C1-12

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou

szxjzhou@163.com

2016.9.17

第十三集：高斯混合模型

- 高斯混合(Mixture of Gaussians)
- 朴素贝叶斯混合(Mixture of Naive Bayes)
- 因子分析(Factor Analysis)
- 高斯分布的性质(Properties of Gaussians)

第十三集：高斯混合模型

- 高斯混合
  - 在上一节中推导出了EM算法的一般形式，现在将在这个一般形式的基础上推导高斯混合模型的结论
    - 在EM算法中将重复迭代下面两步直至收敛
      - 1、E-step：更新每个样本x对应类标z的分布Q
        - 此时已知参数θ，用贝叶斯公式即可得到新的分布Q

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta)$$

      - 2、M-step：更新参数θ的取值
        - 此时已知类标z的分布Q，用极大似然法即可得到新的参数θ

$$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

第十三集：高斯混合模型

- 高斯混合
  - 高斯混合模型根据之前所述，其类标z的概率p(z)满足多项式分布，而p(x|z)满足混合高斯分布
    - 对于EM算法的E-step来说需要在已知参数θ的情况下更新每个样本x对应类标z的分布Q，而这通过贝叶斯公式可以很容易地计算出来

$$w_j^{(i)} = Q_i(z^{(i)} = j) = P(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma)$$

$$p(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)} = l; \mu, \Sigma)p(z^{(i)} = l; \phi)}$$

由于此时多项式分布参数φ和高斯分布参数μ,Σ均已知，因此直接代入贝叶斯公式即可计算出结果

    - 而对于EM算法的M-step来说需要在已知类标z的分布Q的情况下用极大似然法计算出参数θ

$$\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} Q_i(z^{(i)} = j) \log \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

这里面需要计算的参数θ有多项式分布参数φ和高斯分布参数μ,Σ

第十三集：高斯混合模型

- 高斯混合
  - 对于多项式分布参数φ的计算

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot \phi_j}{w_j^{(i)}}$$

现在需要用极大似然法求出参数φ的权重，也就是需要对上面这个式子对于参数φ求偏微分并令其偏微分为 0.观察这个式子发现，log 的后面是个分数，分子是对应的高斯分布概率和对应的多项式分布概率的乘积。经过对数 log 运算后乘法变两项相加，而左边一项只有高斯分布的参数μ,Σ，仅右边一项有多项式分布参数φ.因此左边一项对于参数φ求偏微分后直接为 0，可不考虑，只需考虑右边一项的偏微分，即只考虑：

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j$$

但我们不能直接令其对于参数φ求偏微分为 0，因为我们还有一个限制条件：

$$\sum_{j=1}^{k} \phi_j = 1$$

因此我们需要引入 Lagrange 乘子将这两个式子合并为下面这个 Lagrange 函数

$$\mathcal{L}(\phi) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta(\sum_{j=1}^{k} \phi_j - 1)$$

接下来我们暂时不理会 Lagrange 乘子β的值是什么，将其看成一个参数，令函数对于参数φ求偏微分的结果为 0，可得：

$$\nabla_{\phi j}\mathcal{L}(\phi) = \nabla_{\phi j}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} \omega_j^{(i)} \log \phi j + \beta\left(\sum_{j=1}^{k} \phi j - 1\right)\right] = \sum_{i=1}^{m} \frac{\omega_j^{(i)}}{\phi j} + \beta = 0$$

由此可得参数φj 的表达式：

$$\phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)}}{-\beta}$$

现在需要解决 Lagrange 乘子β的值，现在要用上最后一个条件$\sum_{j=1}^{k}\phi j = 1$

$$1 = \sum_{j=1}^{k} \phi j = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} \omega_j^{(i)}}{-\beta} = \frac{m}{-\beta} \Leftrightarrow -\beta = m$$

由此可得：

$$\phi_j := \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)}$$

第十三集：高斯混合模型

- 高斯混合
  - 对于高斯分布参数μ的计算

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot\phi_j}{w_j^{(i)}}$$

现在需要用极大似然法求出参数μ的权重，也就是需要对上面这个式子对于参数μ求偏微分并令其偏微分为0.观察这个式子发现，log 的后面是个分数，分子是对应的高斯分布概率和对应的多项式分布概率的乘积。经过对数 log 运算后乘法变两项相加，而左边一项只有高斯分布的参数μ,Σ，仅右边一项有多项式分布参数φ.因此右边一项对于参数μ求偏微分后直接为 0，可不考虑，只需考虑左边一项的偏微分，即：

$$\nabla_{\mu l}\sum_{i=1}^{m}\sum_{j=1}^{k}\omega_j^{(i)}\log\frac{\frac{1}{(2\pi)^{n/2}\left|\Sigma_j\right|^{1/2}}exp\left(-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T\Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)\right)}{\omega_j^{(i)}}$$

$$=\nabla_{\mu l}\sum_{i=1}^{m}\sum_{j=1}^{k}\omega_j^{(i)}\left(\log\frac{1}{(2\pi)^{n/2}\left|\Sigma_j\right|^{1/2}}-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T\Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)+\log\phi j-\log\omega_j^{(i)}\right)$$

$$=\nabla_{\mu l}\sum_{i=1}^{m}\sum_{j=1}^{k}\omega_j^{(i)}\left(-\frac{1}{2}\left(x^{(i)}-\mu_j\right)^T\Sigma_j^{-1}\left(x^{(i)}-\mu_j\right)\right)$$

由此可进一步推知：

$$-\nabla_{\mu_l}\sum_{i=1}^{m}\sum_{j=1}^{k}w_j^{(i)}\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)$$

$$=\frac{1}{2}\sum_{i=1}^{m}w_l^{(i)}\nabla_{\mu_l}2\mu_l^T\Sigma_l^{-1}x^{(i)}-\mu_l^T\Sigma_l^{-1}\mu_l$$

$$=\sum_{i=1}^{m}w_l^{(i)}\left(\Sigma_l^{-1}x^{(i)}-\Sigma_l^{-1}\mu_l\right)$$

由此参数μ的值为：

$$\mu_l:=\frac{\sum_{i=1}^{m}w_l^{(i)}x^{(i)}}{\sum_{i=1}^{m}w_l^{(i)}}$$

第十三集：高斯混合模型

- 高斯混合
  - 由此可得最终的高斯混合模型的算法
    - 重复迭代下面两步直至收敛
      - 1、E-step：在已知参数θ的情况下更新每个样本x对应类标z的分布Q

$$w_j^{(i)} = Q_i(z^{(i)} = j) = P(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma)$$

$$p(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma)p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)} = l; \mu, \Sigma)p(z^{(i)} = l; \phi)}$$

