

基于信息扩散原理的估计理论 方法及其抗差性*

王新洲

(武汉测绘科技大学科技处,武汉市珞喻路 129 号,430079)

摘 要 介绍了适合小样本情况下估计概率密度函数的信息扩散理论,导出了参数估计的信息扩散估计法——ISE(information spread estimation) 理论和实践表明,ISE 法可以在小样本情况下估计母体的概率密度函数。该方法不仅具有良好的抗差性,而且具有方法简便、便于应用等优点。

关键词 信息扩散原理;参数估计;密度函数

分类号 P207

在实际参数估计中,人们往往不是选用最小二乘估计,就是选用抗差估计。由于不知道观测向量究竟服从什么分布,所以不论选用哪种估计,都有一定的风险。如采用最小二乘估计,当观测向量不服从正态分布时,其结果可能并非最优,甚至面目全非;而采用抗差估计,当观测向量服从正态分布时,估计结果可能不如最小二乘估计。尽管根据专业知识、实际经验和假设检验,可使风险的概率尽量减小,但却不可能做到没有风险。

事实上,由于观测条件千差万别,在不同观测条件下得到的观测向量很可能服从不同的分布。如果对任一具体观测条件下得到的观测向量,在参数估计之前就能估计出它所服从的具体分布,然后再根据此分布进行参数估计,就可避免上述风险。基于这一思想,本文提出了基于信息扩散原理的估计理论。

1 信息扩散原理简介

1.1 信息扩散

设母体 K 的概率密度函数为 $f(x)$, W 为给定的来自母体 K 的样本。当由 W 不能完全精确地认识 $f(x)$ 时,称 W 对 K 来说是非完备的^[1]。根据非完备的样本 W 对母体 K 的认识,必然不确切,对有关物理规律的解释也含糊不清。但当增加样本点,使 W 趋于或达到完备时,则据 W 对 K 的认识就会趋于或达到清晰。由此可见, W 从非完备到完备,具有一种过渡趋势。当 W 非完备时,这种趋势表现在 W 的样本点上,就是每一个样本点

都有发展成多个样本点的趋势,使每一个样本点都充当“周围未出现之样本点的代表”。因为“周围”的边界是不清楚、模糊和富有弹性的,所以每一个样本点所提供的包括周围影响在内的信息总体是一个模糊信息。由此可知,当 W 非完备时,其过渡性导致它具有模糊不确定性。这一模糊不确定性体现在每个样本点都具有一定的影响域,以显示它们来自非完备样本。

由于非完备样本 W 的每个样本点 w_i 均可作为其“周围”的代表,这就意味着 w_i 的出现不再仅仅是提供它的观测值那一点上的信息,它同时还提供了关于“周围”情况的信息。当然,它对样本点上所提供的信息量大于它对“周围”点上所提供的信息量。设它对样本点上所提供的信息量为 1,则它对“周围”点上所提供的信息量小于 1。

当 W 非完备时, w_i 只是“周围”的代表,设 w_i 的观测值为 l_i , W 在 l_i 点提供的信息应被周围点所分享。而周围各点所分享到的信息与其属于“ l_i 点周围”的程度有关。显然,越靠近 l_i 的点,属于“ l_i 点周围”的程度越高,从 l_i 分享到的信息就越多;反之,越远离 l_i 的点,属于“ l_i 点周围”的程度就越低,从 l_i 分享的信息也就越少。我们称从 l_i 所分享到的这种信息为从 l_i 扩散来的信息,而将 l_i 点的信息被周围点分享的过程称为信息扩散过程,简称信息扩散。

1.2 信息扩散原理与扩散估计

信息扩散原理^[1]如下:

设 $W = \{w_1, w_2, \dots, w_n\}$ 是知识样本, V 是基础论域,记 w_i 的观测值为 v ,设 $x = h(v - w_i)$,则

W 非完备时,存在函数 $_ (x)$,使 v_i 点获得的量(值为 1 的信息可按 $_ (x)$ 的量值扩散到 l ,且扩散所得的原始信息分布 $Q(x)=\sum_{j=1}^n _ (x)=\sum_{j=1}^n (h(v-v_j))$ 能更好地反映 W 所在总体的规律,这一原理称为信息扩散原理。根据这一原理对母体概率密度函数的估计称为扩散估计。

