第四章　SVD

奇异值分解 (SVD) 实际上是数学专业内容，但它现在已经渗入到不同的领域中。SVD 的过程不是很好理解，因为它不够直观，但它对矩阵分解的效果却非常好。比如，Netflix（一个提供在线电影租赁的公司）曾经就悬赏 100 万美金，如果谁能提高它的电影推荐系统评分预测准确率提高 10% 的话。令人惊讶的是，这个目标充满了挑战，来自世界各地的团队运用了各种不同的技术。最终的获胜队伍”BellKor’s Pragmatic Chaos” 采用的核心算法就是基于 SVD [2].

4.1　线性变换几何意义 (The geometry of linear transformations)

让我们来看一些简单的线性变换例子，以 2×2 的线性变换矩阵为例，首先来看一个较为特殊的对角矩阵：

$$M=\begin{bmatrix}3 & 0\\ 0 & 1\end{bmatrix}$$

从几何上讲，M 是将二维平面上的点 (x,y) 经过线性变换到另外一个点的变换矩阵，

$$M\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}3 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}3x\\ y\end{bmatrix}$$

对角矩阵 M 变换的效果如图4.1所示，变换后的平面仅仅是沿 X 水平方面进行了拉伸 3 倍，y 垂直方向是并没有发生变化。

注记 4.1. 值得注意的是：经对角矩阵变换，相互垂直的网格还是相互垂直的，只是在某些方向上做了伸缩变换.

现在考虑对称矩阵

$$M=\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}. \tag{4.1}$$

这个对称矩阵产生的变换效果如图4.2所示

这种变换效果看起来非常的奇怪，在实际环境下很难描述出来变换的规律（这里应该是指无法清晰辨识出旋转的角度，拉伸的倍数之类的信息)。还是基于上面的对称矩阵，假设我们把左边的平面旋转 45 度角，然后再进行对称矩阵 M (4.1) 的线性变换，效果如下图4.3所示：

看起来是不是有点熟悉? 对! 经过对称矩阵 M (4.1) 线性变换后，跟前面的对角矩阵的功能是相同的，都是将网格沿着一个方向拉伸了 3 倍.

注记 4.2. 值得注意的是:

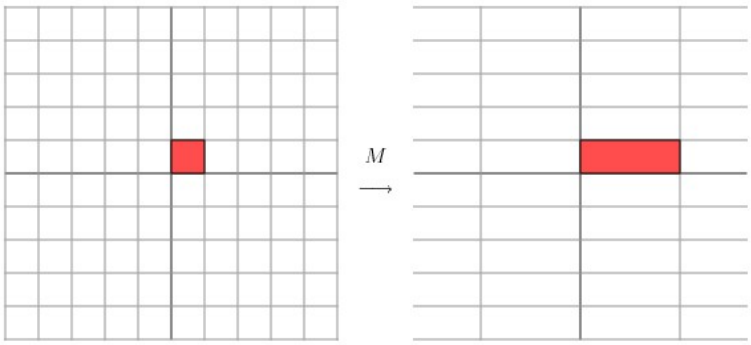

图 4.1: 对角矩阵变换的效果图.

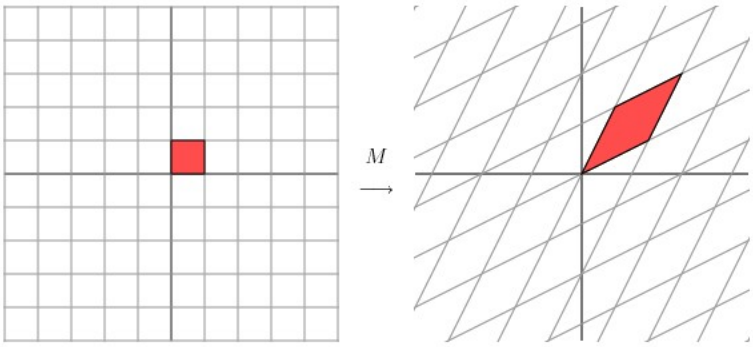

图 4.2: 对称矩阵变换的效果图.

