高阶高斯分布迭代的云模型及其数学性质研究

刘玉超[1,2],马于涛[2,3],张海粟[4],陈桂生[2]
(1.清华大学 计算机科学与技术系,北京 100084;2.中国电子设备系统工程公司研究所,北京 100141;
3.武汉大学 软件工程国家重点实验室,湖北武汉 430072;4.中国人民解放军理工大学 指挥自动化学院,江苏南京 210007)

摘 要: 云模型通过二阶高斯分布研究不确定性,它产生的云滴分布具有尖峰肥尾特性,呈现出幂率衰减.社会学和经济学的研究发现,由于在演化过程中具有偏好依附的特点,许多实际数据呈现出尖峰肥尾的特性,本文试图通过高阶高斯分布迭代产生的高阶云模型的数学性质研究,探寻高斯分布与尖峰肥尾分布之间的联系.基于高斯分布迭代构造具有尖峰肥尾特性的概率分布,通过基于高阶高斯分布迭代的云模型刻画更多的不确定性现象,分析高阶高斯分布迭代的典型参数,与云模型参数进行对比分析,为雾化后的逆向云发生器求解提供了新的手段,同时也为高阶云模型的逆向求解过程提供了方法.

关键词: 高斯分布; 峰度; 肥尾; 云模型
中图分类号: TP302.7 **文献标识码:** A **文章编号:** 0372-2112 (2012)10-1913-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.10.001

Study on Characters of Cloud Model Based on High-Order Gaussian Distribution with Iterations

LIU Yu-chao[1,2], MA Yu-tao[2,3], ZHANG Hai-su[4], CHEN Gui-sheng[2]
(1. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*;
2. *Institute of China Electronic System Engineering Corporation, Beijing* 100141, *China*;
3. *State Key Lab of Software Engineering, Wuhan University, Wuhan, Hubei* 430072, *China*;
4. *College of Command Automation, PLA University of Science and Technology, Nanjing, Jiangsu* 210007, *China*)

Abstract: Cloud model uses two-order Gaussian distribution to produce cloud drops whose distribution displays high kurtosis and fat tail with power-law decay. In sociology and economics, many phenomena have been found to share high kurtosis and fat tail because of preferential attachment in evolution processes. This paper tries to investigate the relation between Gaussian distribution and fat-tail distribution, and to construct fat-tail distribution based on Gaussian distribution with iterations to depict more uncertain phenomena. The comparative study on parameters of high-order Gaussian distribution with iterations and cloud model can provide a new thought and method for the research of reverse solution of high-order cloud model.

Key words: Gaussian distribution; kurtosis; fat tail; cloud model

1 引言

在客观世界中许多现象都服从或者近似服从高斯分布,例如正常生产条件下的产品质量指标、随机测量误差、同一生物群体的某种特征、某地的年平均气温等等.在概率理论的研究历史中,高斯分布的地位举足轻重,高斯分布的密度函数和分布函数有比较简单的数学形式和各种很好的性质,而且是许多重要概率分布的极限分布,这些都使得高斯分布在理论和实际中应用非常广泛.中心极限定理从理论上阐述了产生高斯分布的条件,其简单直观的说明如下:如果决定某一随机变量结果的是大量微小的、独立的随机因素之和,并且每一个因素的单独作用相对均匀的小,没有一种因素可起到压倒一切的主导作用,那么这个随机变量一般近似于高斯分布.

定义 1[1] 若随机变量 X 的概率密度函数为:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

则称 X 为高斯分布,记为 $X\sim N(\mu,\sigma^2)$.其中,μ 和 σ^2 分别是高斯分布的期望和方差,分别表征随机变量

收稿日期:2011-07-25;修回日期:2012-06-18
基金项目:国家自然科学基金重点项目(No.61035004);国家自然科学基金面上项目(No.60974086)

的最可能取值以及一切可能取值的离散程度.高斯分布函数具有关于期望对称,然后向两个方向衰减,呈现"中间大、两头小"的特性,约 99.73% 的取值都落在以 μ 为中心的 3σ 区间内,这一性质被称为 3σ 原则,在数理统计中有广泛的应用.

