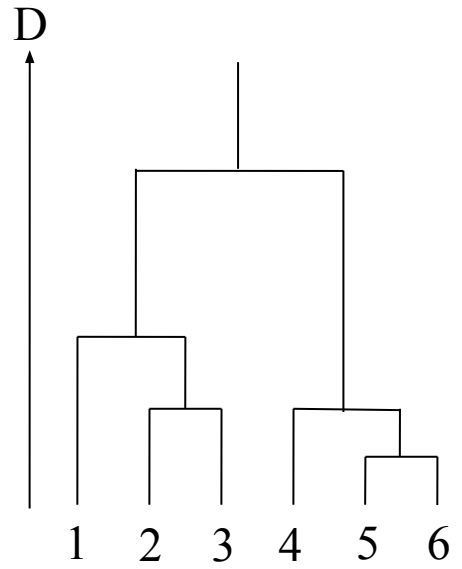


Соединение прогнозируется как активное, поскольку $D_a < D_{na}$

Кластерный анализ

Задача кластерного анализа – изучение внутренней структуры и выявление группировки данных





Дендограмма

Свойства пространства соединений

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

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

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

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}}\sqrt{c_{jj}}}$$ - 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

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

$$S = \underline{X}L$$

$$S = \begin{pmatrix} s_{11} & \ldots & s_R \\ \ldots & \ldots & \ldots \\ s_{N1} & \ldots & s_{NR} \end{pmatrix}$$

Матрица оценок (scores)

$$L = \begin{pmatrix} l_{11} & \ldots & l_{1R} \\ \ldots & \ldots & \ldots \\ l_{M1} & \ldots & l_{MR} \end{pmatrix}$$

Матрица нагрузок (loading)

Вектора ***s*** обычно берутся ортогональными,
т.е. латентные переменные линейно независимы

Метод главных компонент
(PCA – Principal Component Analysis)

Цель метода главных компонент – описание данных минимально возможным количеством латентных переменных

$$\underline{X} = S_0 \Lambda^{\frac{1}{2}} L_0^T$$ - SVD (Singular Value Decomposition) разложение

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix} \lambda_1^{\frac{1}{2}} & & & & & & \\ & \lambda_2^{\frac{1}{2}} & & & & & \\ & & \ddots & & & & \\ & & & \lambda_R^{\frac{1}{2}} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{pmatrix}$$

Матрица сингулярных значений

$$\lambda_i = \frac{1}{N-1} \sum_{j=1}^{N} (s_i^j)^2$$

Метод главных компонент
(PCA – Principal Component Analysis)

$$(\underline{X}^T\,\underline{X})L_0 = \lambda_i L_0$$

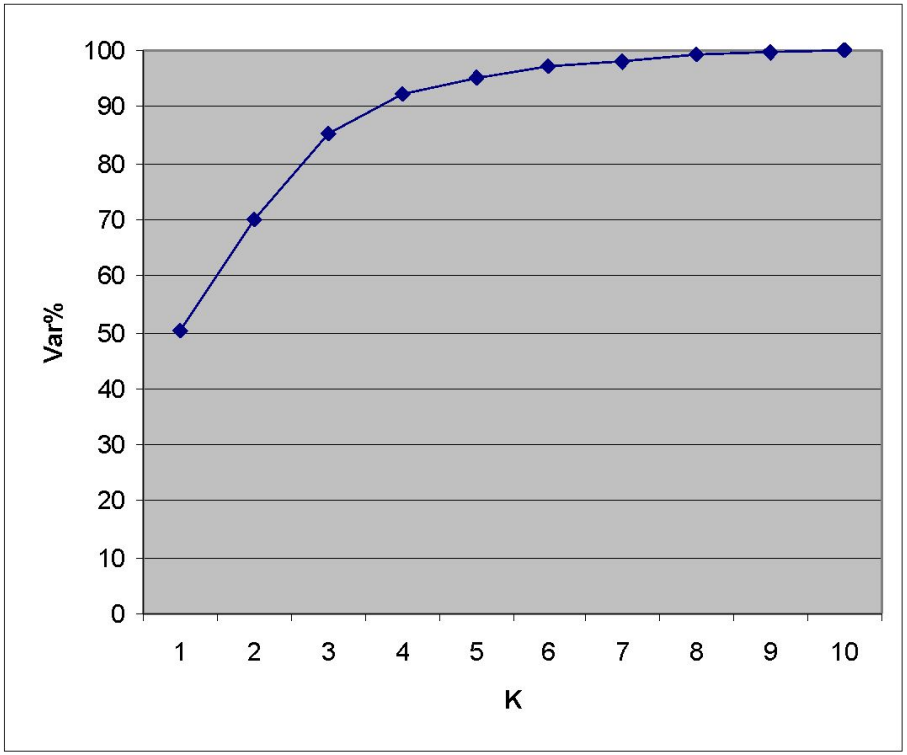
$\boldsymbol{C} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ – матрица ковариаций