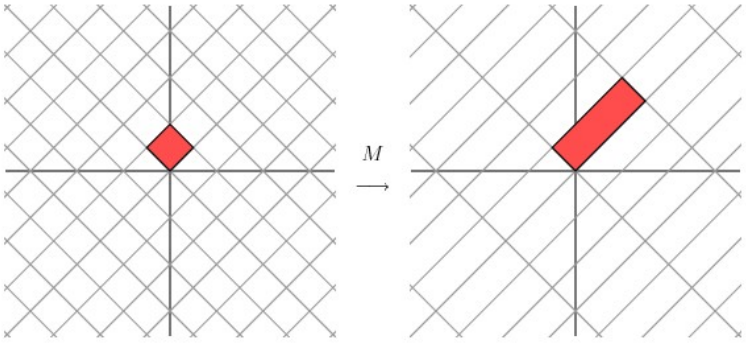

图 4.3: 把左边的平面旋转 45 度角，然后再进行矩阵 M 的线性变换的效果图.

- 旋转 45 度后经过对称矩阵 M (4.1) 变换和对角矩阵变换一样，原来正交的向量还是正交的. 如何证明?

  写出 M 的特征根和特征向量，定义见下面.
  特征值是 1 和 3，对应的特征向量是 $(-\sqrt{2}/2, \sqrt{2}/2)^T$ 和 $(\sqrt{2}/2, \sqrt{2}/2)^T$.

- 对于一个 2×2 的对称矩阵 M，我们总可以先将网格平面旋转一定的角度，然后经该对称矩阵变换，可得到沿两个方向进行拉伸变换. 换句话说，旋转一定角度后，对称矩阵有类似对角矩阵的效果.

数学上可进一步细化:

给定一个对称矩阵 M，我们可以找到一组相互正交 v_i，使得 Mv_i 就是沿着 v_i 方向进行拉伸变换，即

$$Mv_i = \lambda_i v_i.$$

这里的 λ_i 是拉伸尺度 (scalar). 从几何上看，M 对向量 v_i 进行了拉伸映射变换. v_i 称作矩阵 M 的特征向量 (eigenvector)，λ_i 称作为矩阵 M 特征值 (eigenvalue)。这里有一个非常重要的定理

定理 4.3. *实对称矩阵 M 的特征向量 v_i 是相互正交的.*

证明: 设 λ_1, λ_2 是 A 两个不同的特征值，v_1, v_2 分别是其对应的特征向量，有

$$\begin{aligned} Av_1 &= \lambda_1 v_1, \\ Av_2 &= \lambda_2 v_2. \end{aligned}$$

对上面两个式子，分别两边左乘 v_2^T 和 v_1^T，得

$$\begin{aligned} v_2^T A v_1 &= \lambda_1 v_2^T v_1, \\ v_1^T A v_2 &= \lambda_2 v_1^T v_2 = \lambda_2 v_2^T v_1. \end{aligned}$$

由于 $v_2^T A v_1 = (v_2^T A v_1)_T = v_1^T A v_2$，将两式相减可有，

$$0 = v_2^T A v_1 - v_1^T A v_2 = (\lambda_1 - \lambda_2) v 1^T v_2.$$

而 $\lambda_1 \neq \lambda_2$，因此 $v 1^T v_2 = 0$，即 v_1 与 v_2 正交.

也就是说，如果我们用对称矩阵的特征向量来构成网格的话，那么通过 M 矩阵特征值对网格平面进行伸缩的效果跟对 M 矩阵对网格进行线性变换的效果是一样的. 对于线性变换，我们给出几何上一个简单的解释:

注记 4.4. *所谓线性变换，就是对网格在一个方向上进行简单伸缩变换.*

问题 4.5. *对于一般的矩阵 M，我们该怎么做才能让一个原来就是相互垂直的网格平面 (orthogonal grid)，线性变换成另外一个网格平面同样垂直呢?*

让我们考虑一个非对称的矩阵

$$M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \tag{4.2}$$

经过矩阵 (4.2) 变换以后的效果如图4.4. 可以看出该矩阵的一个特征向量是沿着水平方向的 (因为横坐标没有变 $Mv = \lambda v$). 但另一个特征向量不是，故无法构成正交网格.