通常在讨论有关概率问题时,分布函数 $F(x)$ 起着非常重要的作用,但在实际的应用中,有时候 $F(>x)$ 更为重要.例如,在有关可靠性的研究中,可靠性与失效率等概念都同 $F(>x)$ 有关.形象而言,肥尾分布(fat-tailed distribution)就是密度函数"尾巴" $F(>x)$ 比较长的分布,其物理意义是极端事件的概率不为 0.如保险业的大额索赔问题,在 N 次索赔中有一次索赔的额度非常大,以至于其它 $N-1$ 次索赔相对于这次索赔而言都是微不足道的,就需要用肥尾分布来处理.这类问题就是所谓的极端事件问题,如地震、洪水、股灾等.自 20 世纪 60 年代以来,国外出现了许多肥尾分布的研究文献[2~4].肥尾分布有很多等价的数学定义,其中一种定义是相对于高斯分布而言、以四阶中心矩为基础的.四阶中心矩具有峰度(kurtosis)的含义,峰度是统计中描述分布状态的一个重要特征值,用以判断分布曲线相比于高斯分布曲线的尖平程度.

定义 2[5] 随机变量 X 称为是肥尾的,如果 $Kur = E\left[\frac{(X-\mu)^4}{\sigma^4}\right] - 3 > 0$,其中 μ, σ 分别为 X 的期望和标准差.

如果将高斯分布视为常峰态(峰度为 0),分布曲线的形状比高斯分布更高更瘦的称为高峰态,该性质被称为超过或大于标准峰度,否则称为低峰态.峰度建立了高斯分布和尖峰肥尾分布之间的联系,是研究高斯分布的一个重要数学特征.

1995 年李德毅研究员提出云模型[6],通过二阶高斯随机数产生器构建云发生器算法,根据三个数字特征(期望、熵和超熵)自动生成数据样本(称之为云滴),云滴分布表现出很好的数学特性,既能体现高斯分布的"钟形"特征,又能体现出幂率分布的"肥尾"特性.云模型已在智能控制、数据挖掘、性能评价等方面得到成功应用[7~9].

算法:正向云发生器

输入:数字特征 (Ex, En, He),生成云滴的个数 N

输出:N 个云滴 x 及其确定度 y(也可表示为 $\mathrm{Drop}(x_i; y_i), i = 1, \cdots, N$)

算法步骤:

1.生成以 En 为期望,He^2 为方差的一个正态随机熵 $En_i' = \mathrm{NORM}(En, He^2)$;

2.生成以 Ex 为期望,$En_i'^2$ 为方差的一个正态随机数 $x_i = \mathrm{NORM}(Ex, En_i'^2)$;

3.计算 $y_i = \exp\left\{-\frac{(x_i - Ex)^2}{2(En_i')^2}\right\}$ 具有确定度 y_i 的 x_i 成为论域 U 中的一个云滴;

4.重复步骤 1 - 3,直至产生 N 个云滴.

如果不考虑步骤 3 样本对概念确定度的计算,那么整个算法就是一个二阶高斯随机数发生器.在此基础上,模糊集合专家王立新博士提出了 p 阶云模型的数学表示并就相关数学性质进行了讨论[7],1995 年提出的云模型具有三个参数 (Ex, En, He),Ex 表示当前数据样本的期望,当前数据样本的方差是一个随机变量,其期望表示为 En,方差表示为 He,也就是说 He 是一个 2 阶方差,是从不确定性现象出发定义云模型参数.王立新博士认为 p 阶云模型具有的 $p+1$ 个参数应该是 $(He, En_1, En_2, \cdots, En_{p-1}, En_p)$,他认为应该从不确定性现象的本质出发,从初始的参数 He 开始,经过层层迭代形成了以 En_p 为期望的 p 阶不确定性数据样本.受此启发,本文从概率分布的角度,对云模型的数学基础—高阶高斯分布迭代的数学特性进行推导研究,并与云模型进行参数比较,一方面展示云模型基于高阶高斯分布迭代的数学基础,另一方面为云模型的逆向发生器算法提供新的思路.

2 高阶高斯分布迭代

如果将高斯分布看作 1 阶高斯分布,那么 1 阶随机变量 X_1 的概率密度函数表示为:

$$f(x_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right) \tag{2}$$

其中,μ_1 和 σ^2 分别是 1 阶高斯分布的期望和方差.

