高斯混合模型(GMM)

洪青阳 副教授

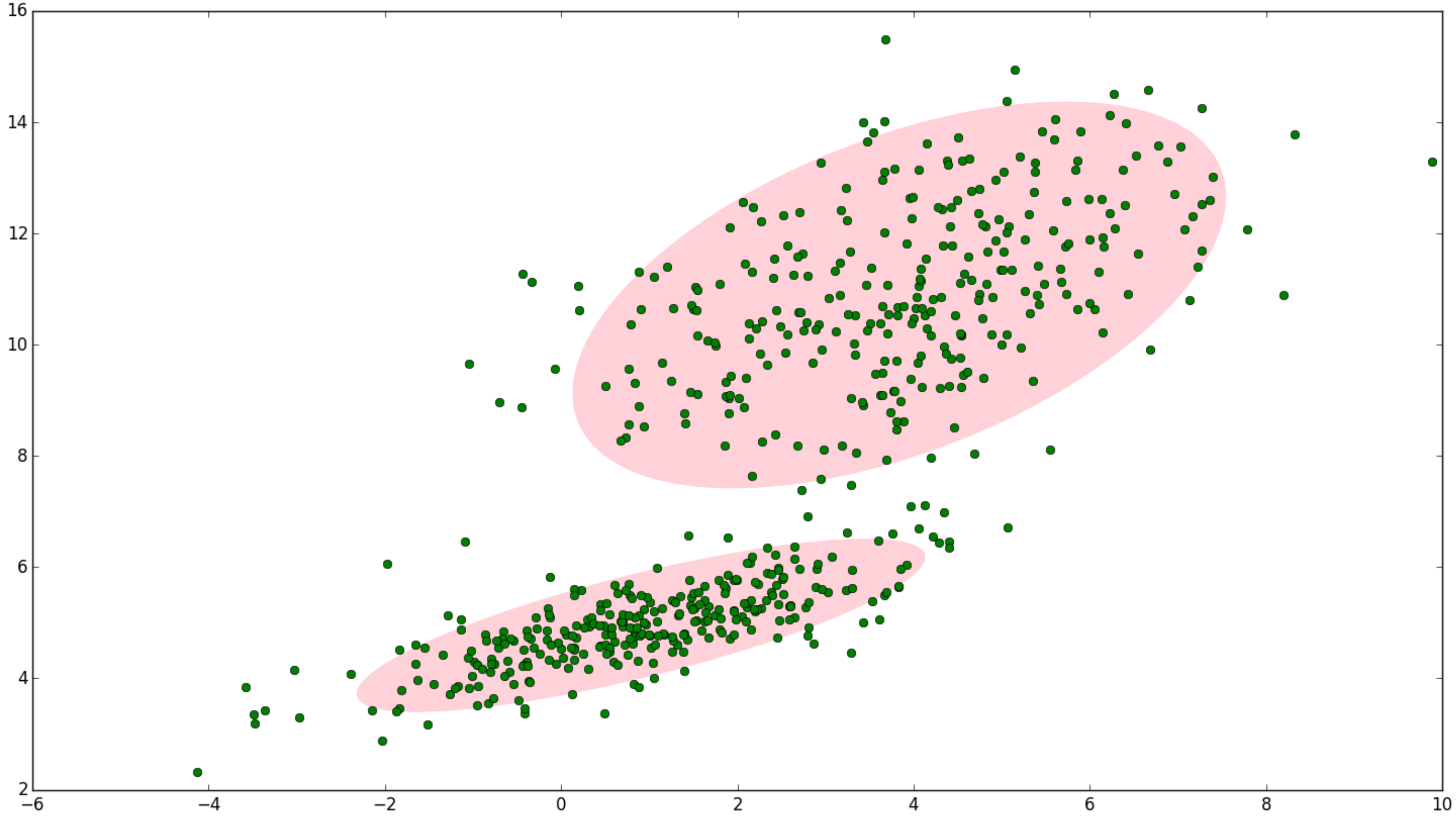
厦门大学信息科学与技术学院
qyhong@xmu.edu.cn

语音信号分布

- 提取后的特征是多维变量：如39维MFCC

0:	-5.954	0.406	-3.967	1.151	-4.128	-0.127	-6.377	-6.841	-8.408	-3.010
	1.127	7.726	52.434	-0.214	-0.619	0.383	-0.689	-1.727	-0.676	-1.324
	0.759	-0.791	0.841	-0.742	-2.830	-0.165	-0.049	0.041	0.030	-0.099
	0.292	0.368	0.174	-0.074	1.162	0.432	0.016	0.145	-0.021	
1:	-5.946	-1.538	-2.467	1.911	-5.919	1.221	-7.645	-1.227	-5.926	0.085
	-5.358	1.309	51.767	-0.177	-0.732	0.511	-1.495	-1.916	0.136	-1.349
	0.698	2.018	2.944	-1.746	-3.288	-0.154	-0.026	0.254	0.037	0.231
	0.755	0.211	0.479	-0.480	0.976	-0.151	0.709	0.939	0.022	
2:	-7.027	-1.717	-2.804	-2.672	-11.870	-4.183	-12.362	-5.854	-13.606	-0.352
