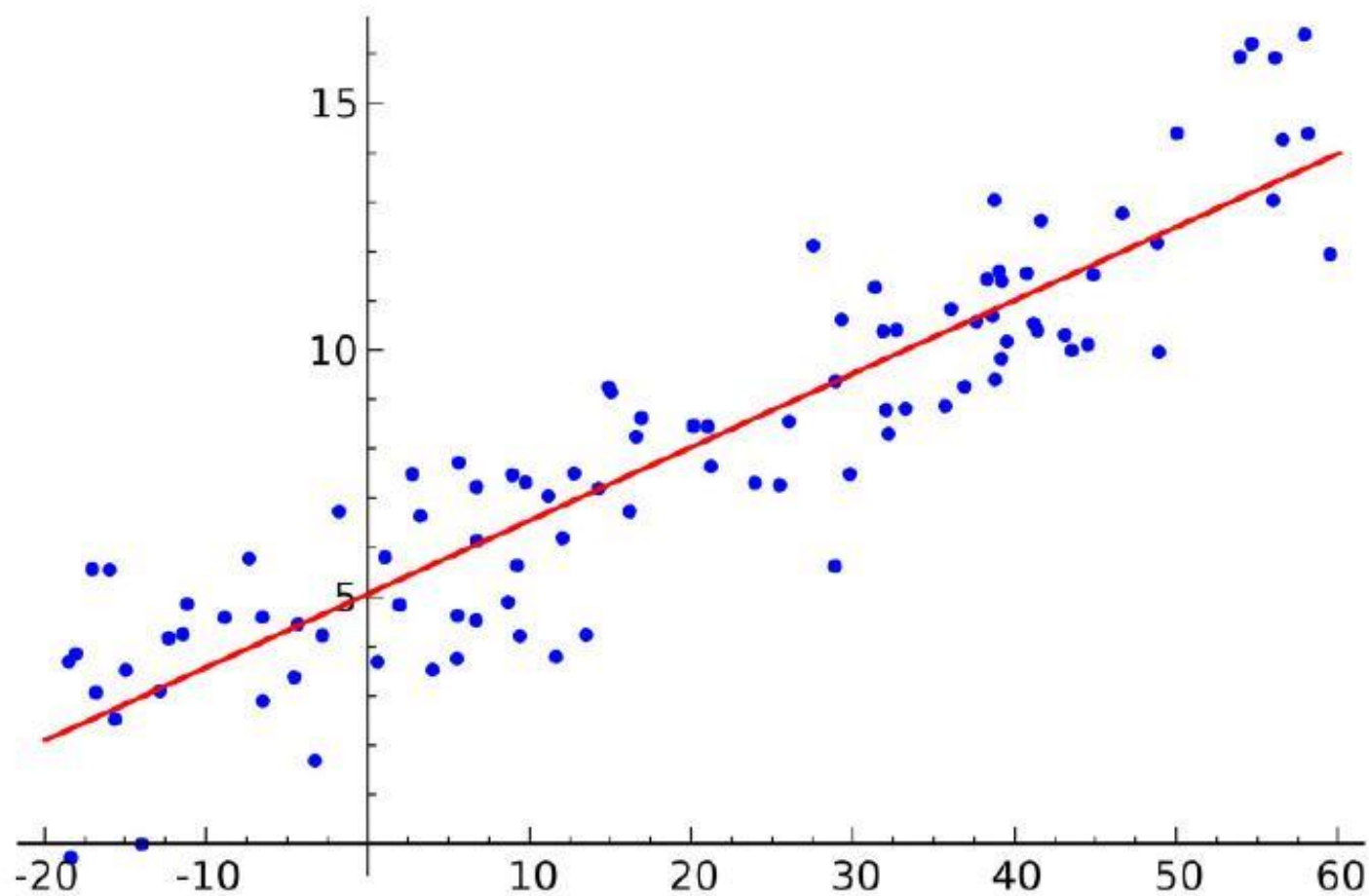


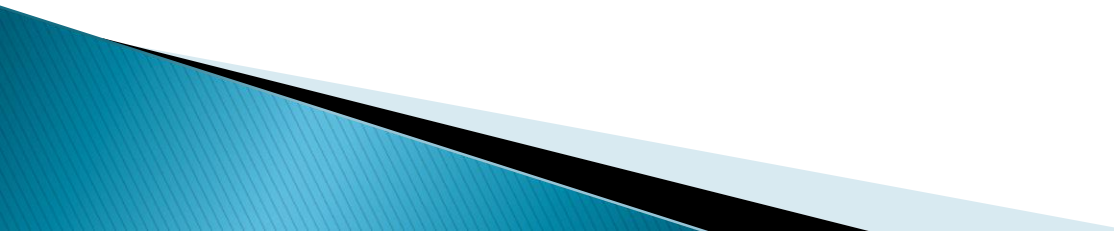
# 第三章 最小二乘数据拟合

- ▶ 拟合直线
- ▶ 多变量线性拟合
- ▶ 非线性曲线拟合
- ▶ 正交多项式拟合

# 引言



# 引言

- ▶ 最小二乘法起源于天文和大地测量的需要
  - ▶ 最小二乘法也称回归分析法，由早期英国生物学家、统计学家高尔顿（F.Gallton）研究父亲和儿子的身高关系所创
  - ▶ 目前最小二乘法已经是数据分析的基本工具之一
- 

# 引言

## ▶ 问题提法

- 已知一组数据  $(x_i, y_i), i = 1, 2, \dots, N$ , 求  $\phi(x)$  满足
$$\phi(x_i) = y_i, i = 1, 2, \dots, N$$

在最小二乘意义下. 即剩余平方和(残差平方和)

$$Q = \|r\|^2 = \sum_{i=1}^N r_i^2 = \sum_{i=1}^N |y_i - \phi(x_i)|^2$$

最小意义下.

- ▶ 类似于插值法, 又有别于插值法. 这里是个矛盾方程组, 只能求最小二乘意义下的解即最小二乘解.
- ▶ 矛盾方程组最小二乘解等价于正规方程组的解

# 拟合直线

## ► 问题

已知一组数据  $(x_i, y_i), i = 1, 2, \dots, N$ , 求拟合直线

