MLAPP-C12

12. 潜在线性模型

12.1 因子分析

混合模型的一个问题在于它们只能使用一个单独的隐变量去生成观测值。特别的，每个观测值只能来自于K个原型中的一个。我们可以认为一个混合模型使用了K个二值隐变量，代表着簇编号的one-hot编码。但是这些变量是互相排斥的，所以模型在表达能力上还是有所局限。

一种替代方案是使用一个实数隐变量向量，**z***i*∈*L*。最简单的先验分布就是使用高斯分布（我们在后面会考虑其他选择）：

（12.1）

如果观测值也是连续的，则**x***i*∈*D*，我们可能使用一个高斯分布作为似然函数。就像在线性回归中的那样，我们假设期望是（隐变量）输入的线性函数，得到：

（12.2）

其中W是一个D×L的矩阵，被称为因子加载矩阵（factor loading matrix），是一个D×D的协方差矩阵。我们令为对角矩阵，因为模型的基本观点是“迫使”去解释相关性，而不是将这个任务交给观测变量xi上。上述的整个模型被称为因子分析（factor analysis,FA）。一种特殊的情况是，这被称为概率主元分析（probabilistic principal components analysis,PPCA）。这个名字的起因在后面将会介绍。

如果L=1，D=2，为球状协方差矩阵，为对角矩阵，则数据生成的过程如图12.1所示。我们将各向同性的高斯分布当作一个“喷雾器”，代表着隐变量zi的先验分布，将其沿着由定义的1维直线上滑动，这将导致一个在二维空间中的被拉长的高斯分布（所以变量之间相关）。

12.1.1 FA是MVN的低秩参数化

FA被认为是一种使用较少参数确定x的联合概率密度模型的方法。为了说明这一点，观察式4.126，推导出来的边缘分布是一个高斯分布：

（12.3）

（12.4）

所以，。不失一般性，我们可以令，因为我们总可以将吸收进。类似的，我们可以令，因为我们总是可以构造一个相关的先验分布，通过定义一个新的权重矩阵，因为

（12.5）

所以我们发现，因子分析使用了一个低秩的分解近似表达观测向量的协方差矩阵：

（12.6）

上式仅包含O(LD)个参数，允许在一个全协方差矩阵和对角协方差矩阵中进行一个灵活的妥协，前者具有O(D2)个参数，后者具有O(D)个参数。需要注意的是，如果我们不限制为对角矩阵，我们可以简单地将它作为全协方差矩阵；然后我们可以设置W=0，在这种情况下，我们就不再需要隐变量了。

每个观测变量的边缘方差可以由确定，其中第一项为共同的因子所导致的方差，第二项被称为独立项（uniqueness），它指定了那个维度的方差项。

12.1.2 潜在因子的推理

12.2 主元分析

如果我们在FA模型中约束，且W为标准正交矩阵。结果表明，如果，该模型将退化成经典的（非概率的）主元分析（principal components analysis,PCA），又被称为Karhunen Loeve变换。的版本被称为概率PCA（PPCA）或者敏感PCA。

为了理解这个结果，我们首先学习经典的PCA。然后将其与SVD联系起来。最后我们再讨论PPCA。

12.2.1 经典PCA：定理陈述

在接下来的理论陈述中表达了经典PCA的综合观点。

定理12.2.1 假设我们需要找到L个线性基向量的正交基wj∈RD，以及对应的分值zi∈RL，从而最小化平均**重构误差**（reconstructioin error）：

（12.26）

其中，W为正交矩阵。等价地，我们可以将上述目标函数写成：

（12.27）

其中Z为N×L的矩阵，zi为它的行向量。为矩阵A的frobenius norm，定义为：

（12.28）

上述模型的最优解为，其中包含经验协方差矩阵（我们假设xi的期望为0）的L个特征向量，这些向量对应前L个大的特征值。更近一步，数据在低维度最优的编码为，它表示数据在由特征向量构成的列空间上的投影。