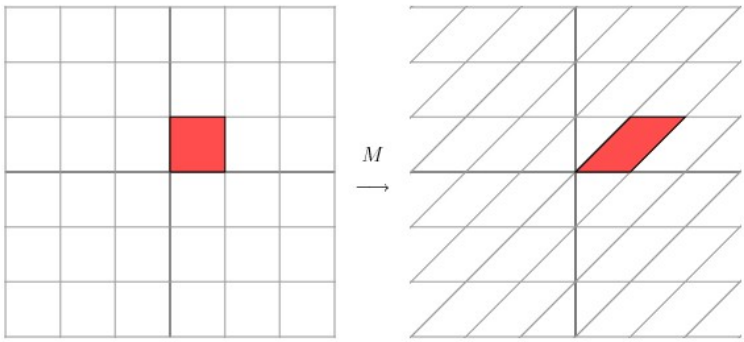

图 4.4: 非对称矩阵 M 线性变换的效果图.

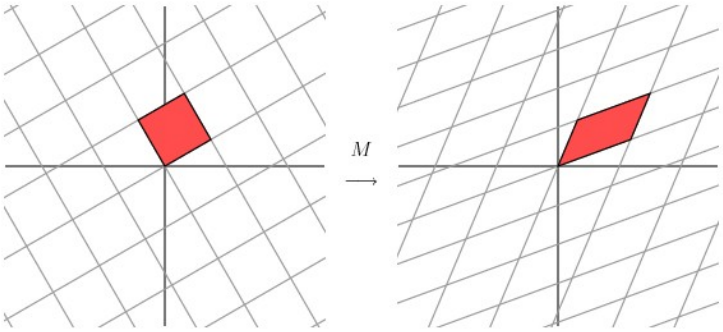

图 4.5: 网格平面旋转 30 度角，然后再进行非对称矩阵 M (4.2) 的线性变换以后的效果图.

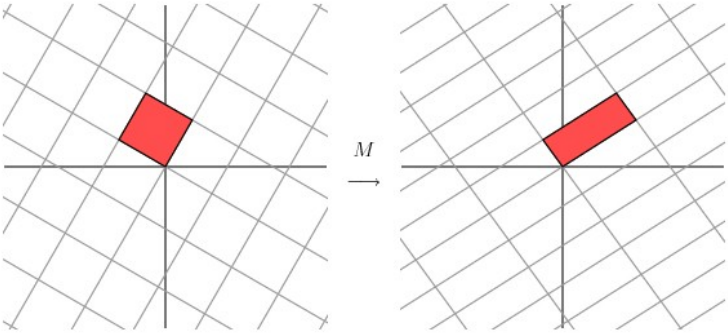

图 4.6: 网格平面旋转 60 度角，然后再进行非对称矩阵 M (4.2) 的线性变换以后的效果图.

把网格平面旋转 30 度角，然后再进行 M (4.2) 同样的线性变换以后的效果，如图4.5所示.

网格平面旋转 60 度角的时候的效果，如图4.6所示:

这个看起来挺不错的样子. 事实上可以精确计算出来的. 设 $x=(x_1,x_2)^T$，$y=(y_1,y_2)^T$ 正交，即

$$x^Ty=x_1y_1+x_2y_2=0.$$

经过非对称矩阵 M (4.2) 的线性变换以后，得到 $\tilde{x}=Ax$ 和 $\tilde{y}=Ay$，且

$$\tilde{x}^T\tilde{y}=x^TA^TAy=(x_1,x_2)\begin{bmatrix}1&1\\1&2\end{bmatrix}\begin{bmatrix}y_1\\y_2\end{bmatrix}=0.$$

这样就可以联立方程:

$$\begin{aligned}x_1y_2+x_2y_1+x_2y_2&=0\\x_1y_1+x_2y_2&=0\end{aligned}$$

从而可以得到

$$\frac{x_1}{x_2}=-\frac{y_1+y_2}{y_2}=-\frac{y_2}{y_1}.$$

在方程 $y_1^2+y_1y_2-y_2^2=0$ 中设 $y_2=1$，计算得到 $y_1=\dfrac{-1-\sqrt{5}}{2}$. 类似得，设 $x_2=1$，计算得到 $x_1=\dfrac{\sqrt{5}-1}{2}$.