$$\begin{aligned} y &= a + bx \\ \text{满足} \quad a + bx_1 &= y_1 \\ a + bx_2 &= y_2 \\ &\dots \end{aligned} \tag{1}$$

$$a + bx_N = y_N$$

在最小二乘意义下. 即剩余平方和(残差平方和)

$$\varphi(a, b) = \sum_{i=1}^N (y_i - (a + bx_i))^2$$

最小之  $a, b$  所确定的直线

# 拟合直线

- ▶ 矛盾方程组(1)可写成矩阵形式 $X\alpha = Y$ 或向量形式 $X_1a + X_2b = Y$

- ▶ 正规方程组:  $X^T X \alpha = X^T Y$

$$0 = \frac{\partial \phi}{\partial a} = -2 \sum_{i=1}^N 1 \times (y_i - (a + bx_i)) \quad Na + \sum x_i b = \sum y_i$$

$$0 = \frac{\partial \phi}{\partial b} = -2 \sum_{i=1}^N x_i \times (y_i - (a + bx_i)) \quad \sum x_i a + \sum x_i^2 b = \sum x_i y_i$$

- ▶ 正规方程组第一个方程三个数据是矛盾方程组第一列与三个列的数对应相乘再求和. 正规方程组第二个方程三个数据是矛盾方程组第二列与三个列的数对应相乘再求和. (可列表进行)
- ▶ 正规方程组总有解:  $\phi \geq 0$ 总取到最小