由于此时多项式分布参数φ和高斯分布参数μ,Σ均已知，因此直接代入贝叶斯公式即可计算出结果

      - 2、M-step：在已知类标z的分布Q的情况下用极大似然法计算出参数θ

$$\phi_j := \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)},$$

$$\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}},$$

$$\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

第十三集：高斯混合模型

- 朴素贝叶斯混合
  - 之前讲到的高斯混合模型适合连续值下的聚类
    - 在分类时，高斯判别分析也被用于解决输入数据x是连续值的情况下的分类问题；而对于输入数据是离散值的情况，提出了朴素贝叶斯算法加以解决
    - 因此对应地，提出了朴素贝叶斯混合模型以解决离散值下的聚类问题，这一方法也被广泛用于文本聚类中
      - 在朴素贝叶斯问题中，我们提出了两种事件模型：一种是多元伯努利事件模型（每一维的特征都是{0,1}二元变量），另一种是多项式事件模型（每一维的特征为多元离散变量）。此处采用的是多元伯努利事件模型对问题进行研究（实际上也很容易推广到多项式事件模型的情况）
      - 首先先给出问题更加明确的定义
        - 由于采用的是多元伯努利事件模型因此样本x中每一维都是{0,1}二元变量（与朴素贝叶斯相同，表示某个词在某个文本中是否出现）
        - 由于没有给出类标因此需要假设隐随机变量z表示类标（与高斯混合模型相同，类标z表示某个样本x属于某个类的概率，这里为了简单起见只假设要聚到{0,1}两个类，其实也很容易推广到聚到多个类的情况）

第十三集：高斯混合模型

- 朴素贝叶斯混合
  - 接下来就之前对问题的定义做出更进一步的分析
    - 由于要聚到{0,1}两个类中，因此可假设类标z服从参数φ二项分布

$$\mathrm{p}\left(z^{(i)}\right)=\left(\phi_z\right)^{z^{(i)}}\left(1-\phi_z\right)^{1-z^{(i)}}$$

    - 与朴素贝叶斯类似，由于存在贝叶斯假设，因此根据链式法则和条件独立可知
      - 条件独立：在给定类标z下，x的每一维特征相互独立

$$\mathrm{p}\left(x^{(i)}\middle|z^{(i)}\right)=\mathrm{p}\left(x_1^{(i)}\middle|z^{(i)}\right)\mathrm{p}\left(x_2^{(i)}\middle|z^{(i)},x_1^{(i)}\right)\mathrm{p}\left(x_3^{(i)}\middle|z^{(i)},x_1^{(i)},x_2^{(i)}\right)\ldots\mathrm{p}\left(x_n^{(i)}\middle|z^{(i)},x_1^{(i)},\ldots,x_{n-1}^{(i)}\right)$$

$$=\mathrm{p}\left(x_1^{(i)}\middle|z^{(i)}\right)\mathrm{p}\left(x_2^{(i)}\middle|z^{(i)}\right)\mathrm{p}\left(x_3^{(i)}\middle|z^{(i)}\right)\ldots\mathrm{p}\left(x_n^{(i)}\middle|z^{(i)}\right)=\prod_{k=1}^{n}\mathrm{p}\left(x_k^{(i)}\middle|z^{(i)}\right)$$

第十三集：高斯混合模型

- 朴素贝叶斯混合
  - 同样地，采用EM算法对问题加以解决
    - 重复迭代下面两步直至收敛
      - 1、E-step：在已知参数θ的情况下更新每个样本x对应类标z的分布Q

$$\omega_j^{(i)} = Q_i\left(z^{(i)}=j\right) = p\left(z^{(i)}=j\middle|x^{(i)}\right) = \frac{p(x^{(i)}|z^{(i)}=j)p(z^{(i)}=j)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)}=l)p(z^{(i)}=l)}$$

$$\mathrm{p}\left(z^{(i)}\right) = \left(\phi_z\right)^{z^{(i)}}\left(1-\phi_z\right)^{1-z^{(i)}} \qquad \mathrm{p}\left(x^{(i)}\middle|z^{(i)}\right) = \prod_{k=1}^{n}\mathrm{p}\left(x_k^{(i)}\middle|z^{(i)}\right)$$

在这里l的值为{0,1}（因为假设要将样本聚成两个类），此时$p(x_k=1|z=0)$, $p(x_k=1|z=1)$和$p(z=1)$的式子均已给出，直接代入即可求得对应结果

      - 2、M-step：在已知类标z的分布Q的情况下用极大似然法计算出参数θ

$$\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i\left(z^{(i)}\right)\log\frac{p\left(x^{(i)},z^{(i)}\right)}{Q_i\left(z^{(i)}\right)} = \sum_{i=1}^{m}\sum_{j=1}^{k} Q_i\left(z^{(i)}=j\right)\log\frac{p\left(x^{(i)}\middle|z^{(i)}=j\right)p\left(z^{(i)}=j\right)}{Q_i\left(z^{(i)}=j\right)}$$

$$= \sum_{i=1}^{m}\sum_{j=0}^{1}\omega_j^{(i)}\log\frac{\prod_{k=1}^{n}p(x_k^{(i)}|z^{(i)}=j)\left(\phi_z\right)^{\mathrm{j}}\left(1-\phi_z\right)^{1-\mathrm{j}}}{\omega_j^{(i)}}$$

