Регресионни методи

Линейна регресия

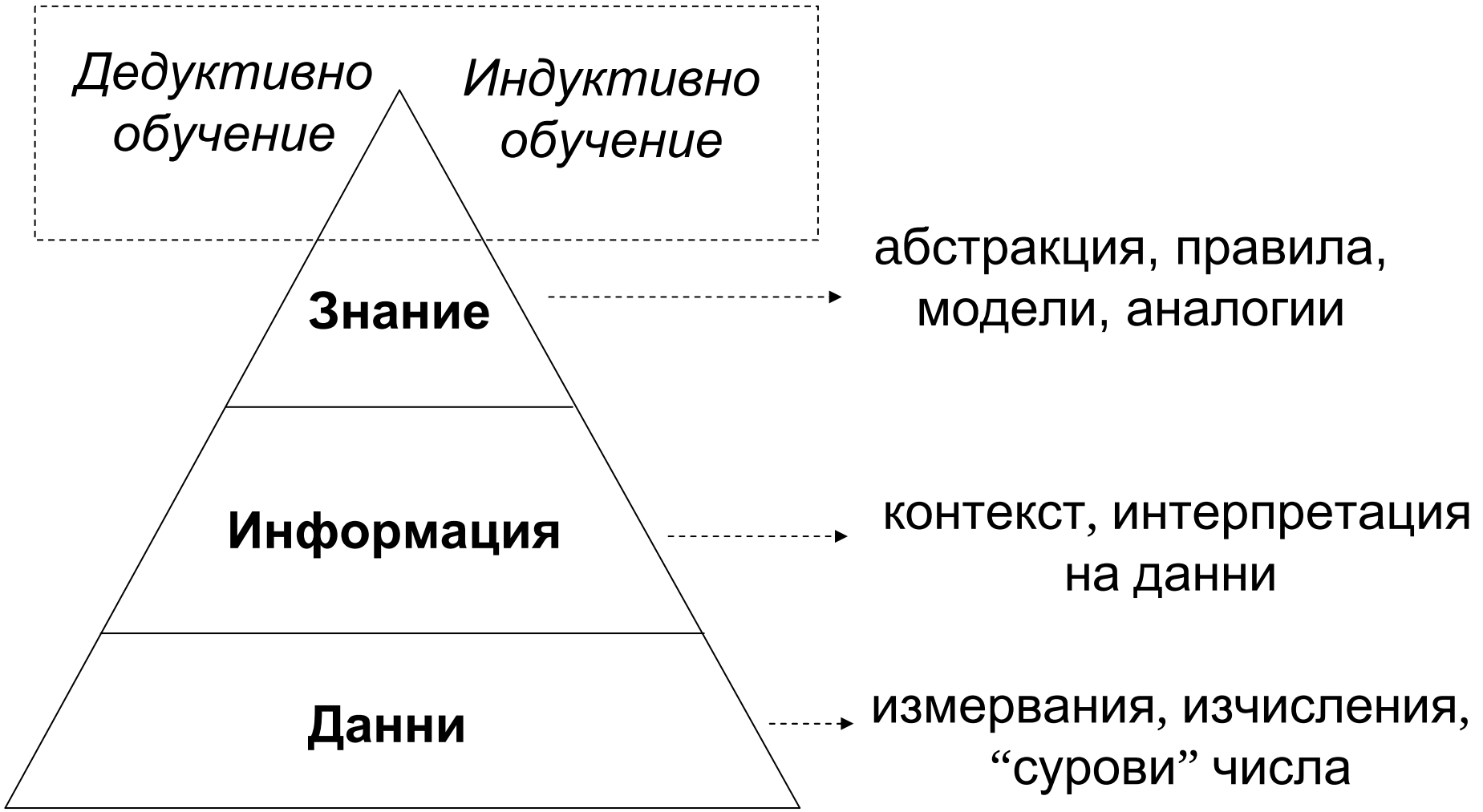
Нелинейна регресия

Статистическа оценка на параметрите

Регресия по главните компоненти

Частична линейна регресия

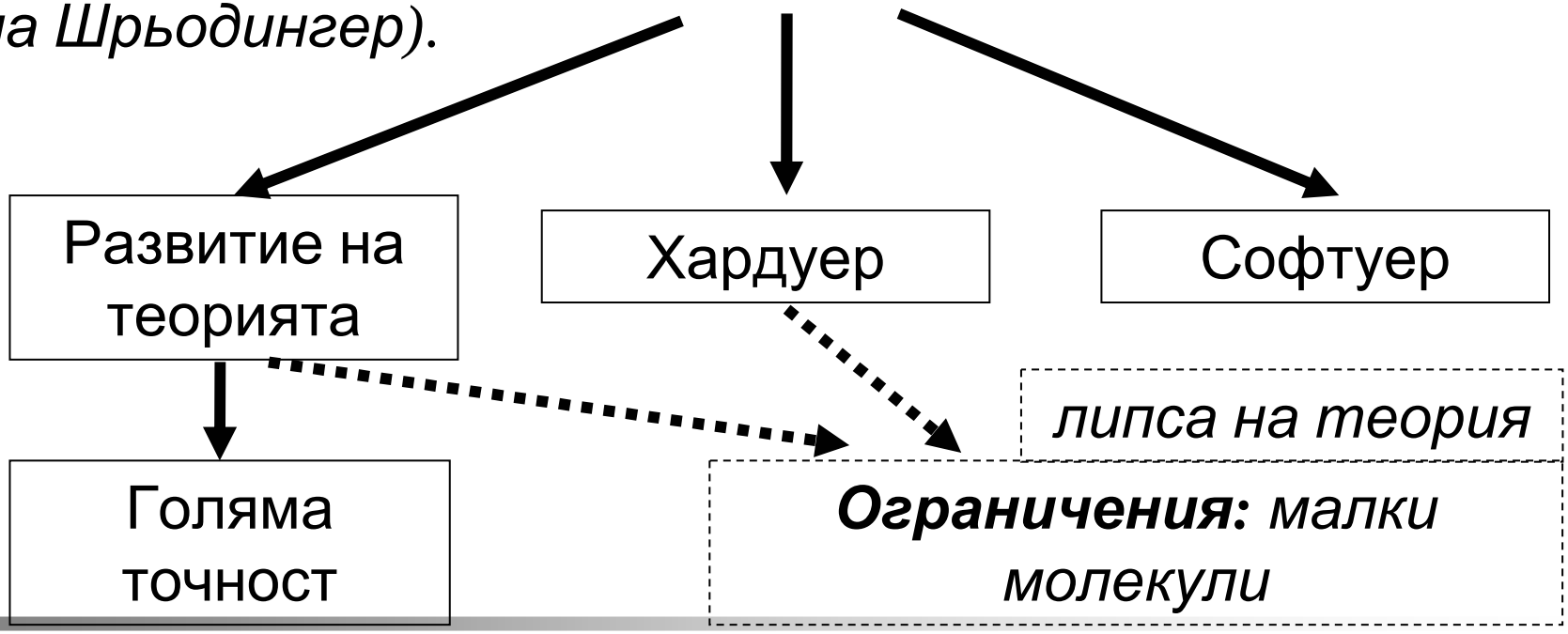