定义 3 若随机变量 X_2 的概率密度函数为:

$$\begin{aligned} f(x_2) &= f(x_2 \mid x_1) P(x_1) \\ &= \int_{-\infty}^{+\infty} f(x_2 \mid x_1) f(x_1) \mathrm{d}x_1 \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi x_1^2}} \exp\left(-\frac{(x_2-\mu_2)^2}{2x_1^2}\right) \\ &\quad \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right) \mathrm{d}x_1 \end{aligned} \tag{3}$$

则称 X_2 为 2 阶高斯分布.其中,μ_2 是 X_2 的期望,x_1 是 1 阶高斯随机变量 X_1 的一次实现,σ^2 是 X_1 的方差.

定义 4 若随机变量 X_3 的概率密度函数为:

$$\begin{aligned} f(x_3) &= f(x_3 \mid x_2) P(x_2) \\ &= \int_{-\infty}^{+\infty} f(x_3 \mid x_2) f(x_2) \mathrm{d}x_2 \\ &= \int_{-\infty}^{+\infty} f(x_3 \mid x_2) f(x_2 \mid x_1) P(x_1) \mathrm{d}x_2 \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_3 \mid x_2) f(x_2 \mid x_1) f(x_1) \mathrm{d}x_1 \mathrm{d}x_2 \end{aligned}$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi x_2^2}}\exp\left(-\frac{(x_3-\mu_3)^2}{2x_2^2}\right) \cdot\frac{1}{\sqrt{2\pi x_1^2}}\exp\left(-\frac{(x_2-\mu_2)^2}{2x_1^2}\right) \cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\mathrm{d}x_2 \tag{4}$$

则称 X_3 为 3 阶高斯分布.其中,μ_3 是 X_3 的期望,x_2 是 2 阶高斯随机变量 X_2 的一次实现,x_1 是 1 阶高斯随机变量 X_1 的一次实现 ,μ_1 是 X_1 的期望,σ^2 是 X_1 的方差.

依此类推,可以给出 p 阶($p\geqslant 2$)高斯分布迭代的定义如下:

定义 5　若随机变量 X_p 的概率密度函数为:

$$\begin{aligned} f(x_p) &= f(x_p\mid x_{p-1})P(x_{p-1}) \\ &= \int_{-\infty}^{+\infty} f(x_p\mid x_{p-1})f(x_{p-1})\mathrm{d}x_{p-1} \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f(x_p\mid x_{p-1})\cdots f(x_2\mid x_1)f(x_1)\mathrm{d}x_1\mathrm{d}x_2 \cdots\mathrm{d}x_{p-1} \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi x_{p-1}^2}}\exp\left(-\frac{(x_p-\mu_p)^2}{2x_{p-1}^2}\right)\cdots \\ &\quad\cdot\frac{1}{\sqrt{2\pi x_1^2}}\exp\left(-\frac{(x_2-\mu_2)^2}{2x_1^2}\right) \\ &\quad\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_{p-1} \end{aligned} \tag{5}$$

则称 X_p 为 p 阶高斯分布迭代.其中,$\mu_i(i=1,\cdots,p)$是 $X_i(i=1,\cdots,p)$的期望,$x_i(i=1,\cdots,p)$是 i 阶高斯随机变量 $X_i(i=1,\cdots,p)$的一次实现,σ^2 是 X_1 的方差.

单纯从数学定义表示上可以看出,对于 p 阶高斯分布迭代 x_p 而言,随着阶数 p 的增加,其方差不断向两极分化,导致 x_p 的两极分化,也就是说变量的实现向靠近期望 μ_p 和远离期望的尾端同时变化,直接形成了尖峰肥尾分布.

3　高阶高斯分布迭代的数字特征

对高阶高斯分布迭代的数字特征进行的计算和推导是研究其数学性质的重要基础,本节主要分析高阶高斯分布迭代的期望、方差、三阶中心矩和四阶中心矩四个数字特征,并计算其峰度.

