数值计算方法：原理、算法和应用
Numerical Methods: Principles, Algorithms and Applications

授课教师：周铁

北京大学数学科学学院

2021 年 11 月 9 日

1. 最小二乘问题
   - 最小二乘函数逼近 (least squares function approximation)
   - 最小二乘问题的法方程组 (normal equation)
   - QR 分解法
   - 矩阵的广义逆 (pseudo-inverse)
   - 矩阵的奇异值分解 (SVD)
   - 再谈广义逆
   - 亏秩最小二乘问题
   - 非线性最小二乘问题
   - 函数最佳逼近

最小二乘函数逼近

一个未知函数 $y(t)$，通过测量知道其在若干点处的近似值 (数据)

$$y_i \approx y(t_i), \quad i = 1, 2, \cdots, m.$$

希望根据这些数据得到 $y(t)$ 的一个近似表达式.
插值：取一组基函数

$$\phi_1(t), \phi_2(t), \phi_3(t), \cdots, \phi_n(t),$$

利用数据构造插值函数 $\varphi(t) \in \text{span}\{\phi_j(t)\}$ 使得

$$\varphi(t_i) = \beta_1\phi_1(t_i) + \beta_2\phi_2(t_i) + \cdots + \beta_n\phi_n(t_i) = y_i, \quad i = 1, 2, \cdots, m. \quad (1.1)$$

由于测量得到的数据有误差，插值方法的出发点(1.1)式就有问题！如果 $y_i \neq y(t_i)$，就没必要一定满足插值条件.
但还是希望利用基函数 $\phi_j(t)$ 的线性组合近似函数 $y(t)$，即

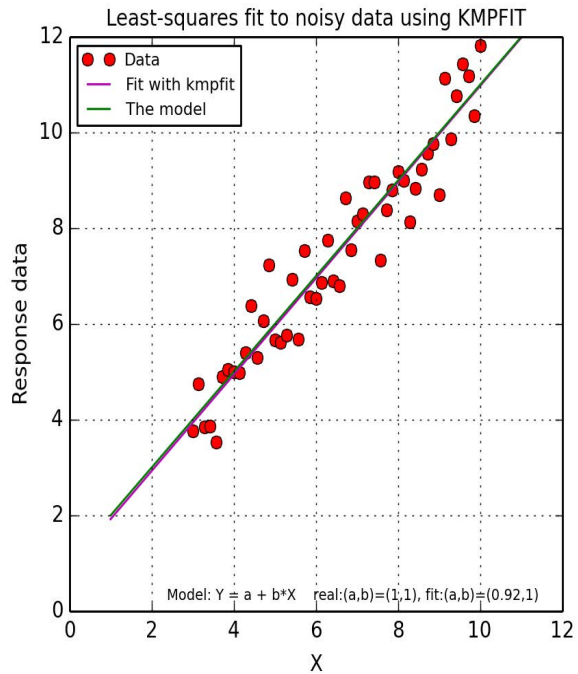
$$y(t) \approx \beta_1\phi_1(t) + \beta_2\phi_2(t) + \cdots + \beta_n\phi_n(t).$$

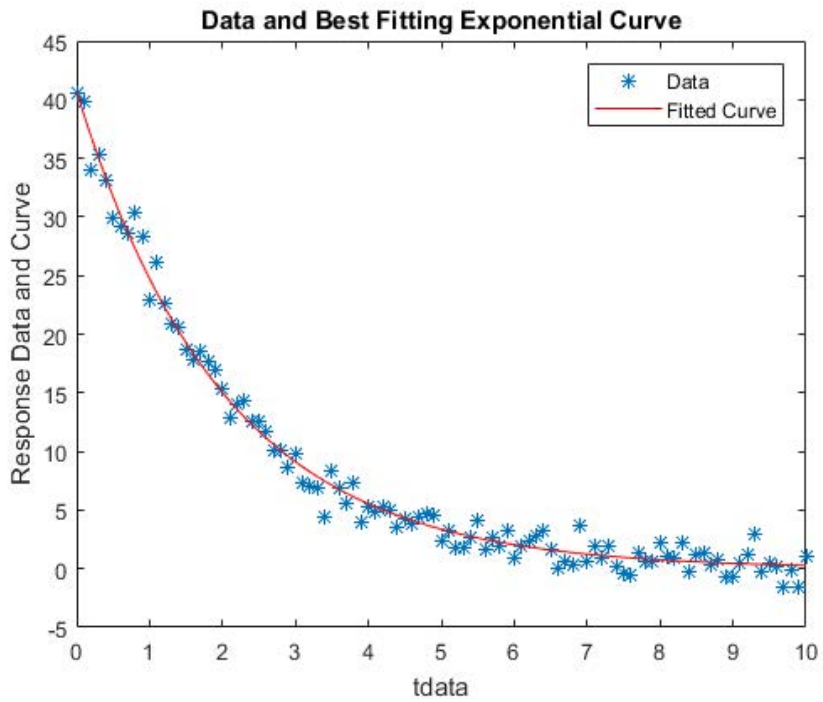
问题：基函数 $\{\phi_j(t)\}$ 和系数 $\{\beta_j\}$ 怎么确定?

基函数的选取

基函数 $\{\phi_j(t)\}$ 的选取—模型选择 (model selection).

1. 直线：$\phi_1(t) = t,\ \phi_2(t) = 1,\ y(t) \approx \beta_1 t + \beta_2$.
2. 多项式：$\phi_1(t) = t^{n-1}, \cdots, \phi_n(t) = 1,\ \ y(t) \approx \sum_{j=1}^{n} \beta_j \phi_j(t)$. $\mathbb{P}_{n-1}$ 的基函数不一定这样取，还有更好的取法.
3. 指数函数：$y(t) \approx Ke^{\lambda t},\ \log y(t) \approx \beta_1 t + \beta_2,\ \beta_1 = \lambda, \beta_2 = \log K$.
4. 三角函数...





系数 $\{\beta_j\}$ 如何确定?

为把插值条件写成向量形式，定义矩阵:

$$X = (x_{ij})_{m\times n}, \quad \text{其中 } x_{ij} = \phi_j(t_i),$$

和列向量:

$$\beta = (\beta_1, \beta_2, \cdots, \beta_n)^T, \quad y = (y_1, y_2, \cdots, y_m)^T,$$

插值条件(1.1)就是要求 β 是线性方程组:

$$X\beta = y \tag{1.2}$$

的解. 如果再引进残差向量:

$$r(\beta) \triangleq X\beta - y,$$

由插值条件确定的 β 等价于要求残差向量 $r(\beta) = 0$.

当数据个数多于基函数个数 $(m > n)$ 时，线性方程组(1.2)是一个超定方程组，很可能无解. 应该抛弃按这种方法确定 β 的思路.

拟合 (fitting) 的思想：不要求残差向量 r 等于零，而是通过解优化问题

$$\min_{\beta\in\mathbb{R}^n} \|r(\beta)\| = \|X\beta - y\|, \quad \text{其中} \|\cdot\| \text{可以取各种范数,}$$

确定 β.

如果范数取为 2-范数，并且平方，就是最小二乘法：

$$\min_{\beta\in\mathbb{R}^n} \|X\beta - y\|_2^2 \tag{1.3}$$

C. F. Gauss 在 1809 年发表的著作中，A-M Legendre 在 1805 年发表的论文中，独立提出了以上最小二乘数据拟合方法.
Gauss 在 1801 年用这种方法成功预测了谷神星 (planetoid Ceres) 的运行轨迹.

如果记二次优化问题(1.3)的解为 β_*，对应的拟合函数为 $\varphi^*(t)$，则有

$$\sum_{i=1}^{m}[\varphi^*(t_i)-y_i]^2 \le \sum_{i=1}^{m}[\varphi(t_i)-y_i]^2,\ \forall\,\varphi(t)\in \mathrm{span}\{\phi_1(t),\cdots,\phi_n(t)\},$$

这就是“最小二乘 (least squares)”名称的由来.

