第九章、数据降维

赵涵

2023 年 5 月 23 日

我们知道，解决过拟合的问题除了正则化和添加数据之外，降维就是最好的方法。降维的思路来源于维度灾难的问题，我们知道n维球的体积为：

$$CR^n \tag{1}$$

那么在球体积与边长为$2R$的超立方体比值为：

$$\lim_{n\to 0}\frac{CR^n}{2^nR^n}=0 \tag{2}$$

这就是所谓的维度灾难，在高维数据中，主要样本都分布在立方体的边缘，所以数据集更加稀疏。降维的方法分为：

1. 直接降维：特征选择；
2. 线性降维：PCA，MDS等；
3. 非线性：流形降维，包括lsomap，LLE等。

1 主成分分析

主成分分析（Principal Component Analysis）中，我们的基本想法是将所有数据投影到一个子空间中，从而达到降维的目标，为了寻找这个子空间，我们基本想法是：

- 所有数据在子空间中更为分散；
- 损失的信息最小，即：在补空间的分量少。

原来的数据很有可能各个维度之间是相关的，于是我们希望找到一组p个新的线性无关的单位基u_i，降维就是取其中的q个基矢。达到两个效果，第一是从相关变到无关，第二是：

1. 最大投影方差；
2. 最小重构距离。

对一个简单的二维坐标变换，我们给出图（1）进行说明： 我们先从第一个特点出发，最大投影方差，我们知道，对于一个数据点，是一个列向量，那么我们方便起见，对这个数据点进行中心化，所谓中心化，就是让这个数据点减去整个数据集对应的均值向量，这样就把所有的数据点平移到了坐标原点附近。即：

$$x^{(i)}-\mu \tag{3}$$

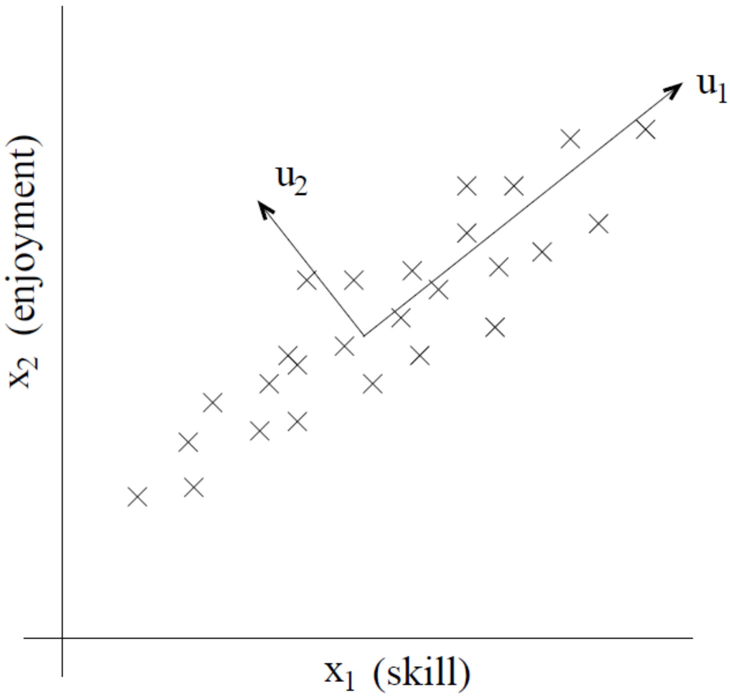

Figure 1: 主成分分析在二维平面直角坐标系下的表示。

在中心化后的数据点，我们对它在新的坐标系下做投影，所谓投影，就是一个向量与一个单位向量u做内积，即：

$$(x^{(i)} - \mu)^T u \tag{4}$$

投影有正有负，所以延续之前的思想，通过平方的方式，消除掉符号的影响。那么对于整个数据集，我们把对第一个方向所有的投影平方求和，得到损失函数：

$$J = \frac{1}{N}\sum_{i=1}^{N}((x^{(i)} - \mu)^T u_1)^2 \tag{5}$$

我们对目标函数进行化简：

$$\begin{aligned} J &= \sum_{i=1}^{N}\frac{1}{N}u_1^T(x^{(i)} - \mu)(x^{(i)} - \mu)^T u_1 && (6)\\ &= u_1^T(\sum_{i=1}^{N}\frac{1}{N}(x^{(i)} - \mu)(x^{(i)} - \mu)^T)u_1 && (7)\\ &= u_1^T S u_1 && (8) \end{aligned}$$