在这里面参数θ包括$p(x_k=1|z=0)$, $p(x_k=1|z=1)$和$p(z=1)$

第十三集：高斯混合模型

- 朴素贝叶斯混合
  - 对于二项分布参数p(z=1)的计算

$$\nabla_{\phi_z}\sum_{i=1}^{m}\sum_{j=0}^{1}\omega_j^{(i)}\log\frac{\prod_{k=1}^{n}p(x_k^{(i)}|z^{(i)}=j)\,(\phi_z)^{\mathrm{j}}(1-\phi_z)^{1-\mathrm{j}}}{\omega_j^{(i)}}=0$$

$$\Leftrightarrow\nabla_{\phi_z}\sum_{i=1}^{m}\sum_{j=0}^{1}\omega_j^{(i)}\left[j\log(\phi_z)+(1-j)\log(1-\phi_z)\right]=0$$

$$\Leftrightarrow\nabla_{\phi_z}\sum_{i=1}^{m}\omega_0^{(i)}\log(1-\phi_z)+\omega_1^{(i)}\log(\phi_z)=0$$

$$\Leftrightarrow\nabla_{\phi_z}\sum_{i=1}^{m}\left(1-\omega_1^{(i)}\right)\log(1-\phi_z)+\omega_1^{(i)}\log(\phi_z)=0\Leftrightarrow\sum_{i=1}^{m}\left(\omega_1^{(i)}-1\right)\frac{1}{1-\phi_z}+\omega_1^{(i)}\frac{1}{\phi_z}=0$$

$$\Leftrightarrow\sum_{i=1}^{m}\left(\omega_1^{(i)}-1\right)\phi_z+\omega_1^{(i)}(1-\phi_z)=0\Leftrightarrow\sum_{i=1}^{m}\omega_1^{(i)}-\phi_z=0\Leftrightarrow\phi_z=\frac{\sum_{i=1}^{m}\omega_1^{(i)}}{m}$$

第十三集：高斯混合模型

- 朴素贝叶斯混合
  - 对于参数$p(x_k=1|z=1)$ 的计算

$$\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i\left(z^{(i)}\right)\log\frac{p\left(x^{(i)},z^{(i)}\right)}{Q_i\left(z^{(i)}\right)}=\sum_{i=1}^{m}\sum_{j=1}^{k} Q_i\left(z^{(i)}=j\right)\log\frac{p\left(x^{(i)}\middle|z^{(i)}=j\right)p\left(z^{(i)}=j\right)}{Q_i\left(z^{(i)}=j\right)}$$

$$=\sum_{i=1}^{m}\sum_{j=0}^{1}\omega_j^{(i)}\log\frac{\prod_{k=1}^{n}p(x_k^{(i)}|z^{(i)}=j)\left(\phi_z\right)^{\mathrm{j}}\left(1-\phi_z\right)^{1-\mathrm{j}}}{\omega_j^{(i)}}$$

先定义如下两个记号：

$$\phi_{k|z=0}=p\left(x_k=1\middle|z=0\right),\phi_{k|z=1}=p\left(x_k=1\middle|z=1\right)$$

因此：

$$\mathrm{p}\left(\mathrm{x}\middle|\mathrm{z}=0\right)=\prod_{k=1}^{n}p\left(x_k\middle|z=0\right)=\prod_{k=1}^{n}\left(\phi_{k|z=0}\right)^{x_k}\left(1-\phi_{k|z=0}\right)^{1-x_k}$$

$$\mathrm{p}\left(\mathrm{x}\middle|\mathrm{z}=1\right)=\prod_{k=1}^{n}p\left(x_k\middle|z=1\right)=\prod_{k=1}^{n}\left(\phi_{k|z=1}\right)^{x_k}\left(1-\phi_{k|z=1}\right)^{1-x_k}$$

$$\nabla_{\phi_{k|z=0}}\sum_{i=1}^{m}\sum_{j=0}^{1}\omega_j^{(i)}\log\frac{\prod_{k=1}^{n}p\left(x_k^{(i)}\middle|z^{(i)}=j\right)\left(\phi_z\right)^{\mathrm{j}}\left(1-\phi_z\right)^{1-\mathrm{j}}}{\omega_j^{(i)}}=0$$

$$\Leftrightarrow\nabla_{\phi_{k|z=0}}\sum_{i=1}^{m}\sum_{j=0}^{1}\omega_j^{(i)}\log\prod_{k=1}^{n}p\left(x_k^{(i)}\middle|z^{(i)}=j\right)=0\Leftrightarrow\nabla_{\phi_{k|z=0}}\sum_{i=1}^{m}\sum_{j=0}^{1}\omega_j^{(i)}\sum_{k=1}^{n}\log p\left(x_k^{(i)}\middle|z^{(i)}=j\right)$$

$$=0\Leftrightarrow\nabla_{\phi_{k|z=0}}\sum_{i=1}^{m}\left[\omega_0^{(i)}\sum_{k=1}^{n}\log p\left(x_k^{(i)}\middle|z^{(i)}=0\right)+\left(1-\omega_0^{(i)}\right)\sum_{k=1}^{n}\log p\left(x_k^{(i)}\middle|z^{(i)}=1\right)\right]=0$$

$$\Leftrightarrow\nabla_{\phi_{k|z=0}}\sum_{i=1}^{m}\left[\omega_0^{(i)}\sum_{k=1}^{n}\log p\left(x_k^{(i)}\middle|z^{(i)}=0\right)\right]=0$$

$$\Leftrightarrow\nabla_{\phi_{k|z=0}}\sum_{i=1}^{m}\left[\omega_0^{(i)}\log\left(\phi_{k|z=0}{}^{1\left\{x_k^{(i)}=1\right\}}\left(1-\phi_{k|z=0}\right)^{1-1\left\{x_k^{(i)}=1\right\}}\right)\right]=0$$

$$\Leftrightarrow\nabla_{\phi_{k|z=0}}\sum_{i=1}^{m}\left[\omega_0^{(i)}1\left\{x_k^{(i)}=1\right\}\log\phi_{k|z=0}+\omega_0^{(i)}\left(1-1\left\{x_k^{(i)}=1\right\}\right)\log\left(1-\phi_{k|z=0}\right)\right]=0$$

$$\Leftrightarrow\sum_{i=1}^{m}\left[\omega_0^{(i)}1\left\{x_k^{(i)}=1\right\}\frac{1}{\phi_{k|z=0}}-\omega_0^{(i)}\left(1-1\left\{x_k^{(i)}=1\right\}\right)\frac{1}{1-\phi_{k|z=0}}\right]=0$$

