

基于第二类统计量的 K 分布参数估计*

孙增国

(华侨大学 计算机科学与技术学院, 福建 厦门 361021)

摘要: 为了高效估计出 K 分布的参数, 提出了对数累积量参数估计方法。基于第二类统计量, 先对 K 分布的概率密度函数进行 Mellin 变换, 从而获得 K 分布的第二类第一特征函数; 然后对第二类第一特征函数进行对数变换, 由此获得 K 分布的第二类第二特征函数; 最后对第二类第二特征函数求导数, 进而获得 K 分布的前两阶对数累积量, 由此可以估计 K 分布的参数。与传统的最大似然估计方法相比, K 分布的对数累积量估计具有解析的表达式, 易于计算。Monte Carlo 仿真表明, 基于第二类统计量的 K 分布对数累积量估计可获得较高的估计精度。

关键词: K 分布; Mellin 变换; 第二类统计量; 对数累积量估计; Monte Carlo 仿真

中图分类号: TP391.9 **文献标志码:** A **文章编号:** 1001-3695(2013)01-0046-03

doi:10.3969/j.issn.1001-3695.2013.01.010

Parameter estimation of K distribution based on second-kind statistics

SUN Zeng-guo

(College of Computer Science & Technology, Huaqiao University, Xiamen Fujian 361021, China)

Abstract: In order to efficiently estimate the parameters of K distribution, this paper proposed the log-cumulant estimator. Based on second-kind statistics, first it derived the second-kind first characteristic function of K distribution by taking the Mellin transformation to the probability density function of K distribution, second obtained the second-kind second characteristic function of K distribution from the logarithmic transformation of the second-kind first characteristic function, and last deduced the first two log-cumulants to estimate the parameters of K distribution by taking the derivative of the second-kind second characteristic function. Compared to the traditional maximum likelihood estimator, the log-cumulant estimator of K distribution with analytical expressions was easy to compute. Monte Carlo simulations demonstrate that the log-cumulant estimator of K distribution based on second-kind statistics achieves high estimation accuracy.

Key words: K distribution; Mellin transformation; second-kind statistics; log-cumulant estimator; Monte Carlo simulations

K 分布广泛用于海洋和地面的建模中, 传统的 Rayleigh 分布可以看做是 K 分布的特例^[1-4]。作为双参数的统计分布, 与单参数的分布 (如 Rayleigh 分布) 相比, K 分布可以更好地描述实验数据^[2]。例如, 当雷达分辨单元较大时, 可以认为分辨单元内包含许多散射体, 并且不存在占主导地位的散射体, 此时, 传统的 Rayleigh 分布可以很好地描述雷达回波。但是, 如果雷达分辨单元的尺寸以及掠射角较小, 在这种情况下, 只有 K 分布才能精确描述雷达回波的统计特性。为了在实际应用中使用 K 分布, 一个关键问题就是如何高效地估计出 K 分布的参数。由于 K 分布的概率密度函数中含有特殊的修正 Bessel 函数, 而修正 Bessel 函数又没有统一的解析表达式, 因此, 传统的最大似然估计方法不能高效地估计 K 分布的参数, 即最大似然估计没有解析的表达形式, 往往需要迭代计算, 而且初值选择和终止条件不易确定。基于第二类统计量, 本文提出了高效估计 K 分布参数的对数累积量方法, 它具有简洁的解析表达式, 不需要迭代计算。Monte Carlo 仿真表明, K 分布的对数累积量估计可以获得较高的估计精度。

1 K 分布

K 分布具有如下的概率密度函数^[3]:

$$f(x) = \frac{4b^{\frac{(\nu+1)}{2}} x^{\nu}}{\Gamma(\nu)} K_{\nu-1}(2x\sqrt{b}) \quad x > 0 \quad (1)$$

其中: $\nu (\nu > 0)$ 是形状参数; $b (b > 0)$ 是尺度参数; $\Gamma(\cdot)$ 是 Gamma 函数; $K_{\nu-1}(\cdot)$ 是第二类 $\nu-1$ 阶修正 Bessel 函数。当 ν 趋向于无穷大时, K 分布变为传统的 Rayleigh 分布。因此, Rayleigh 分布可以看做是 K 分布的特例。图 1 给出了当 ν 和 b 分别取不同值时 K 分布的概率密度函数。显然, ν 决定 K 分布的形状, 当 ν 较小时 (如 $\nu=0.5$), K 分布不再具有单峰特性; b 决定 K 分布向纵坐标轴的压缩程度, b 越大, K 分布向纵坐标轴的压缩程度就越强。

