数据分析与建模方法

Data Analysis and Statistical Modeling

金光 著

国防工业出版社
National Defense Industry Press

数据分析与建模方法

金 光 著

国防工业出版社
·北京·

内容简介

本书面向复杂统计问题求解和统计工程需求，介绍现代统计的基本原理和方法，内容涵盖经典统计、贝叶斯统计、统计学习等统计理论以及计算密集型方法和探索性分析方法，涉及数据特征分析、模型参数推断、回归分析建模和系统状态估计等问题。每章后编配有习题。

本书适合作为高等学校自动控制、管理科学与工程等专业的研究生或高年级本科生教材，也可供从事数据分析与建模、装备试验与评价、随机信号处理等技术专题研究的科技工作者学习与参考。

图书在版编目(CIP)数据

数据分析与建模方法 / 金光著. —北京：国防工业出版社，2013.8

ISBN 978-7-118-09023-9

Ⅰ. ①数… Ⅱ. ①金… Ⅲ. ①统计分析-分析方法 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2013)第 184943 号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号　邮政编码 100048)

国防工业出版社印刷厂印刷

新华书店经售

*

开本 710×1000　1/16　**印张** 17¾　**字数** 348 千字

2013 年 8 月第 1 版第 1 次印刷　**印数** 1—2500 册　**定价** 69.90 元

(本书如有印装错误，我社负责调换)

国防书店：(010)88540777　　发行邮购：(010)88540776

发行传真：(010)88540755　　发行业务：(010)88540717

前　言

现代统计在工程和科学中的地位和作用，在深度和广度两个方面都处于快速发展之中。在深度上，新的模型、方法和工具不断涌现，解决复杂总体和复杂数据下的统计建模问题；在广度上，提出统计工程的概念，强调对现有统计方法和工具的集成，解决复杂系统的论证、分析、运行和控制等问题。

本书作者多年来从事可靠性建模理论方法以及小子样条件下复杂系统试验评估方法的研究，深感现有的专著和教材，在解决复杂的数据分析与建模问题时的不足。这些专著和教材，有的内容深刻，但对基本方法的思想和原理的阐述过于理论化，且覆盖面较窄，缺乏比较有代表性的工程案例，不利于工程人员的理解和应用；有的过于强调方法的实用性，只讨论一些基本的模型和方法，无法满足工程实际问题的复杂性和多样性的要求。

正是在这种需求推动之下，并考虑当前复杂系统评估理论方法的发展和需求，作者撰写了本书。本书涵盖主要的统计原理，如经典统计、统计决策、探索性分析、统计学习等，以及这些原理在解决复杂的参数估计、假设检验、回归分析、状态估计等典型数据分析与建模问题中的基本方法，并以可靠性评估和寿命预测等工程问题中复杂数据类型和复杂模型的处理为例，对这些原理和方法进行说明。与现有专著或教材不同的是，本书注重介绍有关的原理和应用，而省去详细的推导过程；同时对比较新颖的或者教材中较少涉及的内容，进行了必要的详细讨论，力求使读者能够理解提出这些方法的出发点及其解决复杂统计问题的方式。

全书共分为6章，基本涵盖了应用统计方法解决实际数据分析与建模问题过程中涉及的几个步骤，包括数据收集、处理、分析、解释以及从数据中得出结论。第1章介绍经典统计，主要说明经典统计如何看待参数估计和假设检验问题，以及解决这些问题的基本思想方法。第2章和第3章涉及两个重要的工程问题，即回归分析建模和系统状态估计，主要介绍经典统计解决这两个问题时的出发点和处理方式。第4章为统计决策与贝叶斯分析，体现将统计问题作为不确定性情形下的决策问题的处理思想，以及贝叶斯方法如何解决参数估计和假设检验问题。第5章为数据特征分析与图形化，侧重对数据特征和规律性信息的探索性分析，强调不

依赖于数据的概率模型，从数据自身发现有价值的信息。第6章简单介绍统计学习中的支持向量机理论，与经典统计相比，它强调在有限样本量情形下解决判别分析和回归分析等问题。