Придобиване на знания (обучение) в химията



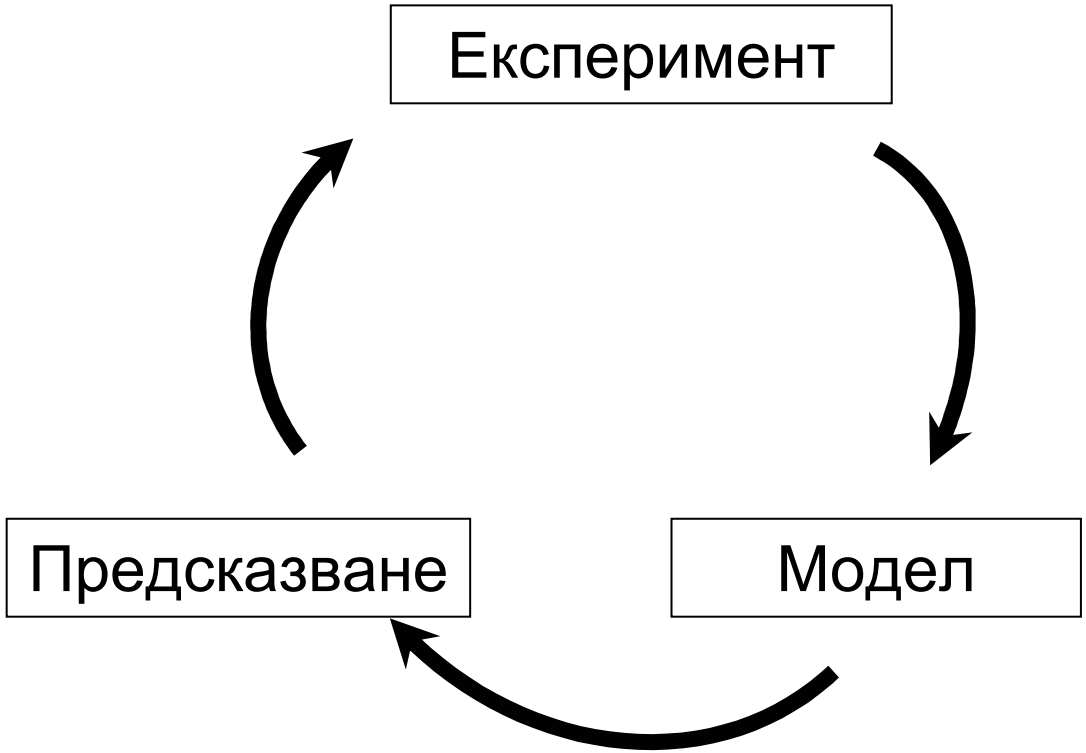
Дедуктивен подход

Необходима е фундаментална теория, която позволява да се правят заключения за изследваните обекти и да се изчисляват желаните свойства.