# 拟合直线

## ▶ 解正规方程组

- 可从第一个方程解出 $a$ ,代入第二个方程解出 $b$ 从而先算出 $b$ 再回代算出 $a$

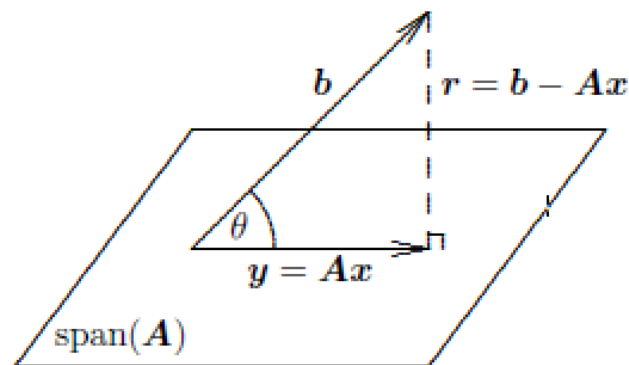
$$a = \frac{1}{N} \sum_{i=1}^N y_i - \frac{1}{N} \sum_{i=1}^N x_i b \quad b = \frac{\sum_{i=1}^N x_i y_i - \frac{1}{N} \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{\sum_{i=1}^N x_i^2 - \frac{1}{N} (\sum_{i=1}^N x_i)^2}$$

# 拟合直线

## 几何解释

最小二乘解是使剩余向量  $\mathbf{r} = \mathbf{Y} - (\mathbf{X}_1 a + \mathbf{X}_2 b)$  长度最小的  $a, b$ .  
几何上  $\mathbf{r}$  最小当且仅当  $\mathbf{r}$  垂直于  $\mathbf{X}_1, \mathbf{X}_2$  构成的平面, 因此有

$$\begin{aligned} \mathbf{X}_1^T (\mathbf{Y} - (\mathbf{X}_1 a + \mathbf{X}_2 b)) &= 0 & \text{即} & & \mathbf{X}_1^T \mathbf{X}_1 a + \mathbf{X}_1^T \mathbf{X}_2 b &= \mathbf{X}_1^T \mathbf{Y} \\ \mathbf{X}_2^T (\mathbf{Y} - (\mathbf{X}_1 a + \mathbf{X}_2 b)) &= 0 & & & \mathbf{X}_2^T \mathbf{X}_1 a + \mathbf{X}_2^T \mathbf{X}_2 b &= \mathbf{X}_2^T \mathbf{Y} \end{aligned}$$



- 正规方程组总有解, 且当  $\mathbf{X}$  列满秩时解唯一



# 拟合直线例

- 已知数据(2,2),(4,11),(6,28),(8,40)求拟合直线解. 用表格计算

$i$	$x_i$	$y_i$	$x_i^2$	$x_i y_i$
1	2	2	4	4
2	4	11	16	44
3	6	28	36	168
4	8	40	64	320
$N=4$ $\Sigma$	20	81	120	536

# 拟合直线例

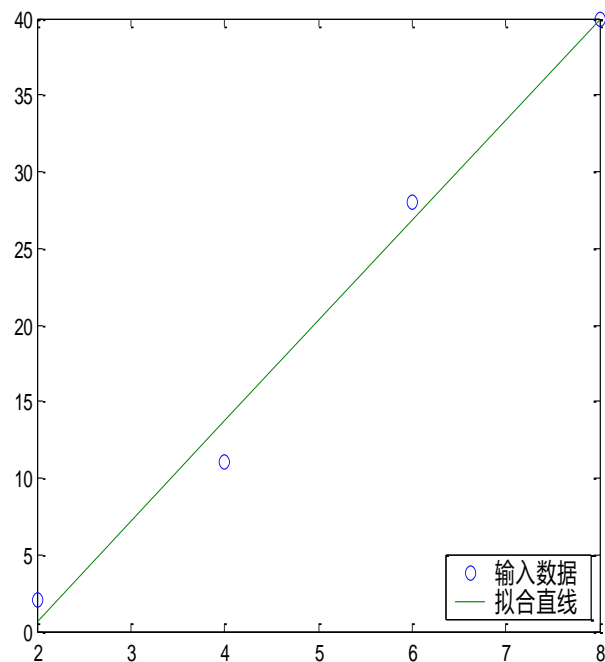
## ▶ 求解线性方程

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 4 & 6 & 8 \end{bmatrix} \begin{bmatrix} 2 \\ 11 \\ 28 \\ 40 \end{bmatrix}$$

