# 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）
5. 一个更好的方法是使用PLDA替换LDA，因为LDA做信道补偿的能力不如PLDA。

### 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：

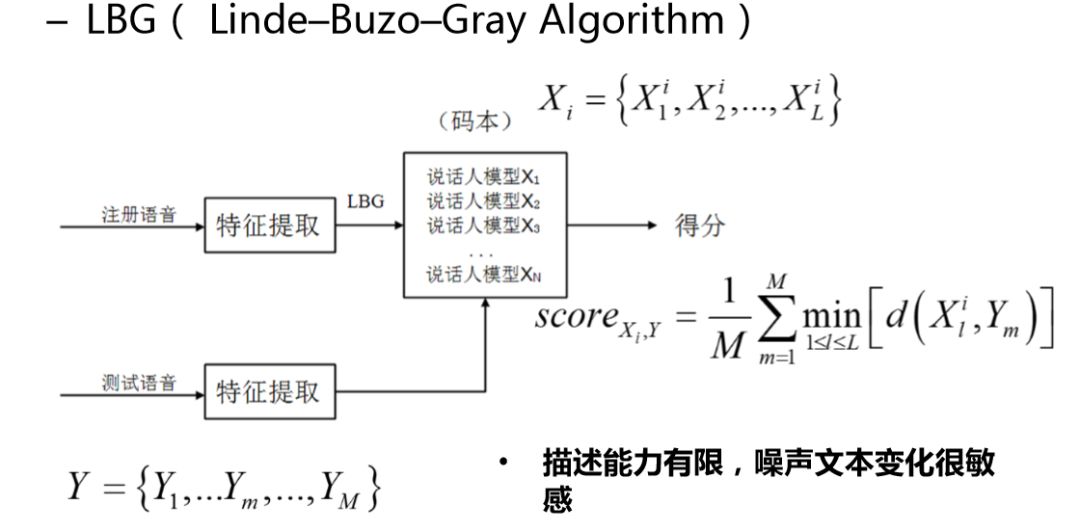
假如已经训练好了说话人模型：一个人对应一个模型。

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：

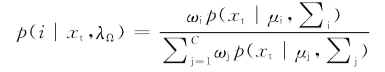
和协方差矩阵。但是协方差矩阵一般设置为对角阵。

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

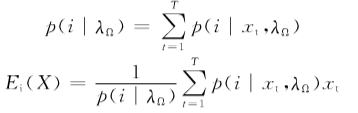
1. 首先计算语音特征向量序列中的各个向量相对于每个UBM混元的概率得分。

IMG_256

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：



3. 利用后验概率计算均值所需要的统计量



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：

IMG_259

自适应IMG_256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。

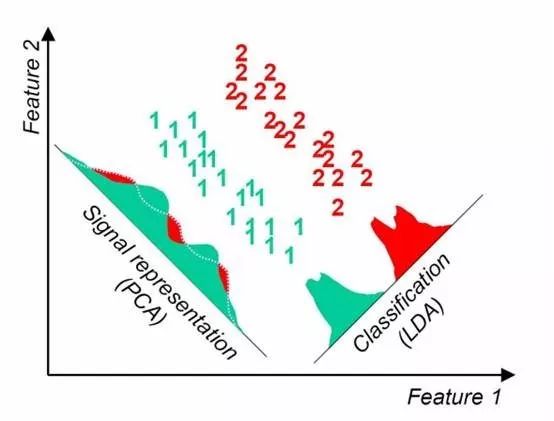
## LDA(线性判别分析)

### 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。

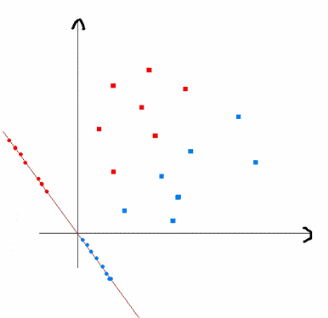
LDA是一种监督学习的降维技术，也就是说它的数据集的每个样本是由类别输出的。PCA是不考虑样本类别输出的无监督降维技术。LDA的思想可以用一句话概括，就是“投影后类内方差最小，类间方差最大”，投影后希望每一种类别数据的投影点尽可能的接近，而不同类别的数据中新之间的距离尽可能的大。



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](http://en.wikipedia.org/wiki/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：

[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455463969.png)

    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：

[](C:\\Documents and Settings\\Administrator\\Local Settings\\Temp\\WindowsLiveWriter-429641856\\supfiles4CEE5\\image[15].png)

     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：

[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455476902.png)

    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)[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455476869.png)

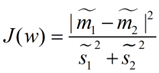
    类别i投影后的中心点为：

[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455473248.png)

    衡量类别i投影后，类别点之间的分散程度（方差）为：

[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/20110108145548391.png)

    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：

[](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455487850.png)

   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.

[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455491720.png)

   带入Si，将J(w)分母化为：

[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455497751.png)

image

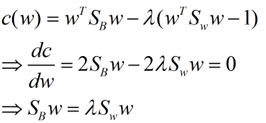
   同样的将J(w)分子化为：

[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455509636.png)

   这样损失函数可以化成下面的形式：

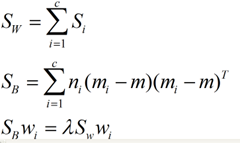
[image](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455509047.png)

   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：

[](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455513997.png)

   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：

[](http://images.cnblogs.com/cnblogs_com/LeftNotEasy/201101/201101081455514488.png)

   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

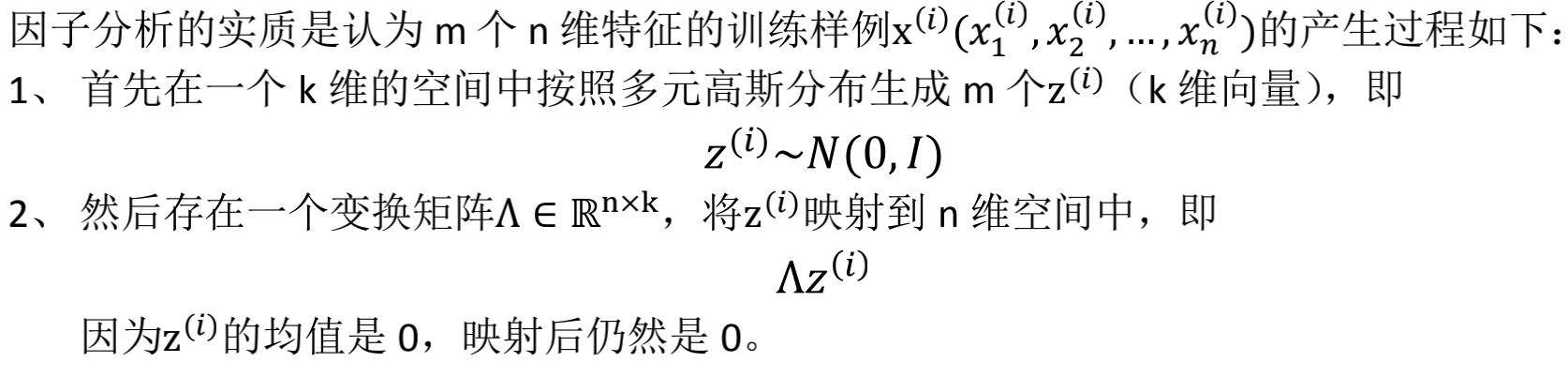
思想：认为高维数据是通过低维数据通过线性变换变换过去的，在不知道是如何变换过去的前提下，使用em算法将原来的低维数据求解出来。