$$\Leftrightarrow\sum_{i=1}^{m}\left[\omega_0^{(i)}1\left\{x_k^{(i)}=1\right\}\left(1-\phi_{k|z=0}\right)-\omega_0^{(i)}\left(1-1\left\{x_k^{(i)}=1\right\}\right)\phi_{k|z=0}\right]=0$$

$$\Leftrightarrow\sum_{i=1}^{m}\left[\omega_0^{(i)}1\left\{x_k^{(i)}=1\right\}-\omega_0^{(i)}\phi_{k|z=0}\right]=0\Leftrightarrow\sum_{i=1}^{m}\left(\omega_0^{(i)}1\left\{x_k^{(i)}=1\right\}\right)-\phi_{k|z=0}\sum_{i=1}^{m}\omega_0^{(i)}=0$$

$$\Leftrightarrow\phi_{k|z=0}=\frac{\sum_{i=1}^{m}\omega_0^{(i)}1\left\{x_k^{(i)}=1\right\}}{\sum_{i=1}^{m}\omega_0^{(i)}}$$

第十三集：高斯混合模型

- 朴素贝叶斯混合
  - 由此可得最终的朴素贝叶斯混合模型的算法
    - 重复迭代下面两步直至收敛
      - 1、E-step：在已知参数θ的情况下更新每个样本x对应类标z的分布Q

$$\omega_j^{(i)} = Q_i\left(z^{(i)}=j\right) = p\left(z^{(i)}=j\middle|x^{(i)}\right) = \frac{p(x^{(i)}|z^{(i)}=j)p(z^{(i)}=j)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)}=l)p(z^{(i)}=l)}$$

$$\mathrm{p}\left(z^{(i)}\right) = \left(\phi_z\right)^{z^{(i)}}\left(1-\phi_z\right)^{1-z^{(i)}} \qquad \mathrm{p}\left(x^{(i)}\middle|z^{(i)}\right) = \prod_{k=1}^{n}\mathrm{p}\left(x_k^{(i)}\middle|z^{(i)}\right)$$

在这里l的值为{0,1}（因为假设要将样本聚成两个类），此时$p(x_k=1|z=0)$, $p(x_k=1|z=1)$和$p(z=1)$的式子均已给出，直接代入即可求得对应结果

      - 2、M-step：在已知类标z的分布Q的情况下用极大似然法计算出参数θ

$$\phi_z = \frac{\sum_{i=1}^{m}\omega_1^{(i)}}{m}$$

$$p\left(x_k=1\middle|z=0\right) = \phi_{k|z=0} = \frac{\sum_{i=1}^{m}\omega_0^{(i)}1\left\{x_k^{(i)}=1\right\}}{\sum_{i=1}^{m}\omega_0^{(i)}}$$

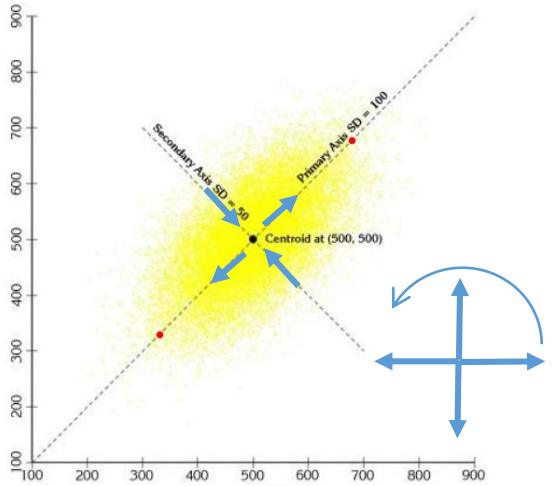
$$p\left(x_k=1\middle|z=1\right) = \phi_{k|z=1} = \frac{\sum_{i=1}^{m}\omega_1^{(i)}1\left\{x_k^{(i)}=1\right\}}{\sum_{i=1}^{m}\omega_1^{(i)}}$$

第十三集：高斯混合模型

- 因子分析
  - 之前讲到用EM算法可以对连续的样本用高斯混合模型进行聚类，但这种聚类建立在样本数目足够多的情况
    - 如果样本数量很少，就拟合出单高斯模型都已经很困难了，更别说是高斯混合模型
      - 特别地，如果样本特别少，只能生成某个子空间（比如说n维的向量，但是样本数m小于n，此时所有样本只能生成R^n空间的子空间）。此时得到的协方差矩阵Σ是个奇异矩阵（因为此时方差可以朝着任意相互垂直的方向进行扩展，从而垂直于子空间的方差将被无限压缩），因此不存在逆，从而无法得到对应的高斯分布（高斯分布中需要计算Σ^{-1}）

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

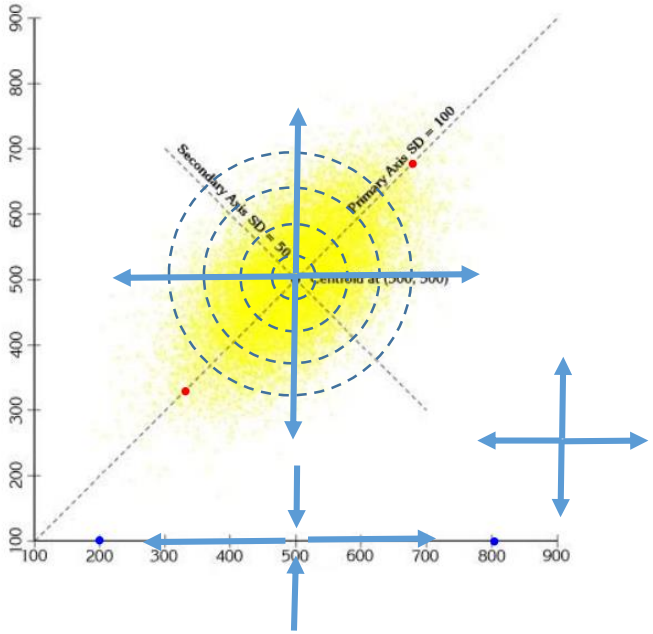
$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T$$