## ▶ 推出

$$a = -12.5, \quad b = 6.55$$

## ▶ 在实际中，还可以利用回归分析的方法分析拟合精度，如回归系数等



# 拟合平面

- 已知一组数据  $(x_{1i}, x_{2i}, y_i), i = 1, 2, \dots, N$ , 求拟合平面

满足

$$y = a_0 + a_1x_1 + a_2x_2$$
$$a_0 + a_1x_{11} + a_2x_{21} = y_1$$
$$a_0 + a_1x_{12} + a_2x_{22} = y_2$$

.....

$$a_0 + a_1x_{1N} + a_2x_{2N} = y_N$$

在最小二乘意义下.即

$$\phi(a_0, a_1, a_2) = \sum_{i=1}^N (y_i - (a_0 + a_1x_{1i} + a_2x_{2i}))^2$$

最小之  $a_0, a_1, a_2$  所确定的平面.

# 拟合平面

## ▶ 正规方程组

$$Na_0 + \sum x_{1i}a_1 + \sum x_{2i}a_2 = \sum y_i$$

$$\sum x_{1i}a_0 + \sum x_{1i}^2a_1 + \sum x_{1i}x_{2i}a_2 = \sum x_{1i}y_i$$

$$\sum x_{2i}a_0 + \sum x_{1i}x_{2i}a_1 + \sum x_{2i}^2a_2 = \sum x_{2i}y_i$$

# 拟合平面

- ▶ 正规方程组的数据与矛盾方程组数据的关系跟拟合直线时类似
- ▶ 如果将矛盾方程组写成  $X\alpha = Y$   
则正规方程组可写成  $X^T X \alpha = X^T Y$

# 拟合平面

## ▶ 正规方程组

- 从正规方程组的第一个方程解出 $a_0$ ,代入第二、三个方程得到 $a_1, a_2$ 的二元一次方程组

$$a_0 = \frac{1}{N} \left( \sum_{k=1}^N y_k - a_1 \sum_{k=1}^N x_{1k} - a_2 \sum_{k=1}^N x_{2k} \right)$$

$$l_{11}a_1 + l_{12}a_2 = l_{1y}$$

$$l_{21}a_1 + l_{22}a_2 = l_{2y}$$

$$l_{ij} = \sum_{k=1}^N x_{ik} x_{jk} - \frac{1}{N} \sum_{k=1}^N x_{ik} \sum_{k=i}^N x_{jk}$$

$$l_{iy} = \sum_{k=1}^N x_{ik} y_k - \frac{1}{N} \sum_{k=1}^N x_{ik} \sum_{k=1}^N y_k$$

$$i, j = 1, 2$$

# $p$ 个自变量的线性拟合

## ▶ $p$ 个自变量的线性拟合

已知一组数据  $(x_{1i}, x_{2i}, \dots, x_{pi}, y_i), i = 1, 2, \dots, N$

求

$$y = a_0 + a_1x_1 + \dots + a_px_p$$

满足

$$a_0 + a_1x_{11} + a_2x_{21} + \dots + a_px_{p1} = y_1$$

$$a_0 + a_1x_{12} + a_2x_{22} + \dots + a_px_{p2} = y_2$$

.....