本书的编写过程得到张金槐教授的鼓励和指导，张教授通读了全书，并提出了很多宝贵意见，作者对张教授的支持表示衷心感谢。本书的出版得到武小悦教授的大力支持，正是由于武教授的认真审查和积极推荐，本书才有幸获得学校专著出版经费资助，在此亦表示衷心感谢。周经伦教授、冯静副教授为本书编写和出版提供了许多宝贵的建议和资料，刘强、厉海涛、肖磊、郝旭东、赵琰、周军等研究生在本书的编写过程中也给予了作者很多帮助，应该说，本书也包含了他们的劳动成果。

本书的完成得到国防科技大学学术著作专项经费与国家自然科学基金（“基于性能退化的可靠性理论方法研究”，编号：71071158）的资助。

由于作者水平有限，本书的选材和文字难免存在不当和疏漏之处，敬请读者不吝批评指正。

编著者

2013年6月

目　录

第1章　经典统计方法

1.1　点估计

设随机变量 X 的分布由含未知参数 θ 的概率分布 $F(x|\theta)$ 描述，$\theta \in \Theta$，Θ 为参数空间。也就是说，已知必存在一个 $\theta_0 \in \Theta$，使得 X 的分布就是 $F(x|\theta_0)$，但不知 θ_0 的具体数值。现在对变量 X 进行观察，得到样本 $\boldsymbol{X}=(X_1,\cdots,X_n)$，希望根据它对真值 θ_0（更一般地，对定义在 Θ 上的取值于 R^k 的函数 $g(\theta)$ 的值 $g(\theta_0)$）作估计，即用 Θ（或 R^k）中的一个点去估计 θ_0（或 $g(\theta_0)$），这就是点估计这个名称的由来。显然，对未知参数或其函数，可能构造无穷多个估计量。统计中解决这一问题的方式，是寻找具有某种最优性质的估计量。

1.1.1　最优估计的意义

我们知道，一致最优的估计量是不存在的。一个整体上很坏的估计量，在局部上也可以具有某种优越性。例如，$\hat{g}(X)=c$，当真值确定为 c 时，无疑是最好的估计；但整体上这是一个很坏的估计。实际上研究点估计问题，一般先把条件（最优性准则）放宽一些，使适合这种最优性准则的估计一般能存在。这可以解释为，我们所寻求的是从某一特定方面看具有最优性的估计量，而不要求它在一切方面都最优。还有另外一种处理办法，即先对估计量的性质作某种特定要求（如无偏性就是最重要、最常见的一种），凡不满足这一要求的估计量都不在我们考虑之列；把一切适合这一要求的估计量的全体记为 $\mathscr{D}$，在 $\mathscr{D}$ 中找一个一致最优或满足其他某种最优性标准的估计量。

无偏估计就是采用上述思路解决最优估计问题的最广为人知的方法。从实际应用的角度看，无偏估计的意义：当估计量经常使用时，它保证了平均来说，即在多次重复的平均意义下，给出接近于真值的估计。如果应用上的要求主要在于这一点，无偏性的要求当然是合理甚至必须的。例如，某工厂生产的产品长期供应给某商店，则无偏性保证了从较长期看，双方是公平的。但不少应用问题没有这种经常性，其中的正、负偏差不能相互抵消，这种情况下无偏性就没有什么意义。所以，一方面无偏估计是一个重要而有用的概念；另一方面应根据问题的性质来估价这个准则的作用。

在所有具有无偏性的估计量中寻找“最好的”估计量，一般以方差评价估计量的优良性质，因此最好的无偏估计就是最小方差无偏估计（MVUE）。具体地，设 g

(θ)为取值于$\boldsymbol{R}^1$的函数，若存在一个无偏估计$\hat{g}$，对$g(\theta)$的任何无偏估计g^*，都有

$$\mathrm{Var}_\theta(\hat{g}(X)) \leqslant \mathrm{Var}_\theta(g^*(X))$$

则称$\hat{g}$为$g(\theta)$的一个最小方差无偏估计。

需要注意的是，无偏估计并非在任何条件下都存在。例如，设X服从参数为p的二项分布，要估计$g(p)=1/p$，设$\hat{g}(x)$为p的无偏估计，记$\hat{g}(i)=a_i$，则