图片中展示了二维数据下高斯分布的均值和方差的关系，其中方差可以朝着任意垂直的方向扩展（协方差部分所起到的效果是将方差扩展的方向进行拉伸旋转）。可以想见如果只有图中所示的两个红点，并采取极大似然法得到方差，那么这个高斯分布一个轴将无限被挤压

第十三集：高斯混合模型

- 因子分析
  - 应对这种情况，一种解决办法是对于协方差矩阵Σ加以限制
    - 由于原来生成某个子空间的问题在于采用极大似然后垂直于子空间的方差将被无限压缩，因此一种解决办法是不采用极大似然法得到协方差矩阵，而直接令协方差矩阵为对角阵（即忽略原样本的协方差信息）



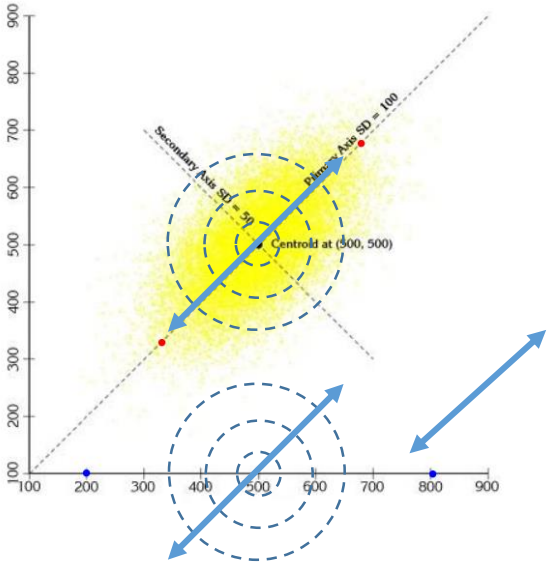
$$\Sigma_{jj} = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

在协方差矩阵为对角阵的情况下将不考虑协方差，而这使得方差只能沿着坐标轴的方向进行扩展。因此只要样本所构成的子平面不垂直于任意某个坐标轴（方向向量），所得到的协方差矩阵就不会是奇异矩阵

但如果样本所构成的子平面垂直于某个坐标轴（方向向量），该坐标轴（方向向量）对应维度的方差为0，由此对角阵中主对角线上有值为0，一定是奇异矩阵（没有对应的逆）

第十三集：高斯混合模型

- 因子分析
  - 由于直接忽略原样本的协方差信息，令协方差矩阵为对角阵不足以应对样本所构成的子平面垂直于某个坐标轴（方向向量）的情况，而对此需要提出更加严格的限制
    - 其中一种方法是进一步限制协方差矩阵的对角线元素，使之完全相同



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

在协方差矩阵为对角阵且主对角线上元素完全相同的情况下将不考虑协方差且方差相同，而这使得方差只能沿着坐标轴的方向进行等比例的扩展。因此只要样本不集中于一点（所有方差值均为0），所得到的协方差矩阵就不会是奇异矩阵

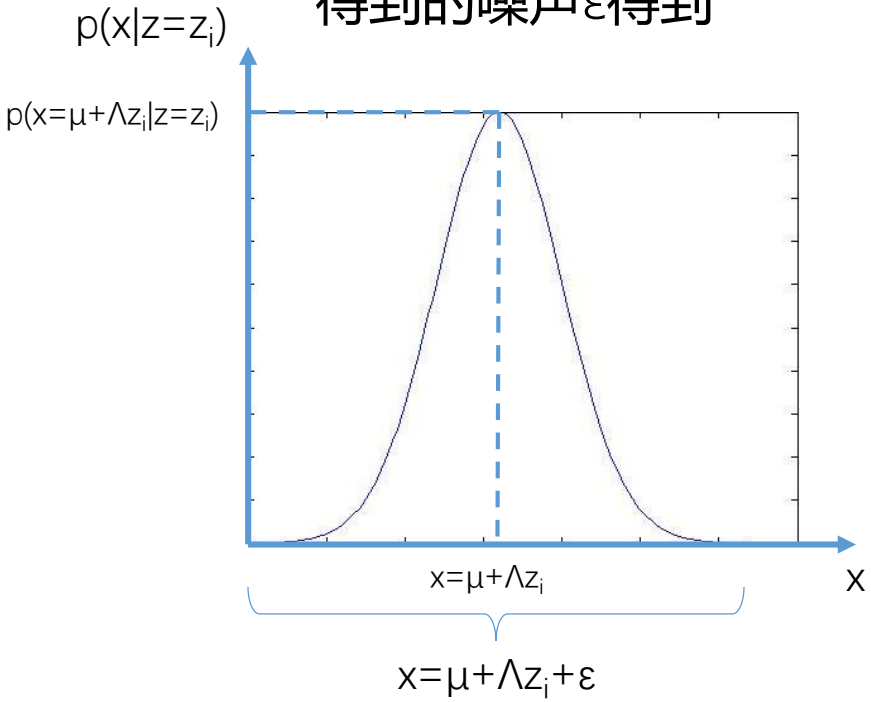
第十三集：高斯混合模型

- 因子分析
  - 之前讲到的对协方差矩阵加以约束的方法确实可以解决样本特别少的情况下对于高斯分布的拟合问题
    - 但这种拟合的代价在于丢失了很多信息（比如协方差信息），而这种信息很可能是有用的。而因子分析模型可以用来解决这个问题，它在保证拟合的有效性的同时尽量避免了有效信息的丢失
    - 因子分析模型的定义如下
      - 因子分析模型也是一种无监督的学习方法，因此只给定了数据样本x，而没有给定类标
      - 在因子分析模型中同样引入了一个隐随机变量z表示类标。但是在这里类标不再属于离散取值的分布（比如：二项分布、多项式分布），而是属于均值为0，协方差矩阵为单位阵I的高斯分布
      - 在因子分析模型中同时假定p(x|z)服从均值为μ+Λz，协方差矩阵为Ψ的高斯分布
        - $X \in R^n$, $z \in R^k$, $\mu \in R^n$, $\Lambda \in R^{n \times k}$, $\Psi \in R^{n \times n}$（一般k<n）