S为数据集的协方差矩阵。根据我们上面说的，要最大化投影方差，我们有：

$$\begin{aligned} \hat{u_1} &= \arg\max u_1^T S u_1 && (9)\\ s.t. & \quad u_1^T u_1 = 1 && (10) \end{aligned}$$

对于含有约束的极值问题，采用拉格朗日乘子法，构造拉格朗日函数，有：

$$\mathcal{L}(u, \lambda) = u_1^T S u_1 + \lambda(1 - u_1^T u_1) \tag{11}$$

对函数参数求导数，并令其等于0，有：

$$\frac{\partial \mathcal{L}}{\partial u_1} = 2Su_1 - \lambda \cdot 2u_1 = 0 \tag{12}$$

我们得到如下的方程：

$$Su_1 = \lambda u_1 \tag{13}$$

这个方程，我们已经很熟悉了，就是对矩阵求特征值与特征向量。

下面看其损失的信息最少这个条件，同样适用系数的平方平均作为损失函数，并最小化：

$$\begin{aligned} J &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=q+1}^{p}((x^{(i)}-\mu)^T u_j)^2 && (14) \\ &= \sum_{j=q+1}^{p} u_j^T S u_j,\ s.t.\ u_j^T u_j = 1 && (15) \end{aligned}$$

同样的：

$$\arg\min_{u_j} L(u_j, \lambda) = \arg\min_{u_j} u_j^T S u_j + \lambda(1 - u_j^T u_j) \tag{16}$$

损失函数最小取在本征值剩下的个最小的几个值。数据集的协方差矩阵可以写成$S = U\Lambda U^T$，直接对这个表达式也可以得到本征矢和特征值。

2 因子分析

2.1 问题引入

之前我们考虑的训练数据中样例$x^{(i)}$的个数m都远远大于其特征个数n，这样不管是进行回归、聚类等都没有太大的问题。然而当训练样例个数m太小，甚至$m << n$的时候，使用梯度下降法进行回归时，如果初值不同，得到的参数结果会有很大偏差（因为方程数小于参数个数）。另外，如果使用多元高斯分布（Multivariate Gaussian）对数据进行拟合时，也会有问题。通过计算，看看会有什么问题：

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \tag{17}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T \tag{18}$$

分别是求均值和协方差的公式，$x^{(i)}$表示样本，共有m个，每个样本n个特征，因此μ是n维向量，Σ是$n \times n$协方差矩阵。当$m << n$时，我们会发现Σ是奇异阵（$|\Sigma| = 0$），也就是说Σ^{-1}不存在，没办法拟合出多元高斯分布，确切的说是我们估计不出来Σ。如果我们仍然想用多元高斯分布来估计样本，就引入了因子分析模型，因子分析模型是建立在多元高斯分布基础上的。换句话说，可以认为是高斯分布为前提的，概率形式的PCA。

2.2 边缘和条件高斯分布

在讨论因子分析之前，先看看多元高斯分布中，条件和边缘高斯分布的求法。这个在后面因子分析的EM推导中有用。假设x是有两个随机向量组成（可以看作是将之前的$x^{(i)}$分成了两部分）：

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{19}$$

其中$x_1 \in \mathbb{R}^r, x_2 \in \mathbb{R}^s$，那么$x \in \mathbb{R}^{r+s}$。假设$x$服从多元高斯分布$x \sim N(\mu, \Sigma)$，其中

$$\mu = \left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right], \Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right] \tag{20}$$

$\mu_1 \in \mathbb{R}^r, \mu_2 \in \mathbb{R}^s$，那么$\Sigma_{11} \in \mathbb{R}^{r\times r}, \Sigma_{12} \in \mathbb{R}^{r\times s}$。，由于协方差矩阵是对称阵，因此$\Sigma_{12} = \Sigma_{21}^T$。整体看来$x_1$和$x_2$联合分布符合多元高斯分布。那么只知道联合分布的情况下，如何求得x_1的边缘分布呢？从上面的μ 和Σ可以看出，$E[x_1] = \mu_1, Cov(x_1) = E[(x_1 - \mu_1)(x_1 - \mu_1)^T] = \Sigma_{11}$，下面我们验证第二个结果：

