当我们有一个来自多个高斯模型的混合的数据集（a mixture of several Gaussians） $x^{(i)} \in R^n$，那么就可以用期望最大化算法（EM algorithm）来对这个混合模型进行拟合。这种情况下，对于有充足数据的问题，我们通常假设可以从数据中识别出多个高斯模型结果。例如，如果我们的训练样本集合规模 m 远远大于数据的维度 n，就符合这种情况。然后来考虑一下反过来的情况，即 $n >> m$。在这样的问题中，可能用单独的一个高斯模型对数据建模都很难，更不用说多个高斯模型的混合了。由于 m 个数据点张成（span）的空间只是一个 n 维空间的低维子空间。如果用高斯模型进行建模，然后还是用常规的最大似然估计来计算均值和方差，得到的则是：

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

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

我们会发现这里的 Σ 是一个奇异矩阵，这也就意味着其逆矩阵 Σ^{-1} 不存在。但是在计算多元高斯分布的常规密度是需要这些变量。还有另一种方法来讲述清楚这个难题，即对参数的最大似然估计会产生一个高斯分布，其概率分布在由样本数据所张成的仿射空间中，对应着一个奇异的协方差矩阵。

通常情况下，除非 m 比 n 大很多，否则最大似然估计得到的均值和方差都会很差。尽管如此，我们还是希望能用已有的数据，拟合出一个合理的高斯模型，而且还希望能够识别出数据中的某些有意义的协方差结构（covariance）。在接下来的内容中，我们首先回顾一个对 Σ 的两个可能的约束，这两个约束条件能让我们使用小规模数据来拟合 Σ，但都不能就我们的问题给出让人满意的解。接下来我们要讨论一下高斯模型的一些特点，这些后面会用得上，具体来说就是如何找到高斯模型的边界和条件分布。最好，我们会讲一下因子分析模型（factor analysis model），以及对应的期望最大化算法（EM algorithm）。

1. Σ 的约束条件

如果我们没有充足的数据来拟合一个完整的协方差矩阵，可以对矩阵空间 Σ 给出某些约束条件。例如，我们可以选择去拟合一个对角的协方差矩阵 Σ。这样，我们可以得到这样一个协方差矩阵的最大似然估计可以由满足如下条件的对角矩阵 Σ 给出：

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

因此，Σ_{jj} 就是对数据中第 j 个坐标位置的方差值的经验估计。由于高斯模型的概率密度的轮廓是椭圆形的，故对角线矩阵 Σ 对于的就是椭圆长轴与坐标轴平行的高斯模型。有时候，我们还要对这个协方差矩阵给出进一步的约束，不仅设为对角的，还要求所有对角元素都相等。这时候，就有 $\Sigma = \sigma^2 I$，其中 σ^2 是我们控制的参数。对这个 σ^2 的最大似然估计则为：

$$\sigma^2 = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

这种模型对应的是密度函数为圆形轮廓的高斯模型（在二维空间中是圆形，在更高维度当中就是球（spheres）或者超球体（hyper-spheres））。

如果我们对数据要拟合一个完整的，不受约束的协方差矩阵 Σ，就必须满足 $m \geq n + 1$，这样才使得对 Σ 的最大似然估计不是奇异矩阵。在上面提到的两个约束条件只下，只要 $m \geq 2$，我们就能获得非奇异的 Σ。

然而，将 Σ 限定为对角矩阵，也就意味着对数据中不同坐标的 x_i, x_j 建模都将是不相关的、而且相互独立。通常，还是从样本数据里面获得某些有趣的相关信息结构比较好；如果使用上面对 Σ 的约束，就可能没办法获取这些信息了。在本章讲义中，我们会提到因子分析模型（factor analysis model），这个模型使用的参数比对角矩阵 Σ 更多，而且能从数据中获得某些相关性信息，但也不能对完整的协方差矩阵进行拟合。

2. 多元高斯模型的边界和条件

在讲解因子分析之前，我们先讨论下一个联合多元高斯分布（a joint multivariate Gaussian distribution）下的随机变量的条件分布（conditional distribution）和边界分布（marginal distributions）。假如我们有一个值为向量的随机变量