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

第十三集：高斯混合模型

- 因子分析
  - 实际上，还可以换个角度来看样本x和类标z的关系
    - 样本x可以被看作是从R^k空间上均值为0，协方差矩阵为单位阵I的高斯分布中抽取得到数据z后通过变换μ+Λz映射到R^n空间，再加上均值为0方差为Ψ的高斯分布中抽取得到的噪声ε得到



注意此处给出的是x|z满足均值为μ+Λz，协方差矩阵为Ψ的高斯分布，并不是说x满足这样的分布（事实上经过这样映射后x很可能不满足均值为μ+Λz，协方差矩阵为Ψ的高斯分布）。x|z表达的是在给定z的情况下对应值为x的概率分布，因此不妨设$z=z_i$，此时$p(x|z=z_i)$满足均值为$\mu+\Lambda z_i$，协方差矩阵为Ψ的高斯分布，由此可得如左图所示的高斯分布图线，从中可以看出x可由看作将数据z后通过变换μ+Λz再加上均值为0方差为Ψ的噪声ε得到

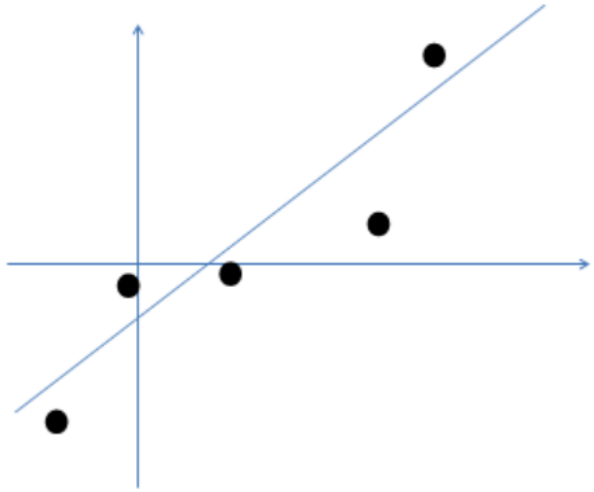
$$\begin{aligned} z &\sim \mathcal{N}(0, I) \\ x|z &\sim \mathcal{N}(\mu + \Lambda z, \Psi). \end{aligned}$$

$$\longrightarrow$$

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

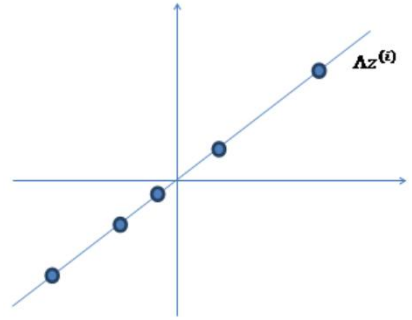
第十三集：高斯混合模型

- 因子分析
  - 现在用一个例子对于样本x和类标z的关系做进一步的说明

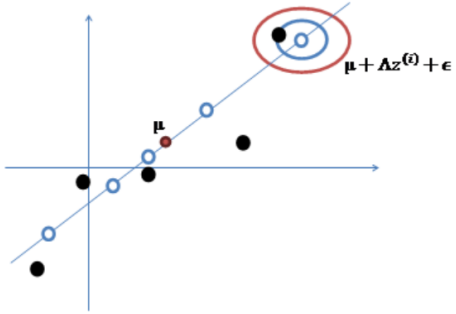


假设有5个R^2空间上的样本点，而这些样本实际上可以通过R^1空间上的均值为0，协方差矩阵为单位阵I的高斯分布中抽取得到数据z得到

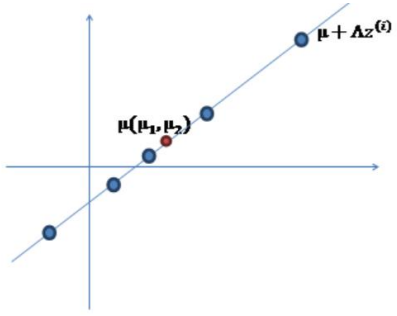
首先在R^1空间上的均值为0，协方差矩阵为单位阵I的高斯分布中抽取得到数据z



其次对于数据z通过矩阵Λ进行变换，映射到R^2空间



接下来将数据加上服从均值为0方差为Ψ高斯分布的噪声ε即可得到原数据样本



接下来将数据加上系数μ，使得直线的中心移动至向量μ处

第十三集：高斯混合模型

- 因子分析
  - 由此可以进一步计算出样本x和类标z的联合概率p(x,z)
    - 由于样本x和类标z均满足高斯分布，由此求其联合概率p(x,z)相当于求其联合高斯分布(joint Gaussian distribution)的均值μ_{zx}和协方差矩阵Σ

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

这里采取的是联合高斯分布的矩阵表达形式，其中z和x均为随机变量

    - 下面需要对此进行联合高斯分布的均值和方差的求解
      - 对于均值μ_{zx}的求解

$$\mathrm{E}[\mathrm{p}(\mathrm{z},\mathrm{x})] = \mu_{zx} = \begin{bmatrix} E[z] \\ E[x] \end{bmatrix} = \begin{bmatrix} \vec{0} \\ E[\mu + \Lambda \mathrm{z} + \varepsilon] \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \mu + \Lambda E[\mathrm{z}] + E[\varepsilon] \end{bmatrix} = \begin{bmatrix} \vec{0} \\ \mu \end{bmatrix}$$