为此，我们得到两个正交得方向是 $x=(\dfrac{\sqrt{5}-1}{2},1)^T$ 和 $y=(\dfrac{-1-\sqrt{5}}{2},1)^T$，其单位向量是 $x=(\dfrac{\sqrt{5}-1}{2},1)^T/\sqrt{\dfrac{5-\sqrt{5}}{2}}=(0.5257,0.8507)^T$ 和 $y=(\dfrac{-1-\sqrt{5}}{2},1)^T/\sqrt{\dfrac{5+\sqrt{5}}{2}}=(-0.8507,0.5257)^T$. 也就是，应该把网格平面旋转 58.28 度就能达到理想的效果

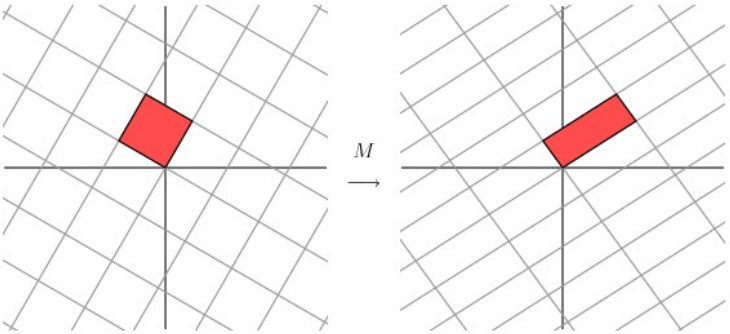

图 4.7: 网格平面旋转 58 度角，然后再进行非对称矩阵 M (4.2) 的线性变换以后的效果图.

4.2 奇异值分解 (Singular value decomposition)

该部分是从几何层面上去理解二维的 SVD:

对于任意的 2×2 矩阵，通过 SVD 可以将一个相互垂直的网格 (orthogonal grid) 变换到另外一个相互垂直的网格。

我们可以通过向量的方式来描述这个事实：首先，选择两个相互正交的单位向量 v_1 和 v_2，使得向量 Mv_1 和 Mv_2 正交. 如图4.8所示：

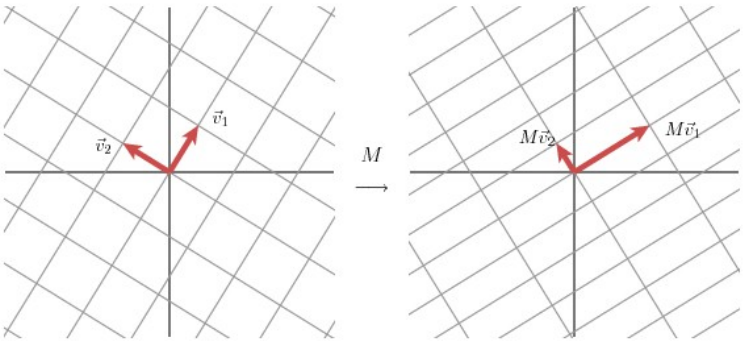

图 4.8：正交向量，经矩阵 M (4.2) 的线性变换以后得到的向量仍然正交.

u_1 和 u_2 分别表示 Mv_1 和 Mv_2 的单位向量，$\sigma_1 u_1 = Mv_1$ 和 $\sigma_2 u_2 = Mv_2$，如图4.9所示. σ_1 和 σ_2 分别表示这不同方向向量上的模，也称作为矩阵 M 的奇异值. 值得注意的是，奇异值不要求 $\sigma_1 u_1 = Mv_1$ 里面的 u_1 和 v_1 一样，但特征值却是要求一样.

