声纹识别

概念

说话人识别，判断一段声音是谁说的。

声纹识别整体流程

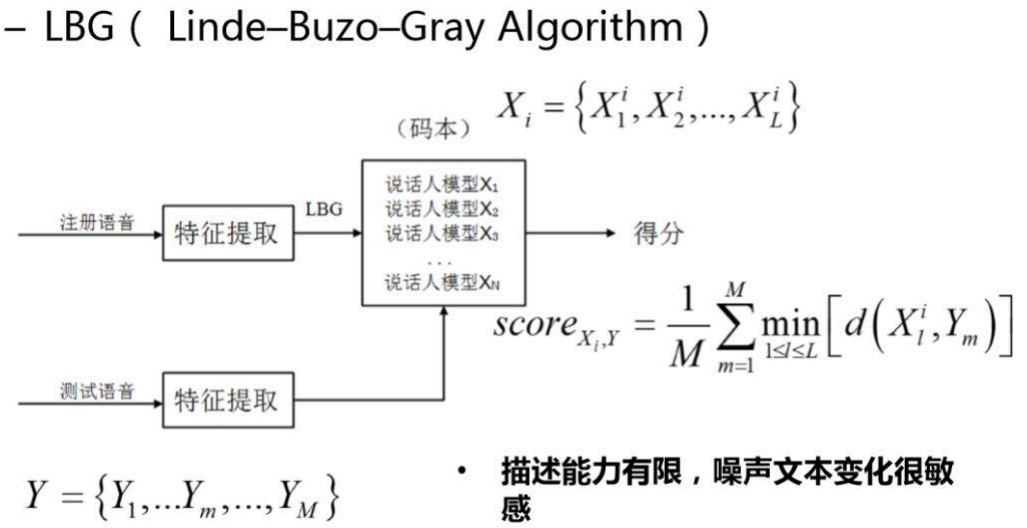
1. 抽取多个人多段话的 mfcc 特征
2. 训练 UBM-GMM 模型，计算每个人的均值超矢量 MAP
3. 获取 i-vector 特征
4. LDA 算法做降维（思路：LDA 是一个有监督的聚类算法，也是降维算法，该算法可以针对特征做降维，也就是我们的目标是说话人分类，但是我们抽取的特征是说话人信息构成的特征和信道信息构成的特征，我们的目标是剔除信道信息，同时考虑 label 就是说话人类别，所以 LDA 算法做降维就是剔除了信道信息，保留了说话人特征）

特征

一般使用 plp 或者 mfcc 做帧的特征抽取，其中帧往往是采样获得的，不是每个都取。

LBG 建模

根据对机器学习的常识，首先就会想到一种方法来做声纹识别：
假如已经训练好了说话人模型：一个人对应一个模型。
伪代码：
1）对帧进行采样
2）对帧进行特征抽取，记为 y1，y2...yn
3）将所有对特征向量代入每个说话人模型，计算所有对 yi 和说话人模型之间对距离之和，距离最小对 score 对应对说话人模型就是这段语音对应对模型。



缺点：
音频文件避免不了各种背景噪音。
就是音频里面往往又很多噪音，这些噪音会影响距离对大小。

高斯混合模型

与语音识别中对 GMM 用法不一样，声纹识别中的 GMM 是对一段语音中的帧会抽取特征，如 mfcc，一段语音会有多帧，将这些帧的语音特征放在一起来训练一个 GMM. 而声学模型是对音素的某一个状态训练 gmm 模型。
如果使用 mfcc 抽取特征，则为 13 维的特征向量。
这个时候高斯分布的维度也是 13 维，而高斯的数量是一个超参数可以调整。
训练高斯混合模型的时候，首先参数需要一个初始值，然后使用 em 算法逐渐收敛。
初始值的选取方法是：
使用 kmeas 对数据进行聚类，假如高斯分布对数量设为 m，则使用 kmeans 聚 m 类。然后对每个类求解高斯模型参数。这就是高斯混合模型对初始值。

GMM–UBM

思想：
其实 UBM 就是 GMM 模型，只是训练的目的不同，GMM 我们希望训练得到一个能够表征说话人音素分布的模型，而 UBM 是希望得到一个通用的模型，简单的说就是能够反应所有人共性的模型，其实某种意义上说就是一个取均值的过程。
操作方法：
对所有人对应的音频混杂在一起训练一个高斯混合模型。这个时候训练出来的高斯混合模型我们理解为“通用模型”
在通用模型上面进行微调，就可以得到每个人的模型。