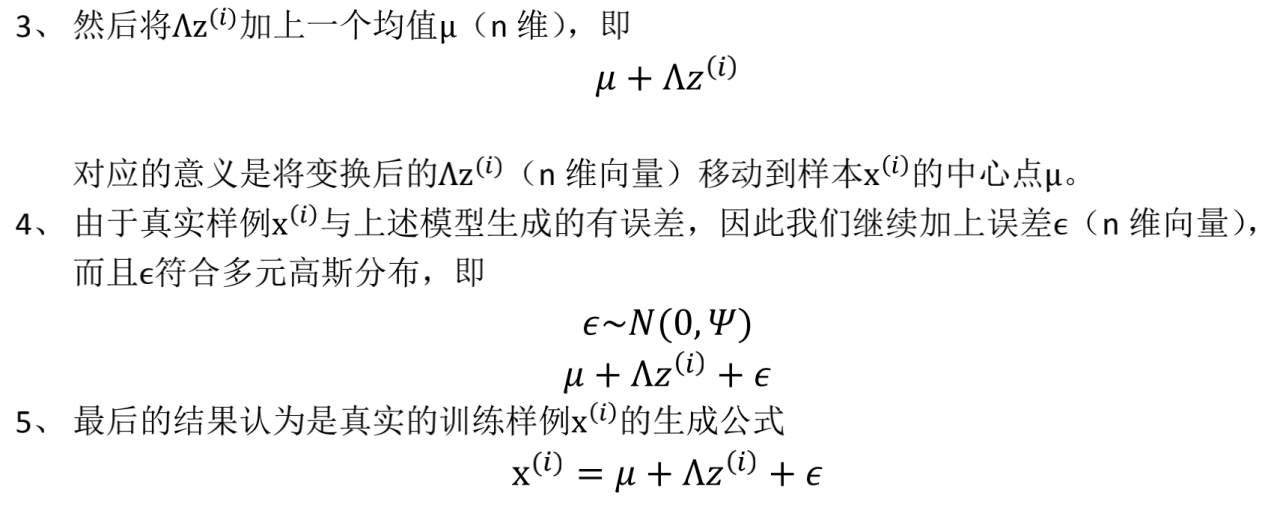
因子分析其实就是认为高维样本点实际上是由低维样本点经过高斯分布、线性变换、误差扰动生成的，因此高维数据可以使用低维来表示。