$$a_0 + a_1x_{1N} + a_2x_{2N} + \dots + a_px_{pN} = y_N$$

在最小二乘意义下.即

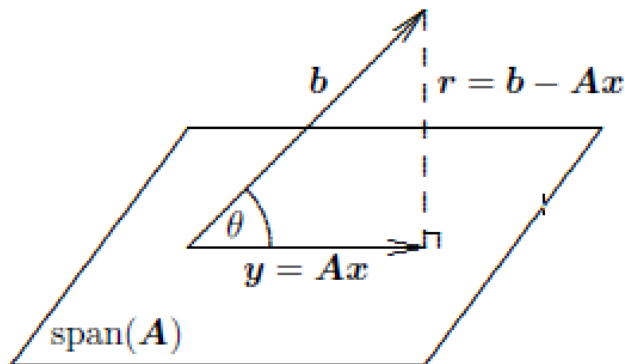
$$\phi(a_0, a_1, \dots, a_p) = \sum_{i=1}^N (y_i - (a_0 + a_1x_{1i} + \dots + a_px_{pi}))^2$$

最小之  $a_0, a_1, a_2, \dots, a_p$

# $p$ 个自变量的线性拟合

## ▶ 正规方程组

- 正规方程组的数据与矛盾方程组数据的关系跟拟合直线时类似
- 还可从正规方程组的第一个方程解出 $a_0$ ,代入第二、三个方程得到 $a_1, a_2, \dots, a_p$ 的 $p$ 元一次方程组
- 如果将矛盾方程组写成  $X\alpha = Y$   
则正规方程组可写成  $X^T X \alpha = X^T Y$





# $p$ 个自变量的线性拟合

- ▶ 求解大型最小二乘矩阵的数值考虑
- ▶ 实际的大型计算中，直接利用正规方程求解会增大条件数，因此一般不推荐使用正规方程求解
- ▶ 可以改变方程，如求解增广方程

$$\begin{bmatrix} I & X \\ X^T & 0 \end{bmatrix} \begin{bmatrix} r \\ \alpha \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

- ▶ 目前最一般的方法是采用对系数矩阵 $X$ 作QR分解求解，其中 $Q$ 是正交矩阵， $R$ 是上三角矩阵。若 $X$ 列满秩，则

$$X = QR \rightarrow \alpha = R^{-1}Q^T b$$

# 非线性曲线拟合

- ▶ 可线性化的非线性曲线拟合
- ▶ 双曲线拟合

双曲线  $\frac{1}{y} = a + \frac{b}{x}$  变换成  $u = a + bv$ . 变量、数据皆变换后作线性拟合

- ▶ 指数曲线拟合

曲线  $y = ae^{bx}$  即  $\log y = \log a + (b \log e)x$  变换成  $u = A + Bv$  可对变换后的变量、数据作线性拟合

# 双曲线拟合例

$i$	$x$	$y$	$x' = \frac{1}{x}$	$y' = \frac{1}{y}$	$x'^2$	$x'y'$
1	2	106.42	0.500 000	0.009 396 73	0.250 000 0	0.004 698 37
2	3	108.20	0.333 333	0.009 242 14	0.111 111 1	0.003 080 71
3	4	109.58	0.250 000	0.009 125 75	0.062 500 0	0.002 281 44
4	5	109.50	0.200 000	0.009 132 42	0.040 000 0	0.001 826 48
5	7	110.00	0.142 857	0.009 090 91	0.020 408 2	0.001 298 70
6	8	109.93	0.125 000	0.009 096 70	0.015 625 0	0.001 137 09
7	10	110.49	0.100 000	0.009 050 59	0.010 000 0	0.000 905 06
8	11	110.59	0.090 909	0.009 042 41	0.008 264 5	0.000 822 04
9	14	110.60	0.071 429	0.009 041 59	0.005 102 0	0.000 645 83
10	15	110.90	0.066 667	0.009 017 13	0.004 444 4	0.000 601 14
11	16	110.76	0.062 500	0.009 028 53	0.003 906 3	0.000 564 28
12	18	111.00	0.055 556	0.009 009 01	0.003 086 4	0.000 500 50
13	19	111.20	0.052 632	0.008 992 81	0.002 770 1	0.000 473 53
$\Sigma$			2.050 883	0.118 266 72	0.537 218 0	0.018 835 17