Квантовата механика е фундаментална теория, която позволява да се опише зависимостта между свойствата на съединенията и 3-измерната им структура *(уравнение на Шрьодингер)*.



Индуктивен подход за обучение



Статистическо описание на модел

При индуктивните методи се работи с извадки от обекти, които представляват генералната съвкупност от всички химични обекти.

Изчислените параметри на модела са статистически оценки на параметрите на ‘истинската закономерност’.

Типичен представител на статистическите методи за моделиране са **регресионните методи.**
Класификационните алгоритми също се характеризират статистически.

Регресионен анализ

Регресионният анализ е статистичеси метод за определяне на връзката между ‘зависима променлива’ **y** и множество независми променливи (дескриптори) $\mathbf{x_1}$, $\mathbf{x_2}$, …,$\mathbf{x_n}$ които в статистически смисъл влияят върху стойностите на целевата променлива **y**.

За целевия параметър и дескрипторите са известни определен брой експериментални стойности.

Зависимостта се описва функционално:

$$\mathbf{y = f(x_1, x_2, \ldots, c_1, c_2, \ldots)}$$

Стойностите на константите **c1, c2, …** определят модела (регресионната зависимост)

Линейна регресия

Съществуват два вида математични модели за описание на връзката между една или повече ‘контролирани независими’ променливи (дескриптори) ***x*** и зависима променлива ***y*** (целева променлива/свойство):

- модели, които са линейни по отношение на параметрите

- модели, които не са линейни по отношение на параметрите

Линейна регресия - примери

Линейните комбинации от няколко променливи:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m = \sum_{i=0}^{m} \beta_i x_i$$

β_i – параметри на модела

x_i – аргументи / дескриптори

Линейна регресия - примери

Полиномите от вида:

$$y = \beta_0 + \beta_1 x + \ldots + \beta_m x^m = \sum_{i=0}^{m} \beta_i x^i$$

също са линейни регресионни методи.

Обърнете внимание, че зависимостта на целевото свойство **y** от дескриптора (аргумента) **x** не е линейна!!!

Линейна регресия – полиномиален модел

$$y = \beta_0 + \beta_1 x + \ldots + \beta_m x^m = \sum_{i=0}^{m} \beta_i x^i$$

$при\ m = 1 \Rightarrow y = \beta_0 + \beta_1 x$ Линеен модел, съответстващ на права линия

при $\boldsymbol{m \geq 2}$ полином от степен $\boldsymbol{m}$ описва връзката между $\boldsymbol{y}$ и $\boldsymbol{x}$

За да се избегнат недоразумения трябва да се отбележи, че линейните регресионни модели не описват само правите линии. Полиномите също принадлежат към групата на линейните регресионни модели. Те се наричат така, защото са линейни (от първа степен) по отношение на параметрите $\boldsymbol{\beta_0 \ldots \beta_m}$

Примери за нелинейна регресия

Често използвани модели, които не са линейни по отношение на параметрите:

експоненциална функция $$y = \beta_1 e^{\beta_2 x}$$

функция на Гаус $$y = \beta_1 e^{-\beta_2 (x-\beta_3)^2}$$

функция на Лоренц $$y = \frac{\beta_1}{1+\beta_2 (x-\beta_3)^2}$$

Линейна регресия

Всички разгледани модели (линейни и нелинейни) могат аналитично да се представят:

$$y = f(x, \beta_0, ..., \beta_m), \text{ където } (\beta_0, ..., \beta_m) \text{ са параметри на модела}$$

Всяко експериментално наблюдение на зависимата променлива y_i може да се представи:

$$y_i = f(x_i, \beta_0, ..., \beta_m) + e_i \qquad i = 1,2,...,n$$

Където n е броят на наблюденията, e_i представлява разликата между стойността предсказана от модела $f(x_i, \beta_0, ..., \beta_m)$ и експерименталното наблюдение y_i.