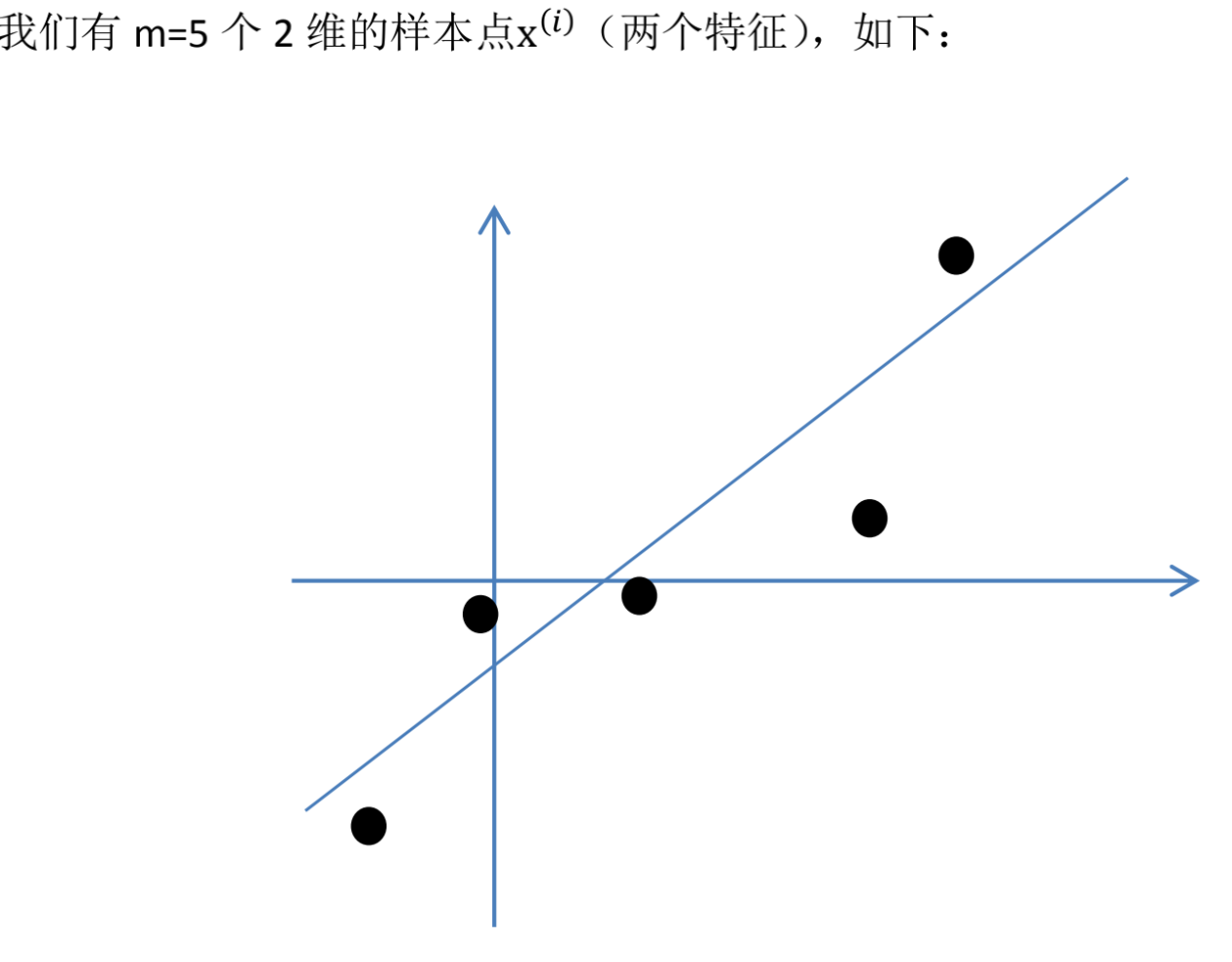
使用方法：

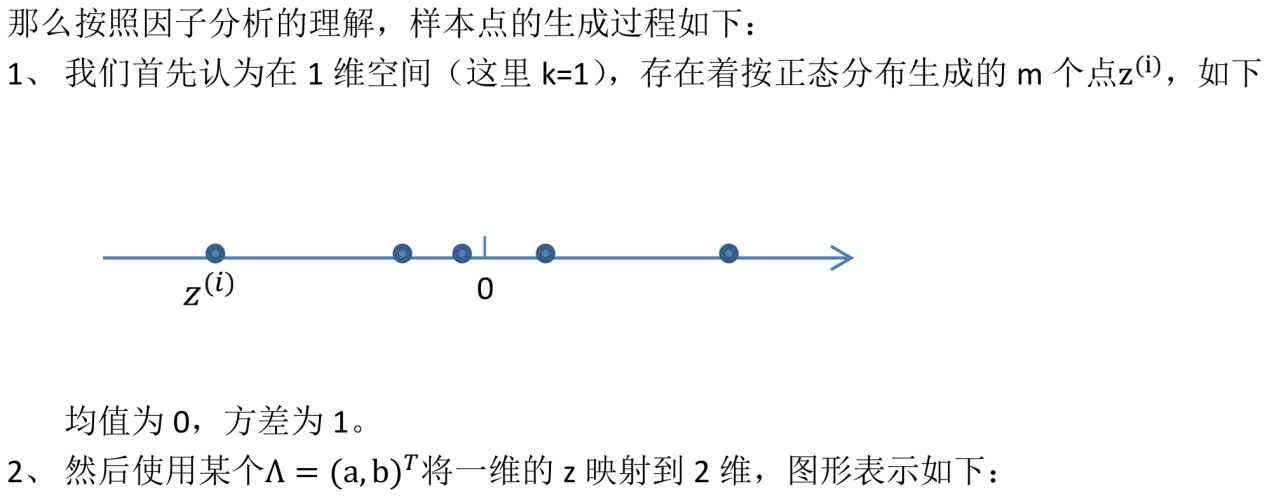
观察变量是xi，低维变量是zi，最终的目标是求将zi表示维xi的函数，因此待求参数就是u 对角矩阵 和w