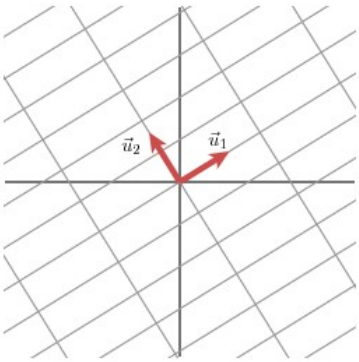

图 4.9：正交向量，经矩阵 M (4.2) 的线性变换以后得到的正交单位向量.

这样我们就有了如下关系式

$$\begin{aligned} Mv_1 &= \sigma_1 u_1 \\ Mv_2 &= \sigma_2 u_2 \end{aligned}$$

现在可以简单描述下经过 M 线性变换后的向量 x 的表达形式. 由于向量 v_1 和 v_2 是正交的单位向量，我们可以得到如下式子：

$$x = (v_1 \cdot x)v_1 + (v_2 \cdot x)v_2$$

这就意味着：

$$\begin{aligned} Mx &= (v1 \cdot x)Mv_1 + (v2 \cdot x)Mv_2, \\ &= (v1 \cdot x)\sigma_1 u_1 + (v_2 \cdot x)\sigma_2 u_2 \\ &= u_1 \sigma_1 v_1^T x + u_2 \sigma_2 v_2^T x \end{aligned}$$

因为对任意 x 都成立，所以

$$M = u_1\sigma_1 v_1^T + u_2\sigma_2 v_2^T.$$

上述的式子经常表示成

$$M = U\Sigma V^T,$$

其中 U 矩阵的列向量分别是 u_1，u_2 ，Σ 是一个对角矩阵，对角元素分别是对应的 σ_1 和 σ_2，V 矩阵的列向量分别是 v_1，v_2.

这就表明任意的矩阵 M 是可以分解成三个矩阵。V 表示了原始域的标准正交基，U 表示经过 M 变换后的 co-domain 的标准正交基，Σ 表示了 V 中的向量与 U 中相对应向量之间的关系. (V describes an orthonormal basis in the domain, and U describes an orthonormal basis in the co-domain, and Σ describes how much the vectors in V are stretched to give the vectors in U.)

4.3 如何获得奇异值分解?(How do we find the singular decomposition?)

事实上我们可以找到任何矩阵的奇异值分解，那么我们是如何做到的呢?

假设在原始域中有一个单位圆，如图4.10所示. 经过 M 矩阵变换以后在 co-domain 中单位圆会变成一个椭圆，它的长轴（Mv_1）和短轴（Mv_2）分别对应转换后的两个标准正交向量，也是在椭圆范围内最长和最短的两个向量.

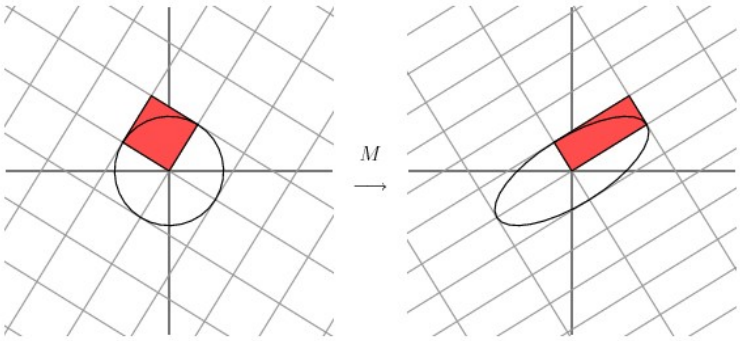

图 4.10: 正交向量，经矩阵 M (4.2) 的线性变换以后得到的正交单位向量.

换句话说，定义在单位圆上的函数 $|Mx|$ 分别在 v_1 和 v_2 方向上取得最大和最小值. 这样我们就把寻找矩阵的奇异值分解过程缩小到了优化函数 $|Mx|$ 上了. 结果发现（具体的推到过程这里就不详细介绍了）这个函数取得最优值的向量分别是矩阵 M^TM 的特征向量. 由于 M^TM 是对称矩阵，因此不同特征值对应的特征向量都是互相正交的，我们用 v_i 表示 M^TM 的所有特征向量. 奇异值 $\sigma_i = |Mv_i|$，向量 u_i 为 Mv_i 方向上的单位向量.