e_i се наричана още остатък

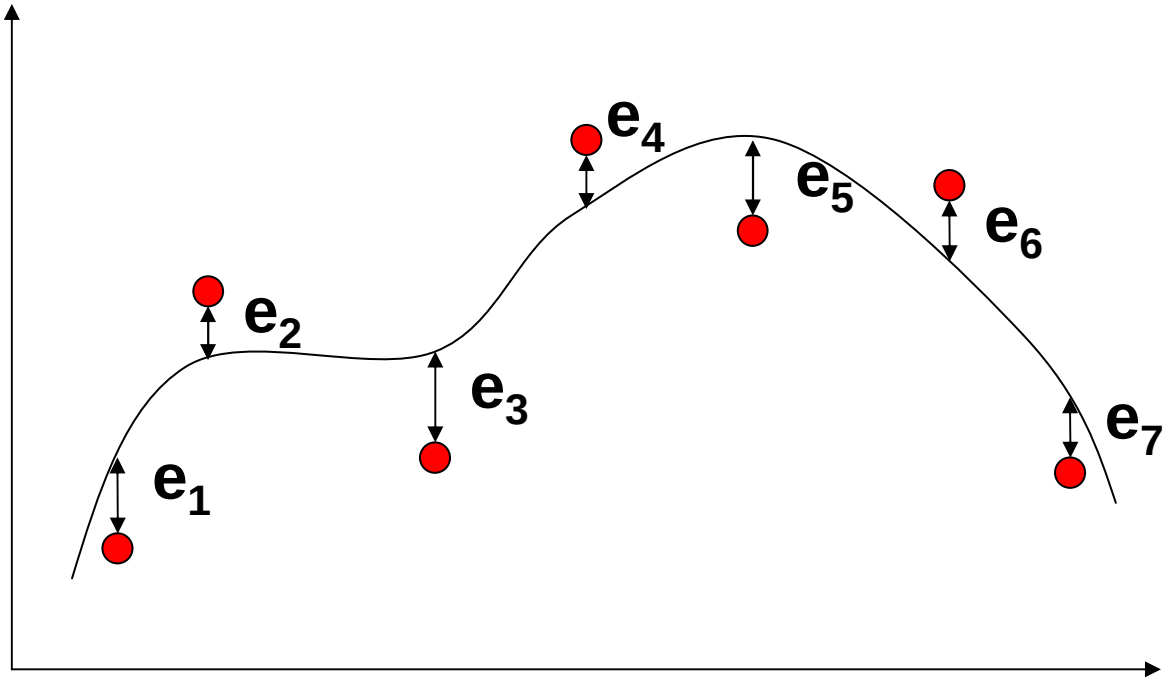
Линейна регресия

Задачата, която се поставя, е намирането на такива стойности на параметрите на модела, за които функцията

$y = f(x, \beta_0, \ldots, \beta_m)$ в максимална степен се доближава до експерименталните точки (x_i, y_i).

Математически това означава да се минимизира сумата

$$\sum_{i=1}^{n} e_i^2$$



Линейна регресия

При определянето на параметрите се прави допускането, че променливите x са контролирани и няма неопределеност по отношение на техните стойности и следователно те нямат принос в разликата между наблюдаваните двойки (x_i,y_i) и модела.

Това допускане е вярно за повечето анализи, когато x се определят експериментално.

Кохато x са определени като молекулни дескриптори също няма неопределеност за стойностите на x, защото изчислителните методи винаги връщат една и съща стойност за даден химичен обект.

Линейна регресия на една променлива

Под линейна регресия на една променлива се разбира случаят, когато в уравнението на регресия участва една единствена независима променлива *x* и всички **параметри на модела са от първа степен**, независимо от сложността на функцията $y = f(x, \beta_0, \ldots, \beta_m)$.

Например, зависимостта:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \frac{\beta_3}{x} + \beta_4 \ln x$$

представлява линейна регресия на една променлива.

Линейна регресия с полином

Нека предположим, че чрез ***n*** експериментални наблюдения (y_1, y_2, …, y_n), получени при ***n*** стойности на независимата променла (x_1, x_2, …, x_n), искаме да създадем модел, който представлява полином от степен ***m***. За определянето на $p = m + 1$ параметъра (β_0 , β_1, …, β_m) е необходимо да се реши система от n уравнения с p неизвестни ($n \geq p$):

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots + \beta_{m2} x_1^m + e_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \ldots + \beta_{m2} x_2^m + e_2 \\
&\cdot \\
&\cdot \\
y_n &= \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \ldots + \beta_{m2} x_n^m + e_n
\end{aligned}
$$

В случая, когато всички експериментални наблюдения **(y_1, y_2, …, y_n)**, се разглеждат равнопоставено, **регресията е непретеглена**

Определяне на параметрите

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots + \beta_{m2} x_1^m + e_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \ldots + \beta_{m2} x_2^m + e_2$$
$$\cdot$$
$$\cdot$$
$$y_n = \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \ldots + \beta_{m2} x_n^m + e_n$$

Видът на тези уравнения и техните решения са доста комплицирани. Затова възниква необходимост от компактен и ясен начин за тяхното представяне и решение.

Това може да се постигне с помоща на матричната алгебра.

Допълнително предимство на матричните уравнения е, че те лесно могат да бъдат подложени на компютърна обработка.