p 阶高斯分布迭代的期望计算如下:

$$\begin{aligned} E(X_p) &= \int_{-\infty}^{+\infty} x_p f(x_p)\mathrm{d}x_p \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} x_p \frac{1}{\sqrt{2\pi x_{p-1}^2}}\exp\left(-\frac{(x_p-\mu_p)^2}{2x_{p-1}^2}\right)\mathrm{d}x_p\cdots \\ &\quad\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_{p-1} \\ &= \mu_p\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1} \\ &\quad\cdots\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1 \\ &= \mu_p \end{aligned} \tag{6}$$

p 阶高斯分布迭代的方差(二阶中心矩)计算如下:

$$\begin{aligned} Var(X_p) &= \int_{-\infty}^{+\infty}(x_p-\mu_p)^2 f(x_p)\mathrm{d}x_p \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(x_p-\mu_p)^2\frac{1}{\sqrt{2\pi x_{p-1}^2}} \\ &\quad\cdot\exp\left(-\frac{(x_p-\mu_p)^2}{2x_{p-1}^2}\right)\mathrm{d}x_p \\ &\quad\cdot\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\cdots \\ &\quad\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_{p-1} \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}x_{p-1}^2\frac{1}{\sqrt{2\pi x_{p-2}^2}} \\ &\quad\cdot\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\cdots \\ &\quad\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_{p-1} \\ &= \int_{-\infty}^{+\infty}\frac{(x_{p-1}-\mu_{p-1})^2}{\sqrt{2\pi x_{p-2}^2}} \\ &\quad\cdot\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots \\ &\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1 \\ &\quad+\int_{-\infty}^{+\infty}\frac{\mu_{p-1}^2}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots \\ &\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1 \\ &= Var(X_{p-1})+\mu_{p-1}^2 = \sum_{i=1}^{p-1}\mu_i^2+\sigma^2 \end{aligned} \tag{7}$$

p 阶高斯分布迭代的三阶中心矩计算如下:

$$\begin{aligned} &E\{[X_p-E(X_p)]^3\} \\ &= \int_{-\infty}^{+\infty}(x_p-\mu_p)^3 f(x_p)\mathrm{d}x_p \\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(x_p-\mu_p)^3\frac{1}{\sqrt{2\pi x_{p-1}^2}}\exp\left(-\frac{(x_p-\mu_p)^2}{2x_{p-1}^2}\right)\cdots \\ &\quad\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_{p-1}\mathrm{d}x_p \\ &= \frac{x_{p-1}^3}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mu_p^3\exp\left(-\frac{\mu_p^2}{2}\right)\mathrm{d}\mu_p\int_{-\infty}^{+\infty}\cdots \\ &\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots \end{aligned}$$

$$\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1=0 \tag{8}$$

p 阶高斯分布迭代的四阶中心矩计算如下：

$$\begin{aligned}
&E\{[X_p-E(X_p)]^4\}\\
&=\int_{-\infty}^{+\infty}(x_p-\mu_p)^4 f(x_p)\mathrm{d}x_p\\
&=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}(x_p-\mu_p)^4\frac{1}{\sqrt{2\pi x_{p-1}^2}}\exp\left(-\frac{(x_p-\mu_p)^2}{2x_{p-1}^2}\right)\cdots\\
&\quad\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\cdots\mathrm{d}x_{p-1}\mathrm{d}x_p\\
&=\int_{-\infty}^{+\infty}(x_p-\mu_p)^4\frac{1}{\sqrt{2\pi x_{p-1}^2}}\exp\left(-\frac{(x_p-\mu_p)^2}{2x_{p-1}^2}\right)\mathrm{d}x_p\\
&\quad\cdot\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots\\
&\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\\
&=\frac{2x_{p-1}^4}{\sqrt{2\pi}}\int_0^{+\infty}\mu_p^4\exp\left(-\frac{\mu_p^2}{2}\right)\mathrm{d}\mu_p\\
&\quad\cdot\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots\\
&\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\\
&=\int_{-\infty}^{+\infty}3x_{p-1}^4\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots\\
&\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1
\end{aligned}\tag{9}$$

$$\begin{aligned}
x_{p-1}^4&=(x_{p-1}-\mu_{p-1}+\mu_{p-1})^4\\
&=[(x_{p-1}-\mu_{p-1})^2+\mu_{p-1}^2+2\mu_{p-1}(x_{p-1}-\mu_{p-1})]^2\\
&=(x_{p-1}-\mu_{p-1})^4+\mu_{p-1}^4+6(x_{p-1}-\mu_{p-1})^2\mu_{p-1}^2\\
&\quad+4\mu_{p-1}^3(x_{p-1}-\mu_{p-1})+4\mu_{p-1}(x_{p-1}-\mu_{p-1})^3
\end{aligned}\tag{10}$$