	0.662	-3.214	51.941	-0.477	-0.357	0.468	-0.780	-0.174	0.758	-0.439
	0.421	3.613	1.950	-0.160	-1.877	-0.277	-0.012	0.422	-0.072	0.372
	0.953	0.267	0.816	-0.479	0.270	-0.617	1.123	1.496	0.036	
3:	-6.302	-2.191	-1.993	-4.415	-9.839	2.581	-10.131	-3.846	4.280	10.383
	-7.370	-3.244	51.911	-0.211	0.520	0.527	0.511	1.272	-0.337	0.630
	-1.472	1.885	-0.470	2.510	1.387	0.002	0.014	0.435	-0.243	0.245
	0.536	-0.053	0.830	-0.331	-1.281	-1.001	1.198	1.287	0.109	

数据在平面空间的分布



模型选择

- 用哪种分布描述？
  - 能有效表征数据（多样本，多维特征）
  - 对新的不确定数据具备泛化能力
- 如何训练模型？

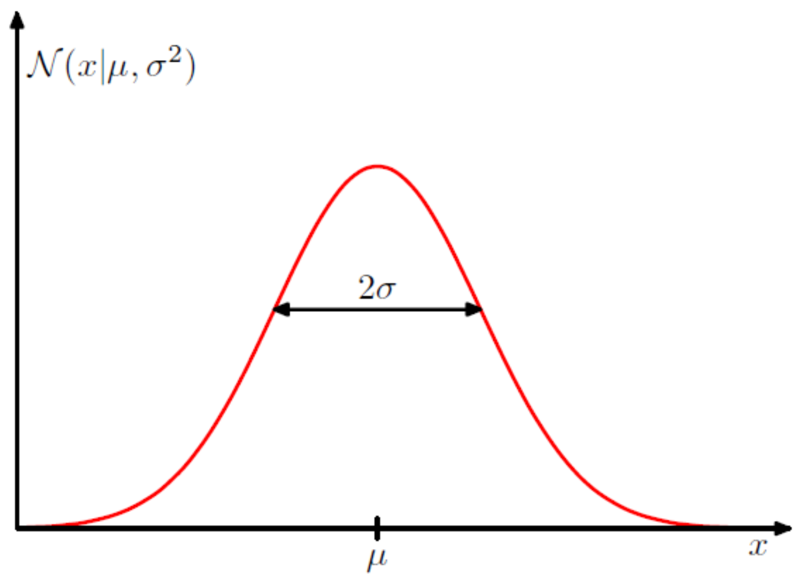
高斯分布(Gaussian Distribution)

- 又称正态分布(Normal Distribution)
- 高斯函数

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

其中μ是均值(mean)，σ^2是方差(covariance).