MAP 自适应过程

虽然高斯混合模型的参数为四个：

$\theta = (w_i, u_i, \sigma_i), i = 1, 2, ..., C$ 和协方差矩阵。但是协方差矩阵一般设置为对角阵。

C 为 GMM 的混合阶数；说话人 X 的训练语音的特征向量序列为 x

1. 首先计算语音特征向量序列中的各个向量相对于每个 UBM 混元的概率得分。

$$p(O \mid \phi) = \frac{1}{(2\pi)^{d/2}\sigma^2}\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(\frac{O_t - \phi}{\sigma}\right)^2\right]$$

2. 对于 UBM 中的任意混元 i，特征向量 xi 对于它的后验分布概率为：

$$p(i \mid x_t, \lambda_\Omega) = \frac{\omega_i p(x_t \mid \mu_i, \sum_i)}{\sum_{j=1}^{C}\omega_j p(x_t \mid \mu_j, \sum_j)}$$

3. 利用后验概率计算均值所需要的统计量

$$p(i \mid \lambda_\Omega) = \sum_{t=1}^{T} p(i \mid x_t, \lambda_\Omega)$$

$$E_i(X) = \frac{1}{p(i \mid \lambda_\Omega)}\sum_{t=1}^{T} p(i \mid x_t, \lambda_\Omega)x_t$$

4. 最后利用上面两个统计量对 UBM 均值进行更新，其对任意混元 i 的均值更新表达式如下：

$$\hat{\mu}_i = \partial_i E_i(X) + (1 - \partial_i)\mu_i$$

自适应∂系数控制着旧估计与新估计之间的均衡，自适应算法就是对 UBM 参数做个微调，使得参数在一定背景的基础下调整到能够表征说话人发音特征，在语音数据不充分的情况下，没有覆盖到的发音特征可以用 UBM 的平均

发音特征来代替。第 2 步公式，反应了当前模型下，第 j 个观测数据，来自第 K 个分模型的概率，称为分模型 K 对观测数据 yj 的响应度。

总结:

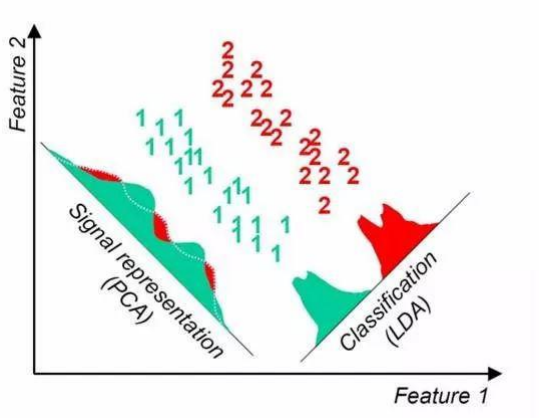
使用所有特征训练一个高斯混合模型（通用模型），使用 MAP 获得每一段语音对应的高斯模型的均值参数。这个均值向量就可以表示这段语音的声纹特征。

LDA(线性判别分析)

LDA 的思想

最终目标是为了求一条线，或者说是一个超平面。使得原来的点在这个超平面上面做一个映射，映射到的点为 lda 之后的结果。这些点有一个特点：类内方差小，类间方差大。
一般的用途是在做数据预处理。也就是将原始的数据转换成这种格式之后再去做其他处理。类似于风控中的 pca。

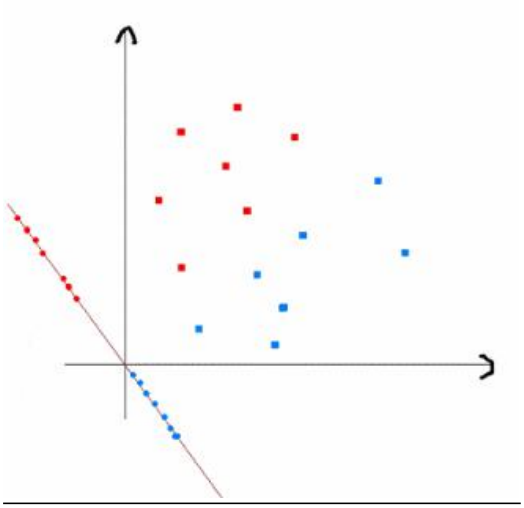
LDA 是一种监督学习的降维技术，也就是说它的数据集的每个样本是由类别输出的。PCA 是不考虑样本类别输出的无监督降维技术。LDA 的思想可以用一句话概括，就是“投影后类内方差最小，类间方差最大”，投影后希望每一种类别数据的投影点尽可能的接近，而不同类别的数据中新之间的距离尽可能的大。