定义 设 $_ (x)$ 为定义在 $(-\infty,\infty)$ 上的一个波雷尔可测函数, $\Delta_n>0$ 为常数, n 为子样容量,则称

$$\hat{f}(l)=\frac{1}{n\Delta_n}\sum_{i=1}^n\left[\frac{l-l_i}{\Delta_n}\right] \tag{1}$$

为母体 K 的概率密度函数 $f(x)$ 的一个扩散估计。式中, $_ (x)$ 称为扩散函数; Δ_n 称为窗宽。

1.3 扩散函数 $_ (x)$ 的确定

由式 (1) 知,基于信息扩散原理的扩散估计的关键是扩散函数 $_ (x)$ 的具体形式。对于不同的 $_ (x)$,可得到不同的扩散估计 $\hat{f}(x)$ 。因此,应在最少假设下确定 $_ (x)$ 的具体形式

综上所述,信息扩散与物理学中的分子扩散极为相似^[2]。所以根据分子扩散理论,采用数学物理方法^[3]来确定扩散函数 $_ (x)$ 。

设扩散函数 $_ (x)$ 是扩散过程完成时刻 $t=t_e$ 时的状态。因此,对于扩散过程中的 t 时刻来讲, x 点的信息状态应为 $_ (x,t)$ 。信息扩散是由于在论域中的某些点没有直接从知识样本获得信息而需要扩散,分子扩散则是当物质分子的空间分布不均匀时而发生扩散。它们的共同之处是都有一种填充空白的功能,而且它们的扩散都是沿着浓度减小的方向发生。因此,可认为信息扩散状态 $_ (x,t)$ 与分子扩散理论中用来表示扩散物质每单位体积内的分子数 d 有相同的物理意义。与分子扩散理论中的流密度 e 和单位长度上浓度的变化率 Ld/Lx 相仿,故引进在 x 点的信息流密度 q 和信息分布梯度 $L_ /Lx$ 的概念。由分子扩散原理中的斐克定律(或扩散定律)知:

$$e=-D(Ld/Lx) \tag{2}$$

式中, D 称为扩散率;负号表示扩散沿浓度减小的方向进行。

为了得到只包含浓度 d 的关系式,可把斐克定律与分子数守恒原理结合起来考虑,如图 1,其体积为 $dV=Sdx$ 。在某一时刻,该体积元内的粒子数为 $ddV=dSdx$ 。其入射通量等于 eS ,出射通量为 $e'S$ 。由粒子数守恒定律知,分子在 dV 内的累积速率等于分子的入射通量与出射通量之差,即

$$\text{累积速率} = eS - e'S = - (e' - e)S =$$

$-(de)S = -(Le/Lx)Sdx$
式中, $de=e'-e$ 是该体积元两端信息流密度之差。

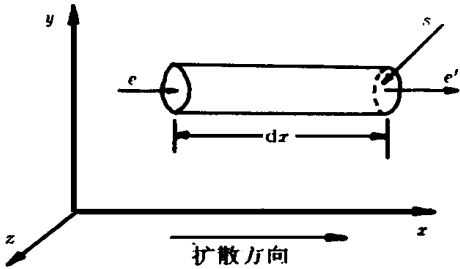


图 1 通过一体积元的分子扩散

Fig. 1 Molecule Spread Through a Volume Element

因为累积速度等于单位时间、单位体积的粒子数的增量 Ld/Lt 与体积的乘积,于是又有:

$$\text{累积速度} = (Ld/Lt)Sdx \tag{4}$$

由式 (3) 和式 (4) 可得:

$$Ld/Lt = -(Le/Lx) \tag{5}$$

顾及式 (2),得

$$Ld/Lt = D(L^2d/Lx^2) \tag{6}$$

由于信息扩散是在封闭系统中进行的,系统内外没有信息交换,信息量的总值在系统内扩散既不会增加,也不会减少,具有守恒性,所以分子扩散可作为信息扩散的一种近似模型。于是,信息扩散函数 $_ (x,t)$ 应近似满足下列偏微分方程:

$$L_ (x,t)/Lt = D(L^2_ (x,t)/Lx^2) \tag{7}$$

利用边界条件,在信息注入点 $x=0$,当 $t=0$ 时有:

$$_ (0,0) = 1$$

而当 $t=0$ 时,在信息注入点的周围点 $x\neq 0$,有:

$$_ (x,0) = 0$$

可解偏微分方程式 (7),对式 (7) 两边进行傅立叶变换^[4],得:

$$\begin{cases} d_ /dt + D\lambda^2_ = 0 \\ _ (\lambda,0) = 1 \end{cases} \tag{8}$$