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$



高斯分布的期望和方差

- 期望

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

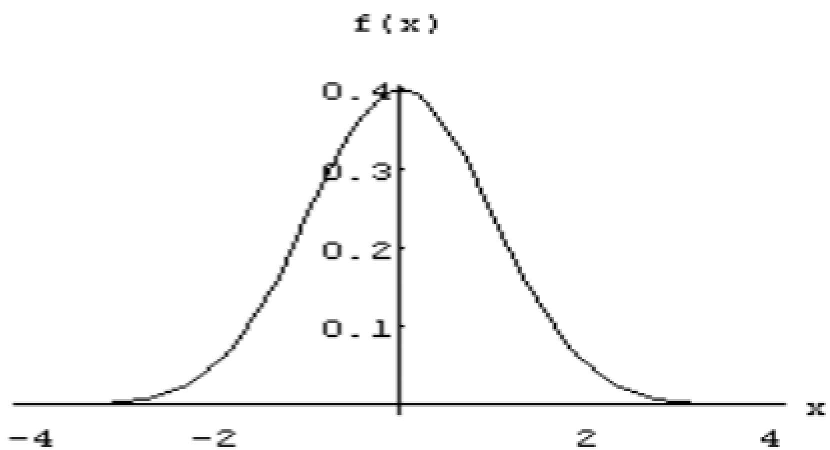
- 方差

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

一维标准正态分布

- 当均值$\mu = 0$，方差$\sigma^2 = 1$，则随机变量x服从标准正态分布。

$$N(x|\mu,\sigma^2) = \frac{1}{(2\pi)^{1/2}} exp\left(-\frac{x^2}{2}\right)$$



二维标准正态分布

$$p(x,y)=p(x)p(y)=N(x|\mu,\sigma^2)N(y|\mu,\sigma^2)$$
$$=\frac{1}{2\pi}exp\left(-\frac{x^2+y^2}{2}\right)$$

把两个随机变量组合成一个随机向量：$\mathbf{v}=[x\ y]^T$

则有

$$p(\mathbf{v})=\frac{1}{2\pi}exp\left(-\frac{1}{2}\mathbf{v}^T\mathbf{v}\right)$$

多维随机变量的高斯分布

D维变量$\mathbf{x}$的高斯分布

$$N(\mathbf{x}|\mu,\sigma^2) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\} = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_d-\mu_d)^2}{{\sigma_d}^2}\right\}$$

其中D维μ是均值， Σ是$D\times D$协方差矩阵， $|\Sigma|$ 是Σ的行列式。

对角矩阵：$|\Sigma|=\prod_{d=1}^{D}{\sigma_d}^2$

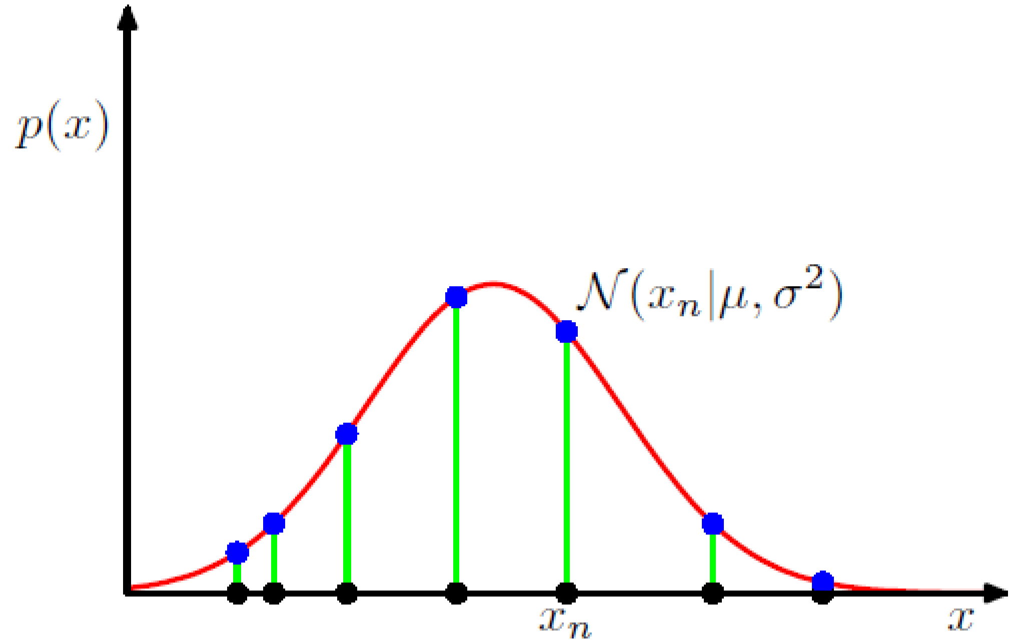
观察系列(多样本)的概率计算

- 给定观察值序列

$$\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$$

- 计算联合概率

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$



对数概率

- 为简化计算，便于计算机处理(防止精度溢出)，一般采用对数概率(似然率)

$$\ln p(\mathbf{x}|\mu,\Sigma) = \sum_{n=1}^{N} \ln(N(x_n|\mu,\sigma^2))$$

$$= \sum_{n=1}^{N} \ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} exp\left\{-\frac{1}{2}\sum_{d=1}^{D}\frac{(x_{nd}-\mu_d)^2}{\sigma_d{}^2}\right\}\right)$$

$$= \sum_{n=1}^{N} \frac{1}{2}\left(-D\ln(2\pi) - \sum_{d=1}^{D}\ln\sigma_d - \sum_{d=1}^{D}\frac{(x_{nd}-\mu_d)^2}{\sigma_d{}^2}\right)$$