LDA 的全称是 Linear Discriminant Analysis (线性判别分析) , **是一种 supervised learning。** LDA 的原理是，将带上标签的数据（点），通过投影的方法，投影到维度更低的空间中，使得投影后的点，会形成按类别区分，一簇一簇的情况，相同类别的点，将会在投影后的空间中更接近。要说明白 LDA，首先得弄明白线性分类器(Linear Classifier)：因为 LDA 是一种线性分类器。对于 K-分类的一个分类问题，会有 K 个线性函数:

$$y_k(x) = w_k^T x + w_{k0}$$

上式实际上就是一种投影，是将一个高维的点投影到一条高维的直线上，LDA 最求的目标是，给出一个标注了类别的数据集，投影到了一条直线之后，能够使得点尽量的按类别区分开，当 k=2 即二分类问题的时候，如下图所示：



红色的方形的点为 0 类的原始点、蓝色的方形点为 1 类的原始点，经过原点的那条线就是投影的直线，从图上可以清楚的看到，红色的点和蓝色的点被**原点**明显的分开了，这个数据只是随便画的，如果在高维的情况下，看起来会更好一点。下面我来推导一下二分类 LDA 问题的公式：

假设用来区分二分类的直线（投影函数)为：

$$y = w^T x$$

LDA 分类的一个目标是使得不同类别之间的距离越远越好，同一类别之中的距离越近越好，所以我们需要定义几个关键的值。

类别 i 的原始中心点为：（Di 表示属于类别 i 的点)$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

类别 i 投影后的中心点为：

$$\widetilde{m}_i = w^T m_i$$

衡量类别 i 投影后，类别点之间的分散程度（方差）为：

$$\widetilde{s}_i = \sum_{y \in Y_i} (y - \widetilde{m}_i)^2$$

最终我们可以得到一个下面的公式，表示 LDA 投影到 w 后的损失函数：

$$J(w)=\frac{|\tilde{m}_1-\tilde{m}_2|^2}{\tilde{s}_1^2+\tilde{s}_2^2}$$

我们**分类的目标是，使得类别内的点距离越近越好（集中），类别间的点越远越好。**分母表示每一个类别内的方差之和，方差越大表示一个类别内的点越分散，分子为两个类别各自的中心点的距离的平方，我们最大化 J(w)就可以求出最优的 w 了。想要求出最优的 w，可以使用拉格朗日乘子法，但是现在我们得到的 J(w)里面，w 是不能被单独提出来的，我们就得想办法将 w 单独提出来。

我们定义一个投影前的各类别分散程度的矩阵，这个矩阵看起来有一点麻烦，其实意思是，如果某一个分类的输入点集 Di 里面的点距离这个分类的中心店 mi 越近，则 Si 里面元素的值就越小，如果分类的点都紧紧地围绕着 mi，则 Si 里面的元素值越更接近 0.

$$S_i=\sum_{x\in D_i}(x-m_i)(x-m_i)^T$$

带入 Si，将 J(w)分母化为：

$$\tilde{s}_i=\sum_{x\in D_i}(w^Tx-w^Tm_i)^2=\sum_{x\in D_i}w^T(x-m_i)(x-m_i)^Tw=w^TS_iw$$

$$\tilde{s}_1^2+\tilde{s}_2^2=w^T(S_1+S_2)w=w^TS_ww$$

同样的将 J(w)分子化为：

$$|\tilde{m}_1-\tilde{m}_2|^2=w^T(m_1-m_2)(m_1-m_2)^Tw=w^TS_Bw$$

这样损失函数可以化成下面的形式：

$$J(w)=\frac{w^TS_Bw}{w^TS_ww}$$

这样就可以用最喜欢的拉格朗日乘子法了，但是还有一个问题，如果分子、分母是都可以取任意值的，那就会使得有无穷解，我们将分母限制为长度为 1（这是用拉格朗日乘子法一个很重要的技巧，在下面将说的 PCA 里面也会用到，如果忘记了，请复习一下高数），并作为拉格朗日乘子法的限制条件，带入得到：

$$c(w) = w^T S_B w - \lambda(w^T S_w w - 1)$$
$$\Rightarrow \frac{dc}{dw} = 2S_B w - 2\lambda S_w w = 0$$
$$\Rightarrow S_B w = \lambda S_w w$$