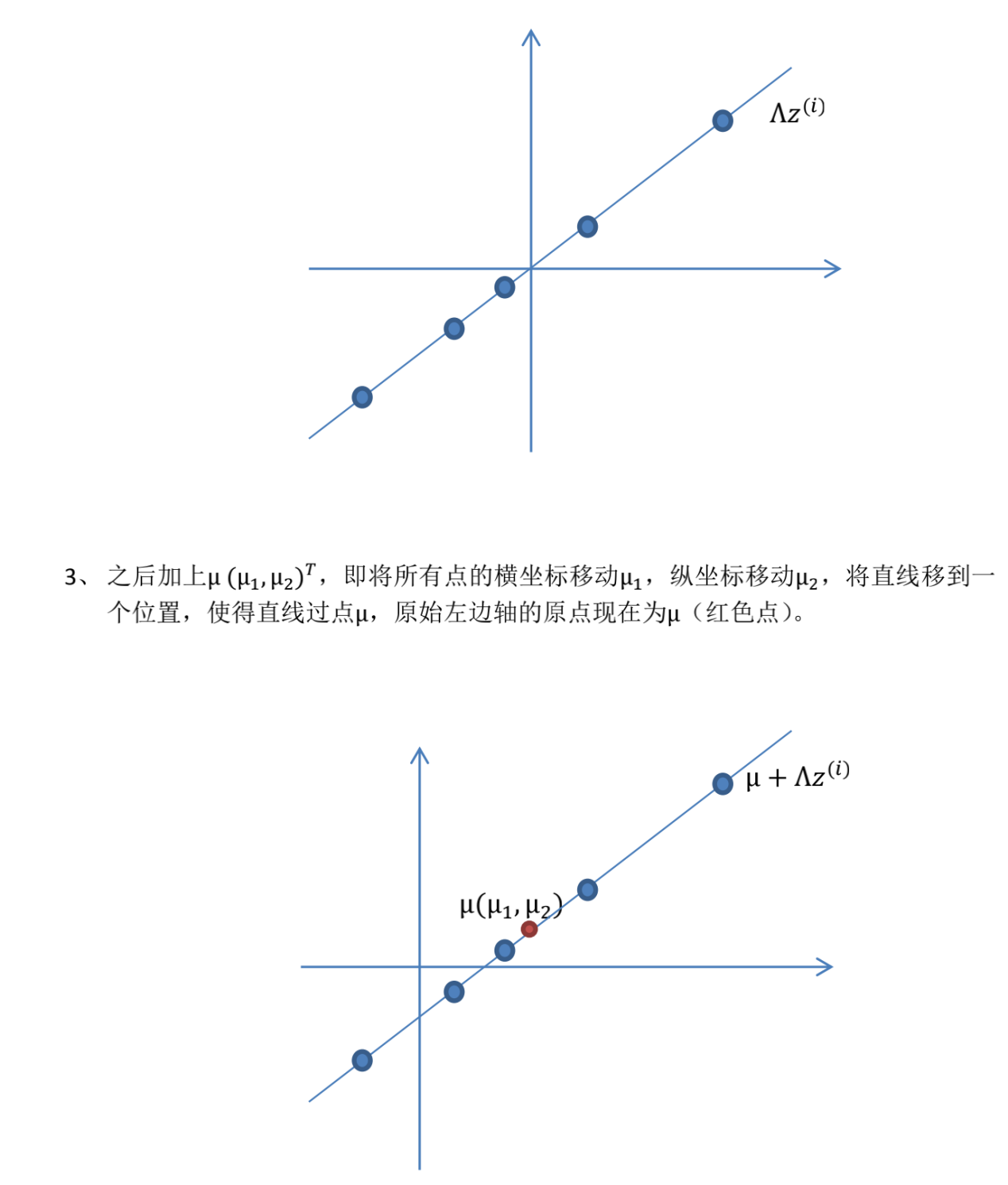
实例：

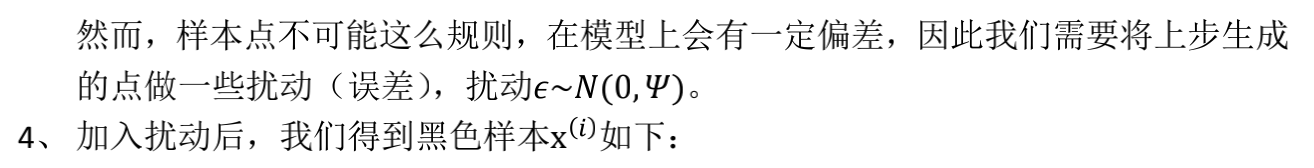


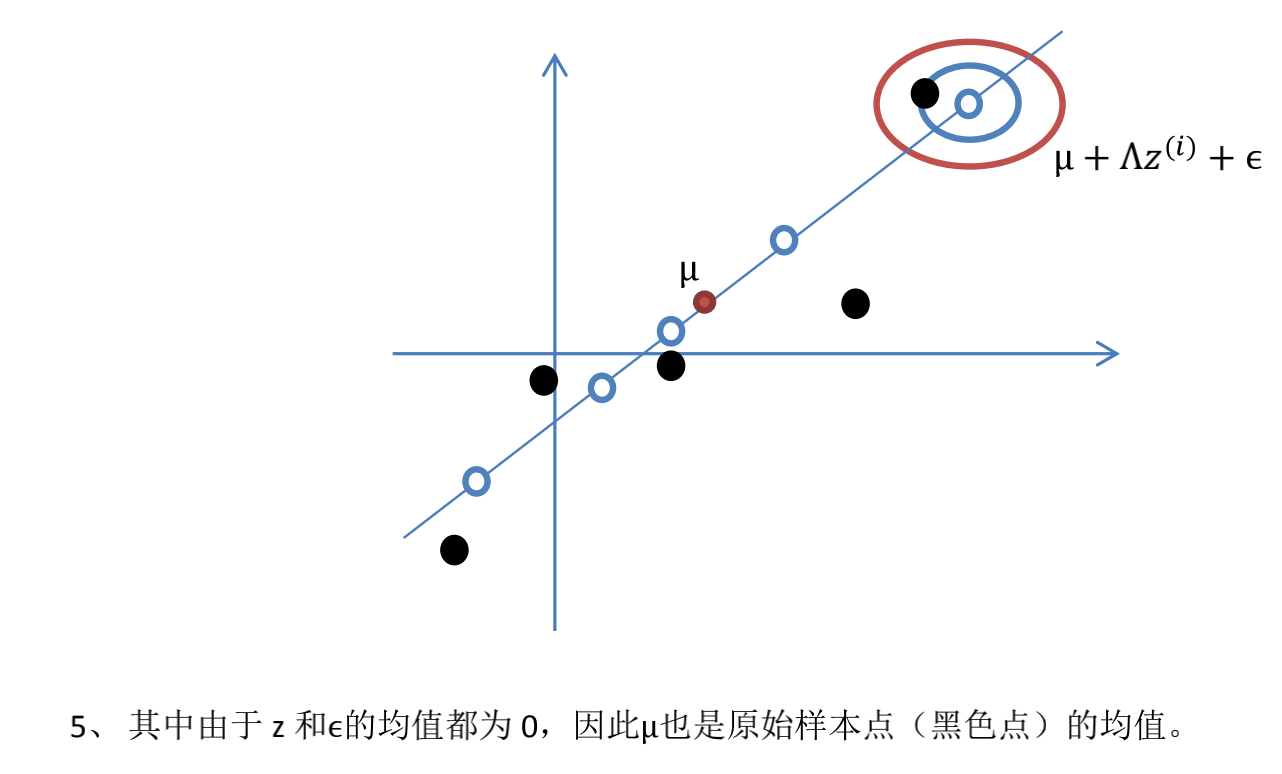