问题 4.6. *为什么 u_i 也是正交的呢?*

Proof. 设 σ_i 和 σ_j 分别是不同两个奇异值，

$$\begin{aligned} Mv_i &= \sigma_i u_i \\ Mv_j &= \sigma_j u_j. \end{aligned}$$

我们看下 $(Mv_i)\cdot(Mv_j)$，并假设它们分别对应的奇异值都不为零. 一方面，由于 v_i 是 M^TM 的特征向量，

$$(Mv_i)\cdot(Mv_j) = v_i^T M^T M v_j = v_i^T(M^TM)v_j = {}_j v_i^T v_j = 0$$

另一方面，我们有

$$(Mv_i)\cdot(Mv_j) = \sigma_i\sigma_j u_i u_j = 0.$$

因此，u_i 和 u_j 是正交的. □

注记 4.7. 根据 Section 4.2，M 表示成

$$M = U\Sigma V^T,$$

其中 U 矩阵的列向量分别是 u_1，u_2 ，Σ 是一个对角矩阵，对角元素分别是对应的 σ_1 和 σ_2，V 矩阵的列向量分别是 v_1，v_2. 从而因为

$$M^TM = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T,$$

V is orthogonal (its columns are eigenvectors of M^TM). 因为

$$MM^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T,$$

U is orthogonal (its columns are eigenvectors of MM^T).

但实际上，这并非是求解奇异值的方法，效率会非常低。这里也主要不是讨论如何求解奇异值.

4.4 应用实例 (Examples)

例子 4.8 (矩阵的秩). 考虑矩阵

$$M = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}.$$

该矩阵的几何效果为图4.11. 可以看出，第二个奇异值为 0，写为

$$M = u_1\sigma_1 v_1^T.$$

换句话说，如果某些奇异值为 0，那么相应项不会出现再 M 的分离中. 为此，M 的秩（图像线性转换的维数）等于非零奇异值的个数.

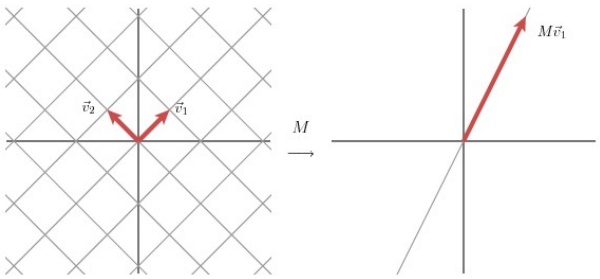

图 4.11: 经矩阵 M (4.2) 的线性变换以后得到的单向量.

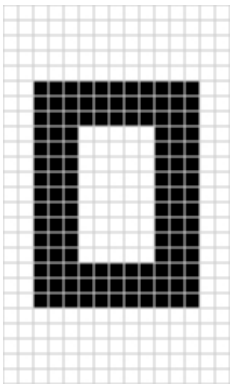

图 4.12: 25×15 由黑白块组成的图像数据.

例子 4.9 (Data compression). 奇异值分解在数据表达上的有效应用. 假设要传送如图4.12的一张 25×15 由黑白块组成的图像数据.

因为构成数据图像只有三种不同的列, 如图4.13 所示, 所以可在压缩的模式下表示数据.

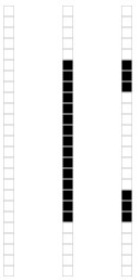

图 4.13: 25×15 由黑白块组成的图像数据只有三种不同的列.

将图像表示成 25×15 的矩阵, 矩阵元素对应着图像的不同像素, 若像素是白色的话, 取 1, 黑色就取 0. 我们得到了一个具有 375 个元素的矩阵, 如图4.14所示

如果我们对矩阵 M 进行奇异值分解, 得到奇异值分别是

$$\sigma_1 = 14.72$$
$$\sigma_2 = 5.22$$
$$\sigma_3 = 3.31$$

矩阵 M 就可以表示成

$$M = u_1\sigma_1 v_1^T + u_2\sigma_2 v_2^T + u_3\sigma_3 v_3^T$$

v_i 具有 15 个元素, u_i 具有 25 个元素, σ_i 对应不同的奇异值, 如上图4.14所示, 我们就可以用 123 个元素来表示具有 375 个元素的图像数据了.