式中, $_ (\lambda,t)=F[_ (x,t)]$ 为对 $_ (x,t)$ 进行傅立叶变换得到的函数。用变量分离法易得:

$$_ (\lambda,t) = e^{-D\lambda^2t} \tag{9}$$

于是:

$$\begin{aligned} _ (x,t) &= F^{-1}[_ (\lambda,t)] = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} _ (\lambda,t) e^{-i\lambda x} d\lambda = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-D\lambda^2t} (\cos\lambda x - i\sin\lambda x) d\lambda = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-D\lambda^2t} \cos\lambda x d\lambda = \end{aligned}$$

$$\frac{1}{2} \frac{1}{D\pi t} \exp\left[-\frac{x^2}{4Dt}\right]$$

令 $e(t)=\frac{1}{2Dt}$,则有:

$$e(x,t)=\frac{1}{2}e(t)\exp\left[-\frac{x^2}{2e^2(t)}\right] \tag{10}$$

由于信息扩散过程是一个抽象的时间过程,可假定这一过程是在一段不太长的时间内完成的。记完成时刻为 t_e ,并记 $e_e=e(t_e)$,则在完成时刻,与信息注入点相距为 x_i 的点所获得的扩散信息为:

$$e(x)=\frac{1}{e_e}\exp\left[-\frac{x^2}{2e_e^2}\right] \tag{11}$$

至此,就得到了信息扩散函数 $e(x)$ 。将式(11)代入式(1),得:

$$\hat{f}(l)=\frac{1}{n\Delta_n}\sum_{j=1}^n\left\{\frac{1}{e_e}\exp\left[-\frac{\left[\frac{l-l_j}{\Delta_n}\right]^2}{2e_e^2}\right]\right\}=\\ \frac{1}{nh}\sum_{j=1}^n\exp\left[-\frac{(l-l_j)^2}{2h^2}\right] \tag{12}$$

式中, $h=e_e\Delta_n$ (13)

由于式(11)所确定的扩散函数与数理统计中的正态分布的密度函数完全一样,所以称由式(12)所得的母体概率密度函数的估计 $\hat{f}(l)$ 为正态扩散估计,并称式(13)所确定的 h 为标准正态扩散的窗宽

1.4 标准正态扩散估计的窗宽 h

由式(12)知,母体概率密度函数的正态扩散估计 $\hat{f}(l)$ 除与观测值 l_k 、子样容量 n (观测值的个数)有关外,还与标准正态扩散的窗宽 h 有关。当观测完成后, l_j 和 n 都是已知量,只有窗宽 h 未知。因此,要根据观测值来估计母体的概率密度函数,首先要确定标准正态扩散的窗宽 h 。为此,根据正态扩散的两点择近原则,给出了 h 的表达式^[1]: 设 $l\in L$,如果存在 $w',w''\in W$,使 $l'<l<l''$,且对任何一个 $w\in W$,如果 $w\neq w',w''$,则必有 $l\notin[l',l'']$ 。于是称 l',l'' 为 l 的最邻近信息注入点。此时窗宽 h 应满足的两点择近原则为:

$$\exp\left[-\frac{(l-l')^2}{2h^2}\right]+\exp\left[-\frac{(l-l'')^2}{2h^2}\right]\geqslant\\ \sum_{l\neq l',l''}\exp\left[-\frac{(l-l_j)^2}{2h^2}\right],l\in L \tag{14}$$