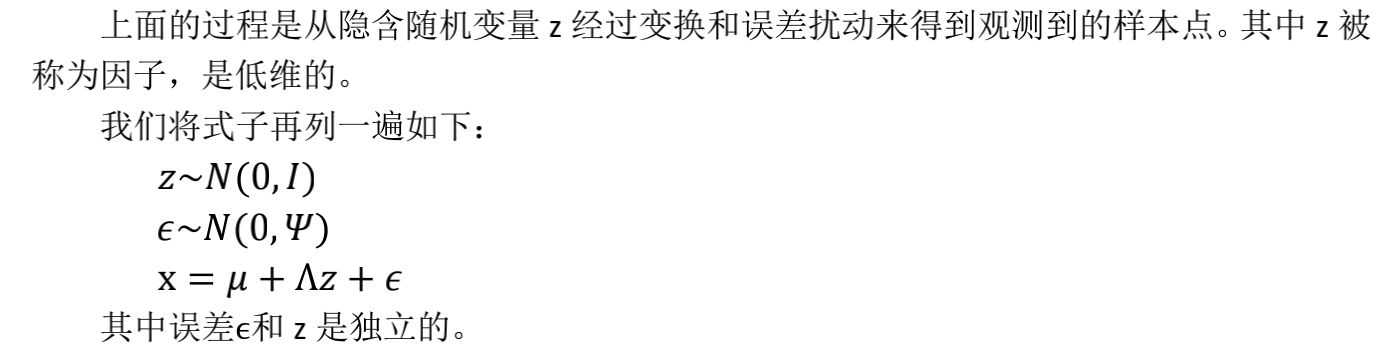


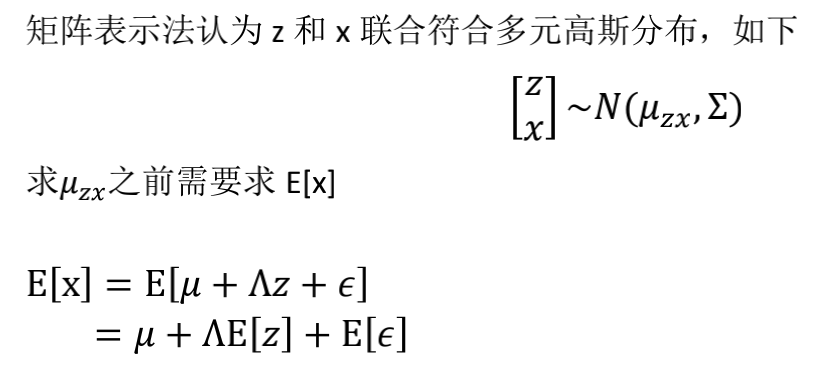


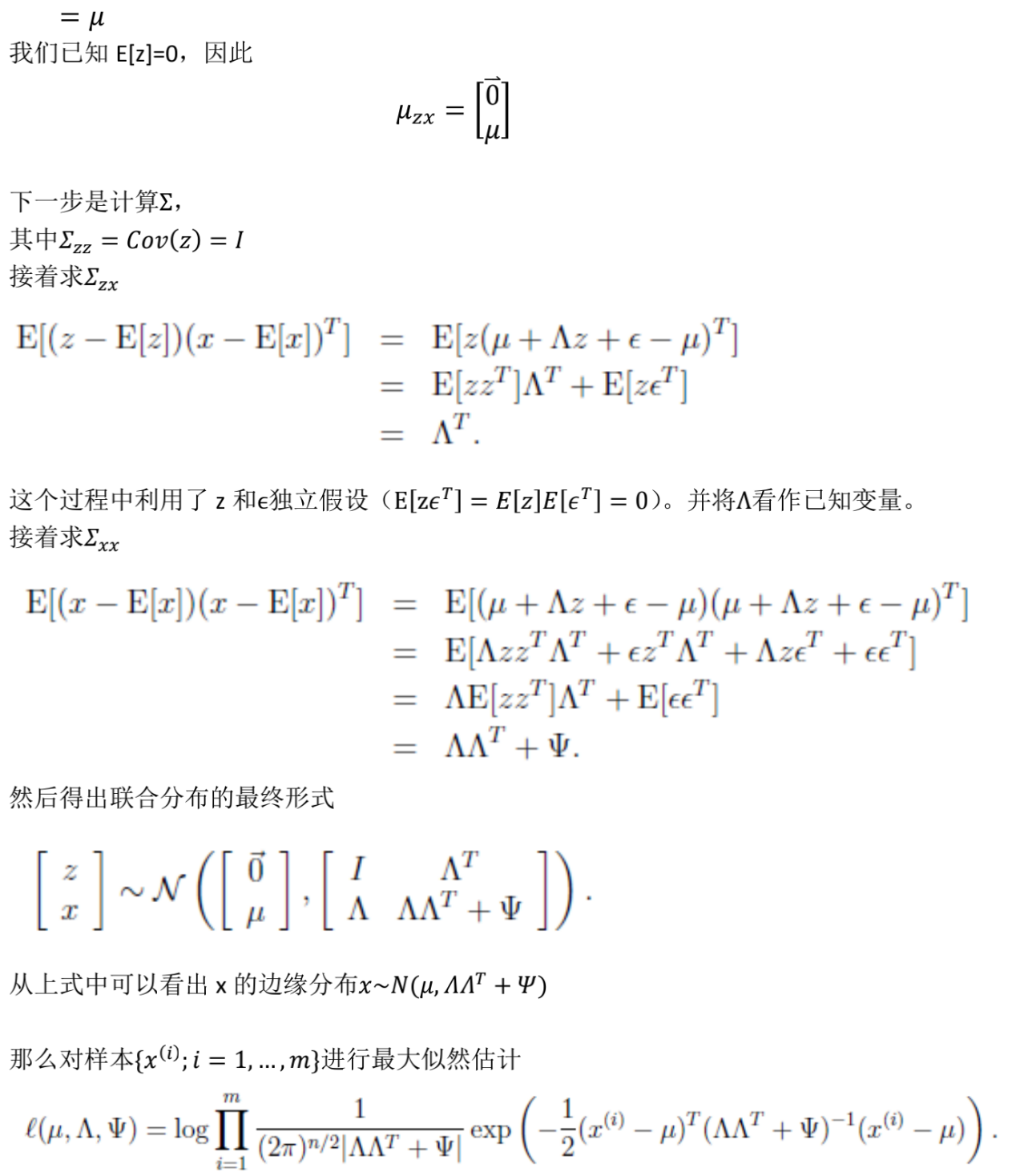