第十三集：高斯混合模型

- 因子分析
  - 下面需要对此进行联合高斯分布的均值和方差的求解
    - 对于协方差矩阵Σ的求解

$$\mathrm{Cov}[\mathrm{p}(\mathrm{z},\mathrm{x})]=\Sigma=\begin{bmatrix}\Sigma_{zz} & \Sigma_{zx}\\ \Sigma_{xz} & \Sigma_{xx}\end{bmatrix}=E\left[\left(\begin{bmatrix}z\\ x\end{bmatrix}-\begin{bmatrix}\vec{0}\\ \mu\end{bmatrix}\right)\left(\begin{bmatrix}z\\ x\end{bmatrix}-\begin{bmatrix}\vec{0}\\ \mu\end{bmatrix}\right)^T\right]=E\begin{bmatrix}(z-\vec{0})(z-\vec{0})^T & (z-\vec{0})(x-\mu)^T\\ (x-\mu)(z-\vec{0})^T & (x-\mu)(x-\mu)^T\end{bmatrix}$$

$$=\begin{bmatrix}I & E[z(\mu+\Lambda\mathrm{z}+\varepsilon-\mu)^T]\\ E[(\mu+\Lambda\mathrm{z}+\varepsilon-\mu)z^T] & E[(\mu+\Lambda\mathrm{z}+\varepsilon-\mu)(\mu+\Lambda\mathrm{z}+\varepsilon-\mu)^T]\end{bmatrix}$$

$$=\begin{bmatrix}I & E[zz^T]\Lambda^T+E[z\varepsilon^T]\\ E[zz^T]\Lambda+E[\varepsilon z^T] & E[\Lambda zz^T\Lambda^T+\varepsilon z^T\Lambda^T+\Lambda z\varepsilon^T+\varepsilon\varepsilon^T]\end{bmatrix}$$

由于z和ε之间相互独立，因此$E[\varepsilon z^T]=E[\varepsilon]E[z^T]=0=E[z]E[\varepsilon^T]=E[z\varepsilon^T]$，由此可得：

$$\mathrm{Cov}[\mathrm{p}(\mathrm{z},\mathrm{x})]=\begin{bmatrix}I & E[zz^T]\Lambda^T+E[z\varepsilon^T]\\ E[zz^T]\Lambda+E[\varepsilon z^T] & E[\Lambda zz^T\Lambda^T+\varepsilon z^T\Lambda^T+\Lambda z\varepsilon^T+\varepsilon\varepsilon^T]\end{bmatrix}=\begin{bmatrix}I & I\Lambda^T\\ I\Lambda & \Lambda I\Lambda^T+\Psi\end{bmatrix}$$

$$=\begin{bmatrix}I & \Lambda^T\\ \Lambda & \Lambda\Lambda^T+\Psi\end{bmatrix}$$

第十三集：高斯混合模型

- 因子分析
  - 由此可得如下的联合高斯分布

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

  - 而事实上除了联合高斯分布p(z,x)外，还可衡量边缘高斯分布(marginal Gaussian distribution)p(x), p(z)和条件高斯分布(conditional Gaussian distribution)p(x|z), p(z|x)
    - 其中边缘高斯分布p(x), p(z)的得到方法比较简单，比如边缘高斯分布p(x)

$$\mathrm{x} \sim \mathcal{N}\left(\mathrm{E(x)}, \mathrm{Var(x)}\right), \mathrm{E(x)} = \mu, \mathrm{Var(x)} = E\left((x-\mu)(x-\mu)^T\right) = \Lambda\Lambda^T + \Psi$$

      - 而在得到p(x)后即可得到对应的极大似然表达式（在下一节中将继续讨论此表达式）

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

第十三集：高斯混合模型

- 高斯分布的性质
  - 下面将对于条件高斯分布的计算做出一般情况下的推导，在推导开始前先做出如下假设

假设x满足高斯分布，即$\mathrm{x}\sim\mathcal{N}(\mu,\Sigma)$，其中

$$\mathrm{x}=\begin{bmatrix}x_1\\x_2\end{bmatrix},E(x)=\mu=\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},Cov(x)=\Sigma=\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}$$

其中不难发现协方差矩阵$Cov(x)=\Sigma$为对称阵，即$\Sigma_{12}=\Sigma_{21}^T$，因此其逆矩阵有如下的性质：

$$\Sigma_{12}^{-1}=\left(\Sigma_{21}^T\right)^{-1}=\left(\Sigma_{21}^{-1}\right)^T$$

由此假设一个协方差矩阵$Cov(x)=\Sigma$的逆矩阵Λ，其定义为：

$$\Lambda=\Sigma^{-1}=\begin{bmatrix}\Lambda_{11}&\Lambda_{12}\\\Lambda_{21}&\Lambda_{22}\end{bmatrix}$$

其中不难得到矩阵Λ也是对称的，因为：

$$\Lambda^T=\begin{bmatrix}\Lambda_{11}^T&\Lambda_{21}^T\\\Lambda_{12}^T&\Lambda_{22}^T\end{bmatrix}=\begin{bmatrix}\Lambda_{11}^T&\left(\Sigma_{21}^{-1}\right)^T\\\left(\Sigma_{12}^{-1}\right)^T&\Lambda_{22}^T\end{bmatrix}=\begin{bmatrix}\Lambda_{11}^T&\Sigma_{12}^{-1}\\\Sigma_{21}^{-1}&\Lambda_{22}^T\end{bmatrix}=\begin{bmatrix}\Lambda_{11}&\Lambda_{12}\\\Lambda_{21}&\Lambda_{22}\end{bmatrix}=\Lambda$$

第十三集：高斯混合模型

- 高斯分布的性质
  - 求解条件高斯分布

$$\mathrm{p}(x_1|x_2)=\frac{p(x_1,x_2)}{\int_{x_1}p(x_1,x_2)\ dx_1}=\frac{1}{\int_{x_1}p(x_1,x_2)\ dx_1}\left[\frac{1}{(2\pi)^{n/2}|\Sigma|}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)\right]$$

由于$\mathrm{p}(x_1|x_2)$指的是在给定x_2的条件下事件x_1发生的概率，因此可以将x_2看成是给定的未知常量，因此原式可变为：

$$\mathrm{p}(x_1|x_2)=\frac{1}{const}\exp\left(-\frac{1}{2}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)^T\begin{bmatrix}\Lambda_{11}&\Lambda_{12}\\\Lambda_{21}&\Lambda_{22}\end{bmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)\right)$$

