谱聚类教程

乌尔里克·冯·卢克斯堡
马克斯普朗克生物控制论研究所
斯佩曼斯特 38, 72076 图宾根, 德国
ulrike.luxburg@tuebingen.mpg.de

本文发表于 Statistics and Computing, 17 (4), 2007。
原始出版物可在www.springer.com。

抽象的

近年来，谱聚类已成为最流行的现代聚类算法之一。它实现简单，可以通过标准线性代数软件有效地解决，并且通常优于传统的聚类算法，如 k-means 算法。乍看之下，谱聚类显得有些神秘，并且根本看不出它为什么起作用以及它到底做了什么。本教程的目标是对这些问题给出一些直觉。我们描述了不同的图拉普拉斯算子及其基本属性，介绍了最常见的谱聚类算法，并通过几种不同的方法从头开始推导这些算法。讨论了不同谱聚类算法的优缺点。

关键词：谱聚类；图拉普拉斯算子

1 简介

聚类是探索性数据分析中使用最广泛的技术之一，应用范围从统计学、计算机科学、生物学到社会科学或心理学。实际上，在处理经验数据的每个科学领域中，人们都试图通过识别数据中的“相似行为”组来获得对数据的第一印象。在本文中，我们想向读者介绍谱聚类算法系列。与“传统算法”相比，例如k-手段或单链接，谱聚类具有许多基本优势。通过谱聚类获得的结果通常优于传统方法，谱聚类实现起来非常简单，并且可以通过标准线性代数方法有效地求解。

本教程是对谱聚类的独立介绍。我们从头开始推导出谱聚类，并就谱聚类为何起作用提出了不同的观点。除了基本的线性代数，读者不需要特定的数学背景。然而，我们并没有试图对所有关于谱聚类的文献进行简明的回顾，这是不可能的，因为关于这个主题的文献太多了。前两节专门逐步介绍谱聚类使用的数学对象：第 2 节中的相似图和第 3 节中的拉普拉斯图。谱聚类算法本身将在第 4 节中介绍。下一个然后用三个部分专门解释为什么这些算法有效。每个部分对应一个解释：第 5 部分描述了一种图形划分方法，第 6 部分描述了随机游走的观点，第 7 部分描述了扰动理论方法。在第 8 节中，我们将研究与谱聚类相关的一些实际问题，并在第 9 节中讨论与谱聚类相关的各种扩展和文献。

2 相似度图

给定一组数据点$X_{1个},...X_n$和一些相似性的概念$秒_{ij}\geqslant$所有数据点对之间为 0$X_{一世}$和X_j，聚类的直观目标是将数据点分成几组，使得同一组中的点彼此相似，不同组中的点彼此不相似。如果我们没有比数据点之间的相似性更多的信息，表示数据的一种很好的方式是*相似图*$G=(V,E)$. 每个顶点$v_{一世}$在此图中代表一个数据点$X_{一世}$. 如果相似性则两个顶点相连$秒_{ij}$对应数据点之间$X_{一世}$和X_j为正值或大于某个阈值，边缘被加权$秒_{ij}$. 现在可以使用相似度图重新表述聚类问题：我们想要找到图的一个分区，使得不同组之间的边具有非常低的权重（这意味着不同聚类中的点彼此不同）并且边组内的点具有高权重（这意味着同一簇内的点彼此相似）。为了能够形式化这种直觉，我们首先要介绍一些基本的图形符号，并简要讨论我们将要研究的图形类型。

2.1 图形符号

让$G=(V,E)$是一个有顶点集的无向图$V=\{v_{1个},...,v_n\}$. 下面我们假设图G是加权的，即两个顶点之间的每条边$v_{一世}$和v_j具有非负权重$w_{ij}\geqslant 0$.加权*邻接矩阵*图的是矩阵$W=(w_{ij})_{i,j=1,...,n}$. 如果$w_{ij}=0$这意味着顶点$v_{一世}$和v_j没有边连接。作为G是无向的，我们需要$w_{ij}=w_{ji}$. 顶点的度数$v_{一世}\in V$定义为

$$d_{一世}=\sum_{j=1}^{n} w_{ij}.$$

注意，事实上，这个和只遍历所有相邻的顶点$v_{一世}$, 对于所有其他顶点v_j重量w_{ij}为 0。*度矩阵了*被定义为具有度数的对角矩阵$d_{1个},...,d_n$
在对角线上。给定一个顶点子集$一个\subset V$,我们表示它的补码$V\backslash A$经过$一个$. 我们定义指标向量$1_{一个}=(f_{1个},...,f_n)'\in \mathbb{R}^n$作为带有条目的向量$f_{一世}=1$如果$v_{一世}\in 一个$和$f_{一世}=0$否则。为了方便我们引入 sh Σ手写符号$w_{一世\in 一个}=\sum_{一世\in 一个} w_{ij}$. 对于两个不一定不相交的集合$一世\in 一个$对于一组索引$\{我\mid v_{一世}\in 一个\}$，特别是在处理像$甲、乙\subset V$我们定义

$$W(甲、乙):=\sum_{一世\in 甲,j\in 乙} w_{ij}.$$

我们考虑两种不同的方法来衡量子集的“大小”$一个\subset V$:

$$|一个|:=一个中的顶点数$$

$$体积(一个):=\sum_{一世\in 一个} d_{一世}.$$

直觉上，$|一个|$测量的大小$一个$通过它的顶点数，而 vol($一个$) 测量的大小$一个$通过对附加到顶点的所有边的权重求和$一个$. 一个子集$一个\subset V$如果图中的任意两个顶点是连通的$一个$可以通过路径连接，使得所有中间点也位于$一个$. 一个子集$一个$如果它是连通的并且如果在中的顶点之间没有连接，则称为连通分量$一个$和$一个$. 非空集$一个_{1个},...,一个_k$形成图的分区，如果$一个_{一世}\cap 一个_j=\emptyset$和$一个_{1个}\cup ...\cup 一个_k=V$.

2.2 不同的相似度图

有几种流行的构造来转换给定的集合$X_{1个},...,X_n$具有成对相似性的数据点$秒_{ij}$或成对距离d_{ij}成图表。构建相似度图时，目标是对数据点之间的局部邻域关系进行建模。

这ε-邻域图：这里我们连接成对距离小于的所有点ε. 由于所有连接点之间的距离大致具有相同的比例（至多ε), 对边缘进行加权不会将有关数据的更多信息合并到图中。因此，ε-邻域图通常被认为是未加权的图。

k- 最近邻图：这里的目标是连接顶点$v_{一世}$有顶点v_j如果v_j是其中之一k-最近的邻居$v_{一世}$. 然而，这个定义导致有向图，因为邻域关系不是对称的。有两种方法可以使这个图无向。第一种方法是简单地忽略边缘的方向，即我们连接$v_{一世}$和v_j如果有无向边$v_{一世}$

是其中之一k-最近的邻居v_j*或者*如果v_j是其中之一k-最近的邻居$v_{一世}$. 结果图就是通常所说的*这k-最近邻图*. 第二种选择是连接顶点 $v_{一世}$和v_j如果两者$v_{一世}$是其中之一k-最近的邻居v_j*和*v_j是其中之一k-最近的邻居 $v_{一世}$. 生成的图形称为*相互的k-最近邻图*. 在这两种情况下，在连接适当的顶点后，我们根据端点的相似性对边进行加权。

全连接图：在这里，我们简单地将所有具有正相似性的点彼此连接起来，然后我们对所有边进行加权$秒_{ij}$. 由于该图应该表示局部邻域关系，因此这种构造仅在相似性函数本身对局部邻域建模时才有用。这种相似度函数的一个例子是高斯相似度函数$秒(X_{一世}, X_j) = 指数(-\|X_{一世}-X_j\|^2_{2个}/(2\sigma^2_{2个}))$, 其中参数$\sigma$控制邻域的宽度。该参数起着与参数类似的作用$\varepsilon$在的情况下$\varepsilon$-邻域图。

上面提到的所有图都经常用于谱聚类。据我们所知，关于相似图的选择如何影响谱聚类结果的问题的理论结果并不存在。关于不同图的行为的讨论，我们参考第 8 节。

3 图拉普拉斯算子及其基本性质

谱聚类的主要工具是图拉普拉斯矩阵。存在一个致力于研究这些矩阵的整个领域，称为谱图理论（例如，参见 Chung，1997）。在本节中，我们要定义不同的图拉普拉斯算子并指出它们最重要的属性。我们将仔细区分图拉普拉斯算子的不同变体。请注意，在文献中没有唯一的约定将哪个矩阵确切地称为“拉普拉斯图”。通常，每个作者只是将“他的”矩阵称为拉普拉斯图。因此，在阅读有关图拉普拉斯算子的文献时需要格外小心。

在下文中，我们总是假设G是具有权重矩阵的无向加权图W，在哪里$w_{ij}=w_{记}\geq 0$. 当使用矩阵的特征向量时，我们不一定假设它们是归一化的。例如，常数向量1个和多个一个1个对于一些一个6个= 0 将被视为相同的特征向量。特征值将始终递增排序，尊重多重性。通过“第一k特征向量”，我们指的是对应于k最小的特征值。

3.1 非归一化图拉普拉斯算子

非归一化图拉普拉斯矩阵定义为

$$大号=D-W。$$

在 Mohar (1991, 1997) 中可以找到对其许多属性的概述。以下命题总结了谱聚类所需的最重要的事实。

提案 1（属性*大号*)*矩阵大号满足以下属性：*

1. 对于每个向量$F\in \mathbb{R}^n$*我们有*

$$F\ '如果=\frac{1个}{2个}\sum_{i,j=1} wij(F_{一世}-F_j)^{2个}.$$

2.大号是对称的和半正定的。

*3. 的最小特征值大号为0，对应的特征向量为常数一向量*1个*.*

4.大号有n非负实值特征值$0=\lambda_{1个}\leqslant \lambda_{2个}\leqslant 。。。\leqslant \lambda_n.$

证明。
第 (1) 部分：根据定义$d_{一世}$,

$$F\ '如果=F\ 'Df-f\ 'Wf=\sum_{一世=1}^{n} d_{一世}F_{一世}^{2}-\sum_{i,j=1}^{n} F_{一世}F_j w_{ij}$$

$$=\frac{1个}{2个}\sum_{一世=1}^{n} d_{一世}F_{一世}^{2个}-2\sum_{i,j=1} F_{一世}F_j w_{ij}+\sum_{j=1} d_j F_j^{2个}=\frac{1个}{2个}\sum_{i,j=1} w_{ij}(F_{一世}-F_j)^{2个}.$$

第 (2) 部分：的对称性*大号*直接从对称性得出*W*和 *丁*. 半正定性是第 (1) 部分的直接结果，它表明*F ‘如果≥*全部为 0$F\in \mathbb{R}^n$. 第 (3) 部分：显而易见。

第 (4) 部分是第 (1) - (3) 部分的直接结果。 2个

请注意，未归一化图拉普拉斯算子不依赖于邻接矩阵的对角线元素*W.* 每个邻接矩阵与*W*在所有非对角线位置上导致相同的非标准化图拉普拉斯算子*大号.* 特别是，图中的自边不会改变相应的图拉普拉斯算子。

非标准化图拉普拉斯算子及其特征值和特征向量可用于描述图的许多属性，参见 Mohar (1991, 1997)。下面是一个对谱聚类很重要的例子：

命题2（连通分量的个数和*大号*) *让G是具有非负权重的无向图。然后是多重性k的特征值*0*的大号等于连接组件的数量一个*1个*,. . . ，一个*$_k$*在图中。特征值的特征空间*0*由指标向量跨越*1个$_{一个1个}$,. . . ,1个$_{一个}$*这些组件。*

k

*证明。*我们从案例入手*k*=1，即图是连通的。假使，假设*F*是一个特征值为0的特征向量。那么我们知道

$$0=F\ '如果=\sum_{i,j=1} w_{ij}(F_{一世}-F_j)^{2个}.$$

作为权重w_{ij}是非负的，这个总和只能消失，如果所有条款$w_{ij}(F_{-世}-F_j)$2个消失。因此，如果两个顶点$v_{-世}$和v_j连接（即，$w_{ij}>0$), 那么$F_{-世}$需要等于F_j. 有了这个论点，我们可以看到F对于可以通过图中的路径连接的所有顶点，需要保持不变。此外，由于无向图中连通分量的所有顶点都可以通过路径连接，F需要在整个连接的组件上保持不变。在仅由一个连通分量组成的图中，我们因此只有一个常量向量1个为特征值为0的特征向量，显然是连通分量的指示向量。

现在考虑的情况k连接的组件。在不失一般性的情况下，我们假设顶点是根据它们所属的连通分量排序的。在这种情况下，邻接矩阵W有分块对角线形式，矩阵也一样*大号*:

$$大号=\begin{bmatrix} 大号_{1个} & & & \\ & 大号_{2个} & & \\ & & \ddots & \\ & & & 大号_k \end{bmatrix}$$

请注意，每个块$大号_{-世}$本身是一个适当的图拉普拉斯算子，即对应于的子图的拉普拉斯算子*一世*-th 连接组件。对于所有块对角矩阵都是这种情况，我们知道*大号*由光谱的并集给出$大号_{-世}$, 以及相应的特征向量*大号*是的特征向量$大号_{-世}$, 在其他块的位置填充 0。作为每个$大号_{-世}$是连通图的拉普拉斯图，我们知道每个$大号_{-世}$具有重数为 1 的特征值 0，对应的特征向量是*一世*-th 连接组件。因此，矩阵*大号*具有与连通分量一样多的特征值 0，对应的特征向量是连通分量的指示向量。

2个

3.2 归一化图拉普拉斯算子

文献中有两个矩阵称为归一化图拉普拉斯矩阵。这两个矩阵彼此密切相关，定义为

$$大号_{符号}:=丁_{-1个/2个}低密度脂蛋白_{-1个/2个}=我\text{-}D_{-1个/2个}西部数据_{-1个/2个}$$

$$大号_{写字}:=丁_{-1个}大号=我\text{-}D_{-1个}W。$$

我们将第一个矩阵表示为$大号_{符号}$因为它是一个对称矩阵，第二个是$大号_{写字}$因为它与随机游走密切相关。下面我们总结了几个属性$大号_{符号}$和$大号_{写字}$. 规范化图拉普拉斯算子的标准参考文献是 Chung (1997)。