因直接从式(14)解算 h 比较困难,因此,顾及 W 的随机性,采用平均距离来代替式(14)中的有关量。设 $l_j(j=1,2,\cdots,n)$ 已排序,并记 $d=l_{j+1}-l_j(j=1,2,\cdots,n-1)$,则平均距离为:

$$d=\frac{1}{n-1}\sum_{j=1}^{n-1}d_j=\frac{b-a}{n-1} \tag{15}$$

式中, $a=\min(l_j)$, $b=\max(l_j)$, $j=1,2,\cdots,n$
为了求得在最不利的情况下也能使式(14)成立的 h ,设 $h=2k+3$,且当 $l=l_{k+3}$ 时,有:

$$1+\exp\left[-\frac{d^2}{2h^2}\right]\geqslant\exp\left[-\frac{d^2}{2h^2}\right]+\\ \sum_{i=1}^k\exp\left[-\frac{(i+1)^2d^2}{2h^2}\right] \tag{16}$$

令 $y=\exp(-d^2/2h^2)$ (17)
则有:

$$y^4+y^9+y^{16}+\cdots+y^{(k+1)^2}-\frac{1}{2}\geqslant 0 \tag{18}$$

再令 $j(y)=y^4+y^9+y^{16}+\cdots+y^{(k+1)^2}-1/2$,则由求解非线性方程的牛顿法

$$y_{k+1}=y_k-j(y_k)/j'(y_k) \tag{19}$$

可求得式(18)的近似根 y_k 。然后,由式(17)并顾及式(15),可得:

$$h=T_d=T(b-a)/(n-1) \tag{20}$$

式中, $T=1/\sqrt{2(-\ln y)}$ (21)

表1列出了 n 为 3~21 时的 T 值。由表1知,当 $n\geqslant 17$ 时, $T\approx 1.420\ 693\ 101$ 。得到 T 后,按式(20)就能很方便地求得标准正态扩散的窗宽 h 。将 h 代入式(12),就可得到母体概率密度函数的估计 $\hat{f}(x)$ 。

表 1 n 取不同值时的 T

Tab. 1 The Value of T About Different n

k	n	y	T
0	3	0.500 000 000	0.849 321 800
0.5	4	0.734 867 275	1.273 982 782
1	5	0.840 896 428	1.698 643 675
1.5	6	0.755 766 749	1.336 252 561
2	7	0.787 172 437	1.445 461 208
2.5	8	0.773 473 322	1.395 189 816
3	9	0.781 191 230	1.422 962 345
3.5	10	0.779 368 401	1.416 278 786
4	11	0.780 613 482	1.420 835 443
4.5	12	0.780 459 404	1.420 269 570
5	13	0.780 576 289	1.420 698 795
5.5	14	0.780 568 361	1.420 669 671
6	15	0.780 574 799	1.420 693 321
6.5	16	0.780 574 501	1.420 692 226
7	17	0.780 574 739	1.420 693 101
7.5	18	0.780 754 739	1.420 693 101
8	19	0.780 574 739	1.420 693 101
8.5	20	0.780 574 739	1.420 693 101
9	21	0.780 574 739	1.420 693 101

2 参数估计方法

由于母体 K 的概率密度函数已根据 n 个子样观测值 l 估计出来了,其表达式为式(12),所以采用极大似然估计可得母体期望。但由于式(12)

为一幂指数多项式,当 n 较大时,似然函数很复杂,所以基于式 (12)的极大似然估计并不可取。为此,应考虑其他方法。限于篇幅,本文仅介绍一维的情况。对于多维情况下的参数估计方法及信息扩散估计的性质等问题有待进一步研究。

2.1 数值方法

对于一维的情况,由数学期望的物理解释^[5]知,随机变量(观测值) l 的数学期望 $E(l)$ 表示质量中心的坐标。因此有:

$$\int_{-\infty}^{\infty} \hat{f}(l) dl = \int_{-\infty}^{\infty} f(l) dl = \frac{1}{2} \tag{22}$$

式中, $\int_{-\infty}^{\infty}$ 为随机变量 l 的数学期望。将式 (12)代入式 (22),得:

$$\int_{-\infty}^{\infty} \sum_{j=1}^n \exp\left[-\frac{(l-l_j)^2}{2h^2}\right] dl = \int_{-\infty}^{\infty} \sum_{j=1}^n \exp\left[-\frac{(l-l_j)^2}{2h^2}\right] dl = \frac{1}{2} \tag{23}$$