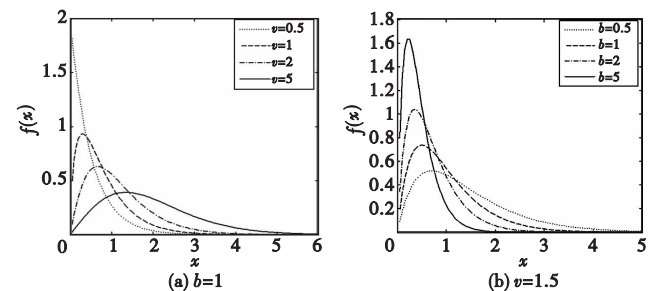


图1 K 分布的概率密度函数

假定 X 是服从 K 分布的随机变量, 其形状参数和尺度参数分别为 ν 和 b 。定义新的随机变量 $Y = \sqrt{b}X$, 可以证明, Y 同

收稿日期: 2012-06-07; **修回日期:** 2012-07-14 **基金项目:** 国家自然科学基金资助项目(61102163, 60805021, 61175121); 国家教育部新世纪优秀人才支持计划资助项目(NCET-10-0117); 福建省自然科学基金资助项目(2012J01271, 2011J01349); 福建省高等学校杰出青年科研人才培育计划资助项目(JA10006); 华侨大学高层次人才科研启动费资助项目(11BS212)

作者简介: 孙增国(1980-), 男, 陕西西安人, 讲师, 博士, CCF 会员, 主要研究方向为雷达信号及图像处理(duffer2000@163.com)。

样服从 K 分布,其形状参数和尺度参数分别为 ν 和 1。也就是说, K 分布可以通过 $\sqrt{b}$ 的乘积进行标准化,其形状参数保持不变,而尺度参数则变为 1。

2 第二类统计量介绍

普通统计量是由傅里叶变换定义的^[5,6]。如果概率密度函数定义在正半轴上,就可以把普通统计量中的傅里叶变换替换为 Mellin 变换,从而定义相应的第二类统计量^[6~9]。对于概率密度函数 $f(x)$ ($0 \leq x < \infty$),其 Mellin 变换定义如下:

$$\Phi(s) = \int_0^{\infty} x^{s-1} f(x) dx \quad (2)$$

$f(x)$ 的 Mellin 变换 $\Phi(s)$ 正是它的第二类第一特征函数。由此, $f(x)$ 的第二类第二特征函数 $\Psi(s)$ 和 r 阶对数累积量 $\bar{k}_r$ 分别定义如下:

$$\Psi(s) = \ln \Phi(s) \quad (3)$$

$$\bar{k}_r = \left. \frac{d^r \Psi(s)}{ds^r} \right|_{s=1} \quad (4)$$

对数累积量可以通过样本对数累积量加以估计。以前两阶对数累积量为例,存在

$$\hat{k}_1 = \frac{1}{M} \sum_{i=1}^M \ln y_i \quad (5)$$

$$\hat{k}_2 = \frac{1}{M-1} \sum_{i=1}^M (\ln y_i - \hat{k}_1)^2 \quad (6)$$

其中: y_i 是样本, M 是样本个数。

3 K 分布的对数累积量估计

K 分布定义在正半轴上,因此可以定义相应的第二类统计量。首先把 K 分布的概率密度函数式(1)代入式(2)中,可得 K 分布的第二类第一特征函数如下:

$$\Phi(s) = \frac{4b^{\frac{(\nu+1)}{2}}}{\Gamma(\nu)} \int_0^{\infty} x^{s+\nu-1} K_{\nu-1}(2x\sqrt{b}) dx \quad (7)$$

使用下面的恒等式^[10]:

$$\int_0^{\infty} x^{\mu} K_{\nu}(ax) dx = 2^{\mu-1} a^{-\mu-1} \Gamma\left(\frac{1+\mu+\nu}{2}\right) \Gamma\left(\frac{1+\mu-\nu}{2}\right) \quad (8)$$

$\text{Re}(1+\mu \pm \nu) > 0, \text{Re}(a) > 0$

K 分布的第二类第一特征函数可以简化为

$$\Phi(s) = \frac{b^{\frac{(1-s)}{2}} \Gamma\left(\frac{s+1}{2}\right) \Gamma\left(\nu + \frac{s-1}{2}\right)}{\Gamma(\nu)} \quad (9)$$