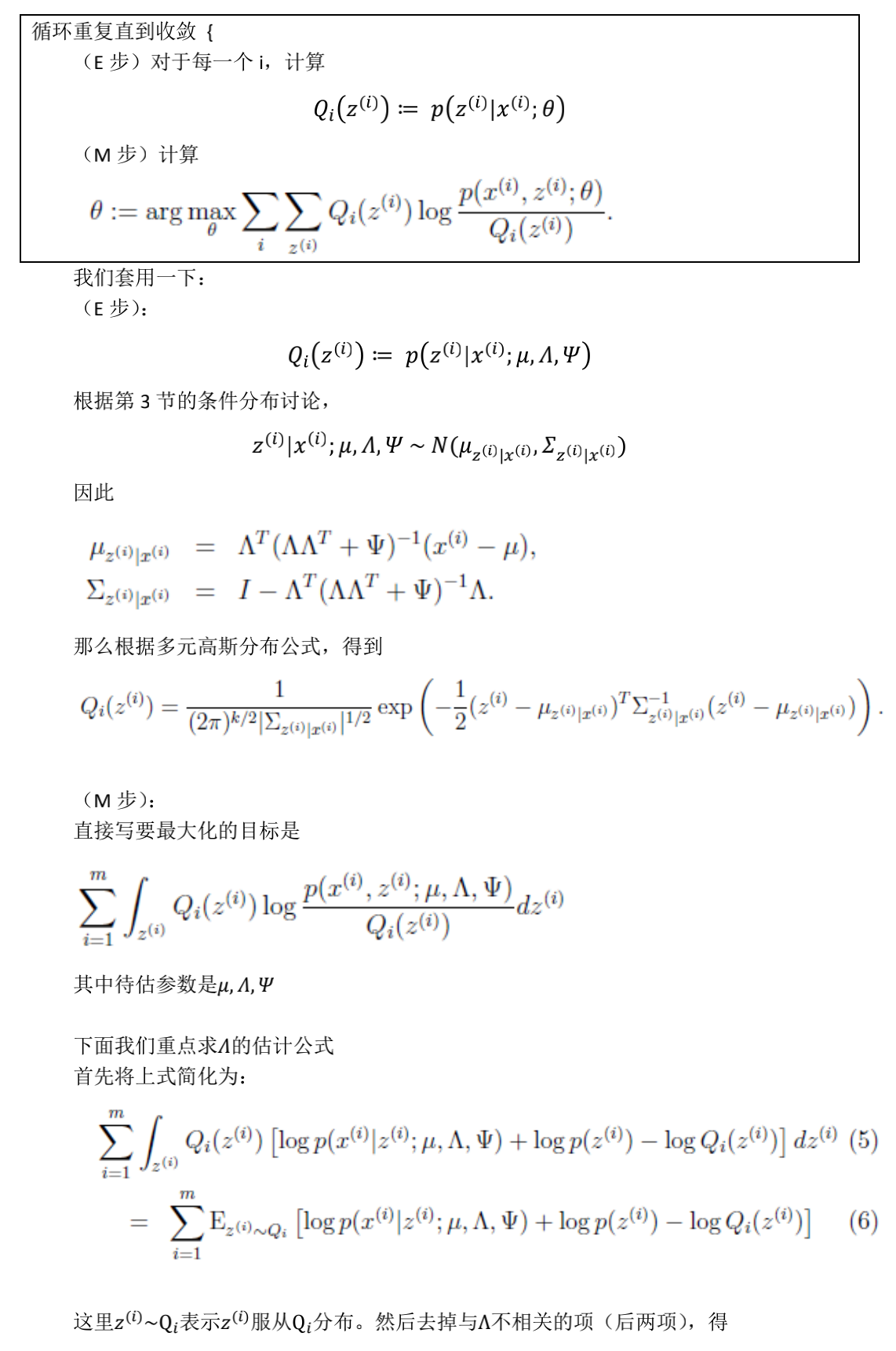


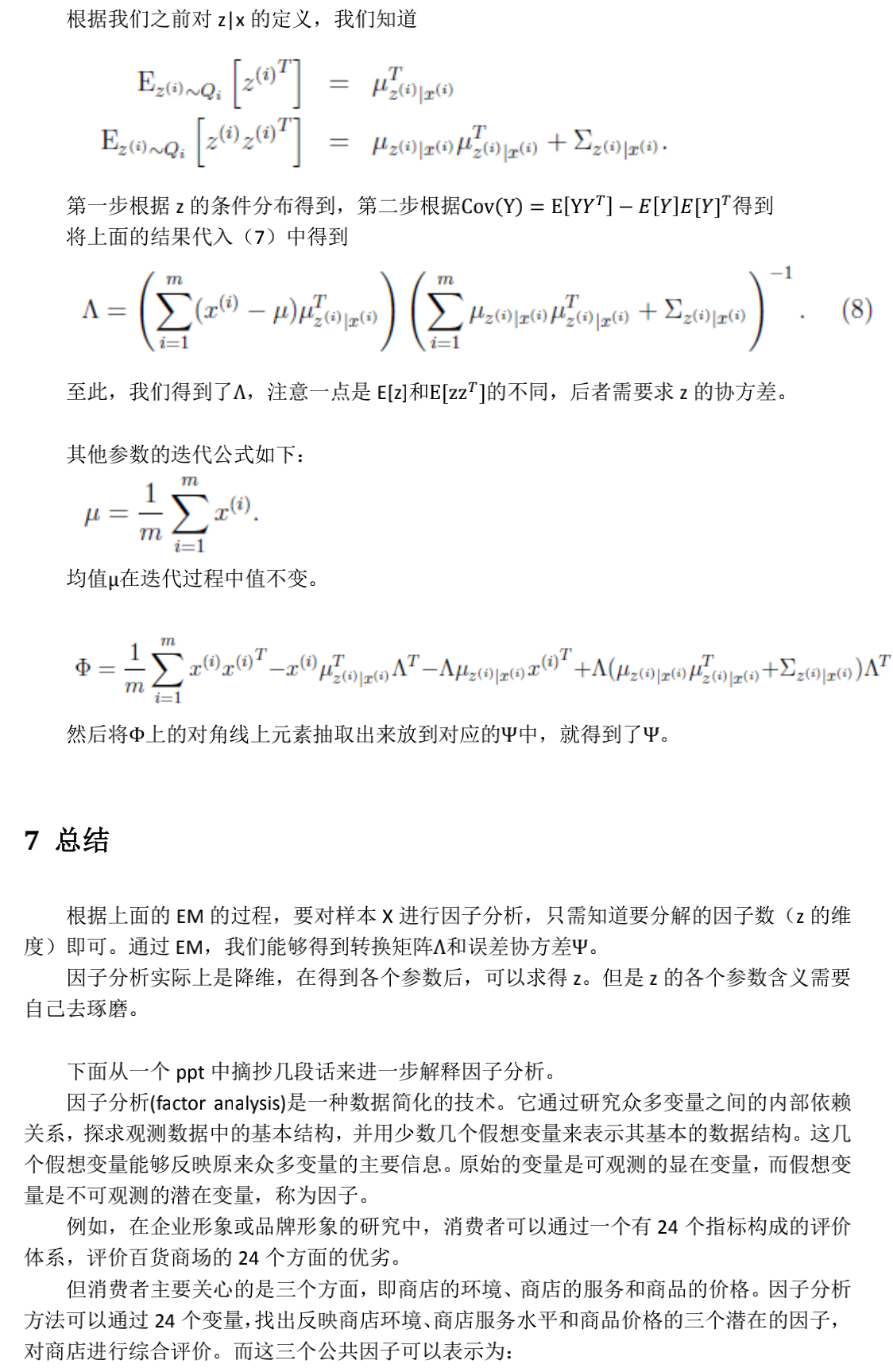
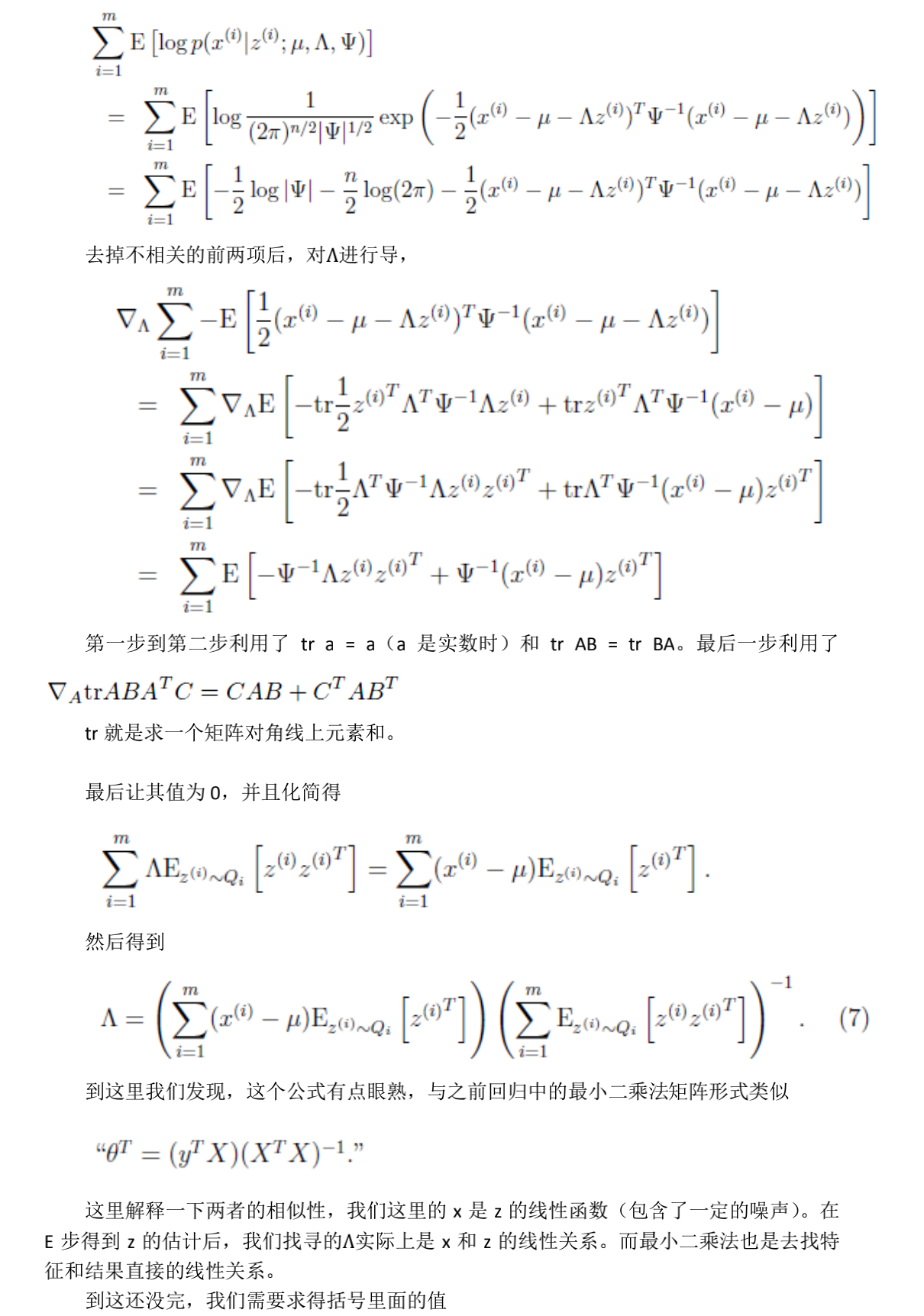
参数求解：

