

文章编号: 1000-7598 (2010) 09-2983-04

基于 K-L 信息距离的多源信息融合法

谢桂华^{1,2}, 张家生²

(1. 江苏大学 理学院, 江苏 镇江 212013; 2. 中南大学 土木建筑学院, 长沙 410075)

摘 要: 为解决小子样条件下岩土参数概率分布推断的难题, 并克服基于专家信息的融合法不可避免地带有主观随意性的弊端, 引入信息论中 K-L 信息距离的概念, 基于先验信息可信度, 提出一种新的多源信息融合方法。利用 K-L 信息距离作为参数分布之间距离的度量, 定义先验分布差异率, 确定融合权重, 进而根据 Bayes 原理得到后验分布, 优化岩土参数分布概率型。工程实例分析结果表明, 该法计算简单, 且克服了推断过程中的主观随意性。计算结果显示该法所得融合分布的方差比已有成果所得方差偏小, 说明该法可实现统计意义上的参数概率型优化, 为岩土参数设计值的合理选取提供了参考。

关 键 词: K-L 信息距离; 多源信息; 融合权重; Bayes 原理

中图分类号: TU 452

文献标识码: A

Multi-source information fusion based on K-L information distance

XIE Gui-hua^{1,2}, ZHANG Jia-sheng²

(1. Faculty of Science, Jiangsu University, Zhenjiang, Jiangsu 212013, China;

2. School of Civil and Architectural Engineering, Central South University, Changsha 410075, China)

Abstract: In order to solve the problem of geotechnical parameters distribution inferring on the condition of small sample and get over the inevitable shortcoming of subjective randomness in those fusion methods based on expert's experience, the notion of K-L information distance in the information field was introduced; and multi-source information fusion method was proposed on the basis of the credibility of prior distributions. Using K-L information distance as a measurement of distances between parameter distributions, "the rate of prior distribution difference" was defined and the fusion weight was determined. Further, the posterior distribution was obtained by using the Bayes principle; and the fitting probability models of geotechnical parameters were optimized. It is shown in the process of a project case that the method suggested is simple while no subjective factors are included in the statistical inference. The results show that the variance of the suggested fusion distribution is smaller than ones of the existing results, which illuminates that an optimum fitting probabilistic model for parameters in statistical sense can be resulted from the proposed method; and it provides a reference for choosing design values of geotechnical parameters reasonably.

Key words: K-L information distance; multi-source information; fusion weight; Bayes principle

1 引 言

在岩土参数统计性质的研究中常常因研究经费和工程采样场地等的限制, 导致试验样本量非常有限, 无法用传统的统计方法确定其概率分布。为解决上述问题, 许多学者深入研究了岩土参数的小样本统计方法^[1-3], 引入 Bayes 原理, 利用已有类似工程的数据获得参数的先验分布, 再在此基础上运用 Bayes 公式分析岩土参数的后验分布概率型和统计量。以上方法架设了先验信息与现场试验信息的桥梁, 为小样本的统计分析和参数优化提供了一条有

效的途径。然而, 这些方法都是针对仅有一种先验分布的情况做出的推断。实际岩土工程中, 先验信息可能是通过不同途径获得的, 由此导致总体分布中的未知参数的先验分布可能具有多个不同的分布形式。为运用 Bayes 公式得到后验分布, 需要解决将多源信息融合为一个综合先验分布的问题^[4]。为此, 张伟等^[5]提出了专家打分的信息融合法、宫凤强等^[6]提出了基于 AHP 的先验分布融合法。专家打分法由专家根据个人经验确定融合权重, 方法简单, 适用面较广, 但结果带有一定的主观性和随意性, 其最终权重很大程度上取决于评价专家的知识

收稿日期: 2009-01-05

第一作者简介: 谢桂华, 女, 1976 年生, 博士, 讲师, 从事岩土工程可靠性方面研究。E-mail: jg97xieguihua@163.com

和经验。基于 AHP 的融合法将复杂问题进行分解, 通过两两对比矩阵计算各因素影响权重, 在一定程度上克服了专家打分法的随意性, 但每层均需进行一致性检验和修正, 涉及因素多, 处理过程较复杂; 文献[7]根据“样本空间距离”来确定融合权重, 具有客观性, 但计算复杂且只适用于正态分布的情况。本文基于先验分布的可信性来确定融合权重, 借助信息理论, 以 K-L 信息距离来量化各先验分布的可信性, 定义分布差异率, 进而计算各先验分布的融合权重, 得到融合分布和后验分布。该法计算简单、客观性强, 适用于各种分布类型。