提案 3（的属性$大号_{符号}$和$大号_{写字}$)*归一化拉普拉斯算子满足以下属性：*

1. *对于每个*$F\in\mathbb{R}^n$*我们有*

$$F\cdot 大号_{符号}F=\frac{1个/\Sigma}{2个}{}_{i,j=1}w_{ij}\left(\frac{F_i}{\sqrt{d_{-世}}}-\frac{F_j}{\sqrt{d_j}}\right)^{2个}.$$

2. *λ 是的特征值大号*$_{写字}$*与特征向量你当且仅当 λ 是大号*$_{符号}$*与特征向量*$w=丁_{1个/2个}$*你。*

3. *λ 是一个特征值大号*$_{写字}$*与特征向量你当且仅当 λ 和你解决广义特征问题鲁=λ杜。*

*4.0是的特征值大号写字与常数一向量*1个 *作为特征向量。*0*是的特征值 大号符号与特征向量丁*1个/2个1个*.*

*5.大号符号和大号写字是半正定的并且有n非负实值特征值*0 = $\lambda_{1个}$ ⩽ 。 . . ⩽ λ_n*.*

*证明。*第(1)部分可以类似于命题1的第(1)部分来证明。
乘以特征值方程可以立即看到部分（2）*大号*符号 *w=λw*和 *丁*-1个/2个
从左边开始代入*你= 丁*-1个/2个*w.*
第（3）部分直接乘以特征值方程*大号写字 你=λu*和 *丁*从左边。第 (4) 部分：第一个陈述是显而易见的*大号*写字1 =0，第二个陈述来自（2）。第（5）部分：关于声明*大号*符号从 (1) 得出，然后是关于*大号写字*从 (2) 得出。

2个

与非归一化图拉普拉斯算子的情况一样，归一化图拉普拉斯算子的特征值 0 的重数与连通分量的数量有关：

命题4（的连通分量数和光谱*大号*符号和*大号*写字) *让G是具有非负权重的无向图。然后是多重性k的特征值*0*两者的大号写字*
*和大号符号等于连接组件的数量一个*1个*,. . . ，一个$_k$在图中。为了大号写字, 本征空间*0*由指标向量跨越*1个*一个-世这些组件。为了大号符号, 本征空间*0*被向量跨越丁*1个/2个1个*一个-世.*

*证明。*证明类似于命题 2 的证明，使用命题 3。 2个

4 谱聚类算法

现在我们想说明最常见的谱聚类算法。对于参考和光谱聚类的历史，我们参考第 9 节。我们假设我们的数据包括*n*“点” $X_{1个}$*,. . . ， X_n*可以是任意对象。我们测量它们的成对相似性*秒$_{ij}$=秒(X-世， Xj)* 通过一些对称且非负的相似性函数，我们将相应的相似性矩阵表示为*小号*= (*秒$_{ij}$*) $_{i,j=1...n}$*.*

非归一化谱聚类

输入：相似度矩阵*小号*∈R$_{n×n}$， 数字*k*要构建的集群。
- 通过第 2 节中描述的方法之一构建相似图。作为其加权邻接矩阵。 让*W*
- 计算非标准化拉普拉斯算子*大号.*
- 计算第一个*k*特征向量*你*1个*,. . . , 你*$_k$的*大号.*
- 让*ü*∈R$_{n×k}$是包含向量的矩阵*你*1个*,. . . , 你*$_k$作为列。
- 为了 *一世*=1个*,. . . , ñ*， 让*是*-世∈R$_k$是对应于 *一世*-第行*ü.*
- 聚集点（*是*-世)-世=1,...,*n*在R$_k$与 *k*-means 算法成簇 *C*1个*,. . . ， C$_k$.

输出：集群*一个*1个*,. . . ， 一个k*和 *一个*-世=*{j| 是*$_j$∈*C*-世*}.*

归一化谱聚类有两种不同的版本，具体取决于哪个归一化的

使用图拉普拉斯算子。我们以两篇热门论文命名这两种算法，更多参考资料和历史请参阅第 9 节。

根据 Shi 和 Malik (2000) 的归一化光谱聚类

输入：相似度矩阵$小号\in\mathrm{R}^{n\times n}$，数字$k$要构建的集群。

- 通过第 2 节中描述的方法之一构建相似图。作为其加权邻接矩阵。让W

- 计算非标准化拉普拉斯算子$大号$.
- 计算第一个k广义特征向量$你_{1个},\ldots,你_k$广义特征问题$鲁=\lambda杜$.

- 让$ü\in\mathrm{R}^{n\times k}$是包含向量的矩阵$你_{1个},\ldots,你_k$作为列。
- 为了$一世$=1个$,\ldots,ñ$，让$是_{-世}\in\mathrm{R}^k$是对应于$一世$-第行$ü$.
- 聚集点（$是_{-世}$）$_{-世=1,\ldots,n}$在R^k与k-means 算法成簇$C_{1个},\ldots,C_k$.

输出：集群$一个_{1个},\ldots,一个_k$和$一个_{-世}=\{j|是_j\in C_{-世}\}$.

请注意，此算法使用的广义特征向量$大号$, 根据命题 3 对应于矩阵的特征向量$大号_{写字}$. 所以实际上，该算法适用于归一化拉普拉斯算子的特征向量$大号_{写字}$，因此称为归一化谱聚类。下一个算法也使用归一化的拉普拉斯算子，但这次矩阵$大号_{符号}$代替$大号_{写字}$. 正如我们将看到的，该算法需要引入其他算法不需要的额外行规范化步骤。原因将在第 7 节中变得清晰。

根据 Ng、Jordan 和 Weiss (2002) 的规范化光谱聚类

输入：相似度矩阵$小号\in\mathrm{R}^{n\times n}$，数字$k$要构建的集群。

- 通过第 2 节中描述的方法之一构建相似图。作为其加权邻接矩阵。让W

- 计算归一化拉普拉斯算子$大号_{符号}$.
- 计算第一个k特征向量$你_{1个},\ldots,你_k$的$大号_{符号}$.
- 让$ü\in\mathrm{R}^{n\times k}$是包含向量的矩阵$你_{1个},\ldots,你_k$作为列。
- 形成矩阵$吨\in\mathrm{R}^{n\times k}$ $\sum_{我知道}$从$ü$通过将行规范化为范数 1，那是设定的$吨_{ij}=你_{ij}/(_k你_{2个}$ $_{1个,2个.}$
- 为了$一世$=1个$,\ldots,ñ$，让$是_{-世}\in\mathrm{R}^k$是对应于$一世$-第行$吨$.
- 聚集点（$是_{-世}$）$_{-世=1,\ldots,n}$与k-means 算法成簇$C_{1个},\ldots,C_k$. 输出：集群$一个_{1个},\ldots,一个_k$和$一个_{-世}=\{j|是_j\in C_{-世}\}$.

除了使用三种不同的图拉普拉斯算子之外，上述所有三种算法看起来都非常相似。在这三种算法中，主要技巧是改变抽象数据点的表示$X_{-世}$到点$是_{-世}\in\mathrm{R}^k$. 由于图拉普拉斯算子的特性，这种表示形式的变化很有用。我们将在下一节中看到这种表示形式的变化增强了数据中的聚类属性，因此可以在新表示形式中轻松检测到聚类。特别是，简单的k-意味着聚类算法可以毫无困难地检测这种新表示中的聚类。不熟悉的读者k-意味着可以在很多地方阅读这个算法

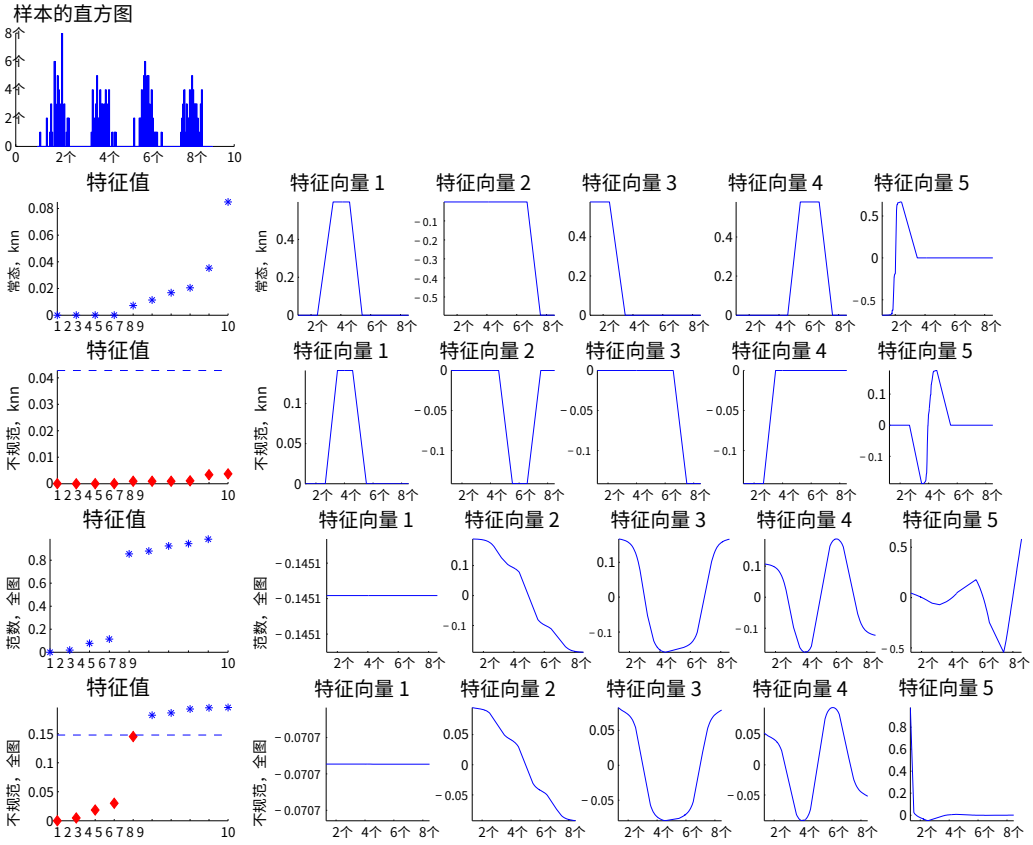

图 1：光谱聚类的玩具示例，其中数据点是从四个高斯分布的混合中提取的R。左上角：数据的直方图。第一行和第二行：的特征值和特征向量*大号*写字和*大号*基于*k*- 最近邻图。第三行和第四行：的特征值和特征向量*大号*写字和*大号*基于全连接图。对于所有图，我们使用高斯核*σ*= 1 作为相似函数。有关详细信息，请参阅文本。

教科书，例如 Hastie、Tibshirani 和 Friedman (2001)。

在我们深入研究谱聚类理论之前，我们想通过一个非常简单的玩具示例来说明其原理。这个例子将在本教程的几个地方使用，我们选择它是因为它非常简单，相关量很容易绘制出来。这个玩具数据集由 200 个点的随机样本组成*X*1个,. . . ， *X*200 ∈R根据四个高斯的混合绘制。图 1 的第一行显示了从该分布中抽取的样本的直方图（*X*-axis 表示一维数据空间）。作为该数据集的相似度函数，我们选择高斯相似度函数*秒*(*X*–世，*X*j) = 指数(−|*X*–世−*X*j|2个/(2*σ*2个)） 和*σ*= 1. 全连接图和10-最近邻图。

作为相似图，我们考虑了在图 1 中我们显示了第一个特征值

和非标准化拉普拉斯算子的特征向量*大号*和归一化的拉普拉斯算子*大号*写字. 也就是说，在我们绘制的特征值图中一世对比*λ*–世(暂时忽略图中未归一化情况下的虚线和特征值的不同形状；它们的含义将在第 8.5 节中讨论）。在特征向量的特征向量图中*你*= (*你*1个,. . . , *你*200) 我们密谋*X*–世对比*你*–世(请注意，在所选示例中*X*–世只是一个实数，因此我们可以在*X*-轴）。图 1 的前两行显示了基于 10 最近邻图的结果。我们可以看到前四个特征值

为 0，并且 corresponding eigenvector 利是ter 指标向量。原因是集群

在 10 最近邻图中形成断开连接的部分，在这种情况下，特征向量如命题 2 和 4 中给出。接下来的两行显示了完全连接图的结果。由于高斯相似函数始终为正，因此该图仅由一个连通分量组成。因此，特征值 0 的重数为 1，第一个特征向量是常数向量。以下特征向量携带有关集群的信息。例如，在未归一化的情况下（最后一行），如果我们将第二个特征向量阈值设置为 0，则 0 以下的部分对应于聚类 1 和 2，而 0 以上的部分对应于聚类 3 和 4。类似地，对第三个特征向量进行阈值分离集群 1 和 4 来自集群 2 和 3，并且对第四个特征向量进行阈值处理将集群 1 和 3 与集群 2 和 4 分开。总而言之，前四个特征向量包含有关四个簇的所有信息。在此图中说明的所有情况下，谱聚类使用*k*-means 在前四个特征向量上很容易检测到正确的四个簇。

5 图切点

聚类的直觉是根据点的相似性将点分成不同的组。对于以相似图形式给出的数据，这个问题可以重述如下：我们想要找到图的一个分区，使得不同组之间的边具有非常低的权重（这意味着不同集群中的点与彼此）并且组内的边具有高权重（这意味着同一簇内的点彼此相似）。在本节中，我们将看到如何将谱聚类导出为此类图划分问题的近似值。