有了观测值 l_j 后,采用数值积分的方法,很容易得到 $\int_{-\infty}^{\infty}$ 的估值 $\hat{\int_{-\infty}^{\infty}}$ 。

2.2 带权平均值法

由数学期望的定义,对于连续型随机变量有:

$$\int_{-\infty}^{\infty} x f(x) dx = \int_{-\infty}^{\infty} x f(x) dx \tag{24}$$

所以其估值为:

$$\hat{\int_{-\infty}^{\infty}} = \int_{-\infty}^{\infty} l \hat{f}(l) dl = \int_{-\infty}^{\infty} l \hat{f}(l) dl \int_{-\infty}^{\infty} \hat{f}(l) dl \tag{25}$$

由于观测值是离散型随机变量,故应按离散型随机变量的数学期望公式计算。令

$$P_i = \hat{f}(l_i) \sum_{i=1}^n \hat{f}(l_i) \tag{26}$$

则由文献 [5]知, $\hat{f}(l) dl$ 的作用与离散型随机变量中的 P_i 相类似。因此,将式 (25)中的积分写成求和,并将 P_i 代替 $\hat{f}(l) dl$,得:

$$\hat{\int_{-\infty}^{\infty}} = \sum_{i=1}^n l_i P_i \sum_{i=1}^n P_i \tag{27}$$

式 (27)是带权平均值公式,由于 $\sum_{i=1}^n P_i = 1$,所以式 (27) 可以简写为:

$$\hat{u} = \sum_{i=1}^n l_i P_i \tag{28}$$

式 (27)与一般的带权平均值的差别仅在于权的确定方法不一样。式 (27)中的权是根据扩散估计得到的概率密度函数所计算的各观测值的概率。以信息扩散原理为基础的估计为信息扩散估计,简记为 ISE。

3 实例

设有来自对数正态分布

$$f(x) = \begin{cases} \frac{1}{0.7x^2} \exp\left[-\frac{(\ln x - 0.02)^2}{0.98}\right], & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{29}$$

的一组子样观测值 l ,见表 2,试用 ISE估计母体的数学期望。

表 2 观测值及其权					
Tab. 2 The Observations and Weights					
i	l_i	P_i	i	l_i	P_i
1	1.872	0.059	11	2.223	0.041
2	3.180	0.011	12	0.179	0.043
3	1.641	0.063	13	1.802	0.061
4	5.981	0.009	14	0.772	0.066
5	1.060	0.064	15	4.100	0.011
6	0.574	0.063	16	4.994	0.010
7	1.026	0.065	17	1.843	0.072
8	0.300	0.050	18	0.695	0.066
9	1.902	0.058	19	1.629	0.063
10	0.876	0.066	20	0.459	0.059

由文献 [6]知,对数正态分布

$$f(x) = \begin{cases} \frac{1}{x^2} \exp\left[-\frac{(\ln x - 0.02)^2}{2 \times 0.98}\right], & x > 0 \\ 0, & x \leq 0 \end{cases}$$

的数学期望为 $E(x) = \exp\{(1/2) \times 0.98 + 0.02\}$,于是,式 (29)所确定的对数正态分布的数学期望为:

$$E(x) = \exp(0.49/2 + 0.02) = 1.3034$$

按 ISE估计,具体步骤如下:

- 1) 根据子样容量 $n = 20$,查表 1得 $T = 1.420693101$
- 2) 计算标准正态扩散的窗宽 h 。由表 2知,最大观测值为 $b = 5.981$,最小观测值为 $a = 0.179$ 。由式 (20)得:
 $h = 1.420693101(5.981 - 0.179)/19 = 0.433835$
- 3) 按式 (12)写出母体概率密度函数的扩散估计:
 $\hat{f}(l) = 0.045977 \cdot \sum_{j=1}^{20} \exp[-2.656570(l - l_j)^2]$ (30)
- 4) 按式 (27)求各观测值的权 $P_i = \hat{f}(l_i)$,见表 2
- 5) 由式 (28)得母体期望的 ISE估计为:
 $\hat{\int_{-\infty}^{\infty}} = \sum_{i=1}^{20} l_i P_i = 1.3195$