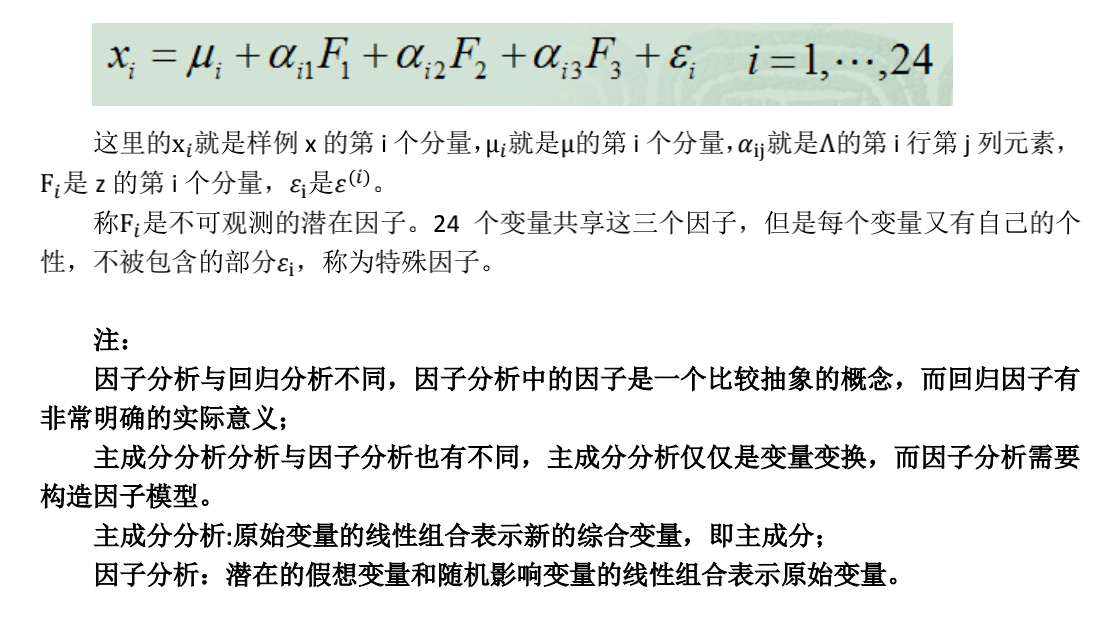












## i-vector

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

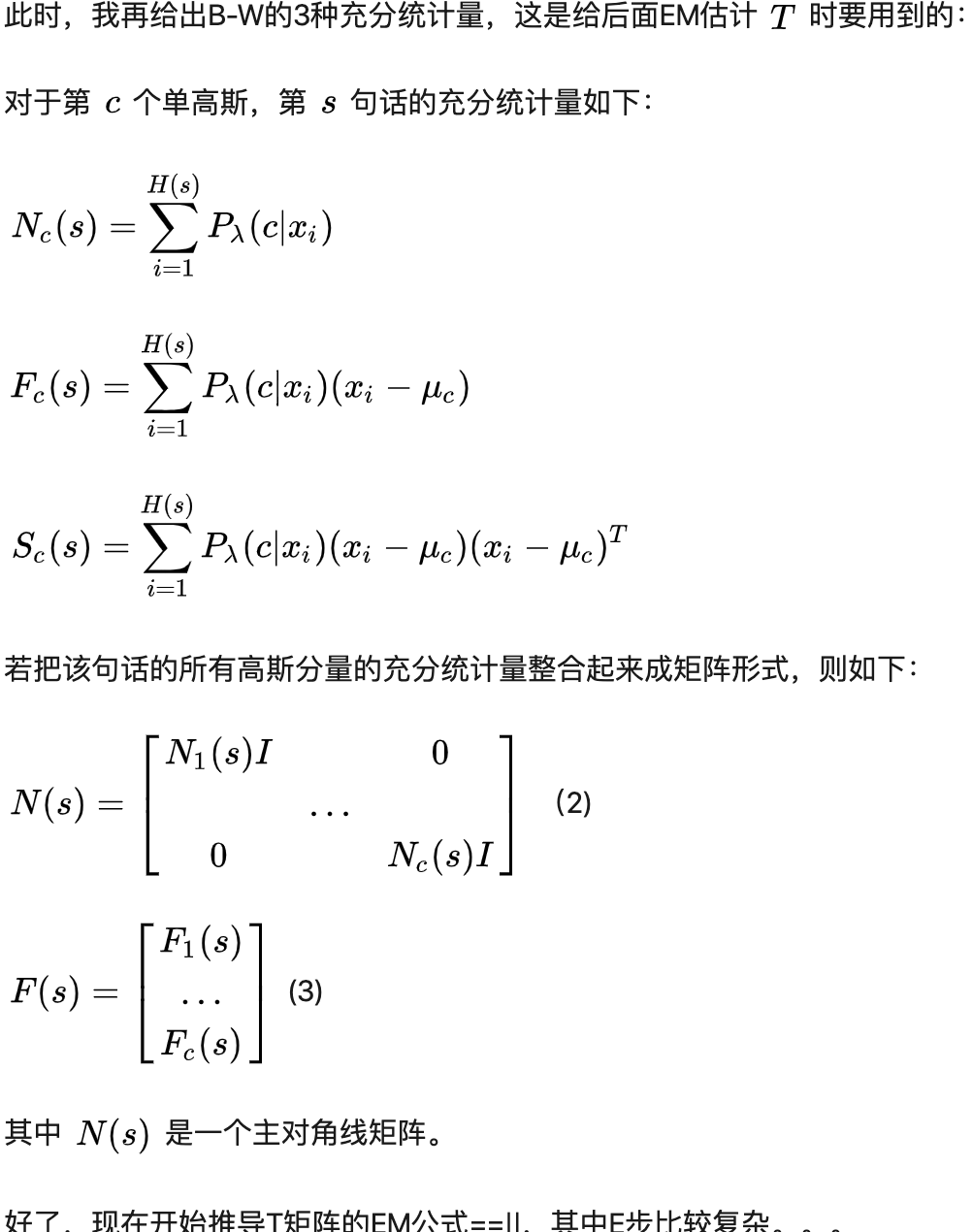
重要公式： IMG_256

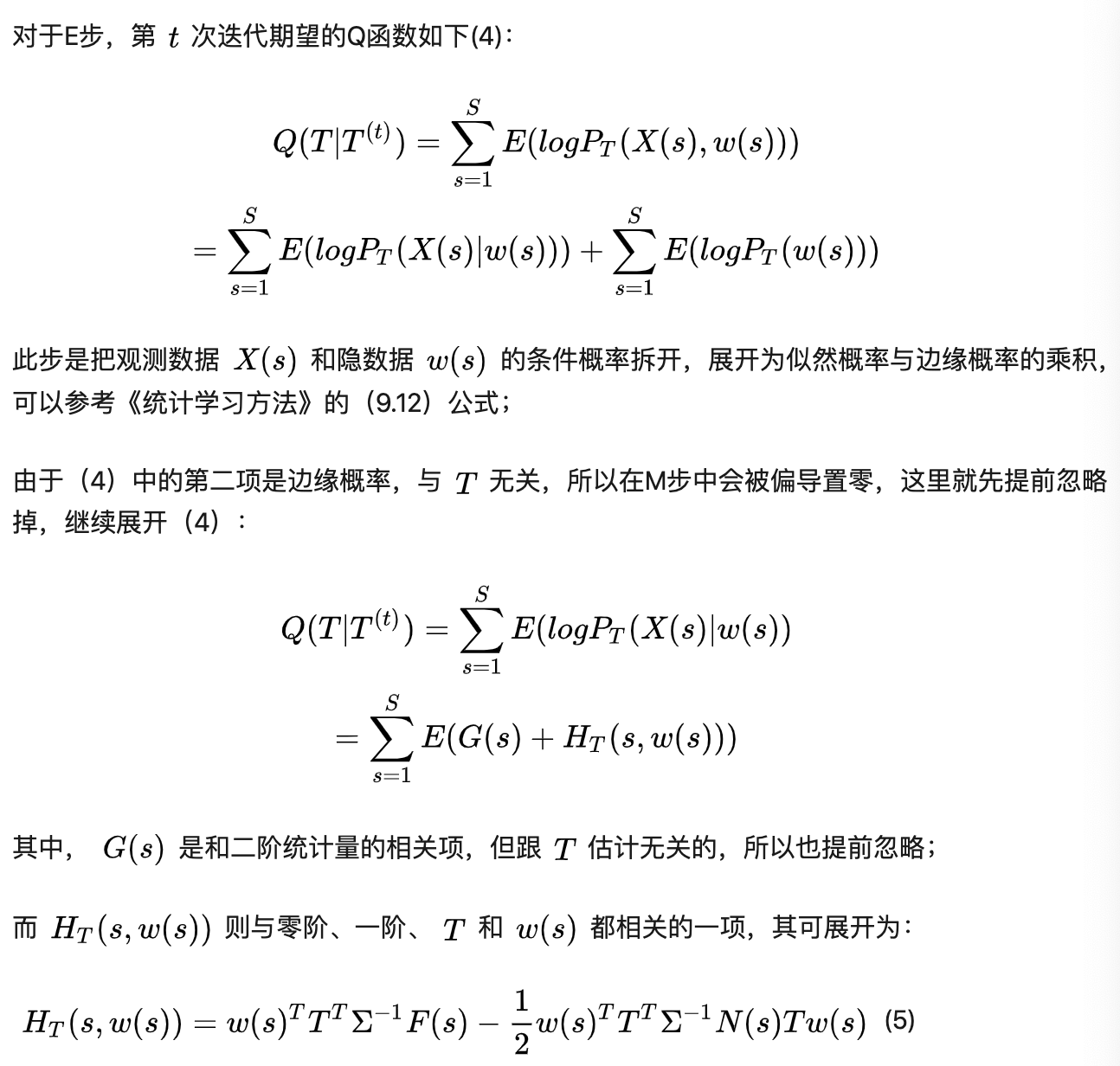
M就是做完UBM-GMM后做了MAP之后的高斯均值超矢量。