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

其中，$x_1 \in R^r, x_2 \in R^s$，因此 $x \in R^{r+s}$。设 $x \sim N(\mu, \Sigma)$，即以 μ, Σ 为参数的正态分布，则这两个参数为：

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

其中，$\mu_1 \in R^r, \mu_2 \in R^s, \Sigma_{11} \in R^{r \times r}, \Sigma_{12} \in R^{r \times s}$，以此类推。由于协方差矩阵对称，所以有 $\Sigma_{21} = \Sigma_{12}^T$。

2.1 边界分布

基于我们的假设，x_1, x_2 是联合多元高斯分布。那么 x_1 的边界分布是什么？不难看出 x_1 的期望 $E[x_1] = \mu_1$，而协方差 $Cov(x_1) = E[(x_1 - \mu_1)(x_1 - \mu_1)^T] = \Sigma_{11}$。为了证明上述等式最右边的等号成立，即等于 Σ_{11}，可以利用一下等式：

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

通过上式的计算，我们可以证明等号成立。故高斯分布的边界分布本身也是高斯分布，所以我们可以给出正态分布 $x_1 \sim N(\mu_1, \Sigma_{11})$ 来作为 x_1 的边界分布。

2.2 条件分布

此外，我们还可以提出另一个问题，给定 x_2 的情况下 x_1 的条件分布是什么呢？通过参考多元高斯分布的定于，就能得到这个条件分布 $x_1|x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})$ 为：

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

> 根据条件概率公式，可以得到条件分布满足一下等式：
>
> $$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{p(x)}{p(x_2)}$$
>
> 其中，$x \sim N(\mu, \Sigma), x_2 \sim N(\mu_2, \Sigma_{22})$。将高斯分布概率公式代入可得条件分布的参数，计算比较复杂。

在下一节对因子分析模型的讲解中，上面这些公式就很有用了，可以帮助寻找高斯分布的条件和边界分布。

3. 因子分析模型

在因子分析模型中，我们制定一个在 (x, z) 上的联合分布，如下所示，其中 $z \in R^k$ 是一个潜在随机变量：

$$
\begin{aligned}
z &\sim N(0, I) \\
x|z &\sim N(\mu + \Lambda z, \Psi)
\end{aligned}
$$

上面的式子中，我们这个模型的参数是向量 $\mu \in R^n$，矩阵 $\Lambda \in R^{n \times k}$，以及一个对角矩阵 $\Psi \in R^{n \times n}$。k 的值通常都选择比 n 小一点的。这样，我们就设想每个数据点 $x^{(i)}$ 都是通过在一个 k 维度的多元高斯分布 $z^{(i)}$ 中采样获得的。然后，通过计算 $\mu + \Lambda z^{(i)}$，就可以映射到实数域 R^n 中的一个 k 维仿射空间。最后，在 $\mu + \Lambda z^{(i)}$ 上加上协方差 Ψ 作为噪声，就得到了 $x^{(i)}$。

反过来，我们也可以使用下面的设定来定义因子分析模型：

$$
\begin{aligned}
z &\sim N(0, I) \\
\epsilon &\sim N(0, \Psi) \\
x &= \mu + \Lambda z + \epsilon
\end{aligned}
$$

其中的 z 和 ϵ 是相互独立的。然后我们来看这个模型定义的分布，其中，随机变量 z 和 x 有一个联合高斯分布：

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

然后我们要找到 μ_{zx}, Σ。

已知 $E[z] = 0$，此外我们可以计算得：

$$\begin{aligned} E[x] &= E[\mu + \Lambda z + \epsilon] \\ &= \mu + \Lambda E[z] + E[\epsilon] \\ &= \mu \end{aligned}$$

综合以上这些条件，就得到了：

$$\mu_{zx} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}$$

下一步就是要找出 Σ，根据分块矩阵可以得到：

$$\begin{aligned} \Sigma &= \begin{bmatrix} E[(z - Ez)(z - Ez)^T] & E[(z - Ez)(x - Ex)^T] \\ E[(x - Ex)(z - Ez)^T] & E[(x - Ex)(x - Ex)^T] \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{zz} & \Sigma_{zx} \\ \Sigma_{xz} & \Sigma_{xx} \end{bmatrix} \end{aligned}$$