优化问题(1.3)的解一定存在吗？它好解吗？定义 n 元函数 $\Phi(\beta)$ 为

$$\begin{aligned}\Phi(\beta_1,\cdots,\beta_n) &= \|X\beta-y\|_2^2\\ &= \sum_{i=1}^{m}[\beta_1\phi_1(t_i)+\cdots+\beta_n\phi_n(t_i)-y_i]^2\end{aligned}$$

这个 n 元二次函数一定有最小值，且最小值点是平衡点，即 β_* 一定是方程组

$$\nabla\Phi(\beta)=0$$

的解. 记:

$$f(t)=\Phi(\beta_*+tv)=\|X(\beta_*+tv)-y\|_2^2,\quad t\in\mathbb{R},\ v\in\mathbb{R}^n, v\neq 0.$$

法方程组 (normal equations)

一元函数 $f(t)$ 在 $t=0$ 处的差商为

$$\frac{f(t)-f(0)}{t}=2(X^T(X\beta_*-y),v)+t(X^TXv,v).$$

由于 $t=0$ 为极小值点，必为平衡点，令 $t\to 0$ 可得，

$$0=\lim_{t\to 0}\frac{f(t)-f(0)}{t}=2(X^T(X\beta_*-y),v),\quad \forall\, v\in\mathbb{R}^n, v\neq 0$$

由 v 的任意性，必有

$$X^T(X\beta_*-y)=0$$

所以只要解 $n\times n$ 线性方程组

$$(X^TX)\beta=X^Ty \tag{2.1}$$

就可得到最小二乘问题的解 β_*.
线性方程组(2.1)称为最小二乘问题的 "**法方程**"，它在实数域一定有解!

法方程组有解的证明：

$$\mathrm{rank}(X^TX) \le \mathrm{rank}(X^TX, X^Ty) \le \mathrm{rank}(X) = \mathrm{rank}(X^TX)$$

如果 X 列满秩，则解存在且惟一，且有

$$\beta_* = (X^TX)^{-1}X^Ty.$$

□

当 X^TX 列满秩时，法方程组(2.1)的系数矩阵为实对称正定阵，可以用 Cholesky 分解方法求解. 其计算量为 $n^2m + \frac{1}{3}n^3 + \mathcal{O}(n^2)$.

1. 计算 $B = X^TX$ 和 $c = X^Ty$;
2. 计算 B 的 Cholesky 分解 $B = LL^T$;
3. 解下三角线性方程组 $Lz = c$;
4. 解上三角线性方程组 $L^T\beta = z$.

但如果问题病态，即 X 的列向量组几乎线性相关时，X^TX 会把数据中的很多信息丢失，这种解法的可靠性堪忧.

```
>> A = [1  1; 5*eps  0; 0  5*eps]
A =
 1.000000000000000   1.000000000000000
 0.000000000000001   0
 0                   0.000000000000001
>> rank(A)
>> B = A.'*A
B =
 1        1
 1        1
>> rank(B)

>> C = sym(A)
>> D = C.'*C
>> rank(D)
```

常用的最小二乘问题数值算法有：

- Cholesky 分解方法
- QR 分解法
- 奇异值分解法

QR 分解法

用 Cholesky 算法解法方程组(2.1)，需要先计算 $B = X^T X$，当 X 为病态矩阵的时候，B 会比 X 更病态.

如果列满秩矩阵 $X \in \mathbb{R}^{m,n}(m > n)$ 具有上三角形式：

$$X = \begin{bmatrix} R \\ O \end{bmatrix}, \quad R \in \mathbb{R}^{n,n} \text{ 是满秩上三角阵；} \quad O \in \mathbb{R}^{m-n,n} \text{ 是零矩阵；}$$

把 $y \in \mathbb{R}^m$ 分裂为 $y = (\tilde{y}^T, \bar{y}^T)^T$，其中 $\tilde{y} \in \mathbb{R}^n$, $\bar{y} \in \mathbb{R}^{m-n}$，则有

$$\|X\beta - y\|_2^2 = \left\| \begin{bmatrix} R\beta \\ O\beta \end{bmatrix} - \begin{bmatrix} \tilde{y} \\ \bar{y} \end{bmatrix} \right\|_2^2 = \|R\beta - \tilde{y}\|_2^2 + \|\bar{y}\|_2^2.$$

所以只要解 $n \times n$ 上三角线性方程组

$$R\beta = \tilde{y}$$

就可以得到最小二乘问题的解.
如果 X 不是上三角形怎么办？初等行变换？

任何列满秩的 $m \times n (m \geq n)$ 矩阵 X 一定有如下 QR 分解：

$$X = Q_{m\times m} \begin{bmatrix} R \\ O \end{bmatrix}_{m\times n} = [Q_1, Q_2]_{m\times m} \begin{bmatrix} R \\ O \end{bmatrix}_{m\times n} \tag{3.1}$$

其中 $Q = [Q_1, Q_2]$ 为 m 阶正交阵，R 是 n 阶上三角阵，对角元素都是正数. 由于正交矩阵左乘一个向量不改变这个向量的 2 范数，所以有

$$\begin{aligned} \|X\beta - y\|_2^2 &= \left\| Q \begin{bmatrix} R \\ O \end{bmatrix} \beta - QQ^T y \right\|_2^2 \\ &= \left\| \begin{bmatrix} R \\ O \end{bmatrix} \beta - Q^T y \right\|_2^2 \\ &= \|R\beta - \tilde{y}\|_2^2 + \|\bar{y}\|_2^2 \end{aligned}$$

这里的 $\tilde{y} \in \mathbb{R}^n$, $\bar{y} \in \mathbb{R}^{m-n}$ 为 $Q^T y$ 的分裂

$$Q^T y = \begin{bmatrix} \tilde{y} \\ \bar{y} \end{bmatrix}$$

为什么可以做 QR 分解(3.1)?

设 $X \in \mathbb{R}^{m,n}$, $\mathrm{rank}(X) = n < m$.
正定阵 X^TX 有 Cholesky 分解：$X^TX = R^TR$，其中 R 为 n 阶非奇异上三角阵，对角元素大于零. 于是有 $(R^{-1})^TX^TXR^{-1} = I_n$，即

$$(XR^{-1})^T(XR^{-1}) = I_n.$$

这表明 $Q_1 = XR^{-1} \in \mathbb{R}^{m,n}$ 的列向量组是标准正交的. 并且

$$X = Q_1R, \quad Q_1 \in \mathbb{R}^{m,n},\ R \in \mathbb{R}^{n,n}.$$

再把 Q_1 扩充成 m 阶正交矩阵 $Q = [Q_1, Q_2]$，就有

$$X = [Q_1, Q_2] \begin{bmatrix} R \\ O \end{bmatrix}.$$

这就是 QR 分解.

```
1 >> [Q,R] = qr(A)
```

最小二乘问题

$$\min_{\beta\in\mathbb{R}^n} \|X\beta - y\|_2^2, \quad X \in \mathbb{R}^{m,n} \text{列满秩}, \quad y \in \mathbb{R}^m$$

的 QR 分解算法:

1. 计算 QR 分解 $X = QR$，其中 $Q \in \mathbb{R}^{m,n}$ 列向量标准正交，$R \in \mathbb{R}^{n,n}$ 是上三角矩阵，对角元素都大于零;
2. 计算 $b = Q^T y$;
3. 解上三角线性方程组 $R\beta = b$.

关键是如何计算 QR 分解?
矩阵的 QR 分解计算很像 LU 分解.
LU 分解是用初等行变换把矩阵化为上三角形;
QR 分解是用正交变换把矩阵化为上三角形.

QR 分解的方法有:

1. Gram-Schmidt 正交化
2. Givens 变换
3. Householder 变换

Householder 变换

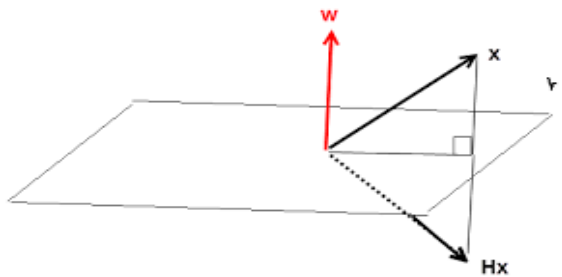
实 Householder 变换 (Householder 矩阵) 是指如下形式的实方阵:

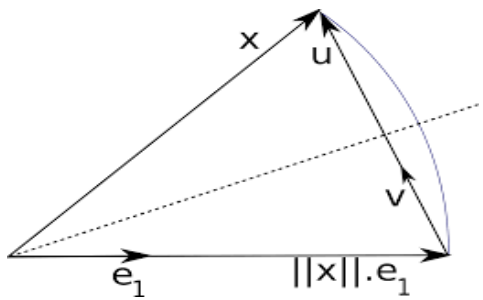
$$H = I - \frac{2ww^T}{(w, w)}, \quad \text{其中 } w \in \mathbb{R}^n \text{ 为非零列向量.}$$

Householder 变换也叫“初等反射矩阵”或“镜像变换”,
它是 A.S. Householder 在 1958 年提出的.