这样的式子就是一个求特征值的问题了。

对于 N(N>2)分类的问题，我就直接写出下面的结论了：

$$S_W = \sum_{i=1}^{c} S_i$$
$$S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T$$
$$S_B w_i = \lambda S_w w_i$$

这同样是一个求特征值的问题，我们求出的第 i 大的特征向量，就是对应的 Wi 了。

因子分析

思想：认为高维数据是通过低维数据通过线性变换变换过去的，在不知道是如何变换过去的前提下，使用 em 算法将原来的低维数据求解出来。
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。
使用方法：
观察变量是 xi，低维变量是 zi，最终的目标是求将 zi 表示维 xi 的函数，因此待求参数就是 u 对角矩阵 和 w
实例：
因子分析的实质是认为 m 个 n 维特征的训练样例$x^{(i)}(x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$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

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

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

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

因为$z^{(i)}$的均值是 0，映射后仍然是 0。

3、 然后将$\Lambda z^{(i)}$加上一个均值μ（n 维），即

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

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

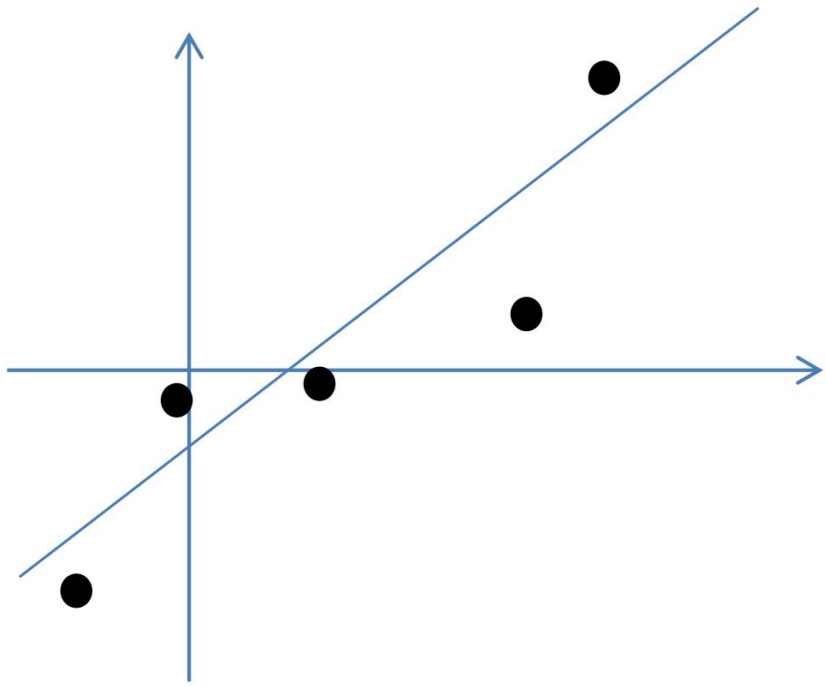
$$\epsilon \sim N(0, \Psi)$$

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

5、 最后的结果认为是真实的训练样例$x^{(i)}$的生成公式

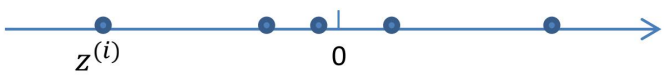
$$x^{(i)} = \mu + \Lambda z^{(i)} + \epsilon$$