$$\begin{aligned}
Cov(x) &= \Sigma && (21)\\
&= \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right] && (22)\\
&= E[(x-\mu)(x-\mu)^T] && (23)\\
&= E\left[\left(\begin{array}{c} x_1 - \mu_1 \\ x_2 - \mu_2 \end{array} \right) \left(\begin{array}{c} x_1 - \mu_1 \\ x_2 - \mu_2 \end{array} \right)^T \right] && (24)\\
&= E\left[\begin{array}{cc} (x_1-\mu_1)(x_1-\mu_1)^T & (x_1-\mu_1)(x_2-\mu_2)^T \\ (x_2-\mu_2)(x_1-\mu_1)^T & (x_2-\mu_2)(x_2-\mu_2)^T \end{array} \right] && (25)
\end{aligned}$$

由此可见，多元高斯分布的边缘分布仍然是多元高斯分布。也就是说$x_1 \sim N(\mu_1, \Sigma_{11})$。

上面$Cov(x)$里面有趣的是Σ_{12}，这个与之前计算协方差的效果不同。之前的协方差矩阵都是针对一个随机变量（多维向量）来说的，而Σ_{12}评价的是两个随机向量之间的关系。比如$x_1 =$身高，体重$x_2 =$性别，收入，那么Σ_{11}求的是身高与身高，身高与体重，体重与体重的协方差。而Σ_{12}求的是身高与性别，身高与收入，体重与性别，体重与收入的协方差，看起来与之前的大不一样，不太常见的求法。

上面求的是边缘分布，让我们考虑一下条件分布的问题，也就是$x_1|x_2$的问题。根据多元高斯分布的定义，$x_1|x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$。我们有：

$$\begin{aligned}
\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) && (26)\\
\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} && (27)
\end{aligned}$$

这是我们接下来计算时需要的公式，这两个公式直接给出，没有推导过程。具体的细节，在我们讲解线性高斯模型时，已经给出了构造性证明。

2.3 因子分析例子

下面通过一个简单例子，来引出因子分析背后的思想。因子分析的实质是认为m个n维特征的训练样本$x^{(i)}(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$的产生过程如下：

1. 首先在一个k维的空间中按照多元高斯分布生成m个$z^{(i)}$(k维向量)，即：

$$z^{(i)} \sim N(0, I) \tag{28}$$

2. 然后存在一个变换矩阵$\Lambda \in \mathbb{R}^{n\times k}$，将$z^{(i)}$映射到$n$维空间中，即：

$$\Lambda z^{(i)} \tag{29}$$

因为$z^{(i)}$的均值是0，映射后仍然是0。

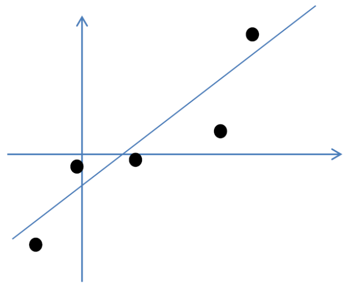

Figure 2: 数据点的分布。

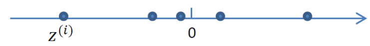

Figure 3: 假设原始数据点的来源。

3. 然后将$\Lambda z^{(i)}$加上一个均值μ（n维），即：

$$\mu + \Lambda z^{(i)} \tag{30}$$

对应的意义是将变换后的$\Lambda z^{(i)}$(n维向量）移动到样本$x^{(i)}$的中心点μ。

4. 由于真实样例$x^{(i)}$与上述模型生成的有误差，因此我们继续加上误差ϵ（n维向量），而且ϵ符合多元高斯分布，即

$$\epsilon \sim N(0, \Psi) \tag{31}$$

$$\mu + \Lambda z^{(i)} + \epsilon \tag{32}$$

5. 最后的结果认为是真实的训练样例$x^{(i)}$的生成公式：

$$x^{(i)} = \mu + \Lambda z^{(i)} + \epsilon \tag{33}$$

通过一种直观方法来解释上述过程：假设我们有$m = 5$个2 维的样本点$x^{(i)}$（两个特征），如图（2）：那么按照因子分析的理解，样本点的生成过程如下：