最大似然(ML)估计

- 对数似然函数

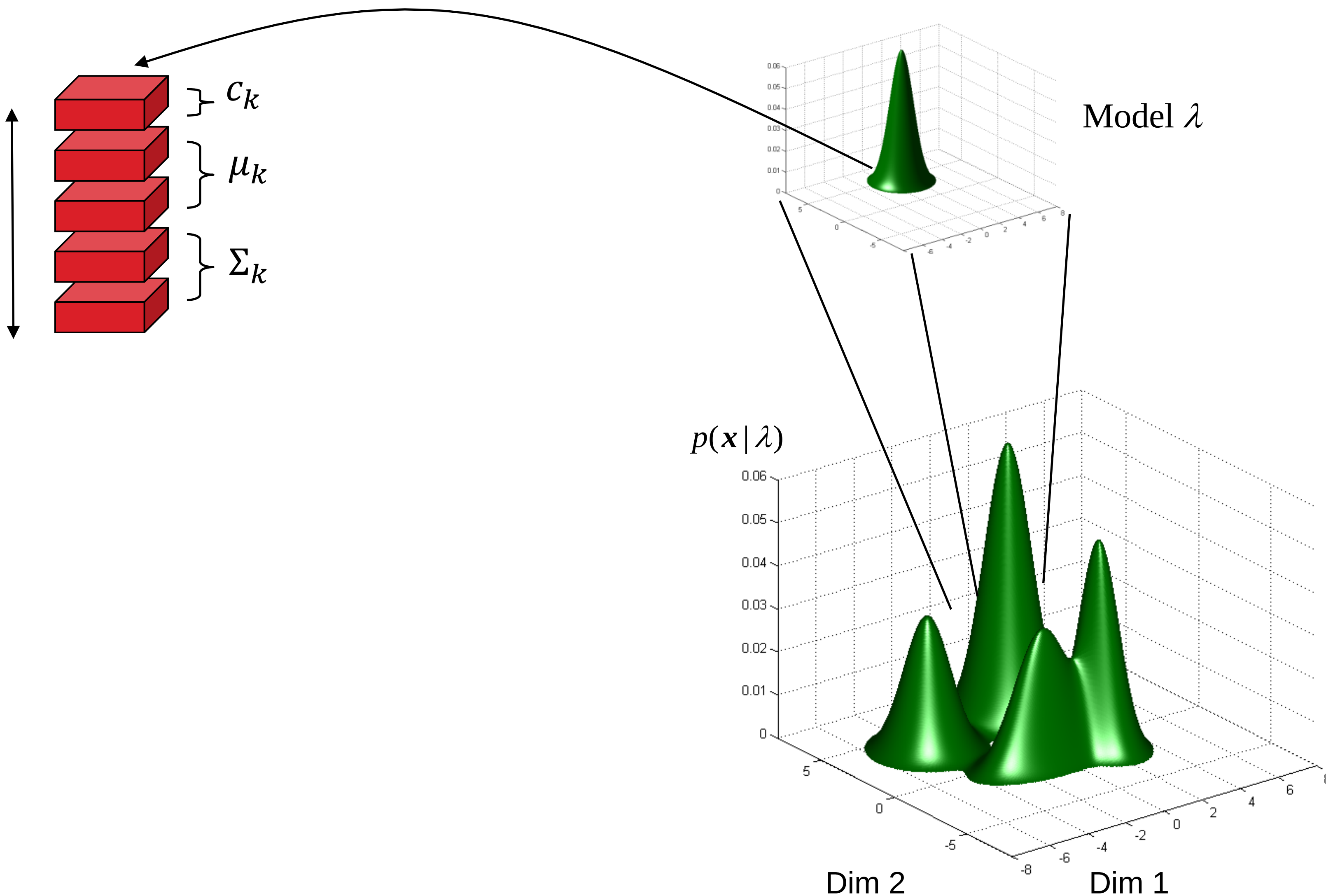
$$\ln p(\mathbf{x}|\mu,\Sigma)=\sum_{n=1}^{N}\frac{1}{2}\left(-D\ln(2\pi)-\ln|\Sigma|-(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)\right)$$

$$=-\frac{ND}{2}\ln(2\pi)-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$$