把式(9)代入式(3)中,可得 K 分布的第二类第二特征函数如下:

$$\Psi(s) = \frac{(1-s) \ln b}{2} + \ln \left(\Gamma\left(\frac{s+1}{2}\right) \right) + \ln \left(\Gamma\left(\nu + \frac{s-1}{2}\right) \right) - \ln \left(\Gamma(\nu) \right) \quad (10)$$

把式(10)代入式(4)中,经过适当化简,可得 K 分布的前两阶对数累积量如下:

$$\bar{k}_1 = -\frac{\ln b}{2} - \frac{C_e}{2} + \frac{\psi(\nu)}{2} \quad (11)$$

$$\bar{k}_2 = \frac{\pi^2}{24} + \frac{\psi(1, \nu)}{4} \quad (12)$$

其中: C_e 是欧拉常数($C_e \approx 0.5772$); $\psi(\cdot)$ 是 Digamma 函数; $\psi(1, \cdot)$ 是 Trigamma 函数(Digamma 函数的一阶导数)。给定一组样本,通过式(5)和(6)计算样本对数累积量 $\hat{k}_1$ 和 $\hat{k}_2$,并用 $\hat{k}_1$ 和 $\hat{k}_2$ 分别代替上面两式中的 $\bar{k}_1$ 和 $\bar{k}_2$,由此,形状参数 ν 可由式(12)估计得出,而尺度参数 b 可由式(11)估计得出。因此,式(11)和(12)就构成了 K 分布的对数累积量估计式。与传统的最大似然估计方法相比, K 分布的对数累积量估计具有简洁的

解析表达式。同时,由于 Trigamma 函数 $\psi(1, \cdot)$ 是严格单调函数,因此,使用常见的数值方法如二分法很容易获得 ν 的精确估计值^[11]。也就是说,由于具有解析表达式以及 $\psi(1, \cdot)$ 的单调性, K 分布的对数累积量估计易于计算,便于实现。

4 Monte Carlo 仿真

Monte Carlo 仿真的一个重要问题就是如何产生服从特定分布的样本。Skolnik 的研究表明, K 分布可由一个传统 Rayleigh 分布所描述的快速波动分量被一个 Gamma 分布所描述的缓慢波动分量所调制而形成^[2]。因此,分别产生服从 Rayleigh 分布和 Gamma 分布的样本,由此形成服从 K 分布的样本。产生服从 K 分布的一组样本,使用对数累积量方法估计出参数,这就完成了一次仿真实验。由于样本产生的随机性,往往需要独立运行多次仿真实验,并把多次仿真结果的均值和标准差作为最终的估计结果,这就是 Monte Carlo 仿真的一般步骤。

K 分布的对数累积量估计首先要通过式(5)和(6)计算样本对数累积量,因此,样本数量是影响对数累积量估计性能的重要因素。图2给出了对数累积量估计的性能随样本数量的变化关系,包括形状参数的估计值和估计标准差,以及尺度参数的估计值和估计标准差。可见,不论是形状参数 ν 还是尺度参数 b ,其估计值均随着样本数量的增加而逐渐趋向真实值,其估计标准差均随着样本数量的增加而逐渐降低。也就是说,较多的样本数量可以获得较高的估计性能。为了克服样本产生的随机性, Monte Carlo 仿真需要独立运行多次,并把多次仿真结果的均值作为最终的参数估计值。因此,运行次数是影响对数累积量估计精度的又一因素。图3给出了对数累积量估计的精度随运行次数的变化关系,包括形状参数的估计值以及尺度参数的估计值。可见,不论是形状参数 ν 还是尺度参数 b ,其估计值总体上均随着运行次数的增加而趋向真实值。也就是说,较多的运行次数可以获得较高的估计精度。在形状参数和尺度参数分别取不同值时,对数累积量估计的 Monte Carlo 仿真结果分别如表1和2所示。其中,括号中的数值为多次仿真结果的标准差。可见,不论形状参数或尺度参数取为何值,对数累积量估计均能获得较高的估计精度。