我们有 m=5 个 2 维的样本点$x^{(i)}$（两个特征），如下：



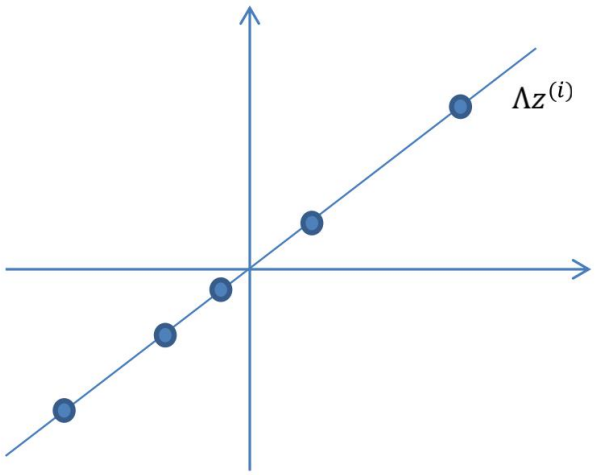
那么按照因子分析的理解，样本点的生成过程如下：

1、 我们首先认为在 1 维空间（这里 k=1），存在着按正态分布生成的 m 个点$z^{(i)}$，如下

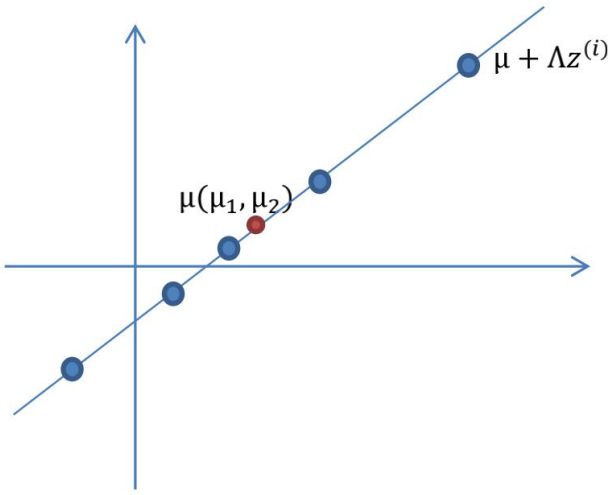


均值为 0，方差为 1。

2、 然后使用某个$\Lambda = (a, b)^T$将一维的 z 映射到 2 维，图形表示如下：

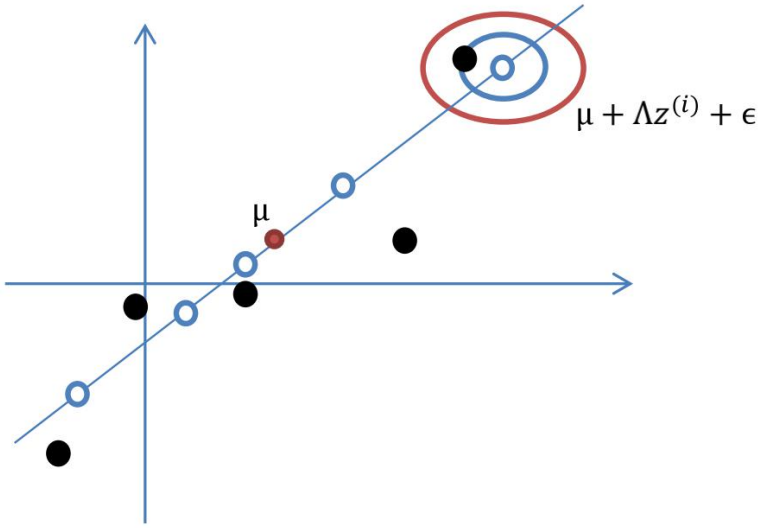


3、 之后加上$\mu\ (\mu_1, \mu_2)^T$，即将所有点的横坐标移动μ_1，纵坐标移动μ_2，将直线移到一个位置，使得直线过点μ，原始左边轴的原点现在为μ（红色点）。



然而，样本点不可能这么规则，在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$\mathbf{x}^{(i)}$如下：



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。

参数求解:

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：

$$z \sim N(0, I)$$

$$\epsilon \sim N(0, \Psi)$$

$$\mathrm{x} = \mu + \Lambda z + \epsilon$$

其中误差ϵ和 z 是独立的。

矩阵表示法认为 z 和 x 联合符合多元高斯分布，如下

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

求μ_{zx}之前需要求 E[x]

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

$= \mu$

我们已知 E[z]=0，因此

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

下一步是计算Σ，

其中$\Sigma_{zz} = Cov(z) = I$

接着求Σ_{zx}

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

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

接着求Σ_{xx}

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

然后得出联合分布的最终形式

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

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$\{x^{(i)}; i = 1, \ldots, m\}$进行最大似然估计