1、首先假设在1维空间（这里$k = 1$存在着按正态分布生成的m个点$z^{(i)}$，如图（3）：

2、然后使用某个$\Lambda = (a, b)^T$将一维的z映射到2维，如图（4）：

3、之后加上$\mu(\mu_1, \mu_2)^T$，即将所有点的横坐标移动μ_1，纵坐标移动μ_2，将直线移到一个位置，使得直线过点μ，原始左边轴的原点现在为μ（红色点），如图（5）。然而，样本点不可能这么规则，在

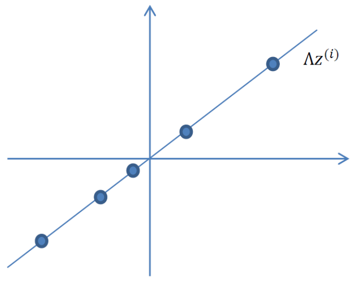

Figure 4: 对假设的数据点进行映射。

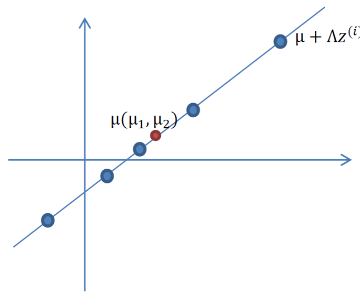

Figure 5: 平移。

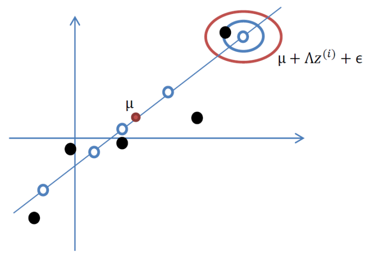

Figure 6: 添加噪声。

模型上会有一定偏差，因此我们需要将上步生成的点做一些扰动（误差），扰动$\epsilon \sim N(0, \Psi)$。

4、加入扰动后，我们得到黑色样本$x^{(i)}$如图（6）：

5、其中由于z和ϵ的均值都为0，因此μ也是原始样本点（黑色点）的均值。由以上的直观分析，我们知道了因子分析其实就是认为高维样本点实际上是由低维样本点经过高斯分布、线性变换、误差扰动生成的，因此高维数据可以使用低维来表示。

2.4 因子分析模型

上面的过程是从隐含随机变量z经过变换和误差扰动来得到观测到的样本点。其中z被称为因子，是低维的。我们将式子再列一遍如下：

$$z \sim N(0, I) \tag{34}$$

$$\epsilon \sim N(0, \Psi) \tag{35}$$

$$x = \mu + \Lambda z + \epsilon \tag{36}$$

其中误差ϵ和z是独立的。随机变量z和x联合概率分布符合多元高斯分布，如下：

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim N(\mu_{zx}, \Sigma) \tag{37}$$

我们要把μ_{zx}和Σ分别求出来。先解$E[x]$：

$$\begin{aligned} E[x] &= E[\mu + \Lambda z + \epsilon] &(38)\\ &= \mu + \Lambda E[z] + E[\epsilon] &(39)\\ &= \mu &(40) \end{aligned}$$

我们已知$E[z]=0$，因此：

$$\mu_{zx} = \begin{bmatrix} \overrightarrow{0} \\ \mu \end{bmatrix} \tag{41}$$

接着计算Σ，其中$\Sigma_{zz}=Cov(z)=I$，接着求Σ_{zx}：

$$\begin{aligned} E[(z-E[z])(x-E[x])^T] &= E[z(\mu+\Lambda z+\epsilon-\mu)^T] &(42)\\ &= E[zz^T]\Lambda^T + E[z\epsilon^T] &(43)\\ &= \Lambda^T &(44) \end{aligned}$$

这个过程中利用了z和ϵ独立假设（$E[z\epsilon^T]=E[z]E[\epsilon^T]=0$）。并将$\Lambda$看作已知变量。接着求$\Sigma_{xx}$：

$$\begin{aligned} E[(x-E[x])(x-E[x])^T] &= E[(\mu+\Lambda z+\epsilon-\mu)(\mu+\Lambda z+\epsilon-\mu)^T] &(45)\\ &= E[\Lambda zz^T\Lambda^T + \epsilon z^T\Lambda^T + \Lambda z\epsilon^T + \epsilon\epsilon^T] &(46)\\ &= \Lambda E[zz^T]\Lambda^T + E[\epsilon\epsilon^T] &(47)\\ &= \Lambda\Lambda^T + \Psi &(48) \end{aligned}$$