将式(10)带入式(9)计算，得到：

$$\begin{aligned}
&E\{[X_p-E(X_p)]^4\}\\
&=3\int_{-\infty}^{+\infty}(x_{p-1}-\mu_{p-1})^4\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots\\
&\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\\
&\quad+3\int_{-\infty}^{+\infty}\mu_{p-1}^4\frac{1}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots\\
&\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\\
&\quad+18\int_{-\infty}^{+\infty}\frac{(x_{p-1}-\mu_{p-1})^2\mu_{p-1}^2}{\sqrt{2\pi x_{p-2}^2}}\exp\left(-\frac{(x_{p-1}-\mu_{p-1})^2}{2x_{p-2}^2}\right)\mathrm{d}x_{p-1}\cdots\\
&\quad\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_1-\mu_1)^2}{2\sigma^2}\right)\mathrm{d}x_1\\
&=3E\{[X_{p-1}-E(X_{p-1})]^4\}+3\mu_{p-1}^4+18\mu_{p-1}^2Var(X_{p-1})\\
&=3^p\sigma^4+6\sigma^2\sum_{i=1}^{p-1}3^{p-i}\mu_i^2+6\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{p-i}\mu_i^2\mu_j^2+\sum_{i=1}^{p-1}3^{p-i}\mu_i^4
\end{aligned}\tag{11}$$

p 阶高斯分布迭代的峰度计算如下：

$$\begin{aligned}
&Kur(X_p)\\
&=\frac{E\{[X_p-E(X_p)]^4\}}{[Var(X_p)]^2}-3\\
&=\frac{3E\{[X_{p-1}-E(X_{p-1})]^4\}+3\mu_{p-1}^4+18\mu_{p-1}^2Var(X_{p-1})}{[\sum_{i=1}^{p-1}\mu_i^2+\sigma^2]^2}-3\\
&=\frac{3^p\sigma^4+6\sigma^2\sum_{i=1}^{p-1}3^{p-i}\mu_i^2+6\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{p-i}\mu_i^2\mu_j^2+\sum_{i=1}^{p-1}3^{p-i}\mu_i^4}{[\sum_{i=1}^{p-1}\mu_i^2+\sigma^2]^2}-3\\
&=3^p\frac{\sigma^4+2\sigma^2\sum_{i=1}^{p-1}3^{1-i}\mu_i^2+2\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{1-i}\mu_i^2\mu_j^2+\sum_{i=1}^{p-1}3^{-i}\mu_i^4}{\sigma^4+2\sigma^2\sum_{i=1}^{p-1}\mu_i^2+2\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}\mu_i^2\mu_j^2+\sum_{i=1}^{p-1}\mu_i^4}-3
\end{aligned}\tag{12}$$

本节高阶高斯分布迭代的数字特征的证明方法和结论与文献[10]中对高阶正态云的数学性质证明相同.

4 高阶高斯分布迭代的参数对峰度的影响分析

从数学定义的表达式中可以看出，p 阶高斯分布迭代具有 $p+1$ 个参数，分别是 $\mu_i(i=1,\cdots,p)$ 和 σ，参数的取值决定了高阶高斯分布迭代反映出来的数学特性.下面针对三种具有简单参数的高阶高斯分布迭代，对其数学性质的变化趋势进行研究，并加以试验分析.

(1)对于 $\mu_i=0(i=1,\cdots,p),\sigma\neq0$ 的情况

$$E(X_p)=\mu_p=0$$

$$Var(X_p)=\sum_{i=1}^{p-1}\mu_i^2+\sigma^2=\sigma^2$$

$$E\{[X_p-E(X_p)]^4\}=3^p\sigma^4$$

$$Kur(X_p)=\frac{E\{[X_p-E(X_p)]^4\}}{[Var(X_p)]^2}-3=\frac{3^p\sigma^4}{\sigma^4}-3=3^p-3$$

峰度随阶数变化趋势如图 1 所示.