由于 z 是一个正态分布，很容易知道 $\Sigma_{zz} = Cov(z) = I$。另外：

$$\begin{aligned} \Sigma_{zx}^T = \Sigma_{xz} &= E[(x - Ex)(z - Ez)^T] \\ &= E[(\mu + \Lambda z + \epsilon - \mu)(z - 0)^T] \\ &= E[zz^T]\Lambda + E[\epsilon z^T] \\ &= \Lambda \end{aligned}$$

同样的方法，我们可以找到 Σ_{xx}：

$$\begin{aligned} \Sigma_{xx} &= E[(x - Ex)(x - Ex)^T] \\ &= E[(\Lambda z + \epsilon)(\Lambda z + \epsilon)^T] \\ &= E[\Lambda z z^T \Lambda^T + \epsilon z^T \Lambda^T + \Lambda z \epsilon^T + \epsilon \epsilon^T] \\ &= \Lambda\Lambda^T + \Psi \end{aligned}$$

故，

$$\begin{bmatrix} z \\ x \end{bmatrix} \sim N(\begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}, \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix})$$

因此，我们还能发现 x 的边界分布为 $x \sim N(\mu, \Lambda\Lambda^T + \Psi)$。所以，给定一个训练样本集合 $\{x^{(i)}; i = 1, \ldots, m\}$，我们可以写成参数的最大似然估计函数的对数形式：

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

为了进行最大似然估计，我们就要最大化上面这个关于参数的函数。但确切地对上面这个方程进行最大化，是很难的，而且我们都知道没有算法能够以封闭形式来实现这个最大化。所以，我们就改用期望最大化算法。

4. 因子分析模型的期望最大化算法

E 步骤的推导很简单。只需要计算出来 $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};\mu,\Lambda,\Psi)$。根据多元高斯模型及其条件分布公式可知：$z^{(i)}|x^{(i)};\mu,\Lambda,\Psi \sim N(\mu_{z^{(i)}|x^{(i)}},\Sigma_{z^{(i)}|x^{(i)}})$，其中：

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

所以，通过对 $\mu_{z^{(i)}|x^{(i)}},\Sigma_{z^{(i)}|x^{(i)}}$ 进行这样的定义，就能得到：

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

接下来就是 M 步骤了，这里需要最大化下面这个关于参数 μ,Λ,Ψ 的函数值：

$$\sum_{i=1}^{m}\int_{z^{(i)}} Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\mu,\Lambda,\Psi)}{Q_i(z^{(i)})}dz^{(i)} \qquad (4)$$

在本文中我们仅对 Λ 进行优化，关于 μ,Ψ 的更新就作为练习留给读者自己进行推导了。把等式（4）简化成下面的形式：

$$\sum_{i=1}^{m}\int_{z^{(i)}} Q_i(z^{(i)})[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi)+\log p(z^{(i)})-\log Q_i(z^{(i)})]dz^{(i)} \qquad (5)$$

$$\sum_{i=1}^{m} E_{z^{(i)}\sim Q_i}[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi)+\log p(z^{(i)})-\log Q_i(z^{(i)})] \qquad (6)$$

上面的等式中，下标 $z^{(i)}\sim Q_i$ 表示的意思是这个期望是关于从 Q_i 中取得的 $z^{(i)}$ 的。在后续的推导过程中，如果没有歧义的情况下，我们就会把这个下标省略掉。观察等式（6），事实上我们只需要最大化第一项，因为第二项 $p(z^{(i)})$、第三项 $Q_i(z^{(i)})$ 不依赖参数；第二项不依赖参数很好理解，$z\sim N(0,I)$；第三项看似含有参数，但参数在 E 步骤中已经固定，不影响结果。故省略部分项目后，我们只需要最大化下式：

$$\begin{aligned}&\sum_{i=1}^{m}E[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi)]\\ &\quad=\sum_{i=1}^{m}E[\log\frac{1}{(2\pi)^{(n/2)}|\Psi|^{1/2}}\exp\big(-\frac{1}{2}(x^{(i)}-\mu-\Lambda z^{(i)})^T\Psi^{-1}(x^{(i)}-\mu-\Lambda z^{(i)})\big)]\\ &\quad=\sum_{i=1}^{m}E[-\frac{1}{2}\log|\Psi|-\frac{1}{2}\log(2\pi)-\frac{1}{2}(x^{(i)}-\mu-\Lambda z^{(i)})^T\Psi^{-1}(x^{(i)}-\mu-\Lambda z^{(i)})]\end{aligned}$$