然后得出联合分布的最终形式：

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim N\left(\begin{bmatrix} \overrightarrow{0} \\ \mu \end{bmatrix} \quad \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T+\Psi \end{bmatrix} \right) \tag{49}$$

从上式中可以看出x的边缘分布$x \sim N(\mu, \Lambda\Lambda^T+\Psi)$。那么对样本$\{x^{(i)}; i=1,2,\ldots,m\}$进行最大似然估计：

$$\ell(\mu,\Lambda,\Psi) = \log\prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2}|\Lambda\Lambda^T+\Psi|} \exp(-\frac{1}{2}(x^{(i)}-\mu)(\Lambda\Lambda^T+\Psi)^{-1}(x^{(i)}-\mu)) \tag{50}$$

由于这里面存在耦合，即求偏导数后，各个参数之间的乘积消除不掉，所以得不到closed form。根据之前对参数估计的理解，在有隐含变量z时，我们可以考虑使用EM来进行估计。

2.5 因子分析的EM估计

首先明确一下各个参数，z是隐变量，μ,Λ,Ψ是待估参数。我们套用EM算法，有：

（E-step）：

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\mu,\Lambda,\Psi) \tag{51}$$

根据之前得到的条件分布讨论，我们有：

$$z^{(i)}|x^{(i)};\mu,\Lambda,\Psi \sim N(\mu_{z^{(i)}|x^{(i)}}, \Sigma_{z^{(i)}|x^{(i)}}) \tag{52}$$

因此我们得到：

$$\begin{aligned} \mu_{z^{(i)}|x^{(i)}} &= \Lambda^T(\Lambda\Lambda^T+\Psi)^{-1}(x^{(i)}-\mu), &(53)\\ \Sigma_{z^{(i)}|x^{(i)}} &= I - \Lambda^T(\Lambda\Lambda^T+\Psi)^{-1}\Lambda. &(54) \end{aligned}$$

根据多元高斯分布，就可以写出隐变量的函数表达式：

$$Q_i(z^{(i)}) = \frac{1}{(2\pi)^{k/2}|\Sigma_{z^{(i)}|x^{(i)}}|^{1/2}} \exp(-\frac{1}{2}(z^{(i)}-\mu_{z^{(i)}|x^{(i)}})^T\Sigma_{z^{(i)}|x^{(i)}}^{-1}(z^{(i)}-\mu_{z^{(i)}|x^{(i)}})) \tag{55}$$

（M-step）：直接写要最大化的目标是：

$$\sum_{i=1}^{m}\int_{z^{(i)}} Q_i(z^{(i)}) \log\frac{p(x^{(i)},z^{(i)};\mu,\Lambda,\Psi)}{Q_i(z^{(i)})} dz^{(i)} \tag{56}$$

其中待估参数是μ, Λ, Ψ。下面我们重点求Λ的估计公式首先将上式简化为：

$$\sum_{i=1}^{m} \int_{z^{(i)}} Q_i(z^{(i)})[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})]dz^{(i)} \tag{57}$$

$$= \sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})] \tag{58}$$

这里$z^{(i)} \sim Q_i$服从Q_i分布。然后去掉与Λ不相关的项（后两项），得：

$$\sum_{i=1}^{m} E[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi)] \tag{59}$$

$$= \sum_{i=1}^{m} E[\log \frac{1}{(2\pi)^{n/2}|\Psi|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{(-1)}(x^{(i)} - \mu - \Lambda z^{(i)}))] \tag{60}$$

$$= \sum_{i=1}^{m} E[-\frac{1}{2}\log|\Psi| - \frac{n}{2}\log(2\pi) - \frac{1}{2}(x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{(-1)}(x^{(i)} - \mu - \Lambda z^{(i)})] \tag{61}$$