例子 4.10 (noise reduction). 前面的例子里除了三个大的奇异值外, 其他奇异值都为零. 一般来说大的奇异值可以指出一些有兴趣的信息来. 下面我们来探索一下拥有小的或者非常小的奇异值的情况。比如, 我们有一张扫描的, 带有噪声的图像, 如图4.15所示.

$$M = \begin{bmatrix}
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&0&0&0&0&0&0&0&0&0&0&0&1&1\\
1&1&0&0&0&0&0&0&0&0&0&0&0&1&1\\
1&1&0&0&0&0&0&0&0&0&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&1&1&1&1&1&0&0&0&1&1\\
1&1&0&0&0&0&0&0&0&0&0&0&0&1&1\\
1&1&0&0&0&0&0&0&0&0&0&0&0&1&1\\
1&1&0&0&0&0&0&0&0&0&0&0&0&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1
\end{bmatrix}$$

图 4.14: 25×15 由黑白块组成的图像数据表示成矩阵.

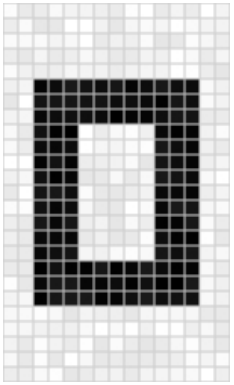

图 4.15: 25×15 由黑白块组成带有噪声的图像数据.

我们采用跟 Example 4.9 相同的处理方式，表示数据为 25×15 矩阵，然后 SVD 分解. 得到图像矩阵的奇异值：

$$\sigma_1 = 14.15$$
$$\sigma_2 = 4.67$$
$$\sigma_3 = 3.00$$
$$\sigma_4 = 0.21$$
$$\sigma_5 = 0.19$$
$$\cdots$$
$$\sigma_{15} = 0.05$$

很明显，前面三个奇异值远远比后面的奇异值要大. 事实上，后面小的奇异值都是由于噪声引入而产生的. 如果我们把后面小的奇异值都抹掉，

$$M = u_1\sigma_1 v_1^T + u_2\sigma_2 v_2^T + u_3\sigma_3 v_3^T.$$

经过这样的奇异值分解重构后，我们得到了一张降噪后的图像4.16.

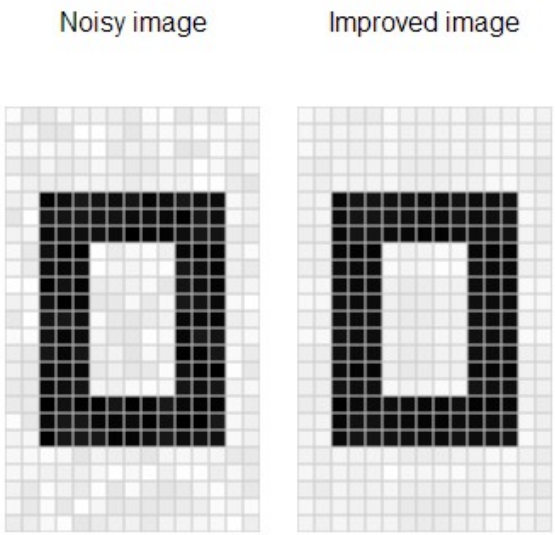

图 4.16: 25×15 由黑白块组成带有噪声的图像数据和去噪后的比较.

例子 4.11 (数据分析 (data analysis)). 我们搜集的数据中总是存在噪声：无论采用的设备多精密，方法有多好，总是会存在一些误差的。如果你们还记得前面的例子提到的，大的奇异值对应了矩阵中的主要信息的话，运用 SVD 进行数据分析，提取其中的主要部分的话，还是相当合理的。

假如我们搜集的数据如图4.17所示：

我们将数据用矩阵的形式表示：

经过奇异值分解后，得到

$$\sigma_1 = 6.04$$
$$\sigma_2 = 0.22.$$

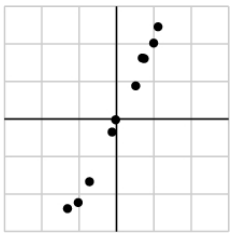

图 4.17: 搜集的数据.