2 先验分布的融合方法

设有 m 个先验信息源 $x_1^{(i)}, x_2^{(i)}, \dots, x_m^{(i)}, i=1, 2, \dots, m$, 其中 m 为验前总体数目; n_i 为第 i 个信息源的样本数。若 θ 为参数空间 $\Theta = \{\theta\}$ 的一个未知参数, $\pi_i(\theta)$ 为第 i 个信息源的先验分布, $\pi(\theta)$ 为融合先验分布, 融合权重为 ω_i , 则融合后的分布可表示为^[8]

$$\pi(\theta) = \sum_{i=1}^m \omega_i \pi_i(\theta); \quad \sum_{i=1}^m \omega_i = 1 \quad (1)$$

设 $x = (x_1, x_2, \dots, x_n)$ 为总体 X 的一个现场样本, 验后分布为 $h(\theta|x)$, 根据 Bayes 公式^[9]:

$$h(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta} = \frac{f(x|\theta)\sum_{i=1}^m \omega_i \pi_i(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta} \quad (2)$$

设任意信息源 i 在取得试验样本 x 后的验后分布为 $\pi_i(\theta|x)$, 同样, 由 Bayes 公式可得

$$\pi_i(\theta|x) = \frac{\pi_i(\theta)f(x|\theta)}{\int_{\Theta_i} \pi_i(\theta)f(x|\theta)d\theta} \quad (3)$$

代入式(2)可得

$$h(\theta|x) = \frac{\sum_{i=1}^m \left(\omega_i \pi_i(\theta|x) \int_{\Theta_i} \pi_i(\theta)f(x|\theta)d\theta \right)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta} \quad (4)$$

令

$$\lambda_i = \frac{\int_{\Theta_i} \pi_i(\theta)f(x|\theta)d\theta}{\int_{\Theta} \pi(\theta)f(x|\theta)d\theta}, \quad (i=1, 2, \dots, m) \quad (5)$$

则

$$h(\theta|X) = \sum_{i=1}^m \omega_i \lambda_i \pi_i(\theta|X) = \sum_{i=1}^m \alpha_i \pi_i(\theta|X) \quad (6)$$

式中: $\alpha_i = \lambda_i \omega_i, i=1, 2, \dots, m$ 。

这表明, 多源信息的融合后验分布为融合前的

各后验分布的加权和, 其权数大小决定于该先验分布的融合权值 ω_i 及该分布的边际密度占整个边际密度的比例 λ_i , 可见多源信息融合的首要工作是合理确定各先验分布的权重 ω_i , 它是 Bayes 公式应用中的一个重要方面。

3 基于 K-L 信息距离确定融合权重

信息论在专家决策中有着广泛的应用, 基于 K-L 信息距离融合先验分布也有一些研究成果^[10], 但已有研究成果主要是针对先验分布的融合权重已知, 求融合分布形式的情况, 鲜有涉及融合权重的确定问题。为确定多源信息的融合权重, 本文研究采用 K-L 信息距离来量化先验分布的可信性, 再据此对先验分布进行融合。

3.1 K-L 信息距离

K-L 信息距离 (Kullback-Leibler 信息距离, 或简称为 Kullback 信息) 的概念是统计学家 Kullback 和 Leibler 于 20 世纪中叶提出来的, 用以判断两个概率分布的接近程度。

定义: 度量空间 (Ω, F) 上的分布 P, Q 之间的 K-L 信息距离定义为^[10]

$$I(Q:P) = \begin{cases} E_Q \left[\ln \left(\frac{dQ}{dP} \right) \right], & (Q \ll P) \\ \infty, & \text{(其他)} \end{cases} \quad (7)$$

式中: Ω 为空间; F 为 Ω 上的 σ 代数; (Ω, F) 为度量空间; P, Q 为 Ω 上的概率分布; $E_Q(\square)$ 为相对于 Q 的数学期望。 dQ/dP 为 Q 对 P 的 Radon-Nikodym 导数; $Q \ll P$ 为 Q 关于 P 绝对连续。