图12.5（a）展示了在D=2和L=1时的情况。对角线为向量w1；被称为第一主元或者第一主方向。数据点xi∈R2被投影到这条直线上，从而得到zi∈R。这是数据在1维空间中的最好的近似。（我们将在12.5（b）讨论这一点）

通常情况下，很难对更高维的数据进行可视化，但如果数据正好是一个图像集合，那么可以很容易地做到这一点。图12.6展示了前三个主向量，将其转换成图片，也就是使用可变数量的基向量对一张图片进行重构。（我们在11.5节讨论了如何L）。

接下来，我们将展示主元方向就是数据呈现方差最大的方向。这就意味着PCA可能会被那些只是因为测量基准而导致方差最大的方向所“误导”。图12.7（a）展示了一个例子，其中垂直方向（体重）使用的范围比水平方向（身高）的范围更大，导致一条不那么（自然）的直线。所以在使用PCA时对数据进行标准化是一个标准的操作，或者我们的所有操作都基于相关系数矩阵而非协方差矩阵。这个好处在图12.7（b）中有所体现。

12.2.2 证明\*

12.2.3 奇异值分解（SVD）

我们已经定义了PCA的解为协方差矩阵的特征向量。然而，有另一种方法得到PCA的解，基于奇异值分解（singular value decomposition，SVD）。基于这个方法我们可以求解非方阵的特征向量。

特别地，任意（实数）N×D的矩阵X可以被分解为：

（12.46）

其中U为N×N的矩阵，其列向量为正交向量（UTU=IN），V为D×D的矩阵，其行和列向量都是正交向量（所以VTV=VVT=ID），S为N×D的矩阵，包含主对角线上的r=min(N,D)个奇异值，矩阵的其他元素为0。U的列向量为左奇异向量，V的列向量为右奇异向量。图12.8（a）给出了案例。

既然最多有D个奇异值（假设N＞D），那么U的最后N-D列则是无关的，因为它们要与0相乘。经济版的SVD（economy sized SVD， thin SVD）避免计算这些不必要的元素。让我们使用表示这种分解。如果N>D，我们有

（12.47）

正如图12.8（a）所表示的。如果N<D，我们有

（12.48）

计算经济版的SVD的时间复杂度为O(NDmin(N,D))。

特征向量与奇异向量之间的联系如下。对于一个任意的实数矩阵X，如果X=USVT，我们有：

（12.49）

其中为包含奇异值平方的对角矩阵。所以

（12.50）

所以的特征向量等于V，即X的右奇异向量，的特征值等于D，奇异值的平方。类似地：

（12.51）

（12.52）

所以的特征向量等于U，即X的左奇异向量。同时，的特征值等于奇异值的平方。我们可以将上述结果汇总如下：

（12.53）

既然矩阵的特征向量不受矩阵的线性变换影响，我们发现X的右奇异向量等于经验协方差矩阵的特征向量。进一步地，的特征值是奇异值平方的尺度化版本。

然而，PCA与SVD的联系则需要进一步深入。根据12.46，我们可以表示一个秩为r的矩阵：

（12.54）

如果奇异值如图12.10所示的迅速趋向于0，我们可以得到一个秩为L的近似矩阵：

（12.55）

这被称为**截断SVD**（truncated SVD）（见图12.8（b））。将一个N×D的矩阵近似为秩为L的矩阵的参数总量为：

NL+LD+L=L(N+D+1)

作为一个例子，考虑一张图12.9（左上）所示的200×320的图片。其中包含64000个数值。我们发现一个秩为20的近似，只需要（200+320+1）×20=10420个参数，它依然是一个好的近似。

结果表明，这种近似的误差为

（12.57）

更进一步，结果表明SVD提供了最好的矩阵的L秩近似（最小化上述Frobenius 范数）。

让我们将其与PCA进行联系，令X=USVT为X的截断SVD。我们知道，。所以：

（12.58）

更进一步，最优的重构方案为，所以我们发现

（12.59）

这与截断SVD近似是完全一样的！这说明PCA是数据的低秩的最好近似。