Householder 变换有如下性质:

- 对称性: $H^T = H$, 正交性: $H^T H = I$, 对合性: $H^2 = I$.
- 反射性: 任意向量 x, Hx 是 x 关于 w 的垂直超平面的镜像反射.





Householder 变换的主要用途：给定一个非零向量，总可以找到一个 Householder 矩阵，把这个向量的若干分量变为零.

给定 $x \in \mathbb{R}^n$ 非零，记 $e_1 = (1, 0, \cdots, 0)^T$，取

$$w = x - \alpha e_1, \quad \text{其中标量} \, \alpha = \begin{cases} \|x\|_2, & x_1 < 0, \\ -\|x\|_2, & x_1 \geq 0. \end{cases}$$

则有

$$Hx = \alpha e_1 .$$

这也就是说，对任意的非零向量 $x \in \mathbb{R}^n$ 都可构造出一个 Householder 变换 H，使 Hx 的后 $n-1$ 个分量为零.

注：α 的符号选取是为了避免相近数相减.

$Hx = \alpha e_1$ 的推导：

$$
\begin{aligned}
Hx &= x - 2\frac{(w,x)}{(w,w)}w \\
&= x - 2\frac{(x-\alpha e_1, x)}{(x-\alpha e_1, x-\alpha e_1)}(x-\alpha e_1) \\
&= x - 2\frac{(x-\alpha e_1, x)}{(x-\alpha e_1, x-\alpha e_1)}(x-\alpha e_1) \\
&= x - 2\frac{\|x\|^2 - \alpha x_1}{\|x\|^2 - 2\alpha x_1 + \alpha^2}(x-\alpha e_1) \\
&= x - 2\frac{\|x\|^2 - \alpha x_1}{2(\|x\|^2 - \alpha x_1)}(x-\alpha e_1) \\
&= x - x + \alpha e_1 \\
&= \alpha e_1
\end{aligned}
$$

用 Householder 变换实现 QR 分解

用 Householder 变换计算 QR 分解与不选主元的 Gauss 消去法很类似，就是利用 Householder 变换逐步将 X 约化为上三角矩阵.

设 $m = 5, n = 4$，并假定已经计算出 Householder 变换 H_1 和 H_2 使得

$$H_2 H_1 X = \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & + & * \\ 0 & 0 & + & * \\ 0 & 0 & + & * \end{bmatrix}.$$

现在看第三列标为"+" 的 3 个元素，首先按照前面介绍的方法确定 Householder 变换 $\tilde{H}_3 \in \mathbb{R}^{3,3}$ 使得

$$\tilde{H}_3 \begin{bmatrix} + \\ + \\ + \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ 0 \end{bmatrix}.$$

然后令

$$H_3 = \begin{bmatrix} I_2 & O \\ \tilde{O} & \tilde{H}_3 \end{bmatrix} \in \mathbb{R}^{5,5},\ \ I_2 \text{ 为 2 阶单位阵},\ O,\ \tilde{O} \text{ 为零矩阵},$$

则有

$$H_3 H_2 H_1 X = \begin{bmatrix} I_2 & O \\ \tilde{O} & \tilde{H}_3 \end{bmatrix} \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & + & * \\ 0 & 0 & + & * \\ 0 & 0 & + & * \end{bmatrix} = \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & * \end{bmatrix}.$$

对于一般的 $X \in \mathbb{R}^{m,n}(m \geq n)$，经过 n 步上述过程可以得到 n 个 m 阶 Householder 矩阵 $H_1, H_2, \cdots, H_n$，使得

$$H_n H_{n-1} \cdots H_1 X = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad R \in \mathbb{R}^{n,n} \text{ 为上三角阵}.$$

于是有

$$X = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$$

列满秩矩阵的广义逆

当 X 列满秩时，法方程 $X^TX\beta = X^Ty$ 的解可以写为

$$\beta = (X^TX)^{-1}X^Ty.$$

记

$$X^\dagger = (X^TX)^{-1}X^T, \quad X^\dagger \text{ 称为 } X \text{ 的}\textbf{广义逆矩阵}.$$

如果 X 是可逆方阵，显然 $X^\dagger = X^{-1}$. 如果 X 是 $m \times n (m > n)$ 矩阵，并且列满秩，则 $X^\dagger$ 只是“左逆”，即

$$X^\dagger X = (X^TX)^{-1}X^TX = I_n.$$

虽然不是右逆，但 $X^\dagger$ 满足：

$$\|XX^\dagger - I_m\|_F = \min_{Z\in\mathbb{R}^{n,m}} \|XZ - I_m\|_F, \text{ 其中 } \|X\|_F = (\Sigma_{i,j}\, x_{ij}^2)^{1/2}.$$

同时 $X^\dagger$ 还是这个优化问题的“最小解” $(\|X^\dagger\|_F = \min)$.

矩阵的奇异值分解 (SVD)

定理 (奇异值分解 (Singular Value Decomposition))

设 $A \in \mathbb{R}^{m,n}$. 则必存在正交阵 $U \in \mathbb{R}^{m,m}$ 和 $V \in \mathbb{R}^{n,n}$，以及对角阵 $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbb{R}^{m,n}$，其中 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$，$p = \min(m, n)$，使得

$$A = U\Sigma V^T \iff U^T A V = \Sigma.$$

U 的列向量 $u_1, \ldots, u_m$ 称为 A 的左奇异向量，V 的列向量 $v_1, \ldots, v_n$ 称为 A 的右奇异向量，σ_i 中大于零的称为 A 的奇异值，奇异值的个数等于 A 的秩.

$$A = U\Sigma V^T \iff AV = U\Sigma, \quad A^T U = V\Sigma^T.$$

记 $r = \mathrm{rank}(A)$, $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r,r}, \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$

$$Av_i = \sigma_i u_i, \ i = 1, \cdots, r, \quad Av_j = 0, \ i = r+1, \cdots, n$$

$$A^T u_i = \sigma_i v_i, \ i = 1, \cdots, r, \quad A^T u_j = 0, \ i = r+1, \cdots, m$$

SVD 的分块矩阵形式：

$$A = [U_1, U_2] \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} [V_1, V_2]^T = U_1 \Sigma_r V_1^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$

其中 $v_1, \ldots, v_r$ 是 A 的行空间的标准正交基，$u_1, \ldots, u_r$ 是 A 的列空间的标准正交基.
不难看出

$$A^T A = \left(U \Sigma V^T\right)^T U \Sigma V^T = V(\Sigma^T \Sigma) V^T, \ \Sigma^T \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2).$$

即 $\sigma_1^2, \ldots, \sigma_n^2$ 是 n 阶实对称矩阵 $A^T A$ 的特征值，v_i 是对应的特征向量.
类似地，

$$AA^T = U \Sigma V^T \left(U \Sigma V^T\right)^T = U(\Sigma \Sigma^T) U^T = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2).$$

即 $\sigma_1^2, \ldots, \sigma_m^2$ 是 m 阶对称矩阵 AA^T 的特征值，u_i 是对应的特征向量.

```
>> s = svd(A)
>> [U,S,V] = svd(A)
>> [U,S,V] = svd(A,'econ')
```