若 Ω 为实数空间, 则 P 和 Q 具有如下形式:

$$P(dx) = p(x)dx; \quad Q(dx) = q(x)dx \quad (8)$$

K-L 信息距离的定义等价于

$$I(Q:P) = I(q:p) = \int_{\Omega} q(x) \ln \left[\frac{q(x)}{p(x)} \right] dx \quad (9)$$

若 Ω 为有限空间, 即 $\Omega = \{x_1, x_2, \dots, x_m\}$, 满足 $P(\{x_i\}) = p_i, Q(\{x_i\}) = q_i, (i=1, 2, \dots, m)$, 则 P 对 Q 的 K-L 信息距离为

$$I(Q:P) = \sum_{i=1}^m q_i \ln \left(\frac{q_i}{p_i} \right) \quad (10)$$

由定义可知, $I(Q:P) \geq 0$, 且仅当 $P=Q$ 时 $I(Q:P)=0$ 。设 $\pi(x), g(x)$ 为先验分布和融合分布的概率密度函数, 则

$$I(\pi : g) = \int \pi(x) \ln \frac{\pi(x)}{g(x)} dx \quad (11)$$

为 $g(x)$ 到 $\pi(x)$ 的 K-L 信息距离。当 $\pi(x)$ 与 $g(x)$ 为离散随机变量的概率密度时,

$$I(\pi : g) = \sum_{i=1}^n \pi(\theta_i) \ln \frac{\pi(\theta_i)}{g(\theta_i)} \quad (12)$$

K-L 信息距离越大, 则两个概率密度函数的差别也越大; 反之, 其差别越小。两者一致时, $I=0$ 。因此, 可以采用 K-L 信息距离来定量描述融合分布与先验分布之间的差异程度。

3.2 融合权重的确定

设先验分布为 $\pi = \{\pi_1, \pi_2, \dots, \pi_n\}$, $I_i(f, \pi_i)$ 为子样分布 f 与第 i 个先验分布 π_i 的 K-L 信息距离, 为根据先验分布的可信性确定融合权重, 定义分布差异率 ζ_i 为

$$\zeta_i = \frac{I_i(f, \pi_i)}{\sum_{j=1}^n I_j(f, \pi_j)} \in [0, 1] \quad (13)$$

差异率 ζ_i 表示第 i 个先验分布 π_i 与子样分布 f 的差异在总差异中的构成比重, 表征了该分布对造成先验分布与子样分布之间差异的“贡献”大小。故各分布的融合权重可采用下式计算:

$$\omega_i = \frac{1 - \zeta_i}{\sum_{i=1}^n (1 - \zeta_i)} = \frac{1 - \zeta_i}{n - 1} \quad (14)$$

式中: ζ_i 为第 i 个先验分布 π_i 的差异率; n 为先验分布的个数。

然后再通过式 (1) 计算融合先验分布。

4 实例分析

山东泰莱盆地粉质黏土分布较广, 厚度相对稳定。该粉质黏土成因类型较多, 其中冲洪积成因的粉质黏土常用作工业民用建筑的基础持力层。为获得该土的抗剪强度, 需对其黏聚力进行统计分析。根据多年的工程设计资料, 共收集了 207 个粉质黏土黏聚力的原始数据, 又在该地区一建筑小区采集到粉质黏土黏聚力的数据 33 个, 数据见文献[3]。

由于原始累计数据的黏聚力离散较大, 按照大小排序, 分为 3 组 (具体分组情况可参考文献[5]), 作为先验分布的 3 个信息源; 来自小区现场的 33 个试验数据作为试验样本。

由 K-S 检验法, 得 3 组先验分布的数据均不拒绝正态分布: $\pi_1 = N_1(44.47, 16.62^2)$; $\pi_2 = N_2(44.54, 16.59^2)$; $\pi_3 = N_3(44.53, 16.50^2)$, 现场样本也符合

正态分布, 其分布函数 $F = N(46.32, 12.73^2)$ 。由于地质成因相同, 可以认为 4 组数据来自同一总体分布。根据式 (13)、(14), 计算各先验分布的融合权重, 计算过程量及结果见表 1。