$$\begin{cases} 13a + 2.050\,883\,b = 0.118\,266\,72 \\ 2.050\,883\,a + 0.537\,218\,0\,b = 0.018\,835\,17 \end{cases}$$

$$y'^* = 0.008\,966 + 0.000\,830\,2\,x'$$

$$y^* = \frac{x}{0.008\,966x + 0.000\,830\,2}$$

# 指数曲线拟合例

$x$	$y$	$u$	$xu$	$x^2$
1	15.3	1.184 7	1.184 7	1
2	20.5	1.311 8	2.623 6	4
3	27.4	1.437 8	4.313 4	9
4	36.6	1.563 5	6.254 0	16
5	49.1	1.691 1	8.455 5	25
6	65.6	1.816 9	10.901 4	36
7	87.8	1.943 5	13.604 5	49
8	117.6	2.070 4	16.563 2	64
$\Sigma$ 36		13.019 7	63.900 3	204

$$\begin{cases} 8a + 36b = 13.019\ 7 \\ 36a + 204b = 63.900\ 3 \end{cases}$$

$$a = 11.44, \quad b = 0.291\ 3$$

$$y = 11.44e^{0.291\ 3x}$$

# 多项式拟合

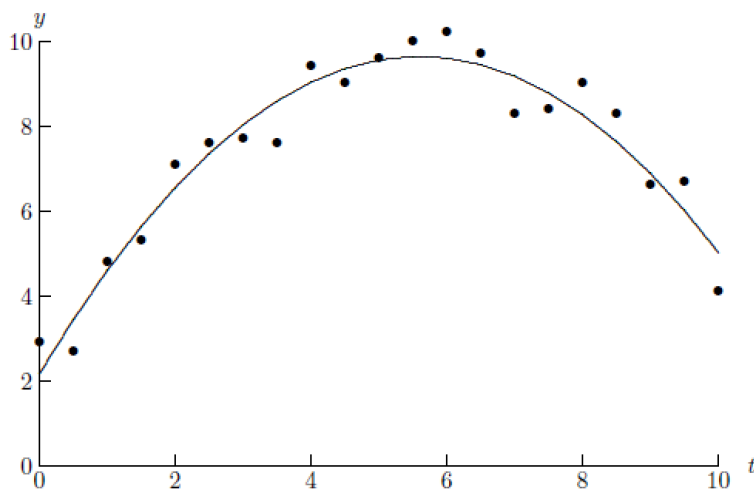
- ▶ 多项式拟合

已知一组数据 $(x_i, y_i), i = 1, 2, \dots, N$ , 求拟合多项式

$$y = a_0 + a_1x + a_2x^2$$

可化成拟合平面去做

- ▶ 对 $p$ 次多项式拟合可化成 $p$ 个自变量的线性拟合去, 亦可由剩余平方和(残差平方和)最小确定



# 多项式拟合

- ▶ 用多项式  $y = a_0 + a_1x + \cdots + a_nx^n$  作数据  $\{(x_i, y_i)\}_{i=1}^m$  拟合，其中  $n \gg 1$  时，正规方程的系数矩阵接近奇异（病态矩阵）
- ▶ 比如，取  $x_i = \frac{i}{m}, i = 1, 2, \dots, m$ ，当  $m > n$  较大时，正规方程的系数矩阵近似为

$$m \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n+1} \\ 1 & \frac{1}{3} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{1}{m} & \cdots & \frac{1}{2n+1} \end{bmatrix}$$