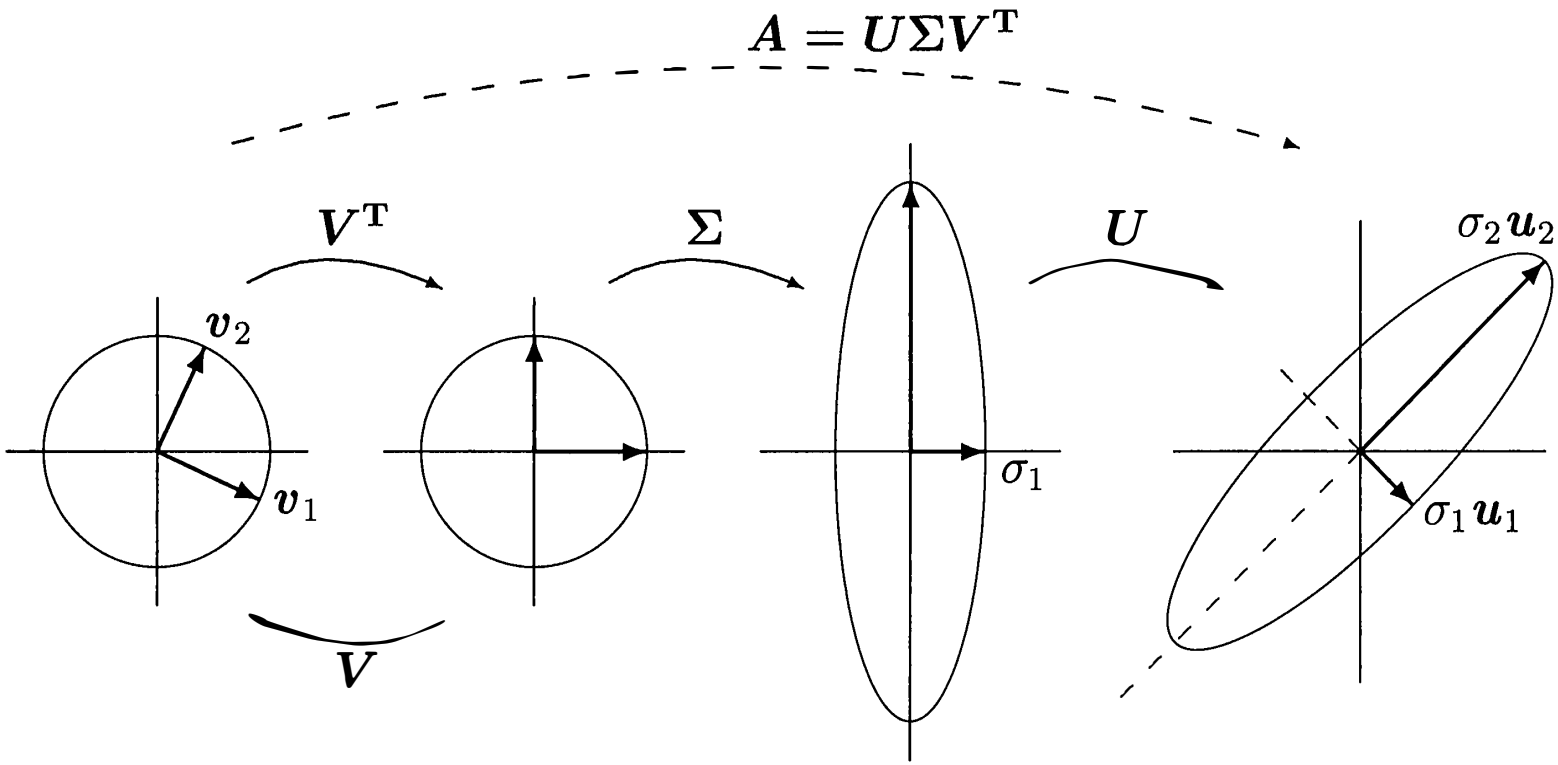


图: 二维奇异值分解的几何意义

用奇异值分解进行图像压缩

```
>> load clown.mat;
>> figure
>> image(X); colormap('gray')
>> k = 3;[U,S,V]=svd(X); figure
>> image(U(:,1:k)*S(1:k,1:k)*V(:,1:k)'); colormap('gray')
```

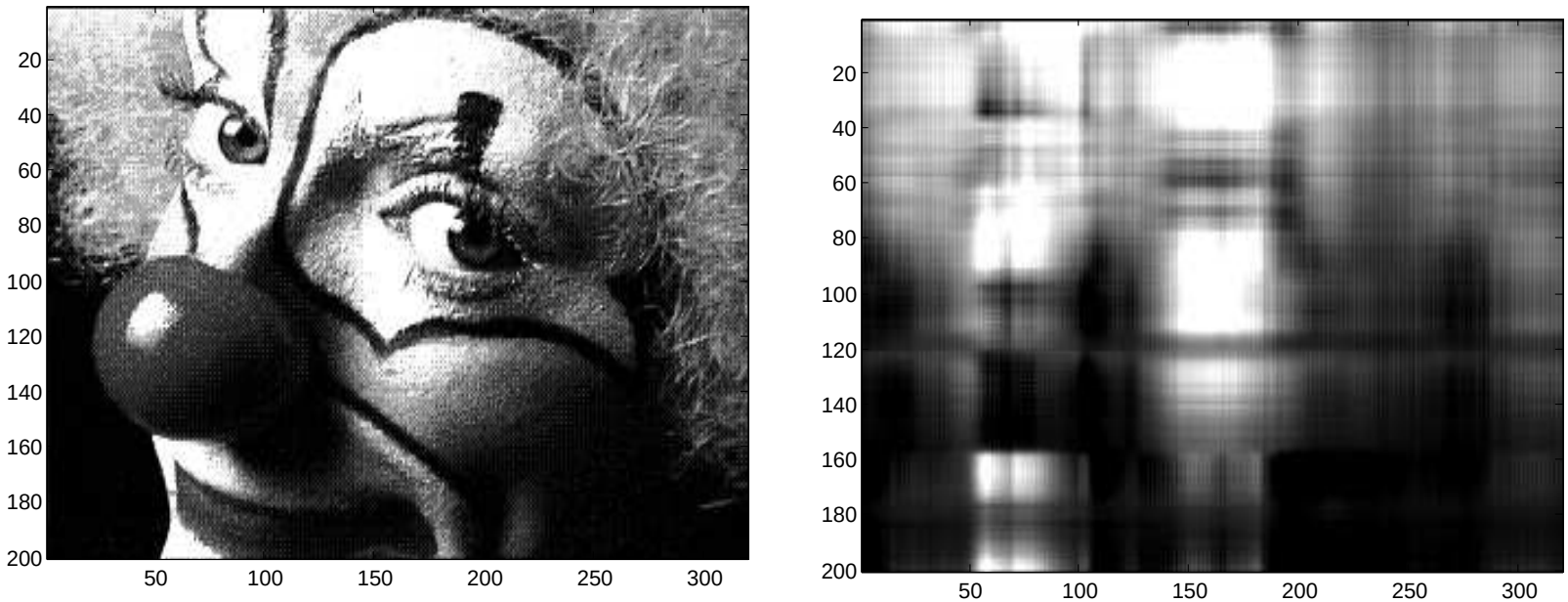


图: 原始图像与 3 个奇异值

```
>> k = 10;
>> figure
>> colormap('gray')
>> image(U(:,1:k)*S(1:k,1:k)*V(:,1:k)')
>> title('k=10')
```