我们先对上面的函数进行关于 Λ 的最大化。可见只有最后的一项依赖 Λ，同时利用下面几个结论：

$$
\begin{aligned}
tra &= a && a \in R \\
trAB &= trBA \\
\nabla_A trABA^TC &= CAB + C^TAB
\end{aligned}
$$

可以得到：

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

令导数为 0，然后简化，就能得到：

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

接下来，求解 Λ，就能得到：

$$
\Lambda = \Big(\sum_{i=1}^{m} (x^{(i)} - \mu) E_{z^{(i)} \sim Q_i}[{z^{(i)}}^T] \Big) \Big(\sum_{i=1}^{m} E_{z^{(i)} \sim Q_i}[z^{(i)} {z^{(i)}}^T] \Big)^{-1} \tag{7}
$$

有一个很有意思的地方需要注意，上面这个等式和用最小二乘线性回归推出的正则方程有密切关系：

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

与之类似，这里的 x 是一个关于 z（以及噪声）的线性方程。考虑在 E 步骤中对 z 已经给出了猜测，接下来就可以尝试来对 x, z 之间的未知线性量 Λ 进行估计。接下来不出意料，我们就会得到某种类似正则方程的结果。然而，这个还是和利用对 z 的 “最佳猜测” 进行最小二乘算法有一个很大的区别。这一点我们很快就会看到了。

为了完成 M 步骤的更新，接下来我们要解出等式（7）当中的期望值。根据 Q_i 的均值和协方差，以及公式 $Cov(Y) = E[YY^T] - E[Y]E[Y]^T$，得到：

$$
\begin{aligned}
E_{z^{(i)} \sim Q_i}[{z^{(i)}}^T] &= \mu_{z^{(i)}|x^{(i)}}^T \\
E_{z^{(i)} \sim Q_i}[z^{(i)} {z^{(i)}}^T] &= \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}
\end{aligned}
$$

把这个公式代入等式（7），就得到了 M 步骤中 Λ 的更新规则：

$$
\Lambda = \Big(\sum_{i=1}^{m} (x^{(i)} - \mu) \mu_{z^{(i)}|x^{(i)}}^T \Big) \Big(\sum_{i=1}^{m} \mu_{z^{(i)}|x^{(i)}} \mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}} \Big)^{-1} \tag{8}
$$

上面这个等式中，要特别注意等号右边这一侧的 $\Sigma_{z^{(i)}|x^{(i)}}$。这是一个 $z^{(i)}$ 的后验分布的协方差，在 M 步骤中必须要考虑到这个后验分布中 $z^{(i)}$ 的不确定性。

最后，我们对参数 μ, Ψ 进行优化，不难发现其中的 μ 为：

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

由于这个值不随着参数的变化而改变（也就是说，和 Λ 的更新不同，这里等式右侧不依赖 $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)};\mu,\Lambda,\Psi)$，其中这个 $Q_i(z^{(i)})$ 是依赖参数的），所以这个只需要计算一次就可以，在算法运行过程中，也不需要进一步更新。类似的，对角矩阵 Ψ 也可以通过计算下面这个式子来获得：

$$\Phi = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}{x^{(i)}}^T - x^{(i)}\mu_{z^{(i)}|x^{(i)}}^T\Lambda^T - \Lambda\mu_{z^{(i)}|x^{(i)}}{x^{(i)}}^T + \Lambda(\mu_{z^{(i)}|x^{(i)}}\mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}})\Lambda^T$$

然后只需要设 $\Psi_{ii} = \Phi_{ii}$（也就是说，设 Ψ 为一个仅仅包含矩阵 Φ 中对角线元素的对角矩阵）。