表 1 先验分布权重计算表
Table 1 Weights of prior distributions

项目	$I_i(f : \pi_i)$	ζ_i	ω_i
π_1	0.038 3	0.340 6	0.329 7
π_2	0.037 7	0.334 8	0.332 6
π_3	0.036 5	0.324 6	0.337 7

注: 表中 $\int_a^b f(x) \ln \frac{f(x)}{\pi_i(x)} dx$; $\frac{I_i(f, \pi_i)}{\sum_{i=1}^n I_i(f, \pi_i)}$; $\frac{1 - \zeta_i}{n - 1}$ 。

由表 1 及式 (1) 得基于信息可信度的融合验前分布为

$$\pi = 0.329 7 \pi_1 + 0.332 6 \pi_2 + 0.337 7 \pi_3 \quad (15)$$

则融合先验分布的平均值和方差分别为 $\mu_0 =$

$$E(\omega_1 \pi_1 + \omega_2 \pi_2 + \omega_3 \pi_3) = \sum_{i=1}^3 \omega_i \mu_i = 44.513 5, \quad \sigma_0^2 =$$

$$D\left(\sum_{i=1}^3 \omega_i \pi_i\right) = \sum_{i=1}^3 \omega_i^2 D(\pi_i) = \sum_{i=1}^3 \omega_i^2 \sigma_i^2 = 9.566 6^2。$$

为进一步说明本文方法的优劣, 将上述结果代入 Bayes 推广公式^[3]以获得后验分布的均值和均方差:

$$\left. \begin{aligned} \mu'' &= \frac{u^*(\sigma')^2 + u'(\sigma^*)^2}{(\sigma')^2 + (\sigma^*)^2} \\ \sigma'' &= \frac{(\sigma'\sigma^*)^2}{(\sigma')^2 + (\sigma^*)^2} \end{aligned} \right\} \quad (16)$$

式中: μ' 、 σ' 分别为先验分布的均值和均方差; μ'' 、 σ'' 分别为后验分布的均值和均方差; μ^* 、 σ^* 为似然函数的均值和均方差。可得后验分布的均值和均方差为 $\mu'' = 45.209 5$, $\sigma'' = 7.647 8$ 。

比较其他文献研究结果, 文献[6]中验后分布参数 $\mu'' = 45.650$, $\sigma'' = 10.094$; 文献[3]中验后分布参数 $\mu'' = 45.643$, $\sigma'' = 10.077$, 可见本文方法所得方差最小, 优化了岩土工程设计参数。

5 结 语

利用历史资料并结合现场试验数据来推断工程参数具有重要的意义: 一方面可以在保证统计推断质量的前提下减少试验次数, 节约试验费用, 降低工程成本。另一方面, 为解决某些试验数据难以取得、参数分布推断困难的特殊工程问题提供了途径, 使合理地确定该类参数的设计值成为可能。

本文将信息论中“K-L 信息距离”的概念引入岩土参数推断中, 基于先验分布的可信度提出了多源信息融合方法。该法克服了基于专家经验融合法的主观随意性, 计算简单, 是一种客观的赋权方法。通过工程实例分析及结果比较可以发现, 该法所得参数方差较小, 较好地优化了岩土参数的概率模型, 为岩土工程可靠性分析和工程决策提供了条件。