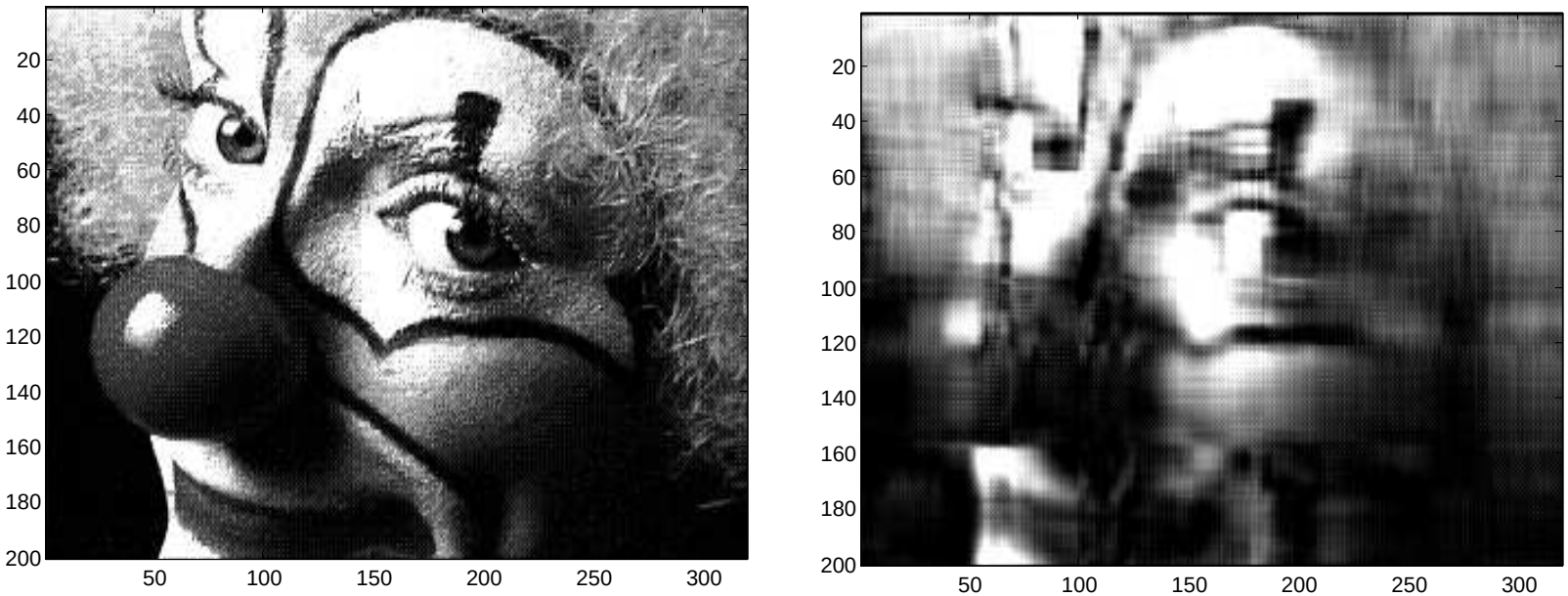


图: 原始图像与 10 个奇异值

```
>> k = 20;
>> figure
>> image(U(:,1:k)*S(1:k,1:k)*V(:,1:k)')
>> colormap('gray');
>> title('k=20')
```

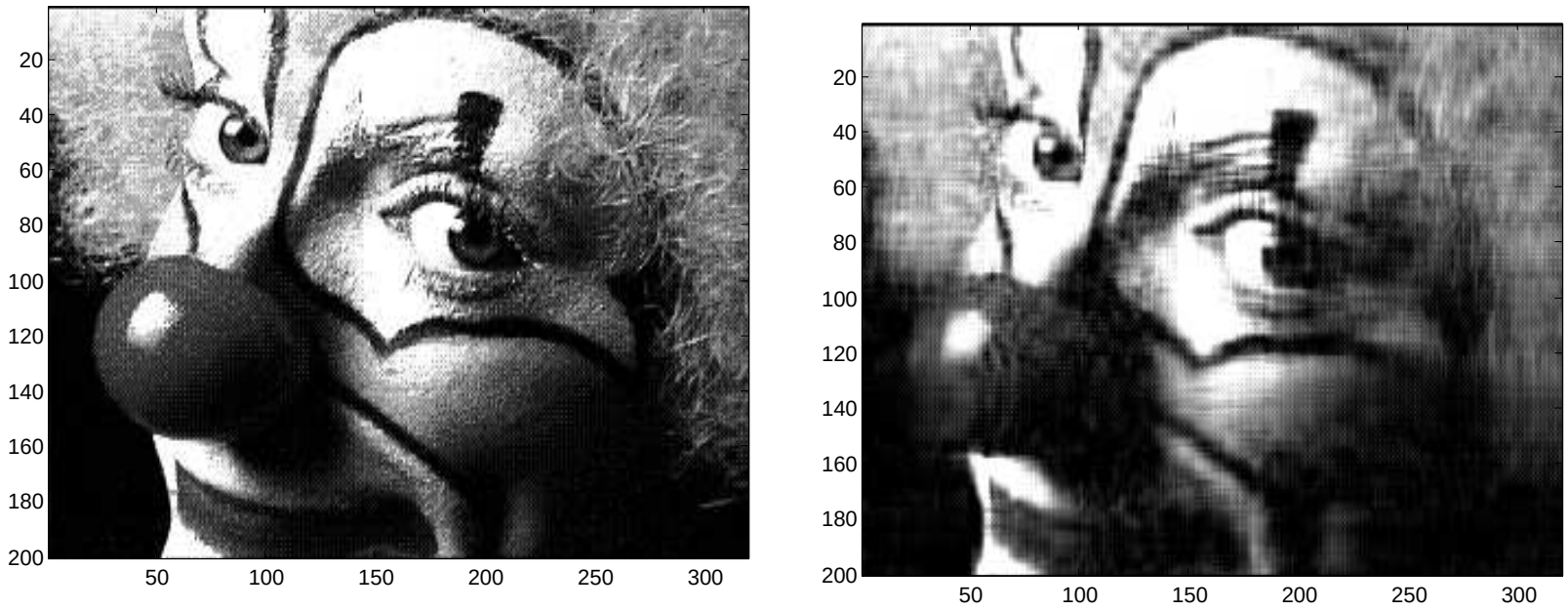


图: 原始图像与 20 个奇异值

```
>> k = 40;
>> figure
>> image(U(:,1:k)*S(1:k,1:k)*V(:,1:k)')
>> colormap('gray')
>> title('k=40')
```

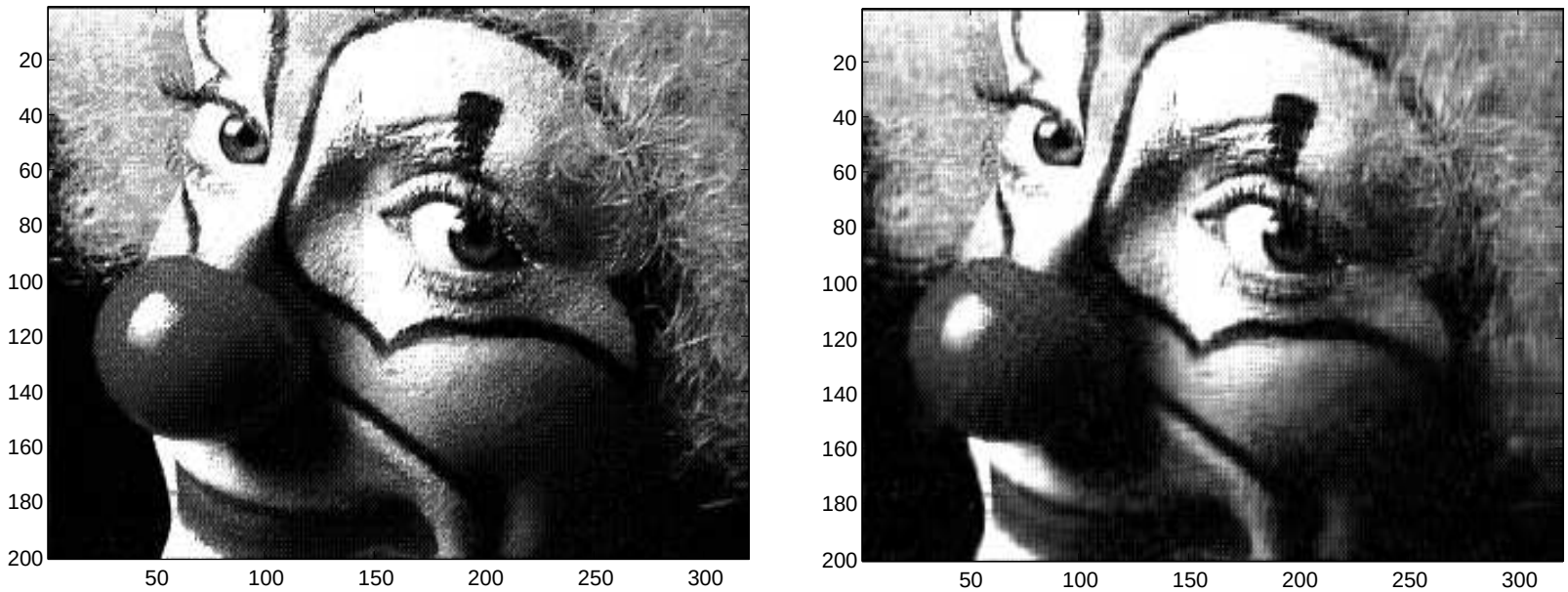


图: 原始图像与 40 个奇异值

亏秩矩阵的广义逆

设 $A \in \mathbb{R}^{m,n}(m \geq n), \mathrm{rank}(A) = r < n.$ A 的奇异值分解为 $U\Sigma V^T$，其中

$$\Sigma = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m,n}, \quad \Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r,r}$$

$\sigma_1 \geq \cdots \geq \sigma_r > 0$ 为 A 的奇异值.