其中：

$$\begin{aligned}&\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)^T\begin{bmatrix}\Lambda_{11}&\Lambda_{12}\\\Lambda_{21}&\Lambda_{22}\end{bmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)\\&=\left[(x_1-\mu_1)^T\Lambda_{11}+(x_2-\mu_2)^T\Lambda_{21}\right](x_1-\mu_1)+\left[(x_1-\mu_1)^T\Lambda_{12}+(x_2-\mu_2)^T\Lambda_{22}\right](x_2-\mu_2)\\&=(x_1-\mu_1)^T\Lambda_{11}(x_1-\mu_1)+(x_2-\mu_2)^T\Lambda_{21}(x_1-\mu_1)+(x_1-\mu_1)^T\Lambda_{12}(x_2-\mu_2)+(x_2-\mu_2)^T\Lambda_{22}(x_2-\mu_2)\\&=\left(x_1^T\Lambda_{11}x_1-x_1^T\Lambda_{11}\mu_1-\mu_1^T\Lambda_{11}x_1+\mu_1^T\Lambda_{11}\mu_1\right)+\left(x_2^T\Lambda_{21}x_1-x_2^T\Lambda_{21}\mu_1-\mu_2^T\Lambda_{21}x_1+\mu_2^T\Lambda_{21}\mu_1\right)\\&+\left(x_1^T\Lambda_{12}x_2-x_1^T\Lambda_{12}\mu_2-\mu_1^T\Lambda_{12}x_2+\mu_1^T\Lambda_{12}\mu_2\right)+\left(x_2^T\Lambda_{22}x_2-x_2^T\Lambda_{22}\mu_2-\mu_2^T\Lambda_{22}x_2+\mu_2^T\Lambda_{22}\mu_2\right)\end{aligned}$$

在这里面除了x_1看成变量外，其余均看成给定的常量，因此原式可化简为：

$$\begin{aligned}&\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)^T\begin{bmatrix}\Lambda_{11}&\Lambda_{12}\\\Lambda_{21}&\Lambda_{22}\end{bmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)\\&=x_1^T\Lambda_{11}x_1-x_1^T\Lambda_{11}\mu_1-\mu_1^T\Lambda_{11}x_1+x_2^T\Lambda_{21}x_1-\mu_2^T\Lambda_{21}x_1+x_1^T\Lambda_{12}x_2-x_1^T\Lambda_{12}\mu_2+const\end{aligned}$$

由于矩阵Λ也是对称的，因此可将原式进一步改写为：

$$\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)^T\begin{bmatrix}\Lambda_{11}&\Lambda_{12}\\\Lambda_{21}&\Lambda_{22}\end{bmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)=x_1^T\Lambda_{11}x_1-2x_1^T\Lambda_{11}\mu_1+2x_1^T\Lambda_{12}(x_2-\mu_2)+const$$

由此得到了：

$$\mathrm{p}(x_1|x_2)=\frac{1}{const}\exp\left(-\frac{1}{2}\left(x_1^T\Lambda_{11}x_1-2x_1^T\Lambda_{11}\mu_1+2x_1^T\Lambda_{12}(x_2-\mu_2)+const\right)\right)$$

现在需要发现这个式子中哪些部分是均值，哪些部分是方差，因此在这里假设$\mathrm{p}(x_1|x_2)\sim\mathcal{N}(\mu_{1|2},\Sigma_{1|2})$，并列出关于参数$\mu_{1|2},\Sigma_{1|2}$的形式：

$$\begin{aligned}\mathrm{p}(x_1|x_2)&=\frac{1}{(2\pi)^{n/2}|\Sigma_{1|2}|}\exp\left(-\frac{1}{2}(x-\mu_{1|2})^T\Sigma_{1|2}^{-1}(x-\mu_{1|2})\right)=\frac{1}{const}\exp\left(-\frac{1}{2}(x-\mu_{1|2})^T\Sigma_{1|2}^{-1}(x-\mu_{1|2})\right)\\&=\frac{1}{const}\exp\left(-\frac{1}{2}\left(x^T\Sigma_{1|2}^{-1}x-\mu_{1|2}^T\Sigma_{1|2}^{-1}x-x^T\Sigma_{1|2}^{-1}\mu_{1|2}+\mu_{1|2}^T\Sigma_{1|2}^{-1}\mu_{1|2}\right)\right)\end{aligned}$$

由于矩阵$\Sigma_{1|2}$是对称的，因此可将原式进一步改写为：

$$\mathrm{p}(x_1|x_2)=\frac{1}{const}\exp\left(-\frac{1}{2}\left(x^T\Sigma_{1|2}^{-1}x-2x^T\Sigma_{1|2}^{-1}\mu_{1|2}+\mathrm{const}\right)\right)$$

$$\mathrm{p}(x_1|x_2)=\frac{1}{const}\exp\left(-\frac{1}{2}\left(x_1^T\Lambda_{11}x_1-2x_1^T\Lambda_{11}\mu_1+2x_1^T\Lambda_{12}(x_2-\mu_2)+const\right)\right)$$

与之前的形式对比可知

$$\Sigma_{1|2}^{-1}=\Lambda_{11},\Sigma_{1|2}^{-1}\mu_{1|2}=\Lambda_{11}\mu_1-\Lambda_{12}(x_2-\mu_2)$$

因此：

$$\Sigma_{1|2}=\Lambda_{11}^{-1}$$

$$\Sigma_{1|2}\Lambda_{11}\mu_{1|2}=\mu_{1|2}=\Sigma_{1|2}\Lambda_{11}\mu_1-\Sigma_{1|2}\Lambda_{12}(x_2-\mu_2)=\mu_1-\Lambda_{11}^{-1}\Lambda_{12}(x_2-\mu_2)$$

在此特别强调一点，分块矩阵与它的逆矩阵分块不是一一对应的（比如这里Λ_{12}不能认为等于Σ_{12}^{-1}）而且根据分块矩阵的求逆公式可得：

$$\begin{bmatrix}A&B\\C&D\end{bmatrix}^{-1}=\begin{bmatrix}M&-MBD^{-1}\\-D^{-1}CM&D^{-1}+D^{-1}CMBD^{-1}\end{bmatrix},M=(A-BD^{-1}C)^{-1}$$

由此可得：

$$\begin{aligned}\Lambda_{11}&=(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\\\Lambda_{12}&=-(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\end{aligned}$$

因此：

$$\mu_{1|2}=\mu_1-\Lambda_{11}^{-1}\Lambda_{12}(x_2-\mu_2)=\mu_1+\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)=\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$$

$$\Sigma_{1|2}=\Lambda_{11}^{-1}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$