$$L_0^T(\underline{X}^T\,\underline{X})L_0 = \begin{pmatrix} \lambda_1 & & & & & & \\ & \lambda_2 & & & & & \\ & & \boxtimes & & & & \\ & & & \lambda_R & & & \\ & & & & 0 & & \\ & & & & & \boxtimes & \\ & & & & & & 0 \end{pmatrix}$$

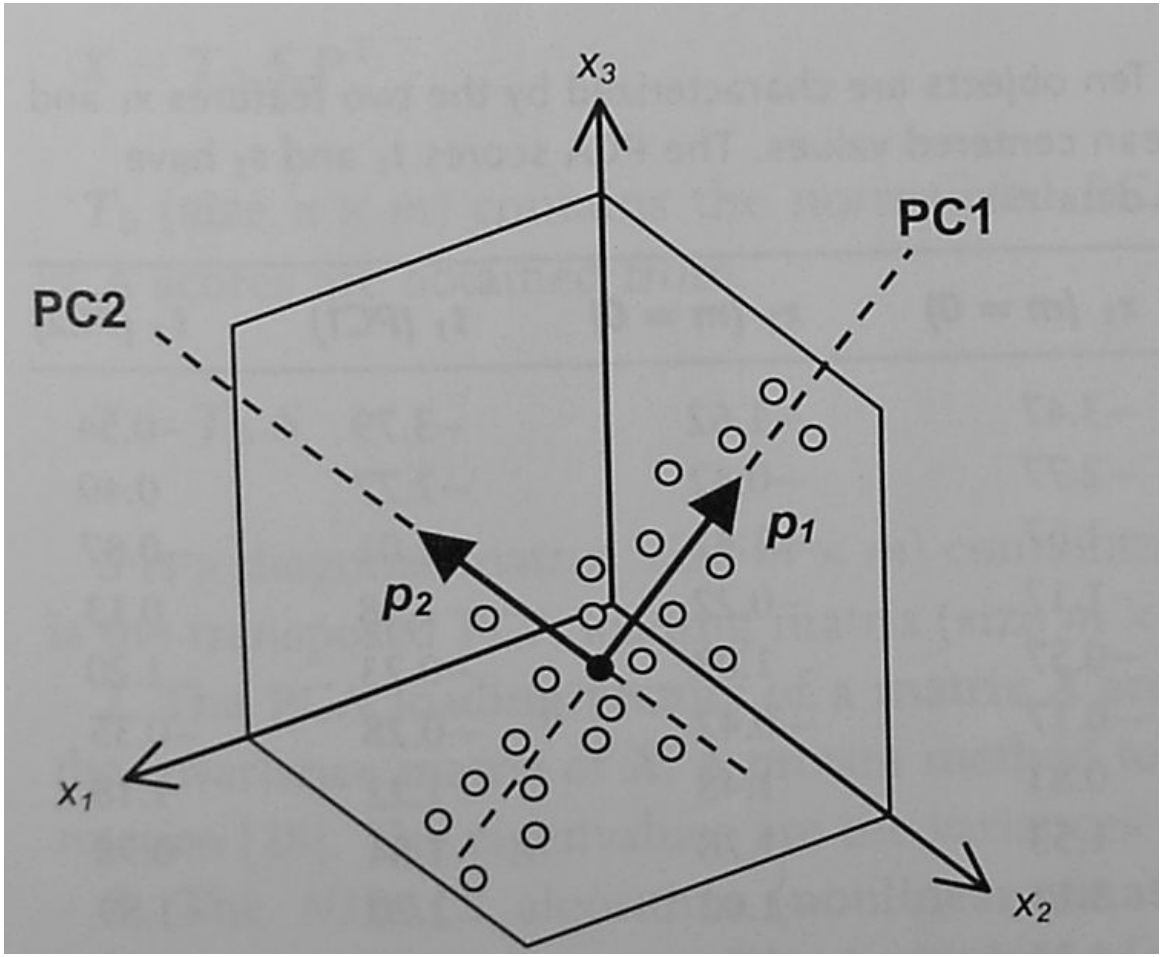
Отбор главных компонент

$$\lambda_1 \geq \lambda_2 \geq \square \geq \lambda_R \geq 0$$

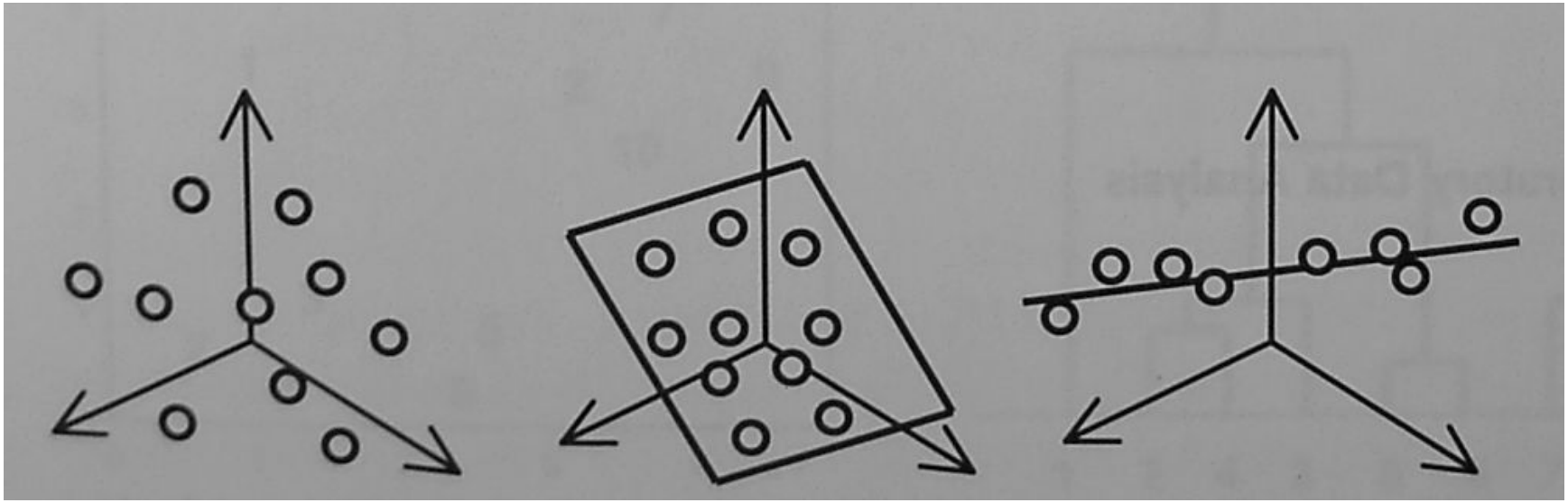
$$Var\% = \frac{100 \cdot \sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{R} \lambda_i}$$