我们称矩阵 $A^\dagger = V\Sigma^\dagger U^T \in \mathbb{R}^{n,m}$ 为 A 的广义逆，其中

$$\Sigma^\dagger = \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n,m}.$$

可以证明:

$$AA^\dagger A = A, \ A^\dagger AA^\dagger = A^\dagger, \ (AA^\dagger)^T = AA^\dagger, \ (A^\dagger A)^T = A^\dagger A. \qquad (6.1)$$

这些关系式后面经常用到!

亏秩最小二乘问题

当 $\mathrm{rank}(X_{m\times n}) = r < n$ 时，最小二乘问题 $\min\limits_{\beta\in\mathbb{R}^n} \|X\beta - y\|_2^2$ 的法方程

$$X^T X\beta = X^T y$$

的解存在但不惟一. 利用亏秩矩阵 X 的广义逆，可以证明下面定理.

定理

设 $X \in \mathbb{R}^{m,n}$, $y \in \mathbb{R}^m$. 亏秩最小二乘问题 $\min\limits_{\beta\in\mathbb{R}^n} \|X\beta - y\|_2^2$ 的通解为

$$\beta = X^\dagger y + (I - X^\dagger X)c, \quad \text{其中 } c \text{ 为 } \mathbb{R}^n \text{ 中任意列向量.}$$

记 $\beta_* = X^\dagger y$，利用(6.1)就有

$$\begin{aligned} X^T X\beta_* &= X^T XX^\dagger y = X^T(XX^\dagger)y = X^T(XX^\dagger)^T y \\ &= (XX^\dagger X)^T y = X^T y. \end{aligned}$$

即 β_* 是法方程的一个特解.

还是利用(6.1)，又有

$$X^TX(I-X^\dagger X)c = X^TXc - X^TXX^\dagger Xc = X^TXc - X^TXc = 0.$$

这说明 $(I-X^\dagger X)c$ 是齐次方程组 $(X^TX)\beta = 0$ 的解.

由于

$$\dim\{(I-X^\dagger X)c \,|\, \forall\, c \in \mathbb{R}^n\} = \mathrm{rank}(I-X^\dagger X),$$

而

$$\begin{aligned}\mathrm{rank}\left(I_n - X^\dagger X\right) &= \mathrm{rank}\left(I_n - V\Sigma^\dagger U^T U\Sigma V^T\right)\\ &= \mathrm{rank}\left(I_n - V\Sigma^\dagger\Sigma V^T\right)\\ &= n - r\end{aligned}$$

所以

$$(I-X^\dagger X)c,\ \ c \in \mathbb{R}^n$$

就是齐次方程组 $(X^TX)\beta = 0$ 的通解.

设 X 的奇异值分解为

$$\begin{aligned} X &= [U_1, U_2] \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \\ &= U_1 \Sigma_r V_1^T \\ &= \sum_{i=1}^{r} \sigma_i u_i v_i^T. \end{aligned}$$

则有

$$X^\dagger = [V_1, V_2] \begin{bmatrix} \Sigma_r^{-1} & O \\ O & O \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} = V_1 \Sigma_r^{-1} U_1^T = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i^T$$

于是

$$\beta_* = X^\dagger y = \sum_{i=1}^{r} \frac{u_i^T y}{\sigma_i} v_i.$$

下面证明在最小二乘问题的所有解中，β_* 是 2 范数最小的解.

通解 $\beta = X^\dagger y + (I - X^\dagger X)c$ 的 2 范数的平方为:

$$\begin{aligned}\|\beta\|_2^2 &= \left(X^\dagger y + (I - X^\dagger X)c,\ X^\dagger y + (I - X^\dagger X)c\right) \\ &= \|X^\dagger y\|_2^2 + 2c^T(I - X^\dagger X)^T X^\dagger y + \|(I - X^\dagger X)c\|_2^2 \\ &= \|X^\dagger y\|_2^2 + \|(I - X^\dagger X)c\|_2^2 \geq \|X^\dagger y\|_2^2\end{aligned}$$

其中第二个等式用到了

$$\begin{aligned}(I - X^\dagger X)^T X^\dagger &= (I - (X^\dagger X)^T)X^\dagger \\ &= X^\dagger - (X^\dagger X)^T X^\dagger \\ &= X^\dagger - (X^\dagger X)X^\dagger \\ &= X^\dagger - X^\dagger = 0.\end{aligned}$$

这些等式又用到了(6.1)式.

非线性最小二乘问题

前面讨论的问题有个特点：基函数 $\phi_j(t)$ 完全是已知的，只有线性组合系数 β_j 未知，$\Phi(\beta_1, \cdots, \beta_n)$ 是二次函数，最后解的方程 (法方程) 是线性的.

如果每个基函数 $\phi_j(t)$ 中还有待定参数，记为 $\phi_j(t, \lambda)$，那要解的优化问题就是

$$\min_{\lambda,\beta} \Phi(\lambda, \beta) = \|X(\lambda)\beta - y\|^2.$$

这不再是二次优化问题. 当然还可以通过联立方程组

$$\begin{cases} \nabla_\lambda \Phi &= 0 \\ \nabla_\beta \Phi &= 0 \end{cases}$$

求解. 此时解的是非线性方程组!

一个例子是

$$y(t) \approx \beta_1 e^{-\lambda_1 t} + \beta_2 e^{-\lambda_2 t}$$

要解的问题是

$$\min_{\lambda_1,\lambda_2,\beta_1,\beta_2} \sum_{i=1}^{m} \left(\beta_1 e^{-\lambda_1 t_i} + \beta_2 e^{-\lambda_2 t_i} - y_i\right)^2$$

上述问题的一般形式可以写成:

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}\|r(x)\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}[r_i(x)]^2, \quad m \geqslant n, \tag{8.1}$$

其中 $r(x):\mathbb{R}^n \to \mathbb{R}^m$ 是 x 的非线性向量值函数. 如果 $r(x)$ 是线性函数, 则问题(8.1)就是前面讨论的线性最小二乘问题.
设 $J(x)$ 是 $r(x)$ 的 Jacobi 矩阵

$$J(x) = \begin{bmatrix} \dfrac{\partial r_1(x)}{\partial x_1} & \cdots & \dfrac{\partial r_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial r_m(x)}{\partial x_1} & \cdots & \dfrac{\partial r_m(x)}{\partial x_n} \end{bmatrix}$$

则 $f(x)$ 的梯度为

$$g(x) := \nabla f(x) = \sum_{i=1}^{m} r_i(x)\,\nabla r_i(x) = J(x)^T r(x) \in \mathbb{R}^n$$

要解问题(8.1)就需要解非线性方程组：$g(x) = 0$.
如果用 Newton 法解，就需要计算 $g(x)$ 对 x 的导数，也就是 $f(x)$ 的 Hessian 矩阵：

$$\begin{aligned} G(x) := \nabla^2 f(x) &= \sum_{i=1}^{m} [\nabla r_i(x)\,\nabla r_i(x)^T + r_i(x)\,\nabla^2 r_i(x)] \\ &= J(x)^T J(x) + S(x), \end{aligned}$$

其中

$$S(x) = \sum_{i=1}^{m} r_i(x)\,\nabla^2 r_i(x) \in \mathbb{R}^{n,n}$$

是包含 $r(x)$ 的分量的二阶偏导数的矩阵.

根据上述推导，解问题(8.1)的 Newton 迭代法为：

$$\begin{aligned} x_{k+1} &= x_k - G(x_k)^{-1} g(x_k) \\ &= x_k - [J(x_k)^T J(x_k) + S(x_k)]^{-1} J(x_k)^T r(x_k). \end{aligned}$$

当 $G(x_k)$ 为正定矩阵时，它具有局部 2 阶收敛速度, 但是 Hessian 矩阵 $G(x)$ 中 2 阶导数项 $S(x)$ 通常难以计算或者计算量很大. 为简化计算，忽略 $S(x)$，或者用 1 阶导数信息近似 $S(x)$. 当 $r(x) = 0$，或者 $r(x)$ 接近于零向量时，$S(x)$ 才可以忽略，这类问题称为小残量问题.
Gauss-Newton 法就是在上述 Newton 迭代法中忽略 $S(x)$ 项，并假设 $J(x_k)$ 列满秩，从而成为

$$x_{k+1} = x_k + p_k, \tag{8.2}$$

其中

$$p_k = -[J(x_k)^T J(x_k)]^{-1} J(x_k)^T r(x_k) \tag{8.3}$$

由此看出

$$-p_k = \arg\min_{p \in \mathbb{R}^n} \|J(x_k)p - r(x_k)\|_2^2$$

这说明 Gauss-Newton 法就是在局部解线性最小二乘问题.