- ▶ 该矩阵是  $n+1$  阶 Hilbert 矩阵，当  $n$  较大时，矩阵非常病态

# 多项式拟合

- 一般的, 数据 $\{(x_i, y_i)\}_{i=1}^m$ 拟合时, 可选取多项式族 $\{P_k(x)\}, k = 0, \dots, n$ , 使得拟合多项式为

$$y = a_0P_0(x) + a_1P_1(x) + \dots + a_nP_n(x) = \sum_{k=0}^n a_kP_k(x)$$

其中 $P_k(x)$ 为 $k$ 次多项式

- 按最小二乘原理, 我们需要最小化

$$r = \sum_{i=1}^m w_i \left( y_i - \sum_{k=0}^n a_k P_k(x_i) \right)^2$$

其中 $w_i$ 为权重

# 多项式拟合

- ▶ 正规方程

$$\sum_{k=0}^n c_{jk} a_k - c_j = 0$$

- ▶ 其中

$$c_{jk} = \sum_{i=1}^m \omega_i P_k(x_i) P_j(x_i)$$
$$c_j = \sum_{i=1}^m \omega_i y_i P_j(x_i)$$



# 加权正交多项式

- 关于节点的加权正交多项式拟合:

$$c_{jk} = \sum_{i=1}^N \omega_i P_k(x_i) P_j(x_i) = 0, k \neq j$$

$$c_{jj} = \sum_{i=1}^N \omega_i P_j^2(x_i) = \sigma_j > 0$$

数据 $(x_i, y_i), i = 1, 2, \dots, N$ 拟合

$$y = a_0 P_0(x) + a_1 P_1(x) + \dots + a_m P_m(x)$$

则有

$$a_j = \sum \frac{\omega_i P_j(x_i) y_i}{c_{jj}}, j = 0, 1, 2, \dots, m$$

# 正交多项式

- ▶ 正交多项式

多项式 $P_k(x)$ , 次数为 $k$ ,  $k = 0, 1, 2, \dots$

$$(P_i, P_j) = \int_a^b P_i(x)P_j(x)dx = 0, \quad i \neq j$$

$$(P_i, P_i) = \int_a^b P_i^2(x)dx = \sigma_i > 0$$

称为 $[a, b]$ 上的正交多项式。

- ▶ 类似于向量空间的内积，函数空间也可以定义内积

# Legendre多项式

## ▶ Legendre多项式

$$P_n(x) = \frac{1}{2^n} \frac{d^n}{dx^n} (x^2 - 1)^n, n = 0, 1, 2, \dots$$