$$\sum_{i=0}^{n} a_i \binom{n}{i} p^i (1-p)^{n-i} = \frac{1}{p}$$

这显然是不可能的。显然，只要$g(p)$是次数大于n的多项式，$g(p)$的无偏估计都不可能存在。

再如，设$X \sim N(\theta,1)$，要估计$g(\theta)=|\theta|$。设$\hat{g}(x)$为无偏估计，则应有

$$E_\theta[\hat{g}(x)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{g}(x) \exp\left(-\frac{(x-\theta)^2}{2}\right) \mathrm{d}x = |\theta|$$

但是由指数族的性质可知，上式中间一项在整个实数集上有各阶连续导数，而$|\theta|$在$\theta=0$处不可导，因此是不可能的。

研究点估计问题的另一个重要方面，而且在历史上也是发展较早的一个方面，就是点估计的大样本理论，即在样本量无限增加时，估计量可能有的性质。大样本理论的重要性在于，首先，要弄清楚一个估计量的性质，必须知道它的分布。在样本量固定时这只在很少情况下能做到。当样本量无限增加时，估计量的分布往往趋向于一个常见的简单分布，这至少在样本量很大时为我们了解估计量的全面性质提供了依据。其次，在样本量固定时，寻求具有某种最优性的估计量是不容易的，其存在性也未必能保证。从大样本理论着眼，为研究种种估计量的性质提供了一个途径。

以样本均值$\bar{X}_n$为例，大数定理表明，在总体方差有限的情况下，当样本量n无限增加时，样本$X_1,\cdots,X_n$的均值$\bar{X}_n$依概率收敛于被估计的总体F的均值$\theta(F)$。也就是说，只要样本量足够大，估计量可以以随意接近于1的概率把$\theta(F)$估计到所需的精度，这个性质叫做估计量的相合性。该定义的相合性有时称为弱相合性。还有所谓强相合性，即几乎处处收敛的概念。

另外，根据中心极限定理，若总体方差σ_F^2不为0，则有

$$\frac{\sqrt{n}(\bar{X}_n - \theta(F))}{\sigma_F} \xrightarrow{L} N(0,1), \quad n \to \infty$$

式中：$\xrightarrow{L}$表示依分布收敛。这个性质叫$\bar{X}_n$的相合渐近正态性，简称CAN估计。

由这个性质，当n很大时，可以近似算出$\bar{X}_n$落在被估计值$\theta(F)$的某个范围内的概率（当然还需对总体方差进行估计）。如果参数空间Θ为欧几里德空间，且上

述收敛性在任何 $\Theta \in \theta$ 的某邻域内有一致性，则称估计量为 θ 的相合一致渐近正态估计，简称 CUAN 估计。

第三个重要的概念是所谓最优渐近正态估计。假设估计量 T 是 CAN 估计，即

$$\sqrt{n}[T(X_1,\cdots,X_n)-\theta] \xrightarrow{L} N(0,v(\theta)),n \longrightarrow \infty$$

并且 $v(\theta)$ 达到 C－R 下界，即

$$v(\theta)=\frac{1}{I(\theta)}$$

式中：$I(\theta)$ 为 Fisher 信息量，则称 T 为 θ 的最优渐近正态估计，简称 BAN 估计。

以上就是最基本的大样本性质。总之，所谓大样本性质，就是当样本量无限增加时与该估计量的分布有关的种种极限性质。显然，这不仅取决于估计量的形式，也与总体的分布有很大关系。

1.1.2　极大似然估计原理

鉴于极大似然估计方法在经典统计中的地位，本节仅介绍极大似然估计方法，对于另一种应用比较广泛的点估计方法——矩估计，请参考其他教材或专著。

极大似然估计的出发点是根据估计的参数所确定的总体产生特定样本数据的概率最大。假设随机变量 $X_1,\cdots,X_n$ 具有联合密度或频率函数 $f(x_1,\cdots,x_n|\theta)$，观测值为 $X_i=x_i(\ i=1,\cdots,n)$，则 θ 的似然函数是 $\boldsymbol{x}=(x_1,\cdots,x_n)$ 的函数，定义为

$$L(\theta)=f(x_1,\cdots,x_n \mid \theta)$$

θ 的极大似然估计（简称 MLE）是使得该似然函数取值最大的 θ，即使得观测数据“概率最大”或“最可能”的 θ。在似然函数具有所需的正则性的条件下，这可通过求解似然方程得到；似然方程通过令似然函数对未知参数的偏导数为 0 获得。