给定一个具有邻接矩阵的相似图*W*, 最简单直接的构造分区的方法 Σ 该图是为了解决mincut问题。要定义它，请记住符号 *W*(*甲*、*乙*) :=
$-世\in,j\in乙 w_{ij}$和 *一个*为了补充 *一个*. 对于给定的数字 *k*的子集，
mincut 方法仅在于选择一个分区 $一个_{1个},\ldots,$ *一个*$_k$最小化

$$切（一个_{1个},\ldots,一个_k) := \frac{1个^k\Sigma}{2个}\ W(AA)_{-世,-世.}$$

这里我们引入因子1/2 为了符号的一致性，否则我们将在切割中计算每条边两次。特别是对于*k*=2，mincut 是一个相对容易的问题，可以有效地解决，参见 Stoer 和 Wagner (1997) 及其中的讨论。然而，在实践中，它通常不会导致令人满意的分区。问题在于，在许多情况下，mincut 的解决方案只是将一个单独的顶点与图中的其余部分分开。当然这不是我们想要在聚类中实现的，因为聚类应该是相当大的点组。绕过这个问题的一种方法是明确要求集合 $一个_{1个},\ldots,$ *一个*$_k$是“相当大”的。对此进行编码的两个最常见的目标函数是 RatioCut（Hagen 和 Kahng，1992）和归一化切割 Ncut（Shi 和 Malik，2000）。在 RatioCut 中，子集的大小 *一个*图的数量是通过它的顶点数来衡量的 /*一个*/，而在 Ncut 中，尺寸是通过其边缘 vol(*一个*). 定义是：

$$比例切割(一个_{1个},\ldots,一个_k) := \frac{1个^k\Sigma}{2个}\ \frac{W(一个_{-世},\ -一个_{-世})}{/一个/} = \sum_{-世=}^{k} \frac{切（一个_{-世,}\ 一个_{-世})}{/一个_{-世}/}$$

$$切割(一个_{1个},\ldots,一个_k) := \frac{1个^k\Sigma}{2个}\ \frac{W(一个_{-世,}\ 一个_{-世})}{体积（一个_{-世}）} = \sum_{-世=1}^{k} \frac{切（一个_{-世,}\ 一个_{-世})}{体积（一个_{-世}）}.$$

请注意，这两个目标函数 Σ 如果集群取一个小值 *一个*$_{-世}$不是太小。特别是-
ular，函数的最小值 $\sum_{-世=1}^{k}(1/$/*一个*$_{-世}$/) 如果全部实现/*一个*$_{-世}$/重合，并且最小的 $\sum_{-世=1}^{k}(1/$体积（*一个*$_{-世}$）) 如果所有 vol(*一个*$_{-世}$）重合。所以这两个目标函数试图实现的是集群是“平衡的”，分别通过顶点数或边权重来衡量。不幸的是，引入平衡条件使得之前解决mincut问题变得简单

成为 NP hard，参见 Wagner 和 Wagner (1993) 的讨论。谱聚类是解决这些问题的轻松版本的一种方法。我们将看到松弛 Ncut 导致归一化谱聚类，而松弛 RatioCut 导致非归一化谱聚类（另请参见 Ding (2004) 的教程幻灯片）。

5.1 近似 RatioCut k=2个

让我们从 RatioCut 的案例开始，k=2，因为放松在这种设置下最容易理解。我们的目标是解决优化问题

$$\min_{一个\subset V} 切割（一个,\bar{一个}）. \tag{1}$$

我们首先以更方便的形式重写问题。给定一个子集 $一个\subset V$我们定义向量 F= ($F_{1个}$,. . . ， F_n) $'\in \mathbb{R}^n$有条目

$$F_{一世}= \begin{cases} \sqrt{\frac{|\bar{一个}|}{|一个||一个|}} & 如果v_{一世}\in 一个 \\ -\sqrt{\frac{|一个|}{|\bar{一个}||一个|}} & 如果v_{一世}\in \bar{一个}。 \end{cases} \tag{2}$$

现在，可以使用非标准化图拉普拉斯算子方便地重写 RatioCut 目标函数。这是由于以下计算：

$$\begin{aligned} F'如果= & \frac{1个}{2个}\sum_{i,j=1} w_{ij}(F_{一世}-F_j)^{2个} \\ = & \frac{1个}{2个}\sum_{一世\in,j\in\bar{一个}} w_{ij}\left(\sqrt{\frac{|\bar{一个}|}{|一个|}}+\sqrt{\frac{|一个|}{|\bar{一个}|}}\right)^{2个} + \frac{1个}{2个}\sum_{一世\in,j\in 一个} w_{ij}\left(-\sqrt{\frac{|\bar{一个}|}{|一个|}}+\sqrt{\frac{|一个|}{|\bar{一个}|}}\right)^{2个} \\ = & 切（一个,\bar{一个}）\left(\frac{|\bar{一个}|}{|一个|}+2+\frac{|一个|}{|\bar{一个}|}\right) \\ = & 切（一个,\bar{一个}）\left(\frac{|一个|+|\bar{一个}|}{|一个|}+\frac{|一个|+|\bar{一个}|}{|\bar{一个}|}\right) \\ = & |五|\cdot 比例切割(一个,\bar{一个}). \end{aligned}$$

此外，我们有

$$\sum_{一世=1}^{n} F_{一世}= \sum_{一世\in 一个}\sqrt{\frac{|\bar{一个}|}{|一个|}} - \sum_{一世\in\bar{一个}}\sqrt{\frac{|一个|}{|\bar{一个}|}} = \sqrt{|一个|}\sqrt{|\bar{一个}|} - \sqrt{|\bar{一个}|}\sqrt{|一个|} = 0.$$

换句话说，向量F如等式（2）中所定义的正交于常数一向量1. 最后，请注意F满足

$$\|F\|^{2个} = \sum_{一世=1}^{n} F_{一世}^{2个} = |一个|\frac{|\bar{一个}|}{|一个|}+|\bar{一个}|\frac{|一个|}{|\bar{一个}|}+2|一个||\bar{一个}| = |一个|+|\bar{一个}|+2|一个||\bar{一个}| = 名词$$

总而言之，我们可以看到最小化 (1) 的问题可以等价地重写为

$$\min_{一个\subset V} 分钟F'如果受制于F\perp 1个，\ F_{一世}如方程式中所定义。(2),\ \|F\|=名词\sqrt{n} \tag{3}$$

这是一个离散优化问题，作为解向量的条目F只允许取两个特定的值，当然它仍然是 NP 难的。在这个设置中最明显的放松是

丢弃离散条件，而是允许$F_{-世}$取任意值R。这导致松弛的优化问题

$$\min_{F\in\mathbb{R}^{n}} F \cdot \text{如果受制于} F\perp 1\text{个}, \|F\|=\sqrt{\text{名词}} \tag{4}$$

根据 Rayleigh-Ritz 定理（例如，参见 Lütkepohl，1997 年的第 5.5.2 节），可以立即看出该问题的解由向量给出F这是对应于第二小特征值的特征向量$大号$(回想一下，最小的特征值$大号$为 0 与特征向量1). 所以我们可以通过的第二个特征向量来近似 RatioCut 的最小值$大号$. 然而，为了获得图的分区，我们需要重新转换实值解向量F将松弛问题转化为离散指标向量。最简单的方法是使用符号F作为指标函数，即选择

$$\begin{cases} v_{-世}\in \text{一个} & \text{如果} F_{-世}\geqslant 0 \\ v_{-世}\in \overline{\text{一个}} & \text{如果} F_{-世}<0. \end{cases}$$

但是，特别是在k >2 下面处理，这个启发式太简单了。大多数谱聚类算法所做的是考虑坐标$F_{-世}$作为点R并将他们分成两组C、C由k-means聚类算法。然后我们将得到的聚类结果传递给底层数据点，也就是我们选择

$$\begin{cases} v_{-世}\in \text{一个} & \text{如果} F_{-世}\in C \\ v_{-世}\in \overline{\text{一个}} & \text{如果} F_{-世}\in \overline{C}_{\circ} \end{cases}$$

这正是*非归一化谱聚类*算法的情况k=2.

5.2 任意近似 RatioCut$k$

一般值情况下RatioCut最小化问题的松弛k遵循与上述类似的原则。给定一个分区 I进入k套 一个$_{1个}$,. . . ， 一个$_{k}$, 我们定义k指标向量 H_{j}= ($H_{1个,j}$,. . . ， $H_{n,j}$) 经过

$$H_{i,j}=\begin{cases} 1个/\sqrt{|\text{一个}_{j}|} & \text{如果} v_{-世}\in \text{一个}_{j} \\ 0 & \text{否则} \end{cases} \qquad (\text{一世}=1个,\ldots,\tilde{n};j=1个,\ldots,k). \tag{5}$$

然后我们设置矩阵$H\in\mathbb{R}^{n\times k}$作为包含那些的矩阵$k$指标向量作为列。观察中的列$H$彼此正交，即$H \cdot H=我$. 在上一节的计算中我们可以看到 类似于

$$H_{-世}Lh_{-世}=\frac{\text{切}(\text{一个}_{-世},\overline{\text{一个}_{-世}})}{|\text{一个}_{-世}|}.$$

此外，可以检查

$$H_{-世}Lh_{-世}=(H \cdot \text{左手})_{二}.$$

结合我们得到的这些事实

$$\text{比例切割}(\text{一个}_{1个},\ldots,\text{一个}_{k})=\sum_{\text{一世}=1}^{k} H_{-世}Lh_{世}=\sum_{\text{一世}=1}^{k}(\text{高低})_{二}=\mathrm{Tr}(H \cdot \text{左手}),$$

其中 Tr 表示矩阵的迹。所以最小化 RatioCut(一个1个,. . . ，一个k) 可以改写为

$$\underset{一个1个，\ldots，一个k}{分钟}\mathrm{Tr}(H\ 'left\ hand')\ 受H\ 'H=我，H如方程式中所定义。(5).$$

与上面类似，我们现在通过允许矩阵的条目来放松问题H取任意实数值。那么松弛的问题就变成了：

$$\underset{H\in\mathbb{R}n\times k}{分钟}\mathrm{Tr}(H\ 'left\ hand')\ 受H\ 'H=我。$$

这是迹最小化问题的标准形式，Rayleigh-Ritz 定理的一个版本（例如，参见 Lütkepohl，1997 年的第 5.2.2.(6) 节）告诉我们解决方案是通过选择 H作为包含第一个的矩阵k的特征向量大号作为列。我们可以看到矩阵 H实际上是矩阵u在第 4 节中描述的非归一化谱聚类算法中使用。我们再次需要将实值解矩阵重新转换为离散分区。如上所述，标准方法是使用k- 表示行上的算法ü. 这导致了第 4 节中介绍的通用非归一化谱聚类算法。

5.3 逼近Ncut

与用于 RatioCut 的技术非常相似的技术可用于导出归一化光谱聚类作为最小化 Ncut 的松弛。在这种情况下k=2 我们定义聚类指标向量F经过

$$F_{一世}=\begin{cases}\sqrt{\frac{体积（一个）}{音量一个}} & 如果v_{一世}\in一个\\ -\sqrt{\frac{体积（一个）}{体积（一个）}} & 如果v_{我}\in世个。\end{cases}\qquad(6)$$

与上面类似，可以检查 (Df) 1 =0,F 'Df=体积（V),和F '如果=体积（V)切割(一个,一个). 因此我们可以用等价问题重写最小化 Ncut 的问题

$$\underset{一个}{分钟}F\ '如果受制于F如（6），自由度⊥1个，\ F\ 'Df=体积（V).\qquad(7)$$

我们再次通过允许放松问题F取任意实数值：

$$\underset{F\in\mathbb{R}n}{分钟}F\ '如果受制于Df⊥1个，\ F\ 'Df=体积（V).\qquad(8)$$

现在我们代入G:= 丁1个/2个F. 替换后，问题是

$$\underset{G\in\mathbb{R}n}{分钟}G\ '丁{-1个/2个}低密度脂蛋白{-1个/2个}G受制于G⊥丁1个/2个1个,\ \|G\|2个=体积（V).\qquad(9)$$

观察那个 丁-1个/2个低密度脂蛋白-1个/2个=大号符号, 丁1个/2个1个是的第一个特征向量大号符号, 和体积 (V)是一个常数。因此，问题 (9) 是标准 Rayleigh-Ritz 定理的形式，及其解G由第二个特征向量给出大号符号. 重新代入F= 丁-1个/2个G并使用命题 3 我们看到 F是的第二个特征向量大号写字，或等效的广义特征向量鲁=λ杜.

对于发现的情况k >2个集群，我们定义指标向量Hj= (H1个,j,. . . ， Hn,j) 经过

$$H_{i,j}=\begin{cases}\sqrt{\frac{1个}{体积（一个）}} & 如果v_{一世}\in一个j\\ 0 & 否则\end{cases}\qquad(一世=1个,\ldots,ñ;j=1个,\ldots,k).\qquad(10)$$

图 2：Guattery 和 Miller (1998) 的蟑螂图。

然后我们设置矩阵H作为包含那些的矩阵k指标向量作为列。观察那个
H 'H=我,H '一世Dh一世=1，和H ' 一世Lh一世=切（一个一世，一个一世)/体积（一个一世). 所以我们可以写出最小化问题切割为

$$\underset{\text{一个}_1,\ \ldots,\ \text{一个}_k}{\text{分钟}}\mathrm{Tr}(H \text{ '左手}) \text{ 受} H \text{ '卫生署=我，} \mathrm{H} \text{ 如 (10).}$$

放宽离散条件并代入吨= 丁1个/2个H我们得到了松弛的问题

$$\underset{\text{吨}\in\mathbb{R}^{n\times k}}{\text{分钟}}\mathrm{Tr}(\text{吨 '丁}^{-1\text{个}/2\text{个}}\text{低密度脂蛋白}^{-1\text{个}/2\text{个}}\text{吨}) \text{ 受吨 '吨=我。} \tag{11}$$

同样，这是由矩阵解决的标准迹线最小化问题吨其中包含第一个k的特征向量大号符号作为列。重新代入H= 丁-1个/2个吨并使用命题 3 我们看到解决方案H由第一个组成k矩阵的特征向量大号写字, 或第一个k的广义特征向量鲁=λ杜. 这产生了根据 Shi 和 Malik (2000) 的归一化谱聚类算法。