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

循环重复直到收敛 {

（E 步）对于每一个 i，计算

$$Q_i\left(z^{(i)}\right) \coloneqq p\left(z^{(i)}|x^{(i)};\theta\right)$$

（M 步）计算

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

}

我们套用一下：

（E 步）：

$$Q_i\left(z^{(i)}\right) \coloneqq p\left(z^{(i)}|x^{(i)};\mu,\Lambda,\Psi\right)$$

根据第 3 节的条件分布讨论，

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

因此

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

那么根据多元高斯分布公式，得到

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

（M 步）：

直接写要最大化的目标是

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

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：

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

$$= \sum_{i=1}^{m} \mathrm{E}_{z^{(i)}\sim Q_i}\left[\log p(x^{(i)}|z^{(i)};\mu,\Lambda,\Psi) + \log p(z^{(i)}) - \log Q_i(z^{(i)})\right] \quad (6)$$

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

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

去掉不相关的前两项后，对Λ进行导，

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

第一步到第二步利用了 tr a = a（a 是实数时）和 tr AB = tr BA。最后一步利用了

$$\nabla_A \mathrm{tr}ABA^TC = CAB + C^TAB^T$$

tr 就是求一个矩阵对角线上元素和。

最后让其值为 0，并且化简得

$$\sum_{i=1}^{m} \Lambda \mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)}z^{(i)T}\right] = \sum_{i=1}^{m}(x^{(i)}-\mu)\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)T}\right].$$

然后得到

$$\Lambda = \left(\sum_{i=1}^{m}(x^{(i)}-\mu)\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)T}\right]\right)\left(\sum_{i=1}^{m}\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)}z^{(i)T}\right]\right)^{-1}. \quad (7)$$

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

$$\text{“}\theta^T = (y^TX)(X^TX)^{-1}.\text{”}$$

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。

到这还没完，我们需要求得括号里面的值

根据我们之前对 z|x 的定义，我们知道

$$\begin{aligned}\mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)^T}\right] &= \mu_{z^{(i)}|x^{(i)}}^T \\ \mathrm{E}_{z^{(i)}\sim Q_i}\left[z^{(i)}z^{(i)^T}\right] &= \mu_{z^{(i)}|x^{(i)}}\mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}.\end{aligned}$$

第一步根据 z 的条件分布得到，第二步根据$\mathrm{Cov}(Y) = \mathrm{E}[YY^T] - E[Y]E[Y]^T$得到
将上面的结果代入（7）中得到

$$\Lambda = \left(\sum_{i=1}^{m}(x^{(i)} - \mu)\mu_{z^{(i)}|x^{(i)}}^T\right)\left(\sum_{i=1}^{m}\mu_{z^{(i)}|x^{(i)}}\mu_{z^{(i)}|x^{(i)}}^T + \Sigma_{z^{(i)}|x^{(i)}}\right)^{-1}. \quad (8)$$

至此，我们得到了Λ，注意一点是 E[z]和$\mathrm{E}[zz^T]$的不同，后者需要求 z 的协方差。

其他参数的迭代公式如下：

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

均值μ在迭代过程中值不变。

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

然后将Φ上的对角线上元素抽取出来放到对应的Ψ中，就得到了Ψ。

7 总结

根据上面的 EM 的过程，要对样本 X 进行因子分析，只需知道要分解的因子数（z 的维度）即可。通过 EM，我们能够得到转换矩阵Λ和误差协方差Ψ。

因子分析实际上是降维，在得到各个参数后，可以求得 z。但是 z 的各个参数含义需要自己去琢磨。

下面从一个 ppt 中摘抄几段话来进一步解释因子分析。

因子分析(factor analysis)是一种数据简化的技术。它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。这几个假想变量能够反映原来众多变量的主要信息。原始的变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因子。

例如，在企业形象或品牌形象的研究中，消费者可以通过一个有 24 个指标构成的评价体系，评价百货商场的 24 个方面的优劣。

但消费者主要关心的是三个方面，即商店的环境、商店的服务和商品的价格。因子分析方法可以通过 24 个变量，找出反映商店环境、商店服务水平和商品价格的三个潜在的因子，对商店进行综合评价。而这三个公共因子可以表示为：

$$x_i = \mu_i + \alpha_{i1}F_1 + \alpha_{i2}F_2 + \alpha_{i3}F_3 + \varepsilon_i \quad i = 1,\cdots,24$$