若诸 X_i 独立同分布（i. i. d. ），则似然函数为

$$L(\theta)=\prod_{i=1}^{n} f(x_i \mid \theta) \tag{1.1.1}$$

由于一个函数和它的对数具有相同的分析性质，实际上多采用对数似然函数，即用似然函数的对数求解极大似然估计，定义为

$$L(\theta)=\sum_{i=1}^{n} \log[f(x_i \mid \theta)] \tag{1.1.2}$$

关于似然函数可以直观地解释如下：若 X 是离散随机变量，则 $L(\theta)=P_\theta(X=x)$。比较参数 θ_1 和 θ_2 的似然，如果

$$P_{\theta_1}(X=x) > P_{\theta_2}(X=x)$$

即观测到的样本值更可能在 $\theta=\theta_1$ 下发生，也就是说，相比 θ_2，θ_1 应该是一个更可能的猜测。对连续随机变量 X，可以利用 X 的邻域作类似的解释，即特定样本取值

发生的概率近似表示为

$$\frac{P_{\theta_1}(x-\varepsilon<X<x+\varepsilon)}{P_{\theta_2}(x-\varepsilon<X<x+\varepsilon)}\approx\frac{L(\theta_1)}{L(\theta_2)}$$

若在 θ_1 下具有更大的概率，则 θ_1 应该更可能。

需要注意的是，不能将似然解释为参数 θ 的概率。

例 1.1 （Hardy-Weinberg 均衡——族群遗传学的基石）族群遗传学是连接孟德尔定律和达尔文进化论的学科，它的特色是利用数学的方法研究受到选择、突变、迁移、近亲交配及其他因素影响的族群基因结构。它肯定了数学在生物学上的作用，而且被公认为数学应用在生物学上唯一成功的例子。

若基因频率是均衡的，则根据 Hardy-Weinberg 定律，基因类型 AA、Aa 和 aa 发生的频率分别为 $(1-\theta)^2$、$2\theta(1-\theta)$ 和 θ^2。1937 年从香港的中国人抽取血型样本，结果如表 1.1.1 所列（其中 M 和 N 分别是红血球和抗原）。

表 1.1.1　血型样本

	血型			
	M	MN	N	合计
频率	342	500	187	1029

有几种方法可以得到参数 θ 的估计。一种容易想到的方法是令 $\theta^2=187/1029$，得到 $\theta=0.4263$。但是这种方法看起来似乎丢掉了其他单元中的信息，因而可能不是一种好的方法。为此把 3 个单元（cell）合起来考虑是合适的，以 X_1、X_2、X_3 表示 3 个单元的计数，令 $n=1029$，则 θ 的对数似然函数为

$$L(\theta)=\log n!-\sum_{i=1}^{3}\log X_i!+X_1\log(1-\theta)^2+X_2\log 2\theta(1-\theta)+X_3\log\theta^2$$

$$=\log n!-\sum_{i=1}^{3}\log X_i!+(2X_1+X_2)\log(1-\theta)+(2X_3+X_2)\log\theta+X_2\log 2$$

令 $L'(\theta)=0$，得到的估计为

$$-\frac{2X_1+X_2}{1-\theta}+\frac{2X_3+X_2}{\theta}=0\Rightarrow\hat{\theta}=\frac{2X_3+X_2}{2n}=0.4247$$

显然，对任何未知参数可以构造无穷多估计量，但是我们关心的是“最好的”估计。关于极大似然估计，下面的结论表明，它在某些意义上具有所要求的优良性。

（1）不变性：设 $\hat{\theta}$ 是未知参数 θ 的 MLE，则对任何函数 $\tau(\theta)$，它的极大似然估计为 $\tau(\hat{\theta})$。