5.4 对放松方法的评论

关于谱聚类的这种推导，我们应该做出几点评论。最重要的是，与精确解相比，松弛问题的解的质量没有任何保证。也就是说，如果一个1个,. . . ， 一个k是最小化 RatioCut 的精确解，并且 乙1个,. . . ,乙k是非归一化谱聚类构造的解，则 RatioCut(乙1个,. . . ,乙k)－比例切割(一个1个,. . . ， 一个k) 可以任意大。在 Guattery 和 Miller (1998) 中可以找到这方面的几个例子。例如，作者考虑了一类非常简单的图，称为“蟑螂图”。这些图基本上看起来像一个梯子，去掉了一些边缘，见图 2。显然，理想的 RatioCut$k$=2 只是通过垂直切割来切割梯子，这样 一个={v1个,. . . , vk, v2个k+1,. . . , v3个k}和 一个={vk+1,. . . , v2个k, v3个k+1,. . . , v4个k}. 这种剪裁与/一个|=|一个|=2个k和切（一个,一个) = 2. 然而，通过研究蟑螂图的非归一化图拉普拉斯算子的第二个特征向量的性质，作者证明了非归一化谱聚类总是通过阶梯水平切割，构造集合乙={v1个,. . . , v2个k}和 乙={v2个k+1,. . . , v4个k}. 这也会导致平衡切割，但现在我们切割k边而不仅仅是 2。所以 RatioCut(一个,一个) = 2/k, 而 RatioCut(乙,乙) = 1。这意味着相对于最优切割，谱聚类得到的RatioCut值为 千/差 2 倍，这是顺序中的一个因素 n. 其他几篇论文研究了由谱聚类构建的聚类的质量，例如 Spielman 和 Teng (1996)（针对非归一化谱聚类）和 Kannan、Vempala 和 Vetta (2004)（针对归一化谱聚类）。通常，已知不存在将平衡图切割近似为常数因子的有效算法。相反，这个近似问题本身可能是 NP 问题（Bui 和 Jones，1992）。

当然，我们上面讨论的松弛并不是唯一的。例如，Bie 和 Cristianini (2006) 推导了导致半定规划的完全不同的松弛，可能还有许多其他有用的松弛。谱弛豫之所以如此吸引人，并不是因为它能带来特别好的解决方案。它的流行主要是因为它导致了一个标准的线性代数问题，它很容易解决。

6 随机游走的观点

解释谱聚类的另一个论据是基于相似度图上的随机游走。图上的随机游走是一个从一个顶点随机跳到另一个顶点的随机过程。我们将在下面看到，谱聚类可以解释为试图找到图的分区，使得随机游走在同一簇内停留很长时间，并且很少在簇之间跳跃。直觉上这是有道理的，特别是结合上一节的图切割解释：具有低切割的平衡分区也将具有随机游走在集群之间跳跃的机会不多的特性。对于一般随机游走的背景阅读，我们参考 Norris (1997) 和 Brémaud (1999)，对于图上的随机游走，我们推荐 Aldous 和 Fill（准备中）和 Lovász (1993)。正式地，$v_{一世}$到顶点v_j与边权重成正比w_{ij}并由$p_{ij}:=w_{ij}/d_{一世}$. 转移矩阵$P=(p_{ij})_{i,j=1,\dots,n}$因此，随机游走的定义为

$$P=丁^{-1个}W_{。}$$

如果图是连通的且非二分的，则随机游走总是具有唯一的平稳分布$\pi=(\pi_{1个},\dots,\pi_n)$ ‘，在哪里$\pi_{一世}=d_{一世}/$体积（V).显然两者之间存在紧密的关系 $大号_{写字}$和P,作为$大号_{写字}=我-P$.作为结果，λ是的特征值$大号_{写字}$与特征向量$你$当且仅当 1-λ是的特征值P与特征向量$你$. 众所周知，图的许多属性可以用相应的随机游走转移矩阵来表示P,参见 Lovász (1993) 的概述。从这个角度来看，最大的特征向量并不奇怪P和最小的特征向量$大号_{写字}$可用于描述图的簇属性。

随机游走和 Ncut

Meila 和 Shi (2001) 观察到 Ncut 和随机游走的转移概率之间的形式等价性。

提案 5 (切割通过转移概率）*让G连接且非双向。假设我们运行随机游走*$(X_{吨})_{吨\in否}$*从...开始*X_0*在平稳分布* π *中。不相交的子集甲、乙⊂,表示为*$P(B|A):=P(X_{1个}\in B|X_0\in一个)$*. 然后：* *作为-为了*

$$切割(一个,\overline{一个})=P(一个|\overline{一个})+P(一个|一个).\overline{}$$

*证明。*首先观察到

$$P(X_0\in斧头_{1个}\in乙)=\sum_{一世\in,j\in乙}P(X_0=i,X_{1个}=j)=\sum_{一世\in,j\in乙}\pi_{一世}p_{ij}$$

$$=\sum_{一世\in,j\in乙}\frac{d_{一世}w_{ij}}{体积（V）d_{一世}}=\frac{1个}{体积（V）}\sum_{一世\in,j\in乙}w_{ij}.$$

使用这个我们得到

$$P(X_{1个} \in B | X_0 \in 一个) = \frac{P(X_0 \in 斧头_{1个} \in 乙)}{P(X_0 \in 一个)}$$

$$= - \frac{1个}{体积（V）} \sum_{一世 \in ,j \in 乙} w_{ij} - \frac{(\quad)}{\text{-体积（一个）}} - 1个 = \frac{\sum_{一世 \in ,j \in 乙} w_{ij}}{体积（一个）}.$$

现在命题直接遵循 Ncut 的定义。 2个

这个命题导致对 Ncut 的一个很好的解释，因此对归一化谱聚类。它告诉我们，当最小化 Ncut 时，我们实际上是在图形中寻找切割，这样随机游走很少从一个至一个反之亦然。

通勤距离

随机游走和图拉普拉斯算子之间的第二个联系可以通过图上的通勤距离来建立。通勤距离（也称为阻力距离）C_{ij}两个顶点之间 $v_{一世}$和 v_j是随机游走从顶点行进所需的预期时间$v_{一世}$到顶点v_j和返回（Lovász，1993 年；Aldous 和 Fill，准备中）。通勤距离有几个很好的特性，这使得它对机器学习特别有吸引力。与图上的最短路径距离相反，如果有许多不同的从顶点到达的短路径，则两个顶点之间的通勤距离会减少$v_{一世}$到顶点v_j. 因此，通勤距离不是仅仅寻找一条最短路径，而是着眼于一组最短路径。在图中由短路径连接且位于图中相同高密度区域的点被认为比由短路径连接但位于图中不同高密度区域的点彼此更接近。从这个意义上说，通勤距离似乎特别适合用于聚类目的。

值得注意的是，可以借助广义逆（也称为伪逆或 Moore-Penrose 逆）计算图上的通勤距离$大号_{\dagger}$图拉普拉斯算子$大号$. 下面我们记$电子_{一世}$= (0,. . .0,1个,0,. . . ,0) 作为$一世$-th 单位向量。定义广义逆$大号$, 回想一下命题 1 的矩阵$大号$可以分解为$大号$=$ü$Λ$ü$ 在哪里$ü$是包含所有特征向量作为列的矩阵，Λ 是具有特征值的对角矩阵$\lambda_{1个}, . . . , \lambda_n$在对角线上。由于至少一个特征值是 0，矩阵$大号$不可逆。相反，我们定义它的广义

逆为$大号_{\dagger}$:=$ü$Λ$_{\dagger}ü$ 其中矩阵Λ$_{\dagger}$是对角元素为 1 的对角矩阵$/\lambda$ 和 0 如果λ=$_{世}$ 图表系列$大号_{\dagger}$可以计算为$升_{\dagger}$ $_{ij}$= $\sum_{k=2}^{n} \frac{1个}{\lambda_k}$ $_{一世如果\lambda_{一世} 6个= 0}$ $_{伊我知道你}$$_{jk}$. 矩阵$大号_{\dagger}$是积极的半定和对称。对于进一步的属性$大号_{\dagger}$参见 Gutman 和 Xiao (2004)。

提案 6（通勤距离）*让G*= (*V*,*E*)*连通的无向图。表示为C*$_{ij}$ *顶点之间的通勤距离v*$_{一世}$*和顶点v*$_j$, *通过大号*$_{\dagger}$= (*升*$_{\dagger}$ $_{ij}$)$_{i,j=1,...,n}$*广义逆的L.然后我们有：*

$$C_{ij} = 体积（V）(升_{\dagger 二-2个升_{ij}} + 升_{伯伯}) = 卷(V)(电子_{一世-电子}) \cdot 大号_{\dagger}(电子_{一世-电子}).$$

这一结果已由 Klein 和 Randic (1993) 发表，其中已通过电气网络理论的方法得到证明。有关使用第一步分析进行随机游走的证明，请参阅 Fouss、Pirotte、Renders 和 Saerens (2007)。在图拉普拉斯算子的帮助下，还存在其他表达通勤距离的方法。例如，根据归一化拉普拉斯算子的特征向量的方法$大号_{符号}$

可以在 Lovász (1993) 中作为推论 3.2 找到，以及一种借助于某些子矩阵的行列式计算通勤距离的方法$大号$可以在 Bapat、Gutman 和 Xiao (2003) 中找到。

提案 6 具有重要的意义。这表明C_{ij}可以被认为是欧几里得 $\sqrt{\quad}$ 图的顶点上的距离函数。这意味着我们可以构建一个嵌入

映射顶点$v_{世}$点图的$z_{世}\in\mathrm{R}_{n}$这样点之间的欧几里得距离$z_{世}$与图表上的通勤距离一致。这工作如下。作为矩阵$大号_{ŧ}$
是半正定和对称的，它在R_{n}(或者更正式地说，它在子空间上导出一个内积R_{n}垂直于矢量1).现在选择 $z_{世}$作为点R_{n}对应于一世-矩阵的第行$ü(\Lambda_{ŧ})_{1个/2个}$. 那么，根据命题 6

并通过建造$大号_{ŧ}$我们有那个〈$z_{世}, z_{j}$〉=电子‘ $_{世}大号_{ŧ}$电子和C_{ij}=体积（V)//$z_{世}-z_{j}$//$_{2个}$.

非归一化谱聚类中使用的嵌入与通勤时间嵌入有关，但不完全相同。在谱聚类中，我们将图的顶点映射到行上$是_{世}$矩阵的 $ü$，而通勤时间嵌入映射行上的顶点$z_{世}$矩阵的 $(\Lambda_{ŧ})_{1个/2个}ü$. 也就是说，与条目相比$是_{世}$, 条目$z_{世}$由的反特征值另外缩放$大号$. 此外，在谱聚类中，我们只取第一个k矩阵的列，而通勤时间嵌入采用所有列。几位作者现在试图证明为什么$是_{世}$和$z_{世}$毕竟并没有太大的不同，并且有点放弃，事实是谱聚类基于欧几里德距离构建聚类$是_{世}$可以解释为基于通勤距离构建图中顶点的集群。但是，请注意，这两种方法可能有很大不同。例如，在图形包含的最佳情况下k断开连接的组件，第一个k的特征值$大号$根据命题 2 为 0，第一个k列的$ü$由聚类指标向量组成。然而，第一k矩阵的列 $(\Lambda_{ŧ})_{1个/2个}ü$仅由零组成，作为第一个$k\Lambda$的对角元素 $_{ŧ}$为 0。在这种情况下，第一个中包含的信息 k列的$ü$在矩阵中完全被忽略（$\Lambda_{ŧ})_{1个/2个}ü$, 以及矩阵的所有非零元素 $(\Lambda_{ŧ})_{1个/2个}ü$可以在列中找到k+1至n在光谱聚类中没有考虑，它会丢弃所有这些列。另一方面，如果底层图是连通的，则不会出现这些问题。在这种情况下，唯一具有特征值 0 的特征向量是常数 1 向量，在这两种情况下都可以忽略它。对应于小特征值的特征向量$\lambda_{世}$

的$大号$然后在矩阵中强调（$\Lambda_{ŧ})_{1个/2个}ü$因为它们乘以$\lambda_{ŧ}$ $_{世}$=1个/$\lambda_{世}$. 在这种情况下——通勤时间嵌入和频谱嵌入做类似的事情可能是真的。

总而言之，通勤时间距离似乎是一个有用的直觉，但如果不做进一步的假设，谱聚类与通勤距离之间的关系就相当松散。这些关系可能会收紧，例如，如果相似函数是严格正定的。然而，我们还没有看到关于这一点的精确数学陈述。

7 微扰理论观点

微扰理论研究矩阵的特征值和特征向量如何变化的问题一个如果我们添加一个小的扰动就会改变H, 即我们考虑扰动矩阵一个:=一个+H. 大多数微扰定理指出特征值或特征向量之间的一定距离一个和一个以常数乘以范数为界H. 该常数通常取决于我们正在查看的特征值，以及该特征值与光谱的其余部分分开的距离（有关正式声明，请参见下文）。谱聚类的理由如下：让我们首先考虑“理想情况”，其中簇间相似性恰好为 0。我们在第 3 节中看到，第一个 k的特征向量$大号$或者$大号_{写字}$是聚类的指标向量。在这种情况下，点$是_{世}\in\mathrm{R}_{k}$

在谱聚类算法中构建的形式为 (0,. . . ,0,1个,0,. . .0) 其中1的位置表示该点所属的连通分量。特别是，所有$是_{世}$属于同一个连通分量重合。这k-means 算法将通过在每个点上放置一个中心点 (0,. . . ,0,1个,0,. . .0) ‘ $\in\mathrm{R}_{k}$. 在“近乎理想的情况”中，我们仍然有不同的集群，但集群之间的相似性不完全为 0，我们认为拉普拉斯矩阵是理想情况下的扰动版本。然后微扰理论告诉我们，特征向量将非常接近理想的指标向量。积分$是_{世}$也许不会

完全符合（0,. . . ,0,1个,0,. . .0）；但这样做会导致一些小的误差项。因此，如果扰动不是太大，则*k*-means 算法仍然会将组彼此分开。

7.1 形式微扰论证

谱聚类扰动方法的正式基础是矩阵扰动理论中的戴维斯-卡汉定理。该定理限制了扰动下对称矩阵的特征空间之间的差异。我们陈述这些结果是为了完整性，但对于背景阅读，我们参考了 Stewart 和 Sun (1990) 的第 V 节和 Bhatia (1997) 的第 VII.3 节。在微扰理论中，子空间之间的距离通常使用“标准角”（也称为“主角”）来测量。要定义主角，让*V*1个和*V*2个是两个*p*-维子空间R*d*，和*V*1个和*V*2个两个矩阵使得它们的列形成标准正交系统*V*1个和*V*2个，分别。然后

余弦 cos Θ−世主角 Θ−世是的奇异值*V*‘ 1个*V*2个. 为了*p*=1、如此定义规范角与角的法线定义一致。也可以定义标准角，如果*V*1个和*V*2个不具有相同的维度，请参见 Stewart 和 Sun (1990) 的第 V 节、Bhatia (1997) 的第 VII.3 节或 Golub 和 Van Loan (1996) 的第 12.4.3 节。矩阵 sin Θ(*V*1个,*V*2个) 将表示对角矩阵，其正弦值位于对角线上。