参考文献

- [1] 张广文, 刘令瑶. 确定随机变量概率分布参数的推广 Bayes 法[J]. 岩土工程学报, 1995, 17(3): 91—94.
- [2] 徐超, 杨林德. 随机变量拟合优度检验和分布参数 Bayes 估计[J]. 同济大学学报, 1998, 26(3): 340—344.
XU Chao, YANG Lin-de. Test of goodness of fit of random variables and Bayesian estimation on distribution parameters[J]. **Journal of Tongji University**, 1998, 26(3): 340—344.
- [3] 王俊杰, 陈爱玖, 姬凤玲, 等. 岩土参数的概率分布拟合及 Bayes 方法优化[J]. 华北水利水电学院学报, 2004, 25(2): 51—54.
WANG Jun-jie, CHEN Ai-jiu, JI Feng-ling, et al. Probability distribution fitting and Bayes method optimization on geotechnical parameters[J]. **Journal of North China Institute of Water Conservancy and Hydroelectric Power**, 2004, 25(2): 51—54.
- [4] JAMES O B. Statistical decision theory and Bayesian analysis(second edition)[M]. New York: Springer-Verlag, Inc.1985.
- [5] 张伟, 李夕兵, 宫凤强. 基于专家信息融合法的岩土参数概率分布推断[J]. 地下空间与工程学报, 2005, 1(6): 929—932.
ZHANG Wei, LI Xi-bing, GONG Feng-qiang. Probabilistic distribution of rock & soil parameters by using expert's information fusion method[J]. **Chinese Journal of Underground Space and Engineering**, 2005, 1(6): 929—932.
- [6] 宫凤强, 李夕兵, 邓建. 基于 AHP 先验分布融合法的岩土参数概率分布推断[J]. 岩土工程学报, 2006, 28(2): 1311—1316.
GONG Feng-qiang, LI Xi-bing, DENG Jian. Probabilistic distribution of geotechnical parameters by using AHP prior distribution fusion method[J]. **Chinese Journal of Geotechnical Engineering**, 2006, 28(2): 1311—1316.
- [7] 宫凤强, 李夕兵, 邓建. 基于多源信息融合法的岩土力学参数概率分布推断[J]. 岩土力学, 2007, 28(3): 599—604.
GONG Feng-qiang, LI Xi-bing, DENG Jian. Inferring probabilistic distribution of rock&soil mechanical parameters by using multiple sources information fusion method[J]. **Rock and Soil Mechanics**, 2007, 28(3): 599—604.
- [8] 冯静. 小子样复杂系统可靠性信息融合方法与应用研究[博士学位论文 D]. 长沙: 国防科学技术大学, 2004.
- [9] 张尧庭, 陈汉峰. 贝叶斯统计推断[M]. 北京: 中国科学出版社, 1991: 18—24.
- [10] 任开军, 吴孟达, 刘琦. 先验分布融合的 Kullback 信息方法[J]. 装备指挥技术学院学报, 2002, 13(4): 90—92.
REN Kai-jun, WU Meng-da, LIU Qi. The combination of prior distributions based on Kullback information[J]. **Journal of the Academy of Equipment Command & Technology**, 2002, 13(4): 90—92.
- [10] 陈立宏, 陈祖煜. 堆石非线性强度特性对高土石坝稳定性的影响[J]. 岩土力学, 2007, 28(9): 1807—1810.
CHEN Li-hong, CHEN Zhu-yu. Effect of nonlinear strength of rockfill on slope stability of high earth-rock dam[J]. **Rock and Soil Mechanics**, 2007, 28(9): 1807—1810.
- [11] 胡卫东, 张国祥. 非线性破坏准则下的边坡稳定塑性极限分析[J]. 岩土力学, 2007, 28(9): 1910—1913.
HU Wei-dong, ZHANG Guo-xiang. Plastic limit analysis of slope stability with nonlinear failure criterion[J]. **Rock and Soil Mechanics**, 2007, 28(9): 1910—1913.
- [12] JANBU N. Soil compressibility as determined by Oedometer and Triaxial Tests[C]//Proc., European Conf. on Soil Mechanics and Foundation Engineering. [S. l.]: [s. n.], 1963, 1: 19—25.
- [13] BARDEN L, KHAYATT A J. Incremental strain rate ratios and strength of sand in the Triaxial Test[J]. **Geotechnique**, 1966, 16(4): 338—357.
- [14] GOODWIN AK, O'NEILI MA, ANDERSON WF. The use of X-ray computer tomography to investigate particulate interactions within opencast coal mine backfills[J]. **Engineering Geology**, 2003, 70: 331—341.
- [15] INDRARATNA B, LACKENBY J, CHRISTIE D. Effect of confining pressure on the degradation of ballast under cyclic loading[J]. **Geotechnique**, 2005, 55(4): 325—328.
- [16] BISHOP A W. Sixth Rankine lecture: the strength of soil as engineering material[J]. **Geotechnique**, 1966, 16(2): 89—130.
- [17] VARADARAJAN A, SHARMA K G, ABBAS S M, et al. Constitutive model for rockfill materials and determination[J]. **International Journal of Geomechanics**, 2006, 4: 226—237.
- [18] Itasca Consulting Group. FLAC2D (Version5.0) User's guide[M]. [S. l.]: Itasca Consulting Group, 2005.

上接第 2972 页