- 求偏导

$$\begin{cases}\dfrac{\partial\ln p(\mathbf{x}|\mu,\Sigma)}{\partial\mu}=\dfrac{1}{\Sigma}\displaystyle\sum_{n=1}^{N}(x_n-\mu)=0\\[2ex]\dfrac{\partial\ln p(\mathbf{x}|\mu,\Sigma)}{\partial\Sigma}=-\dfrac{N}{2\Sigma}+\dfrac{1}{2\Sigma^2}\displaystyle\sum_{n=1}^{N}(x_n-\mu)^2=0\end{cases}\Rightarrow\begin{cases}\mu_{\mathrm{ML}}=\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}x_n\\[2ex]\Sigma_{\mathrm{ML}}=\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(x_n-\mu)^2\end{cases}$$

高斯混合模型(GMM)



Nicolas Malyska, Sanjeev Mohindra, Karen Lauro, Douglas Reynolds, and Jeremy Kepner

高斯混合模型(GMM)

令$\lambda=\{\mu,\Sigma\}$，K阶高斯GMM的概率密度函数如下：

$$P(x|\lambda)=\sum_{k=1}^{K}P(x,k|\lambda)=\sum_{k=1}^{K}P(k)P(x|k,\lambda)=\sum_{k=1}^{K}c_k N(x|\mu_k,\textstyle\sum_k)$$

其中 $\sum_{k=1}^{K}c_k=1$

K 阶高斯GMM是用 K 个单高斯分布的线性组合来描述。

$$N(x|\mu_k,\textstyle\sum_k)=\frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}}\exp\left\{-\frac{(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}{2}\right\}$$

GMM的似然函数

- 根据权重系数c_k限制，加入拉格朗日算子：

$$\ln p(\mathbf{x}|c,\mu,\Sigma) + \lambda\left(\sum_{k=1}^{K} c_k - 1\right)$$

$$= \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\Sigma_k)\right\} + \lambda\left(\sum_{k=1}^{K} c_k - 1\right)$$

- 分别对μ_k,Σ_k,c_k求最大似然函数。

对μ_k求最大似然函数

- 求偏导并令导数为0

$$-\sum_{n=1}^{N}\frac{c_k N(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\Sigma_k)}\Sigma_k(x_n-\mu_k)=0$$

- 两边同除Σ_k，重新整理得到

$$\mu_k=\frac{\sum_{n=1}^{N}\gamma(n,k)x_n}{\sum_{n=1}^{N}\gamma(n,k)}$$

- 其中

$$\gamma(n,k)=\frac{c_k N(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\Sigma_k)}$$

对$\sum_k$求最大似然函数

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(n,k)(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(n,k)}$$

其中

$$\gamma(n,k) = \frac{c_k N(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k, \Sigma_k)}$$

对c_k求最大似然函数

- 求偏导并令导数为0

$$\sum_{n=1}^{N}\frac{N(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K}c_k N(x_n|\mu_k,\Sigma_k)}+\lambda=0$$

得到

$$c_k=\frac{\sum_{n=1}^{N}\gamma(n,k)}{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(n,k)}$$

EM算法估计GMM参数

1.定义高斯数K，对每个高斯设置c_k,μ_k,$\sum_k$ 的初始值；

2.**E-step**

根据当前的c_k,μ_k,$\sum_k$计算后验概率

$$\gamma(n,k)=\frac{c_k N(x_n|\mu_k,\sum_k)}{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\sum_k)}$$

3.**M-step**

根据E-step中计算的$\gamma(n,k)$，更新c_k,μ_k,$\sum_k$

4.计算对数似然函数

$$\ln p(\mathbf{x}|c,\mu,\sum)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K} c_k N(x_n|\mu_k,\sum_k)\right\}$$

5.检查对数似然函数是否收敛，若不收敛，则返回第2步。

思考

- 如何设置c_k,μ_k,$\sum_k$ 的初始值？
- 如何用代码实现GMM？尝试实现MFCC-GMM，实现声音的分类。

Thank you!

Any questions?