Линейна регресия с полином

Въвеждат се следните вектори и матрици:

експериментални стойности

параметри

$$y=\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \quad e=\begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{pmatrix} \quad \beta=\begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_m \end{pmatrix} \quad X=\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^m \\ 1 & x_2 & x_2^2 & \dots & x_2^m \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ 1 & x_n & x_n^2 & \dots & x_n^m \end{pmatrix}$$

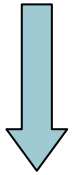
остатъци

матрица на независимите променливи

Линейна регресия с полином

Системата от уравнения може да се представи в матричния вид:

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \ldots + \beta_{m2} x_1^m + e_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \ldots + \beta_{m2} x_2^m + e_2$$
$$\cdot$$
$$\cdot$$
$$y_n = \beta_0 + \beta_1 x_n + \beta_2 x_n^2 + \ldots + \beta_{m2} x_n^m + e_n$$



$$y = X.\beta + e$$

Определяне на параметрите на модела

Параметрите на линейната регресия се определят по **метода на най-малките квадрати**

Критерият е минимизиране на сумата

$$\sum e_i^2$$

записан в матричен вид има смисъл, че квадрата на дължината на вектора $\|\boldsymbol{e}\|$ трябва да бъде минимална:

$$(e_1, e_2, \ldots, e_n)\begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{pmatrix} = e'.e = \| e \|^2$$

Определяне на параметрите на модела

Могат да се направят преобразуванията:

$$y = X.\beta + e$$

$$e = y - X.\beta$$

Следователно:

$$\| e \|^2 = e'.e = (y - X.\beta)'(y - X.\beta)$$

Сумата от остатъците се минимизира чрез приравняване към нула на частните производни спрямо параметрите (β_0 , β_1, …, β_m)

Определяне на параметрите на модела

Извежда се система от линейни уравнения за параметрите на линейната регресия

$$\frac{\partial \| e \|^2}{\partial \beta_i} = 0 \qquad \textbf{за i = 1, 2,…, n}$$

Решението на системата може да се запише и чрез матрични операции върху вектора **y** и матрицата **X**.

$$X \cdot b = y \longrightarrow X' \cdot X \cdot b = X' \cdot y$$

$$\longrightarrow b = \left(X' \cdot X\right)^{-1} \cdot X' \cdot y$$

Определяне на параметрите на модела

b е оценка на вектора **β**, и представлява модела на линейната регресия.

Веднъж получен, моделът се използва за предсказване на стойността $\hat{\mathbf{y}}_0$, съответстваща на дадена стойност на независимата променлива $x = x_0$, за която не е извършено експериментално наблюдение.

$$\hat{y}_0 = b_0 + b_1 x_0 + \ldots + b_m x_o^m$$

Линейна регресия от първа степен

Частен случай на разгледаните уравнения се явява линейната регресия от първа степен ($m = 1 \Rightarrow y = b_0 + b_1 x$):

Съответно:

$$b = \begin{pmatrix} b_o \\ b_1 \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}$$

$$\boxed{(X'.X).b = X'.y}$$

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} . \begin{pmatrix} b_o \\ b_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} . \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}$$

$$(X'.X) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} . \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

Линейна регресия от първа степен

Параметрите $\mathbf{b_0}$ и $\mathbf{b_1}$ се определят от системата:

$$n.b_0 + b_1 \sum x_i = \sum y_i$$

$$b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i$$

Доверителен интервал на параметрите

При условие, че неопределеността на всеки параметър е независима от тази на останалите, вариациите на параметрите $v(b_0)$,…, $v(b_m)$ могат да бъдат изчислени поотделно.

При линейната регресия обаче, **оценяването на един параметър обикновенно зависи от останалите параметри**, а следователно съществува и зависимост между техните неопределености. Информацията необходима за определянето на доверителните интервали на параметрите се съдържа във вариационно-ковариационната матрица:

Вариационно-ковариационна матрица

$$V(b) = \begin{pmatrix} v(b_0) & \text{cov}(b_0, b_1) & \ldots & \text{cov}(b_0, b_m) \\ \text{cov}(b_1, b_0) & v(b_1) & \ldots & \text{cov}(b_1, b_m) \\ . & . & \ldots & . \\ . & . & \ldots & . \\ \text{cov}(b_m, b_0) & \text{cov}(b_m, b_1) & \ldots & v(b_m) \end{pmatrix}$$

Диагоналните елементи на матрицата представляват вариацията на параметрите, а извън диагоналните - ковариацията между всеки два параметъра.

Доверителен интервал на параметрите

Оценка за вариационно-ковариационната матрица може да се определи чрез:

$$V(b) = S_e^2 (X' \cdot X)^{-1}$$

където $\mathbf{S_e^2}$ е оценка на вариацията на експеримен-талната грешка.