（2）一致性：在密度或频率函数 f 具有适当的光滑性条件下，由 i. i. d. 样本得到的极大似然估计是一致收敛的。

（3）渐近正态性：在 f 具有适当光滑性条件下，$\sqrt{nI(\theta_0)}(\hat{\theta}-\theta_0)$ 的概率分布

趋于标准正态分布。

根据上面的结论，还可以进一步知道极大似然估计是渐近无偏和渐近有效的。因此，一般来说极大似然估计比矩估计具有更高的精度，因为矩估计不是渐近有效估计。

通过极大似然求极值的方式，并不总能获得未知参数的极大似然估计，它需要以下条件的保证：首先，要求真值 θ_0 在参数空间 Θ 的内部，否则难以根据偏导数为0求解似然方程；其次，要求密度或频率函数 $f(x|\theta)$ 的支撑集（即使 $f(x|\theta)>0$ 的 x 的取值集合）不依赖于 θ，如 $[0,\theta]$ 上的均匀分布的支撑集依赖于 θ，其位置参数 θ 的极大似然估计也不能通过似然方程获得。

另外，即使在最简单的情形下，极大似然估计也可能失败。例如，对于由两个正态密度通过简单混合形成的密度为

$$f(x;a,\sigma) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)+\frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

式中：参数 (a,σ) 是未知的，用极大似然方法不可能估计出这个密度函数。

事实上，对任何数据 $x_1,\cdots,x_n$ 及任何给定的常数 c_0，总存在一个小的 $\sigma=\sigma_0$，使得对 $a=x_1$，似然函数值超过 c_0，即

$$\begin{aligned} L(a=x_1,\sigma_0) &= \sum_{i=1}^{n}\log f(x_i;a=x_1,\sigma_0) \\ &> \log\left(\frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_0}\right)+\sum_{i=2}^{n}\log\left(\frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}\right) \\ &= -\log\sigma_0-\sum_{i=2}^{n}\frac{x^2}{2}-n\log 2\sqrt{2\pi} > c_0 \end{aligned}$$

也就是说，在这个例子中，似然函数的最大值不存在，因此用极大似然法无法给出参数 (a,σ) 的估计。

1.1.3　数据缺失与 EM 算法

虽然极大似然估计理论比较系统地解决了点估计问题，但是似然函数的解在一般情况下是无法解析获得的。此时一般采用数值算法，典型的有 Newton－Raphson 法、Fisher 的 Scoring 法以及 EM 算法。

EM 算法是解决复杂模型下极大似然估计的一种重要方法，特别适合“缺失数据”（missing data）问题中对参数用 MLE 求解；这里数据缺失包括两种情况：①由于观测过程的限制或问题引起的数据缺失（如聚类问题）；②直接根据观测数据，似然函数极值解析不可求；但若假设缺失数据（隐含变量）的值已知，则似然函数形式很简单。

先以一个例子简单介绍 EM 算法的基本原理。设有 197 只动物分为 4 类，即动物的类别服从多项式分布。观测到的数据为

$$y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

遗传学模型给出的这些动物属于不同类型的概率为

$$\left(\frac{1}{2} + \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi\right)$$

于是观测数据的概率为

$$g(y \mid \pi) = \frac{(y_1 + y_2 + y_3 + y_4)!}{y_1! y_2! y_3! y_4!}\left(\frac{1}{2} + \frac{1}{4}\pi\right)^{y_1}\left(\frac{1}{4} - \frac{1}{4}\pi\right)^{y_2}\left(\frac{1}{4} - \frac{1}{4}\pi\right)^{y_3}\left(\frac{1}{4}\pi\right)^{y_4}$$