定理 7（戴维斯-卡汉）*让啊*∈R*n*×*n是对称矩阵，让* ‖ · ‖ *分别是 Frobenius 范数或矩阵的双范数。考虑* Ã :=一个+*H作为一个扰动的版本一个。让小号*1个⊂R*成为一个区间。用*σ*表示小号*(一个)*的特征值集一个包含在小号*1个*, 通过*V*1个*对应于所有这些特征值的特征空间（更正式地说，*V*1个*是由* σ *引起的光谱投影的图像小号*1个*(一个)*). 用*σ*表示小号*1个*(一个)和*Ṽ1个*的类似数量。定义之间的距离小号*1个*和光谱一个在外面小号*1个*作为*

$$\delta = \text{分钟}\{|\lambda - \text{秒}|; \lambda \text{的特征值一个，} \lambda 6 \in \text{小号}_{1\text{个}}, \text{秒} \in \text{小号}_{1\text{个}}\}.$$

那么距离d(*V*1个, *Ṽ*1个) := ‖ sin Θ(*V*1个, *Ṽ*1个) ‖ *两个子空间之间V*1个*和Ṽ*1个*受限于*

$$d(V_{1\text{个}}, \tilde{V}_{1\text{个}}) \leqslant \frac{\|H\|}{\delta}.$$

有关该定理的讨论和证明，请参见 Stewart 和 Sun (1990) 的第 V.3 节。让我们尝试解密这个定理，为了简单起见，在非标准化拉普拉斯算子的情况下（对于标准化拉普拉斯算子，它的工作方式类似）。矩阵一个将对应于拉普拉斯图 *大号*在图表具有的理想情况下*k*连接的组件。矩阵一个对应于一个扰动的情况，其中由于噪声*k*图中的组件不再完全断开，而是仅由少数权重较低的边连接。我们将这种情况的相应图拉普拉斯算子表示为*L*. 对于谱聚类，我们需要首先考虑*k*的特征值和特征向量*L*. 表示的特征值*大号*经过*λ*1个,. . . *λn*和不安的拉普拉斯算子

*L*经过*λ̃*1个,. . . , *λ̃n*. 选择间隔*小号*1个现在是关键点。我们想这样选择它
都是第一个*k*的特征值*L*和第一个*k*的特征值*大号*包含在*小号*1个. 扰动越小越容易*H*=*L* − *L*本征隙越大|*λk*- *λk*+1|是。如果我们设法找到这样一个集合，那么 Davis-Kahan 定理告诉我们对应于第一个的特征空间*k*理想矩阵的特征值*大号*和第一个*k*扰动矩阵的特征值*L*彼此非常接近，也就是说它们的距离以 ‖ *H* ‖ /*δ*. 然后，由于理想情况下的特征向量在连通分量上是分段常数，因此在扰动情况下也大致如此。“大约”有多好取决于扰动的范数 ‖ *H* ‖ 和距离*δ*之间*小号*1个和 (*k*+1) 的特征向量*大号*. 如果设置*小号*1个已被选为区间[0, *λk*]，然后*δ*与光谱间隙重合|*λk*+1 -*λk*|. 从定理可以看出，这个特征间隙越大，理想情况和扰动情况的特征向量越接近，谱聚类效果越好。下面我们将看到本征间隙的大小也可以用在

不同的上下文作为谱聚类的质量标准，即在选择数字时*k*要构建的集群。

如果扰动*H*太大或特征间隙太小，我们可能找不到集合*小号*1个这样无论是第一个*k*的特征值*大号*和*L*包含在*小号*1个. 在这种情况下，我们需要通过选择集合来做出折衷*小号*1个包含第一个*k*的特征值*大号*, 但可能有一些或多或少的特征值*L*. 然后定理的陈述变得更弱，因为我们要么不比较对应于第一个的特征空间*k*的特征向量*大号*和*L*, 但与第一个对应的特征空间*k*的特征向量*大号*和第一个*k*的特征向量*L*（在哪里*k*是特征值的数量*L*包含在*小号*1个). 或者，它可能发生*δ*变得如此之小，以至于两者之间距离的界限*d*(*V*1个, *Ṽ*1个) 爆炸太多以至于它变得毫无用处。

7.2 关于微扰方法的评论

在使用微扰理论论据来证明基于矩阵特征向量的聚类算法时需要谨慎一些。一般来说，*任何*分块对角对称矩阵具有这样的特性，即存在特征向量的基，这些特征向量在各个块外为零，在块内为实值。例如，基于这个论点，几位作者使用了相似矩阵的特征向量*小号*或邻接矩阵*W*发现集群。然而，在完全分离的簇的理想情况下是块对角线可以被认为是成功使用特征向量的必要条件，但不是充分条件。至少还应满足两个属性：

首先，我们需要确保*命*令的特征值和特征向量是有意义的。在拉普拉斯算子的情况下，这总是正确的，因为我们知道任何连通分量都恰好拥有一个特征值为 0 的特征向量。因此，如果图有*k*连接的组件，我们采取第一个*k*拉普拉斯算子的特征向量，那么我们就知道每个分量只有一个特征向量。但是，对于其他矩阵可能并非如此，例如*小号*或者*W*. 例如，可能是块对角相似矩阵的两个最大特征值*小号*来自同一个街区。在这种情况下，如果我们采取第一个*k*的特征向量*小号*，一些块将出现多次，而另一些块我们将完全错过（除非我们采取某些预防措施）。这就是为什么使用特征向量的原因*小号*或者*W*不鼓励聚类。

第二个属性是，在理想情况下，组件上的特征向量的条目应该“安全地远离”0。假设第一个连接的组件上的特征向量有一个条目*你*1个，*一世*>0 在位置*一世*. 在理想情况下，此条目非零这一事实表明相应的点*一世*属于第一个集群。反过来，如果一个点*j*不属于集群1，那么在理想情况下应该是这样的*你*1个,*j*=0. 现在考虑相同的情况，但数据受到扰动。扰动的特征向量*û*通常不再有任何非零分量；但如果噪声不是太大，那么微扰理论告诉我们条目*û*1个，*一世*和 *û*1个,*j*仍然“接近”它们的原始值*你*1个，*一世*和*你*1个,*j*. 所以两个条目*û*1个，*一世*和*û*1个,*j*会取一些小的值，比如*ε*1个和*ε*2个. 实际上，如果这些值非常小，我们就不清楚我们应该如何解释这种情况。要么我们相信小条目*û*表示这些点不属于第一个集群（然后错误分类第一个数据点*一世*), 或者我们认为这些条目已经表明了类成员并将两个点都分类到第一个集群 (它错误地分类了点*j*).

对于两个矩阵*大号*和*大号*写字，理想情况下的特征向量是指示向量，所以不会出现上述第二个问题。然而，这对于矩阵来说是不正确的*大号*符号，用于 Ng 等人的归一化谱聚类算法。(2002)。即使在理想情况下，该矩阵的特征向量也为*了*1个/2个1个一个.如果顶点的度数相差很大，特别是如果存在度数很低的顶点，则特征向量中的对应项非常小。为了解决上述问题，Ng 等人的算法中的行归一化步骤。(2002) 开始发挥作用。在理想情况下，矩阵*ü*在算法中只有一个

每行非零条目。行归一化后，矩阵*吨*在 Ng 等人的算法中。(2002) 然后由聚类指标向量组成。但是请注意，这在实践中可能并不总是正确。假设我们有$\tilde{u}_{一世,\ 1个}=\varepsilon_{1个}$和$\sqrt{\tilde{u}_{一世,\ 2个}=\varepsilon_{2个}}$. 如果我们现在将一*世*-扔的$\ddot{u}$， 两个都$\varepsilon_{1个}$和$\varepsilon_{2个}$将乘以系数 $1/\varepsilon_{2个}$ $\overline{{}_{1个}+\varepsilon_{2个}}$并且变得相当大。我们现在遇到与上述类似的问题：两个点很可能被归类到同一个集群中，即使它们属于不同的集群。这个论点表明使用矩阵的谱聚类*大号*符号如果特征向量包含特别小的条目，则可能会出现问题。另一方面，请注意，特征向量中的此类小条目仅在某些顶点的度数特别低时才会出现（如特征向量*大号*符号由 $了_{1个/2个}1个_{一个}$).有人可能会争辩说，在这种情况下，数据点无论如何都应该被视为异常值，那么该点最终会出现在哪个集群中并不重要。 $_{一世}$

总而言之，结论是非归一化谱聚类和归一化谱聚类都具有*大号*写字微扰理论方法很好地证明了这一点。归一化谱聚类*大号*符号也可以用微扰理论来证明，但是如果图中包含度数非常低的顶点，则应该更加小心。

8 实用细节

在本节中，我们将简要讨论实际实施谱聚类时出现的一些问题。有几个选择和要设置的参数。但是，本节中的讨论主要是为了提高对发生的一般问题的认识。为了深入研究各种现实世界任务的谱聚类行为，我们参考了文献。

8.1 构建相似度图

为光谱聚类构建相似度图并不是一项微不足道的任务，并且对各种结构的理论含义知之甚少。

相似函数本身

在我们甚至可以考虑构建相似度图之前，我们需要在数据上定义一个相似度函数。因为我们稍后要构建一个邻域图，所以我们需要确保由这个相似函数导出的局部邻域是“有意义的”。这意味着我们需要确保被相似度函数认为“非常相似”的点在数据来源的应用程序中也密切相关。例如，在构建文本文档之间的相似度函数时，检查具有高相似度得分的文档是否确实属于同一文本类别是有意义的。相似性函数的全局“长程”行为对于谱聚类并不那么重要——两个数据点的相似性得分是 0.01 还是 0.001 并不重要，比如说，因为无论如何我们都不会在相似图中连接这两个点。在数据点存在于欧几里德空间中的常见情况下R_{d}, 一个合理的默认候选者是高斯相似函数$秒(X_{一世},\ X_{j})$ = 指数$(-\|X_{一世}-X_{j}\|_{2个}/(2\sigma_{2个}))$ (当然我们需要选择参数σ在这里，见下文）。归根结底，相似度函数的选择取决于数据来自的领域，无法给出一般性建议。

哪种类型的相似度图

必须做出的下一个选择涉及要使用的图形类型，例如*k*-最近的邻居或*ε*-邻域图。让我们使用图 3 中显示的玩具示例来说明不同图形的行为。作为基础分布，我们选择分布$R_{2个}$和

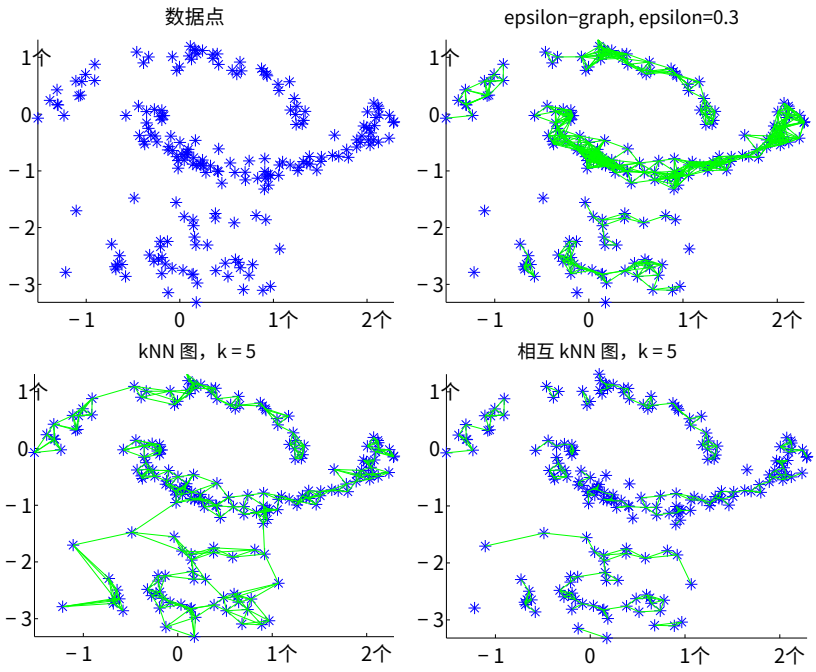

图 3：不同的相似度图，详见正文。

三个星团：两个“月亮”和一个高斯星团。底部卫星的密度被选择为大于顶部卫星的密度。图 3 中的左上面板显示了从该分布中提取的样本。接下来的三个面板显示了此样本的不同相似度图。

在里面ε-neighborhood graph，我们可以看到很难选择一个有用的参数ε. 和 ε=0.3 如图，中间moon上的点已经很紧密的连在一起了，而Gaussian上的点几乎没有连在一起。如果我们有“不同尺度”的数据，那么这个问题总是会发生，即数据点之间的距离在空间的不同区域是不同的。

这k- 另一方面，最近邻图可以连接“不同尺度”的点。我们可以看到低密度高斯中的点与高密度月球中的点相连。这是的一般属性k-最近的邻居图，这可能非常有用。我们还可以看到 k- 如果存在彼此相距相当远的高密度区域，则最近邻图可能会分成几个不相连的组件。本例中的两个卫星就是这种情况。

相互的k- 最近邻图具有这样的特性，即它倾向于连接恒定密度区域内的点，但不连接不同密度的区域。所以相互的 k- 最近邻图可以被认为是“介于”ε-邻域图和 k- 最近邻图。它能够在不同的尺度上起作用，但不会将这些尺度相互混合。因此，相互k- 如果我们想要检测不同密度的集群，最近邻图似乎特别适合。

全连接图经常与高斯相似函数结合使用 $秒(X_{-世}，\ X_j)$ = 指数$(-\|X_{-世}-X_j\|_{2个}/(2\sigma_{2个}))$. 这里的参数$\sigma$起着与参数相似的作用$\varepsilon$在里面$\varepsilon$-邻域图。局部邻域中的点与相对较高的权重相连，而 r 个远离点之间的边具有正的但可以忽略不计的权重。然而，由此产生的