这里的x_i就是样例 x 的第 i 个分量，μ_i就是μ的第 i 个分量，α_{ij}就是Λ的第 i 行第 j 列元素，F_i是 z 的第 i 个分量，ε_i是$\varepsilon^{(i)}$。

称F_i是不可观测的潜在因子。24 个变量共享这三个因子，但是每个变量又有自己的个性，不被包含的部分ε_i，称为特殊因子。

注：

因子分析与回归分析不同，因子分析中的因子是一个比较抽象的概念，而回归因子有非常明确的实际意义；

主成分分析分析与因子分析也有不同，主成分分析仅仅是变量变换，而因子分析需要构造因子模型。

主成分分析:原始变量的线性组合表示新的综合变量，即主成分；

因子分析：潜在的假想变量和随机影响变量的线性组合表示原始变量。

i-vector

问题：如何进行说话人分类?

思路：将变长的语音（语音时长不一样），特征化为一个特定维度的向量。对向量做分类即可。

假设：i-vector 提出说话人和会话差异可以通过一个单独对子空间进行表征。利用这个子空间，可以把一个语音素材上获得的数字矢量，进一步转化为低维矢量。

i-vector 优点：低维

可以作为说话人的特征加入到分类器上进行连续语音识别

可以把 LDA、plda 结合到 i-vector 中。就是方便和其他算法进行结合。

推导过程

重要公式：$$M = m + Tw$$

M 就是做完 UBM-GMM 后做了 MAP 之后的高斯均值超矢量。

m 是均值超矢量

T 是总变化矩阵，是待求参数

w(s)是 R 维的隐变量 i-vector，就是最终要求的 i-vector

此时，我再给出B-W的3种充分统计量，这是给后面EM估计 T 时要用到的：

对于第 c 个单高斯，第 s 句话的充分统计量如下：

$$N_c(s) = \sum_{i=1}^{H(s)} P_\lambda(c|x_i)$$

$$F_c(s) = \sum_{i=1}^{H(s)} P_\lambda(c|x_i)(x_i - \mu_c)$$

$$S_c(s) = \sum_{i=1}^{H(s)} P_\lambda(c|x_i)(x_i - \mu_c)(x_i - \mu_c)^T$$

若把该句话的所有高斯分量的充分统计量整合起来成矩阵形式，则如下：

$$N(s) = \begin{bmatrix} N_1(s)I & & 0 \\ & \dots & \\ 0 & & N_c(s)I \end{bmatrix} \quad (2)$$

$$F(s) = \begin{bmatrix} F_1(s) \\ \dots \\ F_c(s) \end{bmatrix} \quad (3)$$

其中 $N(s)$ 是一个主对角线矩阵。

好了，现在开始推导T矩阵的EM公式==!!，其中E步比较复杂。。。

对于E步，第 t 次迭代期望的Q函数如下(4)：

$$Q(T|T^{(t)}) = \sum_{s=1}^{S} E(logP_T(X(s), w(s)))$$

$$= \sum_{s=1}^{S} E(logP_T(X(s)|w(s))) + \sum_{s=1}^{S} E(logP_T(w(s)))$$

此步是把观测数据 $X(s)$ 和隐数据 $w(s)$ 的条件概率拆开，展开为似然概率与边缘概率的乘积，可以参考《统计学习方法》的（9.12）公式；

由于（4）中的第二项是边缘概率，与 T 无关，所以在M步中会被偏导置零，这里就先提前忽略掉，继续展开（4）：

$$Q(T|T^{(t)}) = \sum_{s=1}^{S} E(logP_T(X(s)|w(s))$$

$$= \sum_{s=1}^{S} E(G(s) + H_T(s, w(s)))$$

其中， $G(s)$ 是和二阶统计量的相关项，但跟 T 估计无关的，所以也提前忽略；

而 $H_T(s, w(s))$ 则与零阶、一阶、 T 和 $w(s)$ 都相关的一项，其可展开为：

$$H_T(s, w(s)) = w(s)^T T^T \Sigma^{-1} F(s) - \frac{1}{2} w(s)^T T^T \Sigma^{-1} N(s) T w(s) \quad (5)$$