去掉不相关的前两项后，对Λ求梯度，有：

$$\nabla_\Lambda \sum_{i=1}^{m} -E[\frac{1}{2}(x^{(i)} - \mu - \Lambda z^{(i)})^T \Psi^{(-1)}(x^{(i)} - \mu - \Lambda z^{(i)})] \tag{62}$$

$$= \sum_{i=1}^{m} \nabla_\Lambda E[-tr\frac{1}{2}{z^{(i)}}^T \Lambda^T \Psi^{-1} \Lambda z^{(i)} + tr{z^{(i)}}^T \Lambda^T \Psi^{-1}(x^{(i)} - \mu)] \tag{63}$$

$$= \sum_{i=1}^{m} \nabla_\Lambda E[-tr\frac{1}{2}\Lambda^T \Psi^{-1} \Lambda z^{(i)} {z^{(i)}}^T + tr\Lambda^T \Psi^{-1}(x^{(i)} - \mu){z^{(i)}}^T] \tag{64}$$

$$= \sum_{i=1}^{m} E[-\Psi^{-1}\Lambda z^{(i)} {z^{(i)}}^T + \Psi^{-1}(x^{(i)} - \mu){z^{(i)}}^T] \tag{65}$$

第一步到第二步利用了$tra = a$（a是实数时）和$trAB = trBA$。最后一步利用了$\nabla_A trABA^TC = CAB + C^TAB^T$。最后让其值为0，并且化简：

$$\sum_{i=1}^{m} \Lambda E_{z^{(i)}\sim Q_i}[z^{(i)} {z^{(i)}}^T] = \sum_{i=1}^{m} (x^{(i)} - \mu) E_{z^{(i)}\sim Q_i}[{z^{(i)}}^T] \tag{66}$$

然后得到：

$$\Lambda = (\sum_{i=1}^{m} (x^{(i)} - \mu) E_{z^{(i)}\sim Q_i}[{z^{(i)}}^T])(\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}[z^{(i)} {z^{(i)}}^T])^{-1} \tag{67}$$

到这里我们发现，这个公式有点眼熟，与之前回归中的最小二乘法矩阵形式类似：

$$\theta^T = (y^T X)(X^T X)^{-1} \tag{68}$$

这里解释一下两者的相似性，我们这里的x是z的线性函数（包含了一定的噪声）。在E步得到z的估计后，我们找寻的Λ实际上是x和z的线性关系。而最小二乘法也是去找特征和结果直接的线性关系。我们继续求得括号里面的值，根据我们之前对$z|x$的定义，我们知道：

$$E_{z^{(i)}\sim Q_i}[{z^{(i)}}^T] = \mu^T_{z^{(i)}|x^{(i)}} \tag{69}$$

$$E_{z^{(i)}\sim Q_i}[z^{(i)} {z^{(i)}}^T] = \mu_{z^i|x^{(i)}} \mu^T_{z^{(i)}|x^{(i)}} + \Sigma_{z^{(i)}|x^{(i)}} \tag{70}$$

第一步根据z的条件分布得到，第二步根据$Cov(Y) = E[YY^T] - E[Y]E[Y]^T$得到将上面的结果代入上式中得到：

$$\Lambda = (\sum_{i=1}^{m} (x^{(i)} - \mu) \mu^T_{z^{(i)}|x^{(i)}})(\sum_{i=1}^{m} \mu_{z^i|x^{(i)}} \mu^T_{z^{(i)}|x^{(i)}} + \Sigma_{z^{(i)}|x^{(i)}})^{-1} \tag{71}$$

至此，我们得到了Λ，注意一点是$E[z]$和$E[zz^T]$ 的不同，后者需要求z的协方差。其他参数的迭代公式如下：

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \tag{72}$$

均值μ在迭代过程中值不变。同样的，我们直接给出Ψ的表达式：

$$\Phi = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}x^{(i)^T} - x^{(i)}\mu^T_{z^{(i)}|x^{(i)}}\Lambda^T - \Lambda\mu_{z^{(i)}|x^{(i)}}x^{(i)^T} + \Lambda(\mu_{z^{(i)}|x^{(i)}}\mu^T_{z^{(i)}|x^{(i)}} + \Sigma_{z^{(i)}|x^{(i)}})\Lambda^T \tag{73}$$

然后将Φ上的对角线上元素抽取出来放到对应的Ψ（$\Psi_{ii} = \Phi_{ii}$）中，进而得到Ψ。

3 小结

降维是解决维度灾难和过拟合的重要方法，除了直接的特征选择外，我们还可以采用算法的途径对特征进行筛选，线性的降维方法以PCA为代表，在PCA中，我们只要直接对数据矩阵进行中心化然后求奇异值分解或者对数据的协方差矩阵进行分解就可以得到其主要维度。

References