为了说明 EM 算法，把数据 y 表示为一个分为 5 类的多项式分布的不完全数据，这 5 类的概率分别为

$$\left(\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi\right)$$

显然这里是把原来的第一类分为了两类。这样，完全数据应该为 $x = (x_1, x_2, x_3, x_4, x_5)$，其中 $y_1 = x_1 + x_2, y_2 = x_3, y_3 = x_4, y_4 = x_5$，完全数据的概率为

$$f(x \mid \pi) = \frac{(x_1 + x_2 + x_3 + x_4 + x_5)!}{x_1! x_2! x_3! x_4! x_5!}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{4}\pi\right)^{x_2}\left(\frac{1}{4} - \frac{1}{4}\pi\right)^{x_3}\left(\frac{1}{4} - \frac{1}{4}\pi\right)^{x_4}\left(\frac{1}{4}\pi\right)^{x_5}$$

通过定义 EM 算法，我们看如何通过 $\pi^{(p)}$ 得到 $\pi^{(p+1)}$，其中 $\pi^{(p)}$ 表示经过 p 次迭代后得到的 π 的估计值，这里 $p = 0, 1, 2, \cdots$。

注意到这里 (x_1, x_2) 的可能取值包括了 $(0, 125), (1, 124), \cdots, (125, 0)$，因此不完全似然函数需要对这些可能的组合求和，而 (x_3, x_4, x_5) 只要用 $(18, 20, 34)$ 简单代替即可。

按 EM 算法，根据给定数据 y 和概率 $\pi^{(p)}$ 估计完整数据 x。由于 (x_3, x_4, x_5) 已知为 $(18, 20, 34)$，所以只要估计 x_1 和 x_2，这里 $x_1 + x_2 = y_1 = 125$。使用当前 π 值，得到

$$x_1^{(p)} = 125\frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{4}\pi^{(p)}}, \quad x_2^{(p)} = 125\frac{\frac{1}{4}\pi^{(p)}}{\frac{1}{2} + \frac{1}{4}\pi^{(p)}}$$

于是假设这就是完全数据中的 x_1 和 x_2 的值，据此估计 π 值，得到更新的估计

$$\pi^{(p+1)} = \frac{x_2^{(p)} + 34}{x_2^{(p)} + 34 + 18 + 20}$$

迭代地执行上述过程，就得到本例的 EM 算法，如表 1.1.2 所列。

表 1.1.2　EM 迭代

p	$\pi^{(p)}$	$\pi^{(p)}-\pi^*$	$(\pi^{(p+1)}-\pi^*)/(\pi^{(p)}-\pi^*)$
0	0.500000000	0.126821498	0.1465
1	0.608247423	0.018574075	0.1346
2	0.624321051	0.002500447	0.1330
3	0.626488879	0.000332619	0.1328
4	0.626777323	0.000044176	0.1328
5	0.626815632	0.000005866	0.1328
6	0.626820719	0.000000779	—
7	0.626821395	0.000000104	—
8	0.626821484	0.000000014	—

从初始值 $\pi^{(0)}=0.5$ 开始，令 $\pi^*=\pi^{(p)}=\pi^{(p+1)}$，其实可以显式地求解二次方程，得到 π 的极大似然估计为

$$\pi^* = (15+\sqrt{53809})/394 \approx 0.6268214980$$

由这个简单的例子看出，EM 算法可以分为两步，即 E 步和 M 步。E 步的作用是求期望，即在给定观测数据的条件下，计算完整似然的期望（随机变量为隐含变量）。其中涉及的计算缺失数据的条件期望，需要利用参数的当前估计值。M 步是求极大值，即求使得完整似然的期望最大的参数，是一个极大值求解问题。通常可以解析求解。EM 算法步骤如下。

（1）选择初始估计 θ_0，重复以下 E 步和 M 步。

（2）E 步：计算

$$L(\theta,\theta_n) = E_{Y|X,\theta_n}[L(X,Y|\theta)] \tag{1.1.3}$$

（3）M 步：

$$\theta_{n+1} = \arg\max_{\theta} L(\theta,\theta_n) \tag{1.1.4}$$

或者取 θ_{n+1} 使得

$$L(\theta_{n+1},\theta_n) > L(\theta_n,\theta_n) \tag{1.1.5}$$

在实际计算过程中，亦可用对数似然函数 l 代替似然函数 L，算法的步骤不变。

EM 算法的理论分析涉及算法收敛性、EM 算法输出结果是否依赖于参数的初始值、收敛速度、潜在变量期望计算的难易程度等问题，此处不再详细讨论。下面看一个稍微复杂一些的问题，并采用 EM 算法估计模型参数。