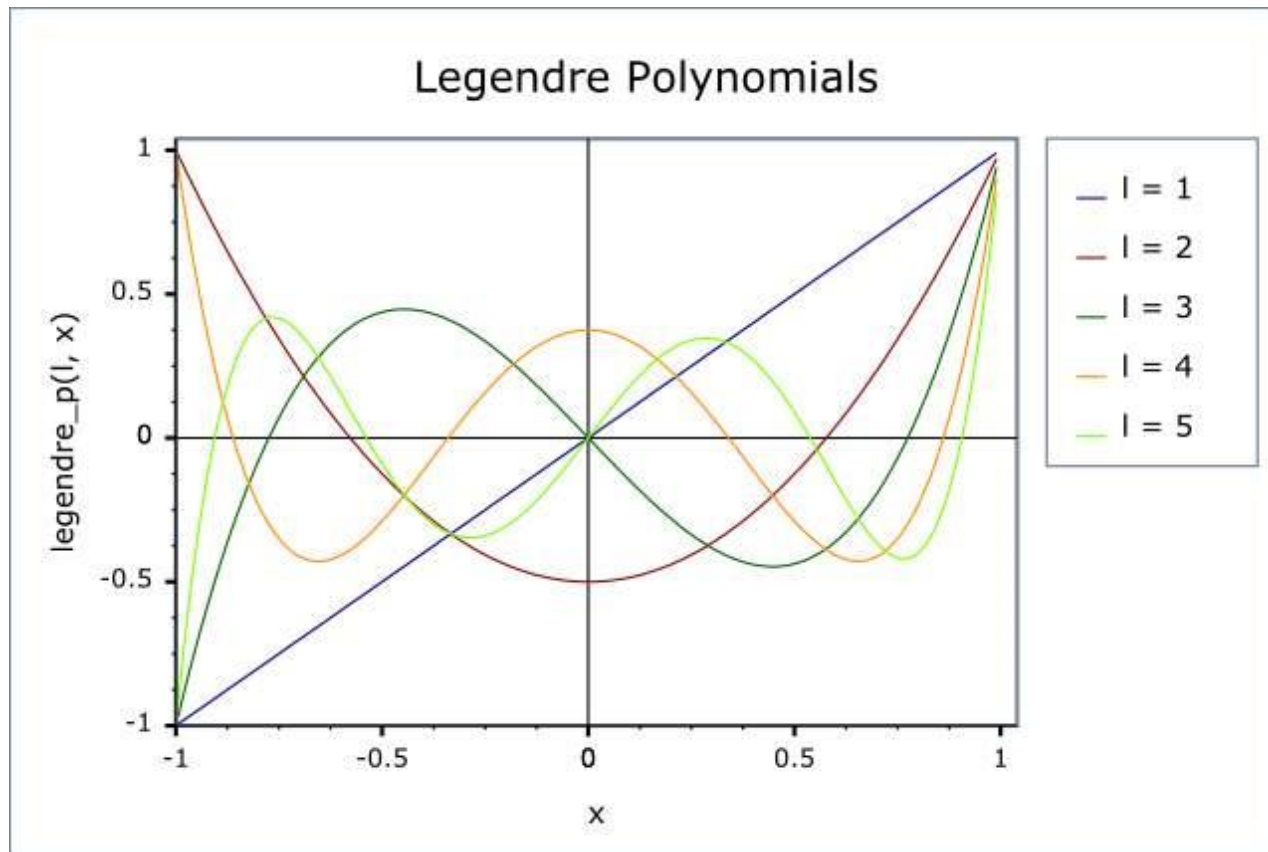
1.  $P_n(x)$ 是 $[-1, 1]$ 上正交多项式

$$(P_i, P_j) = \int_{-1}^1 P_i(x) P_j(x) dx = 0, i \neq j$$

$$(P_i, P_i) = \int_{-1}^1 P_i^2(x) dx = 2/(n+2) > 0$$

2.  $(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x),$   
 $n = 1, 2, \dots, P_0(x) = 1, P_1(x) = x$

# Legendre多项式



# 加权正交多项式

## ▶ 加权正交多项式

多项式 $P_k(x)$ , 次数为 $k$ ,  $k = 0, 1, 2, \dots$

$$(P_i, P_j) = \int_a^b \omega(x) P_i(x) P_j(x) dx = 0, \quad i \neq j$$

$$(P_i, P_i) = \int_a^b \omega(x) P_i^2(x) dx = \sigma_i > 0$$

称为 $[a, b]$ 上关于权函数 $\omega(x)$ 的正交多项式

# Chebyshev多项式

- ▶ Chebyshev多项式（第一类）

$$T_n(x) = \cos(n \arccos(x)) \quad -1 \leq x \leq 1, n = 0, 1, 2, \dots$$

1.  $T_n(x)$ 在 $[-1, 1]$ 上关于权函数 $(1 - x^2)^{-\frac{1}{2}}$ 正交

$$(T_i, T_j) = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} T_i(x) T_j(x) dx = 0, \quad i \neq j$$

$$(T_i, T_i) = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} T_i^2(x) dx = \begin{cases} \pi, & i = 0 \\ \pi/2, & i \neq 0 \end{cases}$$

2.  $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),$   
 $n = 1, 2, \dots, T_0(x) = 1, T_1(x) = x$