如果说(4)式是EM的最核心，那(5)式可以说是i-vector推导中的最重要，把(5)代入到(4)中，并展开如下：

$$Q(T|T^{(t)}) = \sum_{s=1}^{S} E(\widehat{w}^{(t)}(s)^T T^T \Sigma^{-1} F(s))$$
$$- \frac{1}{2} \sum_{s=1}^{S} E(\widehat{w}^{(t)}(s)^T T^T \Sigma^{-1} N(s) T \widehat{w}^{(t)}(s)^T)$$

上式(6)中关键的部分有两个，分别是 $w(s)$ 的后验密度函数的均值与协方差：

均值 $E(\widehat{w}^{(t)}(s)^T) = \sigma^t(s)^{-1} T^T \Sigma^{-1} F(s)$ (7)

协方差 $E(\widehat{w}^{(t)}(s)^T \cdot \widehat{w}^{(t)}(s)) = \sigma^t(s)^{-1} + E(\widehat{w}^{(t)}(s)^T) \cdot E(\widehat{w}^{(t)}(s)^T)^T$ (8)

其中， $\sigma^t(s) = I + T^T \Sigma^{-1} N(s) T$ (9)

截至到现在，E步就完成了，简化得到(6)式。

对于M步，对(6)式求极大：

$T^{t+1} = argmax Q(T|T^{(t)})$ (10)

对 $T^{(t)}$ 求偏导，并求极值点：

$$\sum_{s=1}^{S} \Sigma^{-1} F(s) E(\widehat{w}^{(t)}(s)^T) -$$
$$\frac{1}{2} \sum_{s=1}^{S} \Sigma^{-1} N(s) \cdot 2T^{t+1} \cdot E(\widehat{w}^{(t)}(s)^T \cdot \widehat{w}^{(t)}(s)) = 0$$

化简，把第二项挪到等号右边；除了 T^{t+1} 这项外，把其它一堆东西除到等号左边，就可以求到 $t+1$ 次迭代的T矩阵：

$$T^{(t+1)} = \frac{\sum_{s=1}^{S} \Sigma^{-1} F(s) E(\widehat{w}^{(t)}(s)^T)}{\sum_{s=1}^{S} \Sigma^{-1} N(s) E(\widehat{w}^{(t)}(s)^T \cdot \widehat{w}^{(t)}(s))}$$ (11)

通常会迭代T矩阵5～6次认为收敛，最后把 T 与 m 代入到(1)中，即可用于提取i-vector

i-vector 相对于 ubm-gmm 做 map 后接 LBG 有啥优点?

说话人 GMM 均值超矢量所在的空间分为 本征空间、信道空间、残差空间

最理想的目标：抽取出跟说话人本身相关的特征，去掉和信道相关的特征。

说话人空间、信道空间
i-vector 采用一个空间来代替这两个空间。这个新的空间可以成为全局差异空间，既包含了说话者之间的差异又包含了信道间的差异。
I-vector 是通过高斯超向量基于因子分析而得到的，是通过高斯超向量基于因子分析得到的，是基于单一空间的跨信道算法，该空间既包含了说话人空间的信息也包含了信道空间信息。
相当于用因子分析方法将语音从高维空间投影到低维。

信道补偿

I-vector 中既包含说话者信息又包含信道信息，而我们只关心说话者信息。
由于信道信息的存在对说话人识别产生了干扰，，信道补偿就是减少信道信息对说话人识别产生的干扰。

PLDA

LDA 相当于有监督的 pca
PLDA 相当于因子分析
因子分析，本质上就是将观察变量表示为假想变量之间的线性组合。

$$\mathbf{x}_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}$$

在声纹识别中，Xij 表示第 i 个人的第 j 条语音。
这个观测变量是两个低维变量通过线性映射之后构成的。
跟因子分析的目标一样，我们最终的目标是 miu 和 F 和 G 和 sigma
其中 F 是：每一列相当于类间空间的特征向量
G 是：每一列相当于类内空间的特征向量。
hi 可以看作是 Xij 在说话人空间中的特征表示。
模型求解方法和因子分析一样。
终极目标是：获得了 F 矩阵。意味着就真正做到了信道补偿。这个时候得到的语音的特征表示仅仅和说话人相关了。