例 1.2　设样本数据来自于混合正态分布随机变量 X

$$X = (1-Y)\cdot X_0 + Y\cdot X_1$$

式中：$Y\in\{0,1\}$ 为指示变量，$p(y=1)=\pi$；$X_0\sim N(\mu_0,\sigma_0^2)$、$X_1\sim N(\mu_1,\sigma_1^2)$ 为正态分布随机变量。

混合正态分布如图 1.1.1 所示，其中 $\mu_0=1.6$，$\mu_1=1.75$，$\sigma_0=0.05$，$\sigma_1=0.10$，$\pi=0.60$。

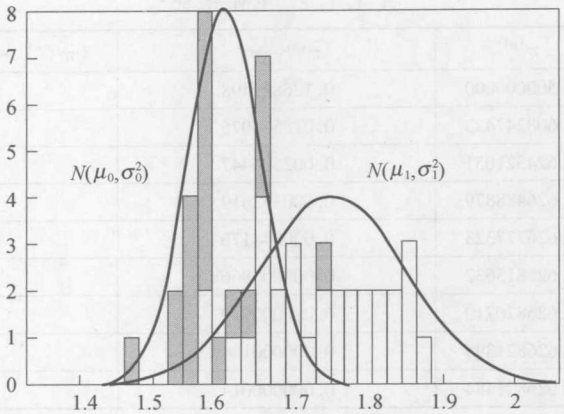

图 1.1.1　混合正态分布

设样本数据为 X,待估参数为 $\theta=(\mu_0,\sigma_0,\mu_1,\sigma_1,\pi)$。易知 X 的概率分布密度为

$$f_X(x)=(1-\pi)\cdot\phi_0(x)+\pi\cdot\phi_1(x)$$

记 $Z_i=(X_i,Y_i)$,$Z=(Z_1,Z_2,\cdots,Z_n)$,由于指示变量 Y 的值未知,故似然函数为

$$L(\theta;Z)=\log\prod_{i=1}^{n}f(X_i;\theta)=\log\prod_{i=1}^{n}\Big[\sum_{Y_i}f(X_i,Y_i;\theta)P(Y_i)\Big]$$

$$=\sum_{i=1}^{n}\log[(1-\pi)\phi_0(x_i)+\pi\phi_1(x_i)]$$

该似然函数涉及求和的对数运算,计算极值困难,为此考虑采用 EM 算法。

首先,若指示变量 Y 的值也已知,则可以得到完整的似然函数

$$L(\theta;Z)=\log\prod_{i=1}^{n}f(X_i,Y_i;\theta)=\sum_{i=1}^{n}\log(f(X_i,Y_i;\theta))$$

$$=\sum_{i=1}^{n}\log(f(X_i\mid Y_i,\theta)\cdot f(Y_i\mid\theta))$$

$$=\sum_{i=1}^{n}\log(\pi_{Y_i}\cdot\Phi_{Y_i}(X_i\mid\theta_{Y_i}))$$

由于指示变量 Y 的值未知,计算完整似然函数对 Y 的期望以去掉其中的变量 Y,则

$$l(\theta,\theta^t)=E_{Y\mid X,\theta^t}[l(\theta;Z)\mid X,\theta^t]$$

$$=\int_{y\in Y}l(\theta;Z)f(y\mid X,\theta^t)\,\mathrm{d}y$$

其中,根据贝叶斯公式获得 Y 的分布,即

$$f(y_i\mid X_i,\theta^t)=\frac{f(X_i\mid y_i,\theta^t)f(y_i\mid\theta^t)}{f(X_i\mid\theta^t)}=\frac{\pi_{y_i}\phi_{y_i}(X_i)}{(1-\pi)\phi_0(X_i)+\pi\phi_1(X_i)}$$

$$f(y\mid X,\theta^t)=\prod_{i=1}^{n}f(y_i\mid X_i,\theta^t)$$