Доверителен интервал на параметрите

При условие, че моделът добре описва експериментално наблюдаваните точки, оценката $\mathbf{S_e}$ може да се определи:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p}$$

n - брой наблюдения

p - брой параметри

оценката за вариационно-ковариационната матрица се получава:

$$V(b) = \frac{\sum_{i=1}^{n} e_i^2}{n - p} (X' \cdot X)^{-1}$$

Доверителен интервал на параметрите

Ако приемем, че грешките от измерванията са нормално разпределени и независими една от друга, може да се определи доверителният интервал, в който със зададената статистическа сигурност (например 0.95) се намира истинската стойност на всеки един параметър:

$$b_i \pm t_{n-p}^{0.95} \sqrt{v(b_i)}$$

Съответно за всяка стойност на независимата променлива x_0, може да се определи доверителният интервал на предсказаната от модела стойност $\hat{y}_0$:

$$\hat{y}_0 \pm t_{n-p}^{0.95} . s_e . \sqrt{X_0^{'} . \left(X^{'} . X\right)^{-1} . X_0}$$

Многопроменлива линейна регресия

В направените до момента разглеждания, променливата y зависи само от една единствена независима променлива x. Разбира се това е само един частен случай. От практична гледна точка много по-реална е ситуацията, при която величината y зависи от множество от m променливи (x_1, x_2,…, x_m). Тази зависимост, примерно може да представлява уравнението:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Или в общ вид:

$$y = f(x_1, x_2, ..., x_m, \beta_0, \beta_1, ..., \beta_m)$$

Многопроменлива линейна регресия

За да определим оценките (b_1, b_2, …, b_m) на параметрите (β_0, β_1, …, β_m), трябва да бъдат проведени n експериментални наблюдения на зависимата променлива y (y_1, y_2,…, y_n), за n различни комбинации от m независими променливи ($x_{1,1}$,…, $x_{m,1}$,…, $x_{1,n}$,…, $x_{m,n}$) ($n \geq m$).

Както вече беше разгледано, всяко наблюдение y_i, може да бъде изразено като сума от стойността получена чрез модела (f_i) и съответния остатък (e_i):

$$y = f\left(x_{1,i}, x_{2,i}, ..., x_{m,i}, \beta_0, \beta_1, ..., \beta_m\right) + e_i \quad (i = 1,2,...,n)$$

Съответно параметрите (β_0, β_1, …, β_m), трябва да се оценят по такъв начин, нормата на вектора с остатъците $||e||$ да е минимална.

Многопроменлива линейна регресия

В практиката често се използва многопроменлива линейна регресия, при която регресионното уравнение е от вида:

$$y = \beta_0 + \beta_1.x_1 + ... + \beta_m.x_m$$

Това е възможно най-лесният вариант, защото всички независими променливи (x_i) са от първа степен и отсъстват комбинации от променливи ($x_i.x_j$). В този конкретен случай броят на параметрите ще бъде равен на $p = m + 1$, а съответно наблюденията y_i могат да бъдат представени:

$$y_1 = \beta_0 + \beta_1.x_{1,1} + \beta_2.x_{2,1} + ... + \beta_m.x_{m,1} + e_1$$

$$y_2 = \beta_0 + \beta_1.x_{1,2} + \beta_2.x_{2,2} + ... + \beta_m.x_{m,2} + e_2$$

.

.

$$y_n = \beta_0 + \beta_1.x_{1,n} + \beta_2.x_{2,n} + ... + \beta_m.x_{m,n} + e_n$$

Многопроменлива линейна регресия

Тази система от уравнения може да се представи като матрично уравнение:

$$y = X.\beta + e$$

което е аналогично на уравнението на линейна регресия на една променлива.

Където y, e и β са вече разгледаните вектори на измерванията, остатъците и съответно на параметрите. Единствено матрицата на независимите променливи X е зададена по различен начин:

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & x_{3,1} & \cdots & x_{m,1} \\ 1 & x_{1,2} & x_{2,2} & x_{3,2} & \cdots & x_{m,2} \\ . & & & & & \\ . & & & & & \\ 1 & x_{1,n} & x_{2,n} & x_{3,n} & \cdots & x_{m,n} \end{pmatrix}$$

Многопроменлива линейна регресия

Следователно оценките на параметрите b, могат да се определят с помощта на метода на най-малките квадрати аналогично на линейната регресия на една променлива:

$$\mathrm{b} = \left(\mathrm{X}'.\mathrm{X}\right)^{-1}.\mathrm{X}'.\mathrm{y}$$

Вариационно-ковариационната матрица на параметрите се изчислява също по аналогичен начин:

$V(b) = s_e^2 \left(X'.X\right)^{-1}$ където $s_e^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-p}$

n - брой наблюдения

p - брой параметри

Доверителни интервали:

$$b_i \pm t_{n-p}\sqrt{v(b_i)} \qquad \hat{y}_0 \pm t_{n-p}.s_e.\sqrt{\mathrm{x}_0'.\left(X'.X\right)^{-1}.\mathrm{x}_0}$$

Претеглена линейна регресия

В случая, когато уравнението на регресия се извежда на базата на n наблюдения при стойности на независимата променлива *x*, вариращи в голям интервал, е удачно използването на т. нар. претеглена линейна регресия. Това се налага от обстоятелството, че в този случай дисперсията на зависимата променлива *y* може да не е константна величина в изследвания динамичен диапазон (хетероскедастичен модел). Следователно, отклоненията, *e*, между модела и измерванията, ще имат по-малка ‘тежест’ когато измерването е свързано със значителна случайна грешка спрямо отклоненията, определени за измервания с малка случайна грешка. Това налага въвеждането на претеглящ фактор, обратно пропорционален на стандартното отклонение на измерванията.