为了便于比较,用最小二乘估计计算母体期望的
估值为 $\hat{\mu}_{LSE} = \sum_{i=1}^{20} l_i / 20 = 1.8554$

两种估计的真误差为:

$$\begin{aligned}\Delta_{ISE} &= \hat{\mu}_{ISE} - \mu = 0.0161 \\ \Delta_{LSE} &= \hat{\mu}_{LSE} - \mu = 0.5520 \\ \Delta_{LSE} / \Delta_{ISE} &= 34.2857\end{aligned}$$

可见,信息扩散估计的真误差远小于最小二乘估计的真误差

4 信息扩散估计的抗差性

假设观测值中存在 3 个粗差: $l_1 = 18.720, l_{19} = 16.290, l_{20} = 10.459$,其余 17 个观测值同表 2 所列.按信息扩散估计得:

$$\begin{aligned}h &= 1.420693101(18.720 - 0.179) / 19 = \\ &1.386372 \\ \hat{f}(l) &= 0.014388 \cdot \\ &\sum_{j=1}^{20} \exp[-0.260142(l - l_j)^2] \\ \hat{\mu}_{ISE} &= \sum_{i=1}^{20} l_i P_i = 1.7658\end{aligned}$$

此时的最小二乘估计为:

$$\hat{\mu}_{LSE} = \sum_{i=1}^{20} l_i / 20 = 3.9309$$

由以上计算可见,当观测值中存在如此之大的多个粗差时,信息扩散估计结果仍然较好,而最小二乘估计严重偏离真值,这说明信息扩散估计具有很强的抗差能力.在观测数据受粗差污染时,

信息扩散估计优于最小二乘估计.这是因为,当观测值中出现粗差时,这组观测值不再服从原来的分布.信息扩散估计在参数估计之前先估计出了这种分布,这样就考虑了粗差信息.更主要的原因是,采用了观测值 l_i 的概率作为权.由式(12)知,权 $P_i = \hat{f}(l_i) = C_2 \sum_{j=1}^n \exp(-C_2(l_i - l_j)^2)$.当 l_i 为粗差时, $l_i - l_j (i \neq j)$ 必然较大,因此 $\exp \sum_{j=1}^n [-C_2(l_i - l_j)^2]$ 必然较小,从而导致 P_i 较小.这样就抑制了粗差的影响,使之具有抗差性.另外,因为 $0 < C_2 \leq P_i$,所以 P_i 恒大于 0.这表明,即使 l_i 为粗差, l_i 中也会含有一定量的有用信息. P_i 不等于 0,就意味着可从粗差观测值 l_i 中提取有用信息.

参 考 文 献

1 黄崇福,王家鼎.模糊信息优化处理技术及其应用.北京:北京航空航天大学出版社,1995.42~64

2 李镇敌.普通物理学.北京:清华大学出版社,1993.406~411

3 苏煜城.数学物理方程.北京:气象出版社,1998.49~60

4 董加礼,曹旭东,史明仁.数学模型.北京:北京工业大学出版社,1990.57~60

5 浙江大学数学系高等数学教研组.概率论与数理统计.北京:人民教育出版社,1979.101~111

6 现代工程数学手册编委会.现代工程数学手册.武汉:华中理工大学出版社,1987.80~81

The Theory, Method and Robustness of the Parameter Estimation
Based on the Principle of Information Spread

Wang Xinzhou

(Projects and Consulting Office, W TUSM, 129 Luoyu Road, Wuhan, China, 430079)

Abstract It is supposed that the observations obey a certain distribution in all the methods of parameter estimation to this day. But in the practical surveying, due to the different observation conditions, the observations obtained under different observation conditions may obey different distributions. So the supposition is not conform to reality. The idea that conform to reality should be that the distribution of the observation is estimated based on the observations before estimated parameter. According to the idea, this paper introduces the theory of information spread at first. Then the method of information spread estimation (ISE) has been presented. The theory and calcultion indicate that ISE not only has fine robustness, but also it is very simple and convenient for application.

Key words principle of information spread; parameter estimation; density function