Главные компоненты



Определение размерности данных



K=3 K=2 K=1

Графики оценок и нагрузок

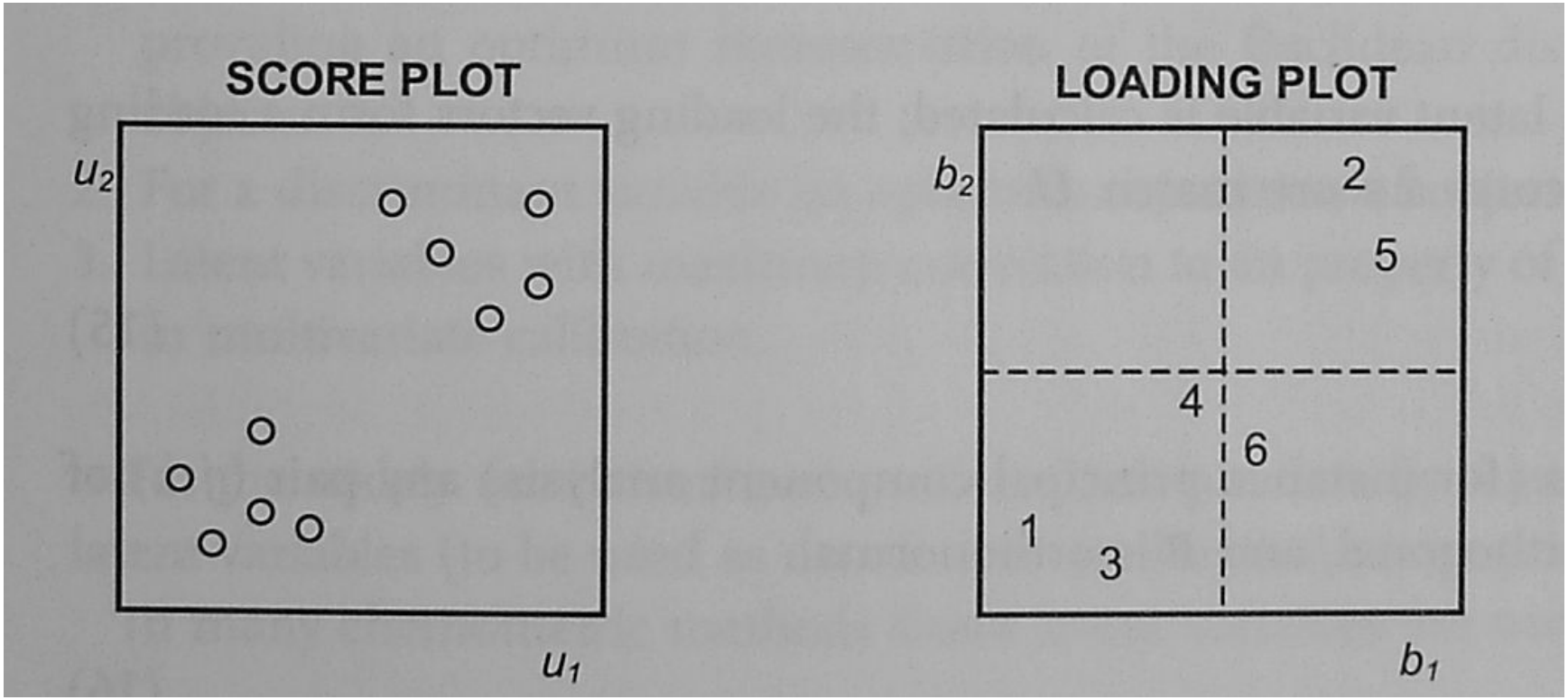


График в координатах главных оценок

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

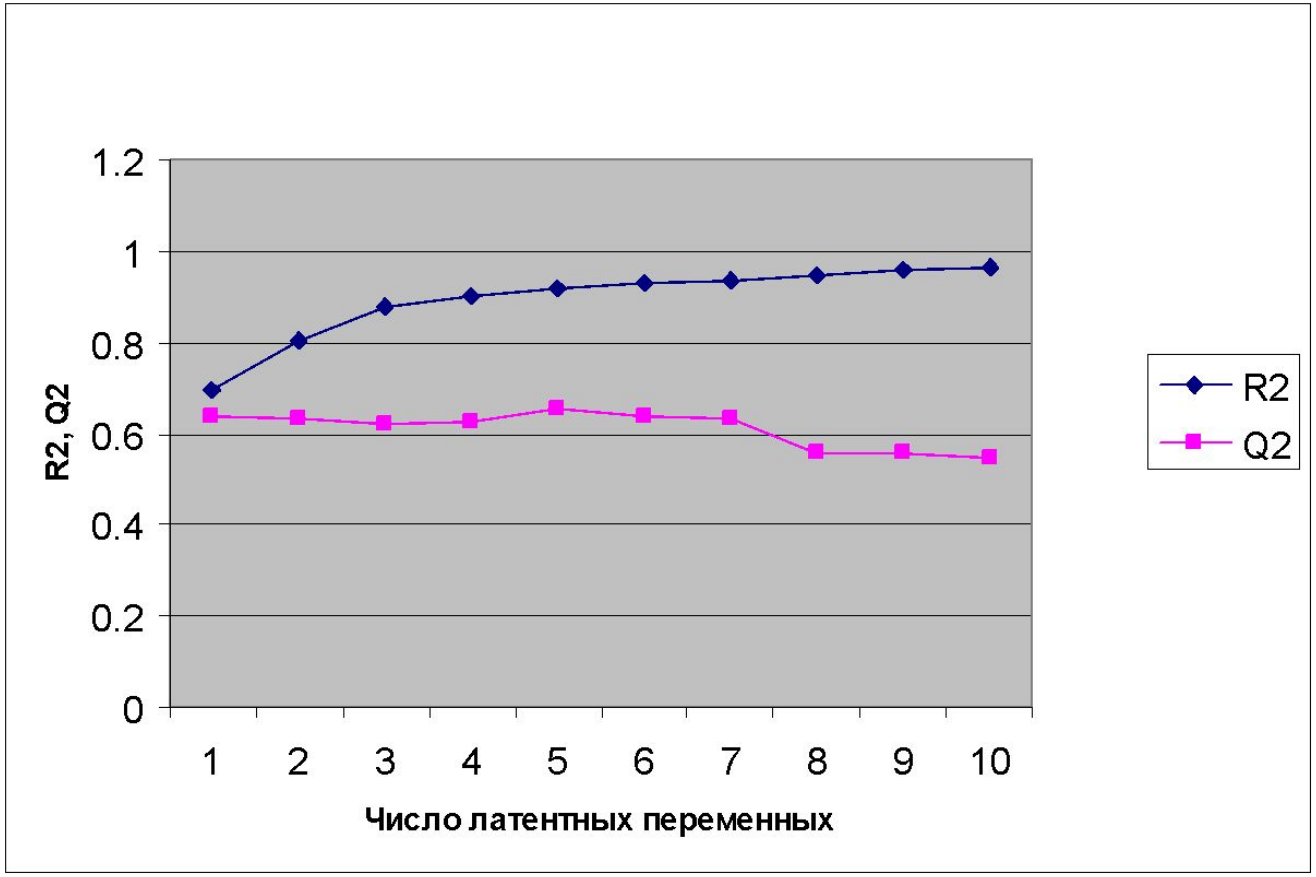
Метод частичных наименьших квадратов (PLS – Partial Least Squares)

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

$$y^j = \sum_{k=1}^{K} a_k s_k^j \qquad s_k^j = \sum_{k=l}^{M} l_{ik} \underline{x}_i^j$$

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

Определение оптимального числа латентных переменных



Оптимальное число латентных переменных - 5

Резюме метода PLS

- Один за одним отбираются латентные переменные, максимально коллинеарные с векторомиси 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