Gauss-Newton 法的每次迭代有如下两步:

(1) 解 $J(x_k)^T J(x_k) p_k = -J(x_k) r(x_k)$ 得 p_k;
(2) 计算 $x_{k+1} = x_k + p_k$.

Gauss-Newton 法的每步迭代只需要计算 $r(x)$ 的 1 阶导数，即 Jacobi 矩阵 $J(x_k)$, 如果它列满秩，法方程组的系数矩阵 $J(x_k)^T J(x_k)$ 还是对称正定的.

在 Gauss-Newton 法中，如果 $J(x_k)$ 不是列满秩的，p_k 的计算就会遇到困难. 实际上 $J(x_k)$ 不列满秩的情形经常发生. 因为希望目标函数

$$f(x) = \frac{1}{2}\|r(x)\|_2^2$$

的值 $f(x_{k+1})$ 要比前一步 $f(x_k)$ 有所下降，处理这种情况的最简单办法是将 p_k 取为负梯度向量.

$$p_k = -g_k = -J(x_k)^T r(x_k), \tag{8.4}$$

还可以引进参数 $\mu_k > 0$，并用下列方程把 (8.3)式和(8.4)联系起来:

$$(J(x_k)^T J(x_k) + \mu_k I) p_k = -J(x_k)^T r(x_k). \tag{8.5}$$

希望找出一个自动调整参数 μ_k 的方法，使得在用

$$p_k = -\left(J(x_k)^T J(x_k) + \mu_k I\right)^{-1} J(x_k)^T r(x_k) \tag{8.6}$$

前进时，所得新点

$$x_{k+1} = x_k + p_k \tag{8.7}$$

处的函数值有较大幅度的下降.

上述对 Gauss-Newton 法的改进是 Kenneth Levenberg(1944) 和 Donald Marquardt(1963) 提出的，所以称为 Levenberg-Marquardt 方法 (简称 LM 方法)，其基本步骤可总结如下:

初始步: 给出 x_0 和 $\mu_0 > 0$, $k := 0$.
第一步: 给出 x_k，计算 $g_k = J(x_k)^T r(x_k)$，若 $\| g_k \| < \varepsilon$，则停止.
第二步: 按某一原则确定 μ_k，然后求解线性方程组:
$$\left(J(x_k)^T J(x_k) + \mu_k I\right) p_k = -J(x_k)^T r(x_k).$$
第三步: 置 $x_{k+1} = x_k + p_k$, 令 $k := k + 1$，转第一步.

例：数据产生于函数：

$$y(x) = 0.12e^{-0.213x} + 0.54e^{-0.17x}\sin(1.23x)$$

加 %5 的正态分布噪声.
拟合模型为

$$y(x) = a_1 e^{-a_2 x} + a_3 e^{-a_4 x}\sin(a_5 x)$$

```
>> x=0:0.1:10;
>> y=0.12*exp(-0.213*x)+0.54*exp(-0.17*x).*sin(1.23*x) + ...
 0.05*randn(size(x));
>> f=@(a,x)a(1)*exp(-a(2)*x)+a(3)*exp(-a(4)*x).*sin(a(5)*x);
>> options = optimoptions('lsqcurvefit',     ...
 'Algorithm','levenberg-marquardt');
>> [a,res]=lsqcurvefit(f,[1,1,1,1,1],x,y)
>> x1=0:0.02:10;
>> y1=f(a,x1);
>> plot(x1,y1,x,y,'o','LineWidth',2)
```

函数最佳逼近的基本概念

考虑一元实函数组成的线性空间 $C[a,b]$，在其中可以定义 3 种常用的范数：

$$\|f(x)\|_\infty = \max_{x\in[a,b]} |f(x)|$$

$$\|f(x)\|_1 = \int_a^b |f(x)|\,dx$$

$$\|f(x)\|_2 = \left(\int_a^b |f(x)|^2\,dx\right)^{1/2}$$

以及一种内积：

$$(f(x),g(x)) = \int_a^b f(x)g(x)\,dx.$$

显然 2 范数可以用这种内积定义

$$\|f(x)\|_2 = \sqrt{(f(x),f(x))}$$

所谓函数最佳逼近，就是对给定的 $f(x) \in C[a,b]$，在一个由简单函数组成的有限维子空间 $V \subset C[a,b]$ 中寻找一个元素 $p(x)$，使得在某种范数下，误差达到最小

$$\|f(x) - p(x)\| = \min_{v(x) \in V} \|f(x) - v(x)\|$$

不同的范数对应不同的函数逼近方法.

- ∞-范数:

  $$\|f(x) - p(x)\|_\infty = \min_{v(x) \in V} \max_{x \in [a,b]} |f(x) - v(x)|$$

  称为连续函数的最佳一致逼近.
- 1-范数:

  $$\|f(x) - p(x)\|_1 = \min_{v(x) \in V} \int_a^b |f(x) - v(x)| dx$$

- 2-范数:

  $$\|f(x) - p(x)\|_2 = \min_{v(x) \in V} \left(\int_a^b |f(x) - v(x)|^2 dx \right)^{1/2}$$

  称为连续函数的最佳平方逼近.

连续函数的最佳平方逼近

设 $f(t) \in C[a,b]$，$V_n = \mathrm{span}\{\phi_1(t), \phi_2(t), \phi_3(t), \cdots, \phi_n(t)\} \subset C[a,b]$，其中的 $\phi_j(t)$ 为多项式，三角函数等简单函数，并且线性无关.
V_n 中的任意一个函数可以写成：

$$v(t) = \sum_{j=1}^{n} x_j \, \phi_j(t)$$

其中 $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

最佳平方逼近问题：求 $p(t) \in V$，使得

$$\begin{aligned} \|f(t) - p(t)\|_2^2 &= \min_{v(t) \in V_n} \int_a^b |f(x) - v(t)|^2 \, dt \\ &= \min_{v(t) \in V_n} \int_a^b \left[\sum_{j=1}^{n} x_j \, \phi_j(t) - f(t) \right]^2 dt \end{aligned}$$

引进一个 n 元函数

$$\begin{aligned}\Phi(x) &= \Phi(x_1, x_2, \ldots, x_n) \\ &:= \int_a^b \left[\sum_{j=1}^{n} x_j\,\phi_j(t) - f(t)\right]^2 dt \\ &= \left\|\sum_{j=1}^{n} x_j\,\phi_j(t) - f(t)\right\|_2^2\end{aligned}$$

显然 $\Phi(x)$ 是定义在整个 $\mathbb{R}^n$ 上的 n 元 2 次非负实值函数，它一定有最小值，且最小值点 x_* 是平衡点，即 x_* 一定是方程组

$$\nabla\Phi(x) = 0$$

的解. 下面需要计算偏导数

$$\frac{\partial\Phi}{\partial x_k}, \quad \text{for } k = 1, 2, \ldots, n$$

对于任意事先取定的 $k \in \{1, 2, \ldots, n\}$，有

$$\begin{aligned}
\frac{\partial \Phi}{\partial x_k} &= \frac{\partial}{\partial x_k}\left(\sum_{j=1}^{n} x_j \left(\phi_j, \sum_{l=1}^{n} x_l \phi_l\right) - 2\sum_{j=1}^{n} x_j (f, \phi_j) + (f, f)\right) \\
&= \frac{\partial}{\partial x_k}\left(\sum_{j=1}^{n}\sum_{l=1}^{n} x_j x_l (\phi_j, \phi_l) - 2\sum_{j=1}^{n} x_j (f, \phi_j)\right) \\
&= 2x_k(\phi_k, \phi_k) + 2\sum_{j=1, j\neq k}^{n} x_j(\phi_j, \phi_k) - 2(f, \phi_k)
\end{aligned}$$

所以

$$\frac{\partial \Phi}{\partial x_k} = 0 \iff \sum_{j=1}^{n} x_j(\phi_j, \phi_k) = (f, \phi_k), \quad k = 1, 2 \ldots, n$$

这是一个关于 $x_1, x_2, \ldots, x_n$ 的线性方程组，称为法方程组.