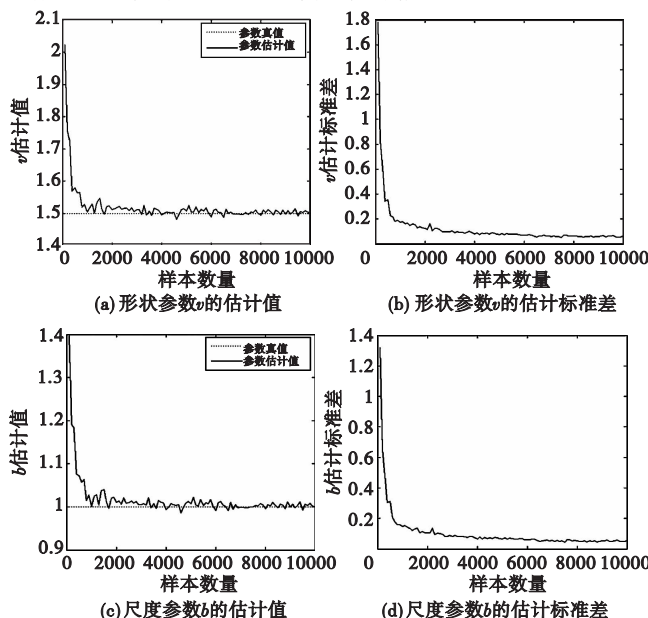


图2 对数累积量估计的性能随样本数量的变化关系
($\nu=1.5, b=1$, 对于不同的样本数量,独立运行次数均为100)

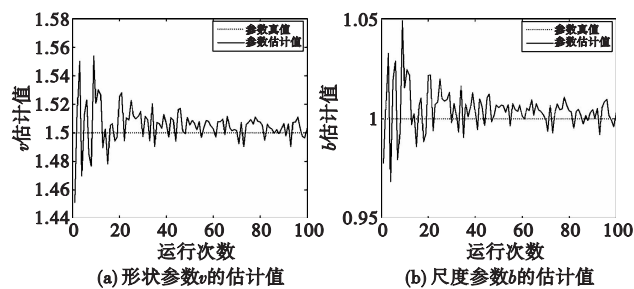


图3 对数累积量估计的精度随运行次数的变化关系
($\nu=1.5$, $b=1$, 对于不同的运行次数, 样本数量均为10 000)

表1 对数累积量估计的 Monte Carlo 仿真结果($b=1$, 独立运行 100 次, 每次的样本数量均为 10000)

参数真值	$\hat{\nu}$	$\hat{b}$
$\nu=0.5$	0.500 2 (0.008 6)	1.000 9 (0.038 1)
$\nu=1$	1.004 7 (0.024 2)	1.006 1 (0.033 8)
$\nu=1.5$	1.502 5 (0.053 9)	1.002 3 (0.047 5)
$\nu=2$	1.997 1 (0.102 2)	0.997 7 (0.063 4)
$\nu=5$	5.000 5 (0.649 6)	1.000 4 (0.138 2)

表2 对数累积量估计的 Monte Carlo 仿真结果($\nu=1.5$, 独立运行 100 次, 每次的样本数量均为 10000)

参数真值	$\hat{\nu}$	$\hat{b}$
$b=0.5$	1.496 8 (0.046 2)	0.499 0 (0.019 1)
$b=1$	1.501 1 (0.053 6)	0.999 9 (0.044 7)
$b=1.5$	1.501 3 (0.060 3)	1.503 8 (0.079 0)
$b=2$	1.501 6 (0.054 5)	2.003 4 (0.093 2)
$b=5$	1.502 5 (0.051 3)	5.004 1 (0.231 0)

5 结束语

为了高效地估计出 K 分布的参数, 本文提出了基于第二类统计量的对数累积量估计方法。首先对 K 分布的概率密度函数进行 Mellin 变换, 从而获得 K 分布的第二类第一特征函数; 然后对第二类第一特征函数进行对数变换, 从而获得 K 分布的第二类第二特征函数; 最后对第二类第二特征函数求前两阶导数, 进而获得 K 分布的对数累积量估计式。与传统的最

大似然估计方法相比, 对数累积量估计具有简洁的解析表达式, 易于计算。Monte Carlo 仿真表明, 对数累积量估计可以获得较高的估计精度。因此, 基于第二类统计量的对数累积量估计是 K 分布参数估计的高效方法。作为强有力的数学工具, 第二类统计量同样可以用于其他常见分布(如对数-正态分布)的参数估计中, 这是下一步的研究内容。

参考文献:

- [1] 张红, 王超, 张波, 等. 高分辨率 SAR 图像目标识别[M]. 北京: 科学出版社, 2009.
- [2] SKOLNIK M I. 雷达系统导论[M]. 3 版. 左群声, 徐国良, 马林, 等译. 北京: 电子工业出版社, 2006.
- [3] WARD K D, TOUGH R J A, WATTS S. Sea clutter: scattering, the K distribution and radar performance[M]. London: The Institution of Engineering Technology, 2006.
- [4] 皮亦鸣, 杨建宇, 付毓生, 等. 合成孔径雷达成像原理[M]. 成都: 电子科技大学出版社, 2007.
- [5] PAPOULIS A, PILLAI S U. 概率、随机变量与随机过程[M]. 保铮, 冯大政, 水鹏朗, 译. 西安: 西安交通大学出版社, 2004.
- [6] 孙增国, 韩崇昭. 基于拖尾分布的高分辨率合成孔径雷达图像建模[J]. 物理学报, 2010, 59(2): 998-1008.
- [7] ACHIM A, KURUOGLU E E, ZERUBIA J. SAR image filtering based on the heavy-tailed Rayleigh model[J]. IEEE Trans on Image Processing, 2006, 15(9): 2686-2693.
- [8] NICOLAS J M. Alpha-stable positive distributions: a new approach based on second kind statistics[C]//Proc of European Signal Processing Conference. Toulouse: EURASIP, 2002: 197-200.
- [9] TISON C, NICOLAS J M, TUPIN F, et al. A new statistical model for Markovian classification of urban areas in high-resolution SAR images[J]. IEEE Trans on Geoscience and Remote Sensing, 2004, 42(10): 2046-2057.
- [10] GRADSHTEYN I S, RYZHIK I M. Table of integrals, series, and products[M]. 北京: 世界图书出版公司, 2007.
- [11] PRESS W H, TEUKOLSKY S A, VETTERLING W T, et al. Numerical recipes in C[M]. Cambridge: Cambridge University Press, 1994.

(上接第 45 页)

4 结束语

本文依据聚类假设的思想, 提出了一种基于聚类核的半监督支持向量机分类方法。该算法根据高密度区域的样本在聚类中被分在同一类中的特性, 采用 K -均值聚类算法对样本集进行多次聚类的方法构造聚类核, 将无标记样本的信息融入分类器的训练过程中。该方法能够有效利用无标记样本信息, 而且实验结果也证实了无标记样本的加入可以提高分类的精度。但本方法涉及到支持向量机中参数的选择、 K -均值聚类算法的有效性, 并且随着数据量的增大, 运行时间也会随之增加。因此, 如何更有效地进行参数选择和 K -均值聚类, 同时降低时间复杂度仍然有待研究。

参考文献:

- [1] SCHOLKOPF B, SMOLA A J. Learning with kernel: support vector machines, regularization, optimization and beyond[M]. Cambridge, MA: MIT Press, 2002: 190-195.
- [2] JOACHIMS T. Transductive inference for text classification using support vector machines[C]//Proc of the 16th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1999: 200-209.
- [3] CHAPPELLE O, SINDHWANI V, KEERTHI S. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research, 2008, 9(2): 203-233.
- [4] CHAPPELLE O, ZIEN A. Semi-supervised classification by low density

separation[C]//Proc of the 10th International Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann Publishers, 2005: 57-64.

- [5] CHAPPELLE O, CHI M, ZIEN A. A continuation method for semi-supervised SVMs[C]//Proc of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006: 185-192.
- [6] COLLOBERT R, SINZ F. Large scale transductive SVMs[J]. Journal of Machine Learning Research, 2006, 7(8): 1687-1712.
- [7] ZHU Xiao-jin. Semi-supervised learning literature survey TR1530[R]. Wisconsin: Dept. of Computer Sciences, University of Wisconsin, 2008: 13-16.
- [8] SERGIOS T, KONSTANTIONS K. Pattern recognition[M]. [S. l.]: Academic Press, 2008: 577-582.
- [9] CHAPPELLE O, WESTON J, SCHOLKOPF B. Cluster kernels for semi-supervised learning[C]//Proc of the 16th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2003: 321-328.
- [10] CHAPPELLE O, SCHÖLKOPF B, ZIEN A. Semi-supervised learning[M]. Massachusetts: MIT Press, 2006: 348-351.
- [11] TUIA D, CAMPS-VALLS G. Urban image classification with semisupervised multiscale cluster kernels[J]. IEEE Journal of Selected topics in Applied Earth Observations and Remote Sensing, 2011, 4(1): 65-74.
- [12] The Berkeley segmentation dataset and benchmark [EB/OL]. [2011-02-20]. <http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/>.
- [13] GULSHAN V, ROTHER C, CRIMINISI A, et al. Geodesic star convexity for interactive image segmentation[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2010: 3129-3136.