F一个

相似矩阵不是稀疏矩阵。

作为一般性建议，我们建议与k- 最近邻图作为首选。使用起来很简单，导致稀疏邻接矩阵W，并且根据我们的经验，与其他图表相比，它更不容易受到参数选择不当的影响。

相似度图的参数

一旦决定了相似图的类型，就必须选择其连接参数 k或者ε， 分别。不幸的是，几乎没有任何理论结果可以指导我们完成这项任务。一般来说，如果相似度图包含的连通分量多于我们要求算法检测的聚类数量，那么谱聚类将平凡地将连通分量作为聚类返回。除非完全确定那些连接的组件是正确的集群，否则应该确保相似图是连接的，或者只包含“很少”的连接组件和很少或没有孤立的顶点。关于如何实现随机图的连通性有很多理论结果，但所有这些结果都只在样本量的限制下成立$n \to \infty$. 例如，众所周知，对于n从一些底层密度中提取的数据点具有连接支持R_d， 这k- 最近邻图和相互k-如果我们选择，最近邻图将被连接k按照日志的顺序（n）（例如，Brito、Chavez、Quiroz 和 Yukich，1997）。类似的论点表明参数ε在里面ε-邻域图必须选择为（$\log（n）/n)_d$以保证在极限范围内的连通性 (Penrose, 1999)。尽管具有理论上的意义，但所有这些结果并不能真正帮助我们进行选择k在有限样本上。

现在让我们给出一些经验法则。使用时k-最近邻图，那么应该选择连通性参数，使得生成的图是连通的，或者至少具有比我们想要检测的簇少得多的连通分量。对于小型或中型图形，可以“步行”进行尝试。对于非常大的图，第一个近似值可能是选择 k按照日志的顺序（n), 正如渐近连通性结果所建议的那样。

对于相互k- 最近邻图，我们不得不承认我们有点迷失了经验法则。互惠互利的优势k- 与标准图相比的最近邻图k- 最近邻图是它倾向于不连接不同密度的区域。虽然如果存在由单独的高密度区域引起的清晰聚类，这可能会很好，但在不太明显的情况下这可能会造成伤害，因为图中不连贯的部分将始终被谱聚类选择为聚类。一般而言，可以观察到相互k- 最近邻图的边比标准图少得多k- 相同参数的最近邻图k. 这建议选择k对共同体来说要大得多 k-最近的邻居图比一个人会做的标准k- 最近邻图。但是，要利用相互的属性k- 最近邻图不连接不同密度的区域，因此有必要允许图中有几个“有意义的”断开连接的部分。不幸的是，我们不知道有任何通用的启发式方法来选择参数k这样就可以实现。

为了ε-neighborhood graph，建议选择ε这样生成的图就安全地连接起来了。确定最小值ε图连接的地方很简单：必须选择 ε作为数据点上全连接图的最小生成树中最长边的长度。后者可以通过任何最小生成树算法轻松确定。但是，请注意，当数据包含异常值时，该启发式方法将选择ε如此之大以至于即使异常值也与其余数据相关联。当数据包含几个彼此相距很远的紧密集群时，会发生类似的效果。在这两种情况下，ε将选择太大以反映数据最重要部分的规模。

最后，如果将全连接图与可缩放的相似度函数一起使用

本身，例如高斯相似函数，那么应该选择相似函数的尺度，使得结果图具有与相应的相似属性*k*-最近的邻居或 *ε*-邻域图会有。需要确保对于大多数数据点，相似度显着大于 0 的邻居集“不太小也不太大”。特别是，对于高斯相似度函数，经常使用几个经验法则。例如，可以选择*σ*按照一个点到它的平均距离的顺序*k*-th 最近的邻居，其中*k*与上面类似地选择（例如，*k*~日志（*n*）+ 1 ）. 另一种方法是确定*ε*通过上面描述的最小生成树启发式，然后选择*σ*=*ε*. 但请注意，所有这些经验法则都是非常临时的，并且取决于手头的给定数据及其点间距离的分布，它们可能根本不起作用。

一般来说，经验表明谱聚类对相似图的变化及其参数的选择非常敏感。不幸的是，据我们所知，还没有系统的研究来调查相似性图及其参数对聚类的影响，并提出合理的经验法则。上述建议都没有坚实的理论基础。寻找具有理论依据的规则应被视为未来研究的一个有趣且重要的课题。

8.2 计算特征向量

要在实践中实现谱聚类，必须计算第一个*k*潜在大图拉普拉斯矩阵的特征向量。幸运的是，如果我们使用*k*-最近邻图或*ε*-邻域图，那么所有这些矩阵都是稀疏的。存在计算稀疏矩阵的第一个特征向量的有效方法，最流行的方法是幂方法或 Krylov 子空间方法，例如 Lanczos 方法（Golub 和 Van Loan，1996）。这些算法的收敛速度取决于特征间隙（也称为谱间隙）的大小*γk*=∣*λk*-*λk*+1∣. 这个 eigengap 越大，算法计算第一个的速度越快*k*特征向量收敛。

请注意，如果所考虑的特征值之一具有大于一的多重性，则会出现一般问题。例如，在理想情况下*k*断开连接的簇，特征值 0 具有多重性*k*. 正如我们所见，在这种情况下，特征空间由*k*聚类指标向量。但不幸的是，由数值特征求解器计算的向量不一定会收敛到那些特定的向量。相反，它们只是收敛到本征空间的某个标准正交基，并且通常取决于算法收敛到哪个基的实现细节。但是这个

毕竟还不错。请注意，∑*k*集群指示符向量跨越的空间中的所有向量1个 *一个–世*
有形式*你*= 对于一些系数*一个–世*，也就是说，它们在
集群。因此，特征求解器返回的向量仍然编码了有关聚类的信息，然后可以由*k*-意味着重建集群的算法。

8.3 簇数

选择号码*k*聚类问题是所有聚类算法的一个普遍问题，针对这个问题已经设计出各种或多或少成功的方法。在基于模型的聚类设置中，存在从数据中选择聚类数量的合理标准。这些标准通常基于数据的对数似然性，然后可以以频率论或贝叶斯方式处理，例如参见 Fraley 和 Raftery (2002)。在没有或很少对基础模型做出假设的情况下，可以使用大量不同的指标来选择聚类的数量。示例包括临时测量，例如集群内和集群间相似性的比率、信息论标准（Still 和 Bialek，2004 年）、差距统计（Tibshirani、Walther 和 Hastie，2001 年），以及稳定性方法（宾虚，Elisseeff 和 Guyon，2002 年；兰格，罗斯，

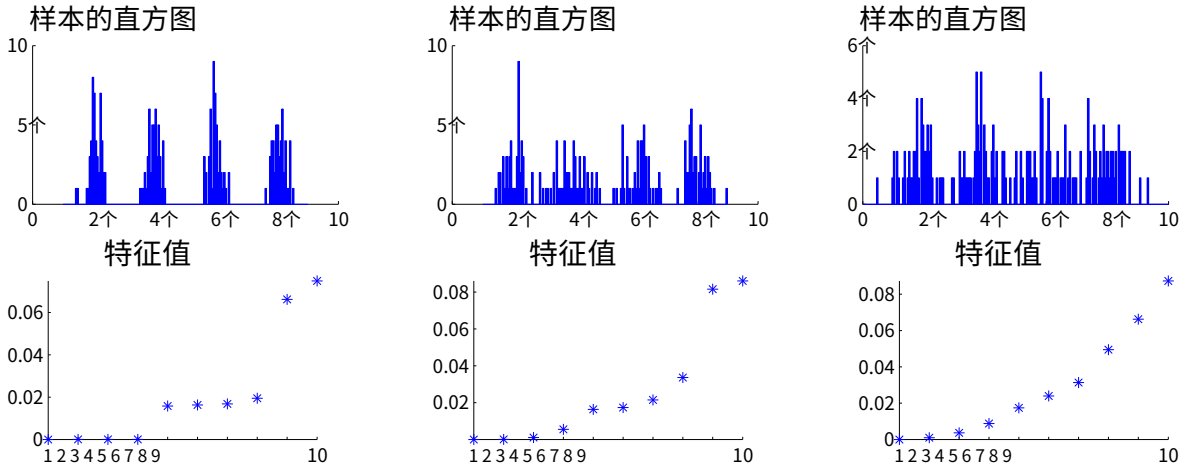

图 4：三个数据集，以及最小的 10 个特征值*大号*写字. 有关详细信息，请参阅文本。

布劳恩和布曼，2004 年；本大卫、冯卢克斯堡和帕尔，2006 年）。当然所有这些方法也可以用于谱聚类。此外，一种专为谱聚类设计的工具是特征间隙启发式算法，它可用于所有三种图拉普拉斯算子。这里的目标是选择号码k这样所有的特征值$\lambda_{1个},\ldots,\lambda_k$很小，但是$\lambda_{k+1}$比较大。这个程序有几个理由。第一个基于微扰理论，我们观察到在理想情况下k完全断开的簇，特征值 0 具有多重性k, 然后 (k+1)第特征值$\lambda_{k+1}>0$. 谱图论可以给出其他解释。在这里，图的许多几何不变量可以在图拉普拉斯算子的第一特征值的帮助下表达或界定。特别是，切割的大小与第一特征值的大小密切相关。有关此主题的更多详细信息，请参阅 Bolla (1991)、Mohar (1997) 和 Chung (1997)。

我们想在第 4 节中介绍的玩具示例中说明本征间隙启发式算法。为此，我们考虑与第 4 节中类似的数据集，但为了改变聚类的难度，我们考虑方差递增的高斯分布。图 4 的第一行显示了三个样本的直方图。我们构建了第 4 节中描述的 10 最近邻图，并绘制了归一化拉普拉斯算子的特征值*大号*写字在不同的样本上（非标准化拉普拉斯算子的结果相似）。第一个数据集由四个分离良好的簇组成，我们可以看到前 4 个特征值近似为 0。然后第 4 个和第 5 个特征值之间存在间隙，即$\lambda_{5个}-\lambda_{4个}$比较大。根据 eigengap 启发式，这个差距表明数据集包含 4 个簇。完全连接图的结果也可以观察到相同的行为（已绘制在图 1 中）。所以我们可以看到，如果数据中的聚类非常明显，则启发式方法效果很好。然而，簇的噪声越大或重叠越多，这种启发式方法的效果就越差。我们可以看到，对于聚类更“模糊”的第二个数据集，第 4 和第 5 个特征值之间仍然存在差距，但检测起来不像之前那样清晰。最后，在最后一个数据集中，没有明确定义的差距，所有特征值之间的差异大致相同。但另一方面，该数据集中的聚类重叠太多，以至于许多非参数算法将难以检测聚类，除非它们对基础模型做出强有力的假设。在此特定示例中，即使对于查看直方图的人来说，也不清楚正确的聚类数量应该是多少。这说明，作为选择聚类数量的大多数方法，如果数据包含非常明显的聚类，则特征间隙启发式算法通常效果很好，但在模棱两可的情况下，它也会返回模棱两可的结果。

最后，注意簇数的选择和邻域图的连通性参数的选择是相互影响的。例如，如果邻域图的连通性参数太小以至于该图分成，比方说，k_0连接组件，然后选择k_0

因为簇的数量是一个有效的选择。然而，一旦邻域图连接起来，就不清楚聚类的数量和邻域图的连通性参数是如何相互作用的。聚类数量的选择和图的连通性参数的选择本身都是难题，据我们所知，关于它们的相互作用没有什么不平凡的。

8.4 的k-意味着一步

我们在第 4 节中介绍的三种谱聚类算法使用k-意味着作为最后一步从特征向量的实值矩阵中提取最终分区。首先，请注意，使用k-表示这一步的算法。事实上，正如我们从谱聚类的各种解释中看到的那样，如果数据包含表达良好的聚类，这一步应该非常简单。例如，在理想情况下，如果簇完全分离，我们知道特征向量$大号$和$大号_{写字}$是分段常数。在这种情况下，所有点$X_{—世}$属于同一个集群$C_{秒}$

被精确映射到样本点$是_{—世}$，即对单位向量$电子_{秒}\in \mathbb{R}^k$. 在这种微不足道的情况下，任何应用于点的聚类算法$是_{—世}\in \mathbb{R}^k$将能够提取正确的簇。

虽然在谱聚类的最后一步中选择什么聚类算法有点随意，但可以争辩说至少点之间的欧几里得距离$是_{—世}$是一个有意义的数量。我们已经看到点之间的欧几里得距离$是_{—世}$与图表上的“通勤距离”相关，在 Nadler、Lafon、Coifman 和 Kevrekidis (2006) 中，作者表明$是_{—世}$也与更一般的“扩散距离”有关。此外，谱嵌入的其他用途（例如，Bolla (1991) 或 Belkin 和 Niyogi (2003)）表明，欧氏距离在$\mathbb{R}^d$是有意义的。

代替k-意味着，人们还使用其他技术从实值表示构造他的最终解决方案。例如，在 Lang (2006) 中，作者为此目的使用了超平面。Bach 和 Jordan (2004) 提出了更高级的特征向量后处理。在这里，作者研究了第一个跨越的子空间k特征向量，并尝试使用分段常数向量尽可能好地近似这个子空间。这也导致最小化空间中的某些欧几里得距离$\mathbb{R}^k$，这可以通过一些加权来完成k-意味着算法。

8.5 应使用哪个图拉普拉斯算子？

与谱聚类相关的一个基本问题是应该使用三个图拉普拉斯算子中的哪一个来计算特征向量。在决定这个问题之前，应该始终查看相似度图的度分布。如果图非常规则并且大多数顶点的度数大致相同，则所有拉普拉斯算子彼此非常相似，并且同样适用于聚类。但是，如果图中的度数分布非常广泛，则拉普拉斯算子会有很大差异。在我们看来，有几个论点提倡使用归一化而不是非归一化的谱聚类，并且在归一化的情况下使用特征向量$大号_{写字}$而不是那些$大号_{符号}$.

不同算法满足的聚类目标

支持归一化谱聚类的第一个论据来自图划分的观点。为简单起见，让我们讨论这个案例k=2. 一般来说，聚类有两个不同的目标：

1. 我们想要找到一个分区，使得不同集群中的点彼此不相似，即我们想要最小化集群间的相似性。在图形设置中，这意味着最小化 cut(一个,一个).

2. 我们想要找到一个分区，使得同一簇中的点彼此相似，即我们想要最大化簇内相似性 W(一个,一个) 和 W(一个,一个).