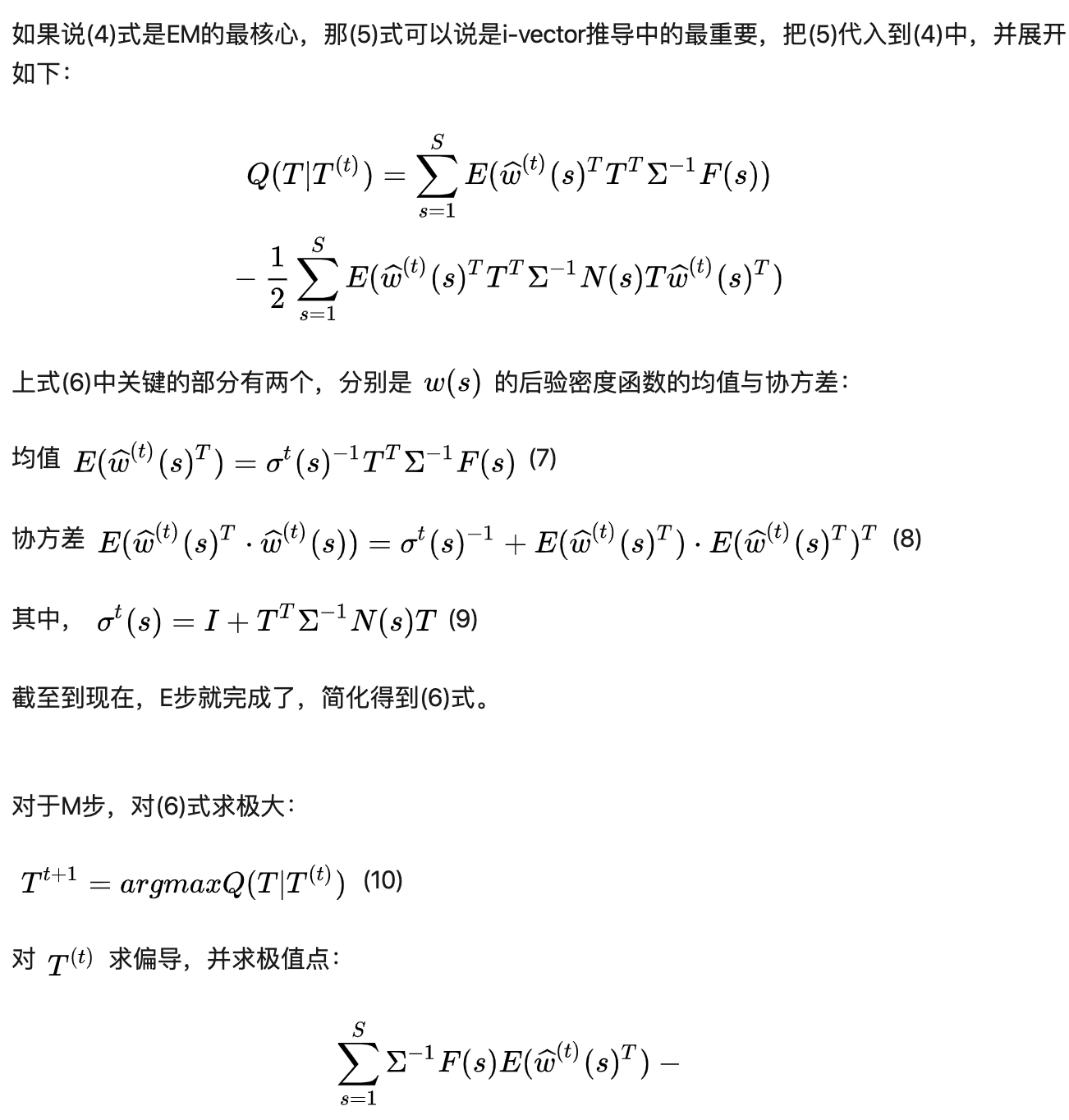
m是均值超矢量

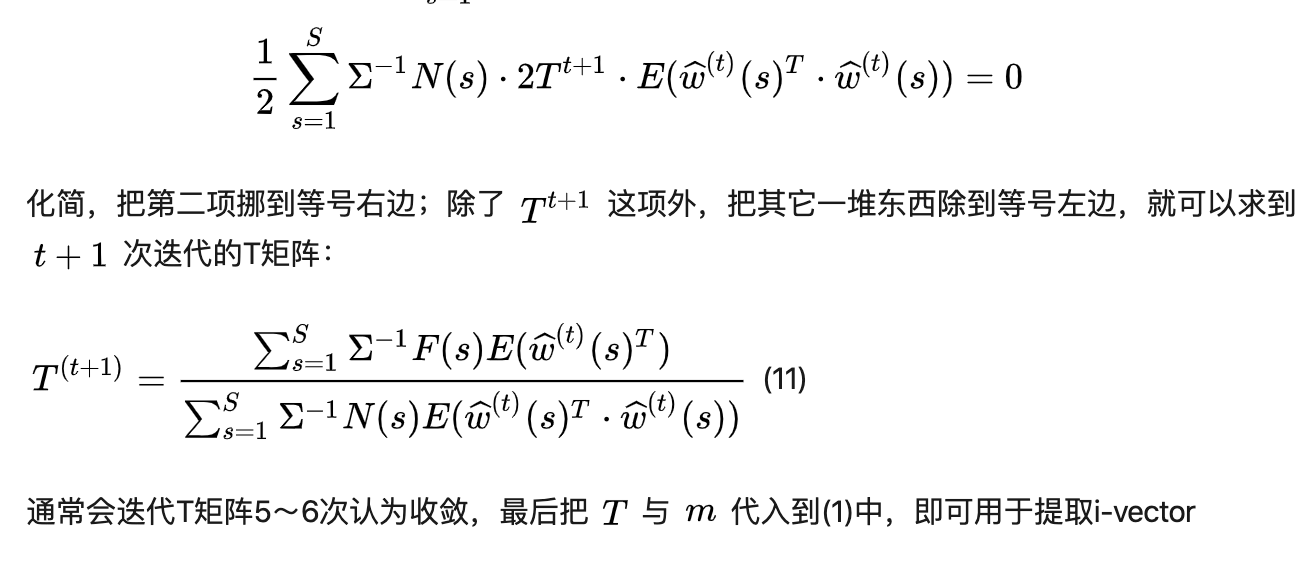
T是总变化矩阵，是待求参数

w(s)是R维的隐变量i-vector，就是最终要求的i-vector









## 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。

IMG_256

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。