第二十八讲：正定矩阵和最小值

本讲我们会了解如何完整的测试一个矩阵是否正定，测试x^TAx是否具有最小值，最后了解正定的几何意义——椭圆（ellipse）和正定性有关，双曲线（hyperbola）与正定无关。另外，本讲涉及的矩阵均为实对称矩阵。

正定性的判断

我们仍然从二阶说起，有矩阵$A=\begin{bmatrix}a&b\\b&d\end{bmatrix}$，判断其正定性有以下方法：

1. 矩阵的所有特征值大于零则矩阵正定：$\lambda_1>0,\ \lambda_2>0$；
2. 矩阵的所有顺序主子阵（leading principal submatrix）的行列式（即顺序主子式，leading principal minor）大于零则矩阵正定：$a>0,\ ac-b^2>0$；
3. 矩阵消元后主元均大于零：$a>0,\ \frac{ac-b^2}{a}>0$；
4. $x^TAx>0$；

大多数情况下使用4来定义正定性，而用前三条来验证正定性。

来计算一个例子：$A=\begin{bmatrix}2&6\\6&?\end{bmatrix}$，在?处填入多少才能使矩阵正定？

- 来试试18，此时矩阵为$A=\begin{bmatrix}2&6\\6&18\end{bmatrix}$，$\det A=0$，此时的矩阵成为半正定矩阵（positive semi-definite）。矩阵奇异，其中一个特征值必为0，从迹得知另一个特征值为20。矩阵的主元只有一个，为2。

  计算x^TAx，得$\begin{bmatrix}x_1&x_2\end{bmatrix}\begin{bmatrix}2&6\\6&18\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix}=2x_1^2+12x_1x_2+18x_2^2$这样我们得到了一个关于$x_1,x_2$的函数$f(x_1,x_2)=2x_1^2+12x_1x_2+18x_2^2$，这个函数不再是线性的，在本例中这是一个纯二次型（quadratic）函数，它没有线性部分、一次部分或更高次部分（Ax是线性的，但引入x^T后就成为了二次型）。

  当?取18时，判定1、2、3都是“刚好不及格”。

- 我们可以先看“一定不及格”的样子，令$? = 7$，矩阵为$A = \begin{bmatrix}2 & 6\\6 & 7\end{bmatrix}$，二阶顺序主子式变为$-22$，显然矩阵不是正定的，此时的函数为$f(x_1, x_2) = 2x_1^2 + 12x_1x_2 + 7x_2^2$，如果取$x_1 = 1, x_2 = -1$则有$f(1, -1) = 2 - 12 + 7 < 0$。

  如果我们把$z = 2x^2 + 12xy + 7y^2$放在直角坐标系中，图像过原点$z(0, 0) = 0$，当$y = 0$或$x = 0$或$x = y$时函数为开口向上的抛物线，所以函数图像在某些方向上是正值；而在某些方向上是负值，比如$x = -y$，所以函数图像是一个马鞍面（saddle），$(0, 0, 0)$点称为鞍点（saddle point），它在某些方向上是极大值点，而在另一些方向上是极小值点。（实际上函数图像的最佳观测方向是沿着特征向量的方向。）

- 再来看一下“一定及格”的情形，令$? = 20$，矩阵为$A = \begin{bmatrix}2 & 6\\6 & 20\end{bmatrix}$，行列式为$\det A = 4$，迹为$trace(A) = 22$，特征向量均大于零，矩阵可以通过测试。此时的函数为$f(x_1, x_2) = 2x_1^2 + 12x_1x_2 + 20x_2^2$，函数在除$(0, 0)$外处处为正。我们来看看$z = 2x^2 + 12xy + 20y^2$的图像，式子的平方项均非负，所以需要两个平方项之和大于中间项即可，该函数的图像为抛物面（paraboloid）。在$(0, 0)$点函数的一阶偏导数均为零，二阶偏导数均为正（马鞍面的一阶偏导数也为零，但二阶偏导数并不均为正，所以），函数在改点取极小值。

  在微积分中，一元函数取极小值需要一阶导数为零且二阶导数为正$\frac{\mathrm{d}u}{\mathrm{d}x} = 0, \frac{\mathrm{d}^2u}{\mathrm{d}x^2} > 0$。在线性代数中我们遇到了了多元函数$f(x_1, x_2, \cdots, x_n)$，要取极小值需要二阶偏导数矩阵为正定矩阵。

  在本例中（即二阶情形），如果能用平方和的形式来表示函数，则很容易看出函数是否恒为正，$f(x, y) = 2x^2 + 12xy + 20y^2 = 2\left(x + 3y\right)^2 + 2y^2$。另外，如果是上面的$? = 7$的情形，则有$f(x, y) = 2(x + 3y)^2 - 11y^2$，如果是$? = 18$的情形，则有$f(x, y) = 2(x + 3y)^2$。

  如果令$z = 1$，相当于使用$z = 1$平面截取该函数图像，将得到一个椭圆曲线。另外，如果在$? = 7$的马鞍面上截取曲线将得到一对双曲线。

  再来看这个矩阵的消元，$\begin{bmatrix}2 & 6\\6 & 20\end{bmatrix} = \begin{bmatrix}1 & 0\\-3 & 1\end{bmatrix}\begin{bmatrix}2 & 6\\0 & 2\end{bmatrix}$，这就是$A = LU$，可以发现矩阵$L$中的项与配平方中未知数的系数有关，而主元则与两个平方项外的系数有关，这也就是为什么正数主元得到正定矩阵。

  上面又提到二阶导数矩阵，这个矩阵型为$\begin{bmatrix}f_{xx} & f_{xy}\\f_{yx} & f_{yy}\end{bmatrix}$，显然，矩阵中的主对角线元素（纯二阶导数）必须为正，并且主对角线元素必须足够大来抵消混合导数的影

响。同时还可以看出，因为二阶导数的求导次序并不影响结果，所以矩阵必须是对称的。现在我们就可以计算$n\times n$阶矩阵了。

接下来计算一个三阶矩阵，$A=\begin{bmatrix}2&-1&0\\-1&2&-1\\0&-1&2\end{bmatrix}$，它是正定的吗？函数$x^TAx$是多少？函数在原点去最小值吗？图像是什么样的？

- 先来计算矩阵的顺序主子式，分别为$2,3,4$；再来计算主元，分别为$2,\frac{3}{2},\frac{4}{3}$；计算特征值，$\lambda_1=2-\sqrt{2},\lambda_2=2,\lambda_3=2+\sqrt{2}$。
- 计算$x^TAx=2x_1^2+2x_2^2+2x_3^2-2x_1x_2-2x_2x_3$。
- 图像是四维的抛物面，当我们在$f(x_1,x_2,x_3)=1$处截取该面，将得到一个椭圆体。一般椭圆体有三条轴，特征值的大小决定了三条轴的长度，而特征向量的方向与三条轴的方向相同。

现在我们将矩阵A分解为$A=Q\Lambda Q^T$，可以发现上面说到的各种元素都可以表示在这个分解的矩阵中，我们称之为主轴定理（principal axis theorem），即特征向量说明主轴的方向、特征值说明主轴的长度。

$A=Q\Lambda Q^T$是特征值相关章节中最重要的公式。