RatioCut 和 Ncut 都通过显式合并 cut(一个,一个) 在目标函数中。然而，关于第二点，两种算法的行为不同。注意

$$W(\text{一个},\text{一个}) = W(\text{一个}, \text{五}) - W(\text{一个},\text{一个}) = \text{卷}(\text{一个}) - \text{切}(\text{一个},\text{一个}).$$

因此，如果 cut(一个,一个) 是小和如果体积（一个) 很大。由于这正是我们通过最小化 Ncut 实现的目标，Ncut 标准实现了第二个目标。通过考虑另一个图形切割目标函数，即由 Ding、He、Zha、Gu 和 Simon (2001) 引入的 MinMaxCut 标准，可以更明确地看出这一点：

$$\text{最小最大切割}(\text{一个}_{1\text{个}}, \ldots, \text{一个}_k) := \sum_{\text{一世}=1}^{k} \frac{\text{切}(\text{一个}_{\text{一世}}, \overline{\text{一个}_{\text{一世}}})}{W(\text{一个}, \overline{\text{一个}_{\text{一世}}})}.$$

与 Ncut 相比，它具有 vol(一个) = 切(一个,一个) + W(一个,一个) 在分母中，MinMaxCut 准则只有 W(一个,一个) 在分母中。在实践中，Ncut 和 MinMaxCut 通常通过类似的切割来最小化，因为好的 Ncut 解决方案将具有较小的切割值（一个,一个) 无论如何，因此分母毕竟没有太大不同。此外，放宽 MinMaxCut 会导致与放宽 Ncut 完全相同的优化问题，即使用特征向量的归一化谱聚类大号写字. 因此，可以通过多种方式看出，归一化谱聚类结合了上述两个聚类目标。

现在考虑 RatioCut 的情况。这里的目标是最大化/一个/和/一个/而不是音量（一个）和体积 (一个). 但/一个/和/一个/不一定与簇内相似性相关，因为簇内相似性取决于边而不是顶点数一个. 例如，想想一个集合一个它有很多顶点，所有顶点彼此之间只有非常低的权重边。最小化 RatioCut 并不试图最大化簇内相似性，并且对于通过非归一化光谱聚类的松弛也是如此。

因此，这是我们要牢记的第一个重点：归一化谱聚类实现了上述两个聚类目标，而非归一化谱聚类仅实现了第一个目标。

一致性问题

对归一化谱聚类的优越性的一个完全不同的论据来自对两种算法的统计分析。在统计设置中，假设数据点$X_{1\text{个}}, \ldots, X_n$
已经根据某种概率分布进行了 iid 采样P在一些底层数据空间 X. 那么最根本的问题就是一致性问题：如果我们绘制越来越多的数据点，谱聚类的聚类结果是否收敛到底层空间的一个有用分区X?

对于两种归一化谱聚类算法，都可以证明确实如此（von Luxburg, Bousquet, and Belkin, 2004, 2005; von Luxburg, Bousquet, and Bousquet, to appear）。在数学上，可以证明当我们取极限$n \to \infty$， 矩阵大号符号强烈收敛

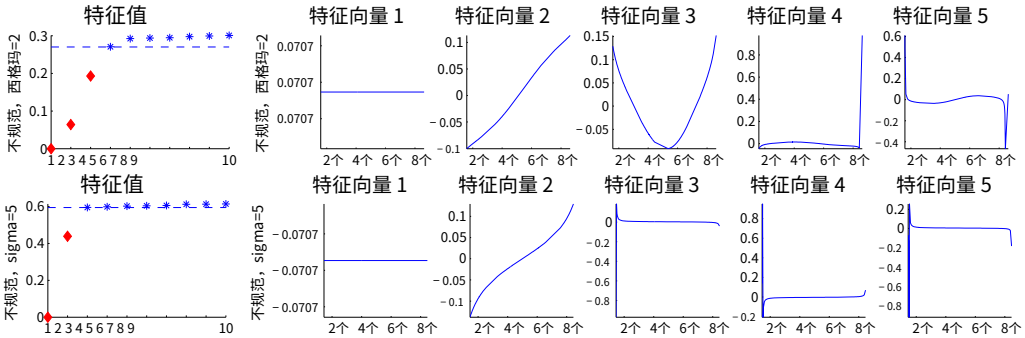

图 5：非标准化光谱聚类的一致性。绘制的是特征值和特征向量*大号*, 对于参数*σ*= 2（第一行）和*σ*= 5（第二行）。虚线表示最小*d*j, 特征值低于 min*d*被绘制为红色菱形，特征值高于最小值*d*被绘制为蓝色星星。有关详细信息，请参阅文本。

给操作员*ü*在空间上*C*(*X*) 上的连续函数*X*. 这种收敛意味着特征值和特征向量*大号*符号收敛于那些*ü*，这又可以转化为关于归一化谱聚类收敛的陈述。可以证明在*X*通过的特征向量*ü*可以解释为类似于谱聚类的随机游走解释。也就是说，如果我们考虑数据空间上的扩散过程*X*, 然后由特征向量引起的划分*ü*是这样的，扩散不会经常在不同的集群之间转换 (von Luxburg et al., 2004)。关于归一化谱聚类的所有一致性声明都适用，对于两者*大号*符号和*大号*写字，在非常温和的条件下，通常在现实世界的应用中得到满足。不幸的是，解释有关这些结果的更多细节超出了本教程的范围，因此我们建议感兴趣的读者阅读 von Luxburg 等人的文章。（出现）。

与归一化谱聚类的明确收敛性陈述相比，非归一化谱聚类的情况要令人不快得多。可以证明，非归一化谱聚类可能无法收敛，或者它可以收敛到构建由数据空间的一个单点组成的聚类的平凡解（von Luxburg et al., 2005, to appear）。在数学上，即使可以证明矩阵 (1*/n*)*大号*本身收敛于某个极限算子*吨*上*C*(*X*) 作为*n* →∞, 这个极限算子的光谱性质*吨*可能非常讨厌，以至于它们阻止了谱聚类的收敛。可以构造示例来表明这不仅是非常大的样本量的问题，而且即使对于小样本量也可能导致完全不可靠的结果。至少可以描述那些问题不发生时的条件：我们必须确保特征值*大号*对应于非归一化谱聚类中使用的特征向量的度数明显小于图中的最小度数。这意味着如果我们使用第一个*k*用于聚类的特征向量，然后*λ*—世

分钟*j*=1,...,*n*d*j*应该适用于所有人一*世*=1个,. . . , *k*.

这种情况的数学原因是对应于大于 min 的特征值的特征向量*d*近似 Dirac 函数，即它们在除一个坐标外的所有坐标中都近似为 0。如果这些特征向量用于聚类，那么它们将特征向量不为零的一个顶点与所有其他顶点分开，我们显然不想构造这样的分区。我们再次参考文献以获得精确的陈述和证明。

为了说明这种现象，再次考虑第 4 节中的玩具数据集。我们考虑基于全连接图的非归一化图拉普拉斯算子的第一个特征值和特征向量，用于不同的参数选择*σ*高斯相似函数（参见图 1 的最后一行和图 5 的所有行）。特征值高于最小值*d*被绘制为蓝色星星，特征值

低于最低限度*d*是 p对tted 作为红色钻石。虚线表示最小*d*j. 一般来说，我们可以看到

与远低于虚线的特征值对应的特征向量是“有用的”特征向量。如果$\sigma=1$（已经绘制在图 1 的最后一行），特征值 2、3 和 4 明显低于最小值d_j, 并且相应的特征向量 2、3 和 4 是有意义的（如第 4 节所述）。如果我们增加参数σ，我们可以观察到特征值趋向于向最小值移动d_j. 如果$\sigma=2$，只有前三个特征值低于最小值d_j（菲图 5 中的第一行），以防万一$\sigma=5$只有前两个特征值低于最小值d_j(图 5 中的第二行）。我们可以看到，一旦特征值接近或高于最小值d_j，其对应的特征向量近似狄拉克函数。当然，那些特征向量不适合构建聚类。在极限为$n\to\infty$，这些特征向量会收敛到完美的狄拉克函数。我们对有限样本案例的说明表明，这种行为不仅发生在大样本量的情况下，而且甚至可以在我们的玩具数据集中的小样本上产生。

非常重要的是要强调这些问题只涉及矩阵的特征向量*大号*，它们不会出现*大号*写字或者*大号*符号. 因此，从统计的角度来看，最好避免非归一化的谱聚类，而改用归一化的算法。

哪个归一化拉普拉斯算子?

查看两种归一化谱聚类算法之间的差异*大号*写字和 *大号*符号, 谱聚类的所有三种解释都支持*大号*写字. 原因是特征向量*大号*写字是聚类指标向量1个一个,而特征向量*大号*符号另外乘以*了*1个/2个，这可能会导致不需要的伪像。作为使用*大号*符号也没有任何计算优势，因此我们提倡使用*大号*写字.

9 展望与延伸阅读

谱聚类可以追溯到 Donath 和 Hoffman (1973)，他们首先建议基于邻接矩阵的特征向量构建图分区。同年，Fiedler（1973）发现图的二分与拉普拉斯图的第二个特征向量密切相关，他建议用这个特征向量对图进行分割。从那时起，谱聚类已在不同社区中多次被发现、重新发现和扩展，例如参见 Pothen、Simon 和 Liou (1990)、Simon (1991)、Bolla (1991)、Hagen 和 Kahng (1992) 、Hendrickson 和 Leland (1995)、Van Driessche 和 Roose (1995)、Barnard、Pothen 和 Simon (1995)、Spielman 和 Teng (1996)、Guattery 和 Miller (1998)。在 Spielman 和 Teng (1996) 中可以找到对谱聚类历史的一个很好的概述。

在机器学习社区中，谱聚类已因 Shi 和 Malik (2000)、Ng 等人的作品而流行起来。(2002)，Meila 和 Shi (2001)，以及 Ding (2004)。随后，谱聚类已扩展到许多非标准设置，例如应用于共聚类问题的谱聚类（Dhillon，2001），具有附加边信息的谱聚类（Joachims，2003）谱聚类与加权内核之间的连接-*k*-均值算法（Dhillon、Guan 和 Kulis，2005 年）、基于谱聚类的学习相似函数（Bach 和 Jordan，2004 年）或分布式环境中的谱聚类（Kempe 和 McSherry，2004 年）。此外，还发现了关于谱聚类与其他算法之间关系的新理论见解。谱聚类和加权核之间的联系*k*-均值算法在 Dhillon 等人中有描述。(2005)。谱聚类和（核）主成分分析之间的关系依赖于这样一个事实，即图拉普拉斯算子的最小特征向量也可以解释为核矩阵（Gram 矩阵）的最大特征向量。这种解释有两种不同的风格：而 Bengio 等人。(2004) 解释矩阵*了*-1个/2个*西部数据*-1个/2个作为内核矩阵，其他作者（Saerens，

Fouss、Yen 和 Dupont，2004 年）解释了 Moore-Penrose 逆函数*大号*或者*大号*符号作为核矩阵。这两种解释都可用于构造（不同的）样本外扩展以进行谱聚类。关于谱聚类的应用案例，近几年在各个科学领域发表的论文数量庞大，无法一一列举。我们鼓励读者使用“谱聚类”一词查询他最喜欢的文献数据库，以了解应用程序的多样性。

谱聚类的成功主要在于它没有对聚类的形式做出强假设。相对于*k*-意味着，在生成的聚类形成凸集（或者更准确地说，位于底层空间的不相交凸集）的情况下，谱聚类可以解决非常普遍的问题，例如螺旋交织。此外，只要我们确保相似图是稀疏的，即使对于大型数据集也可以有效地实现谱聚类。一旦选择了相似图，我们只需要解决一个线性问题，不存在陷入局部极小或多次重新启动算法并使用不同初始化的问题。然而，我们已经提到选择一个好的相似度图并非易事，并且在邻域图参数选择不同的情况下，谱聚类可能会非常不稳定。因此，谱聚类不能充当“黑盒算法”，自动检测任何给定数据集中的正确聚类。但它可以被认为是一个强大的工具，如果小心使用，可以产生良好的效果。

在机器学习领域，图拉普拉斯算子不仅用于聚类，而且还用于许多其他任务，例如半监督学习（例如，Chapelle、Schölkopf 和 Zien，2006 年的概述）或流形重建（例如，贝尔金和 Niyogi，2003 年）。在大多数应用中，图拉普拉斯算子用于对“接近”的数据点（即，w_{ij}大）应该有一个“相似”的标签（即$F_{—世} \approx F_j$）. 一个功能F满足这个假设，如果$w_{ij}(F_{—世}-F_j)_{2个}$

对所有人来说都很小*我, j,* 那是F *:如果*是小。凭借这种直觉，我们可以使用二次形式F *:如果* 作为转导分类问题中的正则化项。另一种解释图拉普拉斯算子使用的方法是通过它们编码的平滑假设。一个功能F它的价值很低F *:如果* 具有这样的特性，即它在数据点密集的区域（即图形紧密连接）仅“一点点”变化，而在低数据区域允许变化更大（例如，改变符号）密度。从这个意义上说，一个小值F *:如果*在半监督学习中编码所谓的“集群假设”，它要求分类器的决策边界应位于低密度区域。

一个经常使用的直觉是，图拉普拉斯算子形式上看起来像一个连续的拉普拉斯算子（这也是“图拉普拉斯算子”这个名字的来源）。要看到这一点，请转换局部相似性
w_{ij}到远处d_{ij}通过关系w_{ij}=1个/$d_{2个}$ $_{ij}$并观察到

$$w_{ij}(F_{—世}-F_j)_{2个} \approx \left(\frac{F_{—世}-F_j}{d_{ij}}\right)_{2个}$$

看起来像一个差商。因此，等式F *:如果*= $\Sigma_{ij} w_{ij}(F_{—世}-F_j)_{2个}$从
命题 1 看起来像是与标准拉普拉斯算子相关联的二次型的离散版本*大号*上R$_n$, 满足

$$\langle G，大号 G\rangle = \int |\nabla 克|^2_{个} dx_。$$

这种直觉在 Belkin (2003)、Lafon (2004)、Hein、Audibert 和 von Luxburg (2005) 的作品中得到了精确体现；M.、Audibert 和 von Luxburg (2007)、Belkin 和 Niyogi (2005)、Hein (2006)、Giné 和 Koltchinskii (2005)。一般来说，证明图拉普拉斯算子是某些连续拉普拉斯算子的离散版本，并且如果拉普拉斯图是在随机采样数据点的相似图上构建的，则它收敛到某个连续拉普拉斯算子（或