Претеглена линейна регресия

Въвежда се матрица на теглата на остатъците:

$$W = \begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{pmatrix}$$

Стандартни тегла: обратно-пропорционални на стандартните отклонения

$$W = \begin{pmatrix} 1/s_1^2 & & & 0 \\ & 1/s_2^2 & & \\ & & \ddots & \\ 0 & & & 1/s_n^2 \end{pmatrix}$$

където s_i^2 е стандартното отклонение на наблюдението y_i

Претеглена линейна регресия

Изчислява се претегления вектор на остатъците

$$W.e = \begin{pmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & . & \\ & & & \\ 0 & & & w_n \end{pmatrix} . \begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{pmatrix} = \begin{pmatrix} w_1.e_1 \\ w_2.e_2 \\ . \\ . \\ w_n.e_n \end{pmatrix}$$

По аналогия с непретеглената линейна регресия, оценките $\boldsymbol{b}$ на действителните параметри на модела $\boldsymbol{\beta}$ се определят чрез минимизиране на квадрата на дължината на вектора $\|\mathbf{We}\|^2$.

Нелинейна регресия

Функцията на модела:

$$y = f\left(x, \beta_0, \beta_1, \ldots, \beta_m\right)$$

е нелинейна функция на параметрите $\{\beta_i\}$

Например модел за хроматограма състояща се от няколко Гаусови ивици:

$$f(x) = \sum_{i=0}^{N} \beta_{1,i} e^{-(x-\beta_{2,i})^2/\beta_{3,i}}$$

Нелинейна регресия

При нелинейната регресия, моделът **не може** да се представи в удобна матрична форма: **y = Xβ+e**

Ако добро начално приближение $(b)_0$ на вектора с параметрите е известно, тогова **f(x)** може да приближи с линейна функция на вектора с разликите:

$$f(x, b + \Delta b) = f(x, b) + \sum_{k=1}^{\infty} \frac{\partial^k f}{\partial b^k} \Delta b^k$$

ако **Δb** е достатъчно малко, членовете от реда на Тейлър за **k > 1**, може да се пренебрегнат.

Нелинейна регресия

Модела става линейна функция на разликите

$\Delta b_i = b_i - b_i^0$

$$f(x, b_1, b_2, ..., b_m) = f(x, b_1^0, b_2^0, ..., b_m^0) + \sum_{i=1}^{m} \frac{\partial f}{\partial b_i} \Delta b_i$$

$$\hat{y}_i = f(x_i, b_1, b_2, ..., b_m)$$

Нелинейна регресия

Ако означим векторът с разлилите между модела и експерименталните стойности с:

$$\Delta y = \begin{pmatrix} y_1 - \hat{y}_{1,0} \\ y_2 - \hat{y}_{2,0} \\ . \\ . \\ y_n - \hat{y}_{n,0} \end{pmatrix}$$

и матрицата с първите производни (Якобияна)

$$V(b) = \begin{pmatrix} \frac{\partial f_1}{\partial b_1} & \frac{\partial f_1}{\partial b_2} & \cdots & \frac{\partial f_1}{\partial b_m} \\ \frac{\partial f_2}{\partial b_1} & \frac{\partial f_2}{\partial b_2} & \cdots & \frac{\partial f_2}{\partial b_m} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ \frac{\partial f_n}{\partial b_1} & \frac{\partial f_n}{\partial b_2} & \cdots & \frac{\partial f_n}{\partial b_m} \end{pmatrix}$$

Нелинейна регресия

Векторът с разликите може да се изрази чрез матрични операции:

$$\Delta y = J \cdot \Delta b$$

Системата се решава се спрямо **Δb**

$$\Delta b = (J' \cdot J) \cdot J' \cdot \Delta y$$

По-точната оценка за параметрите **b** се получава:

$$b = b^0 + \Delta b$$

PCA - анализ на главните компоненти

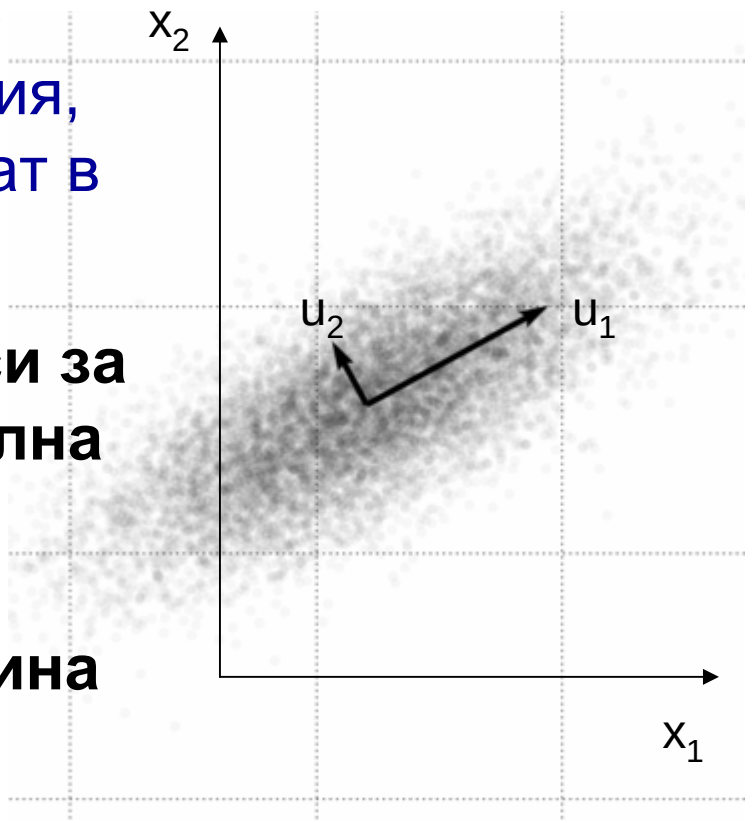
Анализ на главните компоненти PCA (принципен компонентен анализ) е най-често използваният метод за представяне на данните чрез линейна комбинация на латентни променливи.

Анализът на главните компоненти се използва за представяне на данните във векторно пространство с по-малка размерност, визуализация на данните, извличине на най-съществената информация, кластериране и интерпретация на данните директно от експерт чрез визуално наблюдение.

PCA - анализ на главните компоненти

Анализът на главните компоненти е ортогонална линейна трансформация, при която данните се трансформират в нова координатна система,

при която от всички възможни оси за проектиране на данните максимална вариация се получава по направлението на първия главен компонент, следващата по големина вариациа се получава по оста на втория главен компонент и т.н.



Изразяване на първоначалните променливи

$$U' = V.X' \quad \longrightarrow \quad X = U.V$$

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ u_{21} & u_{22} & \dots & u_{2m} \\ . & . & \dots & . \\ . & . & \dots & . \\ u_{n1} & u_{n2} & \dots & u_{nm} \end{pmatrix} . \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ . & . & \dots & . \\ . & . & \dots & . \\ v_{m1} & v_{m2} & \dots & v_{mm} \end{pmatrix}$$

Определяне на натоварващите коефициенти

Коефициентите на главните компоненти (съответните редове от матрицата V) се определат като собствени вектори на ковариационната матрица на данните - C_x

$$C_x.v = \lambda.v$$

Собствените стойности $\{\lambda_k\}$ на C_x съответсват на вариацията на съответния главен компонент: $\lambda_k = var(u_k)$.

Главните компоненти се подреждат според собствените стойности в намаляващ ред: $\lambda_1 > \lambda_2 > ... > \lambda_n$

PCR – регресия по главните компоненти

При PCR метода, като дескриптори се използват галвните компоненти:

$$y = b_0 + b_1 u_1 + b_2 u_2 + \ldots + b_m u_m$$

Може да се използват няколко принципни компоненти за да се получи зависимост с много малък брой параметри

$$y = b_0 + b_1 u_1 + b_2 u_2$$

PCR – регресия по главните компоненти

Не е задължително най-главните компонети се използват в уравнението на линейнара регресия, защото компонентите $\{\mathbf{u}_i\}$ описват вариацията в матрицата с дескрипторите **X**, докато целта на регресионния анализ е да се обясни и вариацията в целевото свойство **Y**.

Обикновенно първите два главни компонента винаги влизат в PCR регресията

PCR – регресия по главните компоненти

При добавяне на нов принципен компонент в регресията, коефициентите на вече участващите компоненти не се променят. Това е свойство е следствие от ортогонал-ността на компонентите.

Като недостатътк на метода може да се отбележи затрудненатa интерпретация на регресиоанния модел в термините на принципните компоненти.

Стъпкова регресия

Автоматична процедура за линейна регресия, при която дескрипторите се избират автоматично измежду голям предварително изчислени дескриптори.

$$y = b_0 + b_1 x_{k_1}$$

$$y = b_0 + b_1 x_{k_1} + b_2 x_{k_2}$$

$$y = b_0 + b_1 x_{k_1} + b_2 x_{k_2} + b_3 x_{k_3}$$

...

$$y = b_0 + b_1 x_{k_1} + b_2 x_{k_2} + b_3 x_{k_3} + b_4 x_{k_4} + \ldots$$

На всяка стъпка се добавят нови променливи, които подобряват статистическите характеристики на модела.

Литература:

1. Handbook of Chemoinformatics: From Data To Knowledge Vol. 3, Ed. J. Gasteiger, WILEY-VCH Verlag GmbH & Co., 2003.

2. D. Massart, B. Vandenginste, S. Deming, Y. Michotte, L. Kaufman. Chemometrics: a textbook. Elsevier, 1988.