法方程组的矩阵形式为：

$$\begin{bmatrix} (\phi_1,\phi_1) & (\phi_2,\phi_1) & \cdots & (\phi_n,\phi_1) \\ (\phi_1,\phi_2) & (\phi_2,\phi_2) & \cdots & (\phi_n,\phi_2) \\ \vdots & \vdots & \ddots & \vdots \\ (\phi_1,\phi_n) & (\phi_2,\phi_n) & \cdots & (\phi_n,\phi_n) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} (f,\phi_1) \\ (f,\phi_2) \\ \vdots \\ (f,\phi_n) \end{bmatrix} \tag{9.1}$$

定理

设 V 为一个实内积空间，其中的内积记为 $\langle\cdot,\cdot\rangle$，$u_1, u_2, \ldots, u_n \in V$. 则 Gram 矩阵

$$G = \begin{bmatrix} \langle u_1, u_1\rangle & \langle u_2, u_1\rangle & \cdots & \langle u_n, u_1\rangle \\ \langle u_1, u_2\rangle & \langle u_2, u_2\rangle & \cdots & \langle u_n, u_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle u_1, u_n\rangle & \langle u_2, u_n\rangle & \cdots & \langle u_n, u_n\rangle \end{bmatrix}$$

非奇异 $(\det(G) \neq 0)$ 的充要条件为 $\{u_1, \cdots, u_n\}$ 线性无关.

上述定理保证了法方程组的解存在且惟一，记其解为

$$x^* = (x_1^*, x_2^*, \ldots, x_2^*)^T,$$

就惟一确定了一个 n 维子空间 V_n 中的函数

$$p(t) = x_1^*\phi_1(t) + x_2^*\phi_2(t) + \cdots + x_x^*\phi_x(t).$$

还需要证明：$p(t)$ 的确是 $f(t) \in C[a, b]$ 的最佳平方逼近.
证明如下：从 n 维向量 x^* 满足法方程组可知

$$\sum_{j=1}^{n} x_j^*(\phi_j, \phi_k) = (f, \phi_k), \quad k = 1, 2 \ldots, n.$$

为增加可读性，内积符号换成 $\langle \cdot, \cdot \rangle$，$\langle \sum_{j=1}^{n} x_j^*\phi_j, \phi_k \rangle = \langle f, \phi_k \rangle$. 所以有

$$\langle \sum_{j=1}^{n} x_j^*\phi_j - f, \phi_k \rangle = 0, \quad k = 1, 2 \ldots, n$$

即

$$\langle p(t) - f(t), \phi_k \rangle = 0, \quad k = 1, 2 \ldots, n$$

由于 $\{\phi_k(t)\}$ 是 V_n 的一组基，上式说明

$$p(t)-f(t)\perp V_n,\quad \text{即 } p(t)-f(t) \text{ 与 } V_n \text{ 中的任意函数都正交,}$$

所以对于 $\forall\, v(t)\in V_n$，都有

$$\begin{aligned}\|v(t)-f(t)\|^2 &= \|v(t)-p(t)+p(t)-f(t)\|^2\\ &= \langle (v(t)-p(t))+(p(t)-f(t)),(v(t)-p(t))+(p(t)-f(t))\rangle\\ &= \|v(t)-p(t)\|^2+\|p(t)-f(t)\|^2+2\langle v(t)-p(t),p(t)-f(t)\rangle\\ &= \|v(t)-p(t)\|^2+\|p(t)-f(t)\|^2\\ &\geq \|p(t)-f(t)\|^2\end{aligned}$$

这就证明了

$$\|p(t)-f(t)\|^2=\min_{v(t)\in V_n}\|v(t)-f(t)\|^2.$$

事实上，最佳平方逼近并不局限于连续函数，只要 $f(t)$ 平方可积就可以. 所有在区间 (a,b) 上平方可积的函数构成一个线性空间，记为 $L^2(a,b)$，它才是可以定义内积 $(f,g)=\int_a^b f(t)g(t)\,dt$ 的自然函数空间.

最佳平方逼近的误差大小应该用 2 范数计算.

$$\|p(t)-f(t)\|_2^2=(p-f,p-f)=(p-f,p)-(p-f,f)$$

由于 $(p-f,p)=0$，所以

$$\begin{aligned}\|p(t)-f(t)\|_2^2&=-(p-f,f)\\&=\|f(t)\|_2^2-(\sum_{j=1}^{n}x_j^*\phi_j,f)\\&=\|f(t)\|_2^2-\sum_{j=1}^{n}x_j^*(\phi_j,f)\end{aligned}$$

即

$$\|p(t)-f(t)\|_2=\sqrt{\|f(t)\|_2^2-\sum_{j=1}^{n}x_j^*(\phi_j,f)}$$

函数的最佳平方逼近多项式

设函数 $f(t) \in L^2[0,1]$，取 $V_n = \mathbb{P}_{n-1}$(所有不高于 $n-1$ 阶的实多项式)，V_n 的基取为单项式

$$\phi_j(t) = t^{j-1}, \quad j = 1, 2, \ldots, n,$$

则有

$$(\phi_k, \phi_j) = \int_0^1 t^{k+j-2}\, dt = \frac{1}{k+j-1}, \quad k, j = 1, 2, \ldots, n.$$

此时的法方程组(9.1)为

$$\begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & \frac{1}{n+1} & \cdots & \frac{1}{2n+1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \int_0^1 f(t)\, dt \\ \int_0^1 f(t)\, t dt \\ \vdots \\ \int_0^1 f(t)\, t^{n-1} dt \end{bmatrix}$$

例子：设 $f(t)=\sqrt{t^2+1}$，求其在 $[0,1]$ 上的一次最佳平方逼近多项式.

$$\int_0^1 f(t)\,dt=\int_0^1 \sqrt{t^2+1}\,dt=\left[\frac{1}{2}\ln(t+\sqrt{t^2+1})+\frac{t}{2}\sqrt{t^2+1}\right]_0^1$$
$$=\frac{1}{2}\ln(t+\sqrt{2})+\frac{1}{2}\sqrt{2}\approx 1.147$$

$$\int_0^1 f(t)t\,dt=\int_0^1 t\sqrt{t^2+1}\,dt=\frac{1}{2}\cdot\frac{2}{3}(1+t^2)^{3/2}\Big|_0^1=\frac{1}{3}(2\sqrt{2}-1)\approx 0.609$$

解方程组

$$\begin{bmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{3}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}1.147\\ 0.609\end{bmatrix}$$

得到 $x_1^*=0.934,\ x_2^*=0.426$，所以一次最佳平方逼近多项式为：

$$p(t)=0.934+0.426\,t$$

其误差为

$$\|p(t)-f(t)\|_2=\sqrt{\int_0^1(1+t^2)\,dt-\sum_{j=1}^{2}x_j^*\int_0^1 t^{j-1}f(t)\,dt}\approx 0.051$$

1. 函数最佳平方逼近的数值解法就是解线性方程组(9.1)，在实内积空间中，Gram 矩阵是对称正定阵，可以用 Cholesky 分解求解.
2. 法方程组(9.1)一般是稠密方程组，数值求解的工作量是 $O(n^3)$.
3. 法方程组很有可能是病态的，例如 $[0,1]$ 区间上的最佳平方逼近多项式，如果采用单项式做基函数，Gram 矩阵是 Hilbert 矩阵.
4. 如果选取正交基函数，Gram 矩阵就是对角阵，可以大大减少计算量.
5. 函数的最小二乘拟合是最佳平方逼近的特殊离散形式.

参考书

1. 徐树方,
   矩阵计算的理论与方法,
   北京大学出版社, 1995.
2. 喻文健,
   数值分析与算法 (第三版),
   清华大学出版社, 2020.
3. James W. Demmel,
   Applied Numerical Linear Algebra,
   SIAM, 1997.
4. Dingyu Xue, YangQuan Chen,
   Scientific Computing with MATLAB,
   Taylor & Francis Inc, 2016.