Laplace-Beltrami 算子）在底层空间上。Belkin (2003) 研究了收敛证明的第一个重要步骤，它处理与离散图 Laplacians 相关的连续算子收敛到 Laplace-Beltrami 算子。Lafon (2004) 将他的结果从均匀分布推广到一般分布。然后在 Belkin 和 Niyogi (2005) 中，作者在具有均匀分布的流形上使用高斯相似函数证明了非归一化拉普拉斯算子的逐点收敛结果。与此同时，海因等人。(2005) 证明更一般的结果，考虑到所有不同的拉普拉斯图*大号*,*大号*写字， 和*大号*符号，更一般的相似函数，以及具有任意分布的流形。在 Giné 和 Koltchinskii (2005) 中，分布和均匀收敛结果在具有均匀分布的流形上得到证明。Hein (2006) 研究了由图 Laplacians 引起的平滑泛函的收敛，并显示了一致的收敛结果。

除了将图拉普拉斯算子应用于最广泛意义上的划分问题外，图拉普拉斯算子还可以用于完全不同的目的，例如图形绘制（Koren，2005）。事实上，图的拓扑和属性与图拉普拉斯矩阵之间的联系比我们在本教程中提到的要紧密得多。现在对最基本的属性有了了解，有兴趣的读者可以自己进一步探索和欣赏该领域的大量文献。

参考

Aldous, D. 和 Fill, J.（准备中）。*可逆马尔可夫链和图上的随机游走。*
在线版本可在 http://www.stat.berkeley.edu/users/aldous/RWG/book.html 获得。Bach, F. 和 Jordan, M. (2004)。学习谱聚类。在 S. Thrun、L. Saul 和 B. Schölkopf
（编辑），*神经信息处理系统进展 16 (NIPS)*(第 305 – 312 页）。马萨诸塞州剑桥市：麻省理工学院出版社。
Bapat, R.、Gutman, I. 和 Xiao, W. (2003)。一种计算阻力距离的简单方法。*Z.*
自然资源。,*58*, 494 – 498。
Barnard, S.、Pothen, A. 和 Simon, H. (1995)。一种稀疏包络缩减的谱算法
矩阵。*数值线性代数及其应用*,*2个*(4), 317 – 334。
贝尔金，M. (2003)。*学习流形的问题*. 芝加哥大学博士论文。Belkin, M. 和 Niyogi, P. (2003)。用于降维和数据表示的拉普拉斯特征图
发送。*神经计算*,*15*(6), 1373 – 1396。
Belkin, M. 和 Niyogi, P. (2005)。迈向基于拉普拉斯流形的理论基础
方法。在 P. Auer 和 R. Meir（编辑）中，*第 18 届学习理论年会 (COLT) 论文集*(第 486 – 500 页）。斯普林格，纽约。
Ben-David, S.、von Luxburg, U. 和 Pál, D. (2006)。对聚类稳定性的清醒认识。在 G. Lugosi
和 H. Simon（编辑），*第 19 届学习理论年会 (COLT) 论文集* （第 5-19 页）。施普林格，柏林。

Bengio, Y.、Delalleau, O.、Roux, N.、Paiement, J.、Vincent, P. 和 Ouimet, M.（2004 年）。学习
特征函数链接谱嵌入和内核 PCA。*神经计算*,*16*, 2197 – 2219. Ben-Hur, A.、Elisseeff, A. 和 Guyon, I. (2002)。一种基于稳定性的结构发现方法
在聚类数据中。在*太平洋生物计算研讨会*(第 6 – 17 页）。Bhatia, R.
(1997)。*矩阵分析*. 斯普林格，纽约。
Bie, TD 和 Cristianini, N. (2006)。图形切割聚类、转换和转换的快速 SDP 松弛
其他组合问题。*JMLR*,*7*, 1409 – 1436。
Bolla, M. (1991)。*多图谱和分类属性之间的关系*(技术再
端口号 DIMACS-91-27）。离散数学和理论计算机科学中心。Brémaud, P. (1999)。*马尔可夫链：吉布斯域、蒙特卡罗模拟和队列*. 纽约：
施普林格出版社。

Brito, M.、Chavez, E.、Quiroz, A. 和 Yukich, J. (1997)。相互 k 最近邻的连通性
聚类和异常值检测中的图形。*统计和概率快报*,*35*, 33 – 42. Bui, TN 和 Jones, C. (1992)。找到好的近似顶点和边缘分区是 NP-hard。
信息。过程。莱特。,*42*(3), 153 – 159。
Chapelle, O.、Schölkopf, B. 和 Zien, A.（编辑）。(2006)。*半监督学习*. 麻省理工学院出版社，剑桥。
钟 F. (1997)。*谱图论*(卷。数学方面的 CBMS 区域会议系列的 92-
数学）。华盛顿数学科学会议委员会。
Dhillon, I. (2001)。使用二分谱图分区对文档和单词进行联合聚类。在
第七届ACM SIGKDD知识发现与数据挖掘（KDD）国际会议论文集(第 269 – 274 页）。纽约：ACM 出版社。
Dhillon, I.、Guan, Y. 和 Kulis, B. (2005)。*统一的内核视图k-意味着，谱聚类，和图划分*(技术报告编号 UTCS TR-04-25）。德克萨斯大学奥斯汀分校。丁，C.（2004 年）。*谱聚类教程。*在 ICML 上发表的演讲。（幻灯片可在
http://crd.lbl.gov/~光盘/光谱/) Ding, C.、He,
X.、Zha, H.、Gu, M. 和 Simon, H. (2001)。图的最小-最大切割算法
分区和数据聚类。在*第一届 IEEE 数据挖掘国际会议 (ICDM) 论文集*(第 107 – 114 页）。美国华盛顿特区：IEEE 计算机协会。Donath, WE 和 Hoffman, AJ (1973)。图划分的下界。*IBM J. Res。*
开发。,*17*, 420 – 425。
M. 菲德勒 (1973)。图的代数连通性。*捷克斯洛伐克数学。J.*,*23*, 298 – 305。Fouss, F.、Pirotte, A.、Renders, J.-M. 和 Saerens, M. (2007)。相似的随机游走计算
图的节点之间的关系在协作推荐中的应用。*IEEE 跨。知识 数据工程*,*19*(3), 355–369。
Fraley, C. 和 Raftery, AE (2002)。基于模型的聚类、判别分析和密度估计。*日本航空航天局*,*97*, 611 – 631。
Giné, E. 和 Koltchinskii, V. (2005)。Laplace-Beltrami 的经验图拉普拉斯近似
运营商：大样本结果。在*第四届高维概率国际会议论文集*(第 238-259 页）。
Golub, G. 和 Van Loan, C. (1996)。*矩阵计算*. 巴尔的摩：约翰霍普金斯大学按。
Guattery, S. 和 Miller, G. (1998)。关于光谱分离器的质量。*SIAM 矩阵杂志*
肛门。申请,*19*(3), 701 – 719。
Gutman, I. 和 Xiao, W. (2004)。拉普拉斯矩阵的广义逆和一些应用。
Bulletin de l'Academie Serbe des Sciences at des Arts（Cl. Math. Natur.）,*129*, 15 – 23. Hagen, L. 和 Kahng, A. (1992)。用于比率切割分区和聚类的新谱方法。
IEEE 跨。计算机辅助设计,*11*(9), 1074 – 1085。
Hastie, T.、Tibshirani, R. 和 Friedman, J. (2001)。*统计学习的要素*. 纽约：施普林格。
海因，M.（2006 年）。基于自适应图的正则化的均匀收敛。在*程序的*
第 19 届学习理论年会 (COLT)(第 50 – 64 页）。斯普林格，纽约。Hein, M.、Audibert, J.-Y. 和 von Luxburg, U. (2005)。从图到流形——弱与强
图拉普拉斯算子的逐点一致性。在 P. Auer 和 R. Meir（编辑）中，*第 18 届学习理论年会 (COLT) 论文集*(第 470 – 485 页）。斯普林格，纽约。Hendrickson, B. 和 Leland, R. (1995)。一种改进的映射谱图划分算法
并行计算。*SIAM J.论科学计算*,*16*, 452 – 469。
约阿希姆斯 (2003)。通过谱图划分的转换学习。在 T. Fawcett 和
N. Mishra（编辑），*第 20 届机器学习国际会议 (ICML) 论文集*（第 290 – 297 页）。AAAI 出版社。
Kannan, R.、Vempala, S. 和 Vetta, A. (2004)。关于聚类：好、坏和光谱。*杂志*
美国计算机协会,*51*(3), 497–515。
Kempe, D. 和 McSherry, F. (2004)。一种用于光谱分析的分散算法。在*诉讼程序*

第 36 届 ACM 计算理论研讨会 (STOC)(第 561 – 568 页）。美国纽约州纽约市：ACM 出版社。

Klein, D. 和 Randic, M. (1993)。阻力距离。*数理化学学报*,*12*, 81 – 95. Koren, Y. (2005)。通过特征向量绘制图形：理论与实践。*计算机和数学*
与应用程序,*49*, 1867 – 1888.
Lafon, S. (2004)。*扩散图和几何谐波*. 耶鲁大学博士论文。
Lang, K. (2006)。修复光谱方法的两个弱点。在 Y. Weiss、B. Schölkopf 和 J. Platt 中
（编辑），*神经信息处理系统的进展 18*(第 715 – 722 页）。马萨诸塞州剑桥市：麻省理工学院出版社。
Lange, T.、Roth, V.、Braun, M. 和 Buhmann, J.（2004 年）。基于稳定性的聚类验证
解决方案。*神经计算*,*16*(6), 1299 – 1323。
Lovász, L. (1993)。图上的随机游走：一项调查。在*组合数学，Paul Erdös 八十岁*(页数
353 – 397）。布达佩斯：János Bolyai Math。社会。
Lütkepohl, H. (1997)。*矩阵手册*. 奇切斯特：威利。
M.、Audibert, J.-Y. 和 von Luxburg, U.（2007 年）。图拉普拉斯算子及其在随机邻域图上的收敛。*JMLR*,
8个, 1325 – 1370。
Meila, M. 和 Shi, J. (2001)。光谱分割的随机游走视图。在*第八届国际*
人工智能与统计研讨会 (AISTATS)。
Mohar, B. (1991)。图的拉普拉斯谱。在*图论、组合学和应用。*
卷。2（卡拉马祖，密歇根州，1988 年）(第 871 – 898 页）。纽约：威利。
Mohar, B. (1997)。图的拉普拉斯特征值的一些应用。在 G. Hahn 和 G. Sabidussi
（编辑），*图对称性：代数方法与应用*(卷。北约 ASI 系列 C 497，第 225-275 页）。克鲁沃。

Nadler, B.、Lafon, S.、Coifman, R. 和 Kevrekidis, I. (2006)。扩散图、光谱聚类
和 Fokker-Planck 算子的特征函数。在 Y. Weiss、B. Schölkopf 和 J. Platt（编辑）中，*神经信息处理系统的进展 18*(第 955 – 962 页）。马萨诸塞州剑桥市：麻省理工学院出版社。

Ng, A.、Jordan, M. 和 Weiss, Y. (2002)。关于谱聚类：分析和算法。在
T. Dietterich、S. Becker 和 Z. Ghahramani（编辑），*神经信息处理系统的进展 14*(第 849 – 856 页）。麻省理工学院出版社。
J. 诺里斯 (1997)。*马尔可夫链*. 剑桥：剑桥大学出版社。
彭罗斯，M. (1999)。最小生成树最长边的强定律。*安。的概率。*，
27(1), 246 – 260。
Pothen, A.、Simon, HD 和 Liou, KP (1990)。用特征向量分割稀疏矩阵
图。*SIAM 矩阵肛门杂志。申请*,*11*, 430 – 452。
Saerens, M.、Fouss, F.、Yen, L. 和 Dupont, P. (2004)。图的主成分分析，
及其与谱聚类的关系。在*第 15 届欧洲机器学习会议 (ECML) 论文集*(第 371 – 383 页）。施普林格，柏林。
Shi, J. 和 Malik, J. (2000)。归一化切割和图像分割。*IEEE 模式汇刊*
分析和机器智能,*22*(8), 888 – 905。
西蒙，H. (1991)。为并行处理划分非结构化问题。*计算系统*
工程,*2个*, 135 – 148。
Spielman, D. 和 Teng, S. (1996)。谱划分工作：平面图和有限元
网格。在*第 37 届计算机科学基础年会（佛蒙特州伯灵顿，1996 年）*(第 96 – 105 页）。加利福尼亚州洛斯阿拉米托斯：IEEE 计算。社会。按。（另请参阅扩展技术报告。）

Stewart, G. 和 Sun, J. (1990)。*矩阵微扰理论*. 纽约：学术出版社。
Still, S. 和 Bialek, W. (2004)。有多少簇？信息论的观点。*神经的*
电脑。,*16*(12), 2483 – 2506。
Stoer, M. 和 Wagner, F. (1997)。一个简单的最小割算法。*J.ACM*,*44*(4), 585 – 591。Tibshirani, R.、
Walther, G. 和 Hastie, T. (2001)。通过以下方式估计数据集中的簇数
差距统计。*J.皇家。国家主义者。社会。乙*,*63*(2), 411 – 423。

Van Driessche, R. 和 Roose, D. (1995)。一种改进的谱二分法及其应用
到动态负载均衡。*并行计算。*,*21*(1), 29 – 48。
von Luxburg, U.、Belkin, M. 和 Bousquet, O.（出场）。谱聚类的一致性。*志
统计学*.（另见技术报告 134，马克斯普朗克生物控制论研究所，2004 年）

von Luxburg, U.、Bousquet, O. 和 Belkin, M. (2004)。关于谱聚类的收敛性
随机样本：归一化的情况。在 J. Shawe-Taylor 和 Y. Singer（编辑）中，*第 17 届学习理论年会
(COLT) 论文集*(第 457 – 471 页）。斯普林格，纽约。von Luxburg, U.、Bousquet, O. 和 Belkin, M.
(2005)。谱聚类的局限性。在 L. Saul 中，
Y. Weiss 和 L. Bottou（编辑），*神经信息处理系统 (NIPS) 的进展 17*（第 857 – 864 页）。马萨诸塞
州剑桥市：麻省理工学院出版社。
Wagner, D. 和 Wagner, F. (1993)。在最小割和图二分之间。在*程序的
第 18 届计算机科学数学基础国际研讨会 (MFCS)*(第 744 – 750 页）。伦敦：施普林格。