-1.03 0.74 -0.02 0.51 -1.31 0.99 0.69 -0.12 -0.72 1.11
-2.23 1.61 -0.02 0.88 -2.39 2.02 1.62 -0.35 -1.67 2.46

由于第一个奇异值远比第二个要大，数据中有包含一些噪声，第二个奇异值在原始矩阵分解相对应的部分可以忽略。经过 SVD 分解后，保留了主要样本点如图4.18所示. 这就意味着所有数据只位于由 u_1 定义的直线上. In that case, the matrix would have rank one meaning that all the data lies on the line defined by u_i.

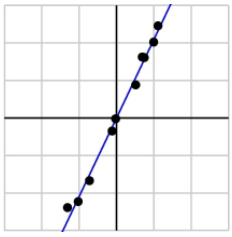

图 4.18: 只保留一个奇异值，这就意味着所有数据只位于由 u_1 定义的直线上.

这个例子也可以视为主成分分析，用特征值来查看数据中的依赖和冗余. This brief example points to the beginnings of a field known as principal component analysis, a set of techniques that uses singular values to detect dependencies and redundancies in data.

类似地，奇异值分解也可以查看数据中的分组. 这就说明了为什么奇异值分解可用来提高 Netflix 电影推荐系统 (Netflix's movie recommendation system). 你对电影的打分会被归类到与你分数类似的其他人组中，从而组里这些人所选的电影会优先推荐给你.

4.5 总结 (Summary)

这篇文章非常的清晰的讲解了 SVD 的几何意义，不仅从数学的角度，还联系了几个应用实例形象的论述了 SVD 是如何发现数据中主要信息的。在 netflix prize 中许多团队都运用了矩阵分解的技术，该技术就来源于 SVD 的分解思想，矩阵分解算是 SVD 的变形，但思想还是一致的。之前算是能够运用矩阵分解技术于个性化推荐系统中，但理解起来不够直观，阅读原文后醍醐灌顶，我想就从 SVD 能够发现数据中的主要信息的思路，就几个方面去思考下如何利用数据中所蕴含的潜在关系去探索个性化推荐系统。也希望路过的各位大侠不吝分享呀。

As mentioned at the beginning of this article, the singular value decomposition should be a central part of an undergraduate mathematics major's linear algebra curriculum. Besides

having a rather simple geometric explanation, the singular value decomposition offers extremely effective techniques for putting linear algebraic ideas into practice. All too often, however, a proper treatment in an undergraduate linear algebra course seems to be missing.

4.6 参考文献推荐

- Gilbert Strang 的书 [3]: ”Linear Algebra and Its Applications” is something of a classic though some may find it to be a little too formal.
- William H. Press 等的经典书 [4]: ”Numercial Recipes in C: The Art of Scientific Computing” yet highly readable. Older versions are available online.
- Dan Kalman 的文章 [5]: ”A Singularly Valuable Decomposition: The SVD of a Matrix”, aims to improve the profile of the singular value decomposition. It also a description of how least-squares computations are facilitated by the decomposition.
- The New York Times 杂志 [6]: ”If You Liked This, You’re Sure to Love That” 值得一读.
- M. Petrou and P. Bosdogianni 的书 [7]: ”Image Processing: The Fundamentals”, 读 page 37-44 - examples of SVD.
- E. Trucco and A. Verri 的书 [8]: ”Introductory Techniques for 3D Computer Vision”, 读 Appendix 6.
- K. Kastleman 的书 [9]: ”Digital Image Processing”, 读 Appendix 3: Mathematical Background.
- F. Ham and I. Kostanic 的书 [10]: ”Principles of Neurocomputing for Science and Engineering”, 读 Appendix A: Mathematical Foundation for Neurocomputing.