4.2.2 高斯变换算法

根据上面介绍的高斯变换参数估计方法，可以得到如下的高斯变换算法。

算法：　高斯变换（GT）

输入：数据样本集 X$\{x_i \mid i=1,2....,N\}$，高斯分量数 M，迭代终止差值ε

输出：M 个高斯分布（μ,σ_i）及幅值a_i

步骤 1：统计计算数据样本集$X\{x_i \mid i=1,2....,N\}$的频度分布

$$h(y_j)=p(x_i),\ \ i=1,2....,N;\ j=1,2,...,N'$$

其中，y 为样本论域空间；

步骤 2：设定 M 个高斯分布的初始值，第 k (k=1,…,M)个高斯分布的初始参数设定为：

$$\mu_k=\frac{k*\max(X)}{M+1}\ ,\quad \sigma_k=\max(X)\ ,\ a_k=\frac{1}{M}$$

步骤 3：定义并计算目标函数

$$J(\theta)=\sum_{i=1}^{N'}[h(y_i)\times\ln\sum_{k=1}^{\boldsymbol{M}}[a_k g(y_i;\mu_k,\sigma_k^2)]],$$

其中

$$g(y_i;\mu_k,{\sigma_k}^2)=\frac{1}{\sqrt{2\pi}\sigma_k}e^{-\frac{(y_i-\mu_k)^2}{2\sigma_k^2}};$$

步骤 4：对第 k (k=1,…M)个高斯分布，根据极大似然估计，计算出该高斯分布的新参数

$$\mu_k=\frac{\sum_{i=1}^{\mathrm{N}}L_k(x_i)x_i}{\sum_{i=1}^{\mathrm{N}}L_k(x_i)},$$

$$\sigma_k^2=\frac{\sum_{i=1}^{\mathrm{N}}L_k(x_i)(x_i-\mu_k)^T(x_i-\mu_k)}{\sum_{i=1}^{N}L_k(x_i)},$$

$$a_k=\frac{1}{N}\sum_{i=1}^{N}L_k(x_i)$$

其中，

$$L_k(x_i)=\frac{a_k g(x_i;\mu_k,\sigma_k^2)}{\sum_{n=1}^{M}(a_n g(x_i;\mu_n,\sigma_n^2))}$$

步骤 5：计算目标函数的估计值

$$J(\theta)=\sum_{i=1}^{N'}[h(y_i)\times\ln\sum_{k=1}^{\boldsymbol{M}}[a_k g(y_i;\mu_k,\sigma_k^2)]]$$

步骤 6：判断目标函数估计值与原目标函数值差异，

如果 $|J(\hat{\theta})-J(\theta)|<\varepsilon$，

输出当前参数估计值；

否则，跳转至步骤 3。

算法的时间复杂度为：

[o (N) + o (N) + o (M*N') + o (3 *M* N') + o (M*N')]*t = o (M*N)

其中，t 是算法的循环次数，M 是高斯分量的个数，N' 是样本数 N 到论域映射。高斯变换为从任何一个数据分布到高斯分布的转换提供了手段，下面通过实验对不同数量高斯分量拟合的误差进行分析。

实验：设计一个由5个高斯分布组成的高斯混合分布，如表4-1所示。

表 4-1　5 个高斯分布参数表

高斯分量	期望	方差	幅值
1	0	0.5	0.2
2	5	1	0.2
3	7	2	0.3
4	8	2	0.1
5	12	2.5	0.2

依此高斯混合分布生成包含10000个样本点的数据集，对此数据集使用高斯变换生成多个高斯分布，将其参数与原来参数比较。

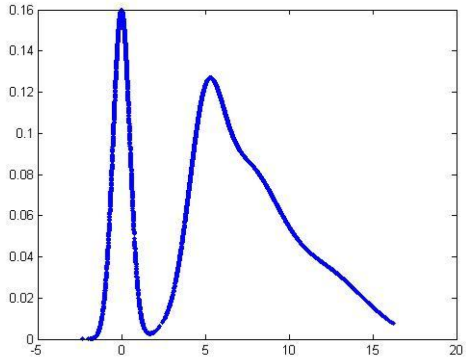

图4-1　5个高斯分布生成的高斯混合分布

实验结果1：使用3个高斯分量进行拟合，估计结果如表4-2和图4-2。

表 4-2　使用 3 个高斯分量估计结果表

高斯分量	**期望**	**方差**	**幅值**
1	11.8340	2.8772	0.2160
2	6.4226	2.7907	0.5922
3	-0.0295	0.4649	0.1918
绝对误差	126.2949		

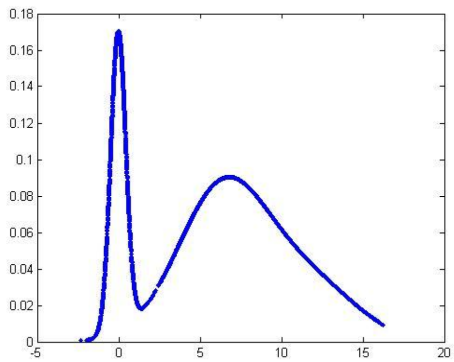

图4-1 利用3个高斯分量的估计结果图