(2)对于 $\mu_i=\mu(i=1,\cdots,p),r=\sigma/\mu$ 的情况

$$E(X_p)=\mu_p=\mu$$

$$Var(X_p)=\sum_{i=1}^{p-1}\mu_i^2+\sigma^2=(p-1)\frac{\sigma^2}{r^2}+\sigma^2$$

$$\begin{aligned}
&E\{[X_p-E(X_p)]^4\}\\
&=3^p\sigma^4+6\sigma^2\sum_{i=1}^{p-1}3^{p-i}\mu_i^2+6\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{p-i}\mu_i^2\mu_j^2+\sum_{i=1}^{p-1}3^{p-i}\mu_i^4\\
&=3^p\sigma^4+\frac{6\sigma^4\sum_{i=1}^{p-1}3^{p-i}}{r^2}+\frac{6\sigma^4\sum_{i=1}^{p-1}(i-1)3^{p-i}}{r^4}+\frac{\sigma^4\sum_{i=1}^{p-1}3^{p-i}}{r^4}
\end{aligned}$$

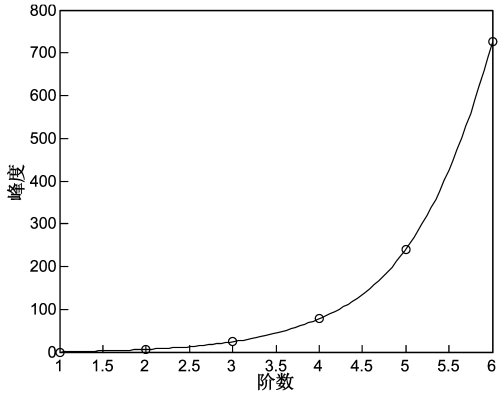

图1　p阶高斯分布迭代的峰度随阶数变化趋势曲线($\mu_i=0,\sigma\neq 0$)

$$= 3^p\sigma^4 + \frac{6\sigma^4\sum_{i=1}^{p-1}3^{p-i}}{r^2} + \frac{\sigma^4\sum_{i=1}^{p-1}(6i-5)3^{p-i}}{r^4}$$

$$Kur(X_p) = \frac{E\{[X_p - E(X_p)]^4\}}{[Var(X_p)]^2} - 3$$

$$= \frac{3^p\sigma^4 + \frac{6\sigma^4\sum_{i=1}^{p-1}3^{p-i}}{r^2} + \frac{\sigma^4\sum_{i=1}^{p-1}(6i-5)3^{p-i}}{r^4}}{[(p-1)\frac{\sigma^2}{r^2} + \sigma^2]^2} - 3$$

$$= \frac{3^p r^4 + 6r^2\sum_{i=1}^{p-1}3^{p-i} + \sum_{i=1}^{p-1}(6i-5)3^{p-i}}{r^4 + (p-1)^2 + 2(p-1)r^2} - 3$$

峰度随阶数变化趋势如图 2 所示.

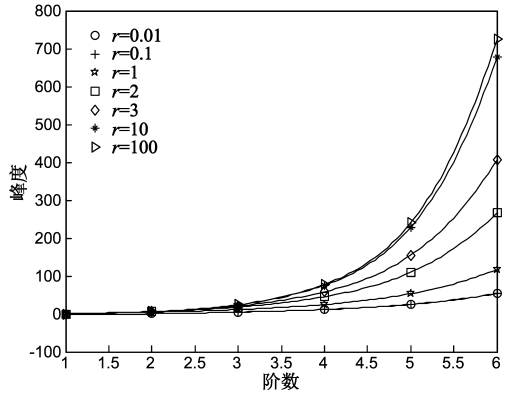

图2　p阶高斯分布迭代的峰度随阶数变化趋势曲线($\mu_i=\mu, r=\sigma/\mu$)

从图 2 可以看出,在 $\mu_i = \mu(i = 1,\cdots,p)$, $r = \sigma/\mu$ 的情况下,随着阶数的增加,高阶高斯分布迭代的峰度不断增加;对于具有相同阶数的高斯分布而言,当 $0.3 < r < 10$ 时,随着 r 的增加,峰度增加显著,而其他情况峰度则趋于稳定,随 r 变化增长不大.

(3)对于 $\mu_i = r^i\sigma(i = 1,\cdots,p)$的情况

$$E(X_p) = \mu_p = r^p\sigma$$

$$Var(X_p) = \sum_{i=1}^{p-1}\mu_i^2 + \sigma^2 = (\sum_{i=1}^{p-1}r^{2i} + 1)\sigma^2 = \sigma^2(\frac{r^{2p}-1}{r^2-1})$$

$$E\{[X_p - E(X_p)]^4\}$$

$$= 3^p\sigma^4 + 6\sigma^2\sum_{i=1}^{p-1}3^{p-i}\mu_i^2 + 6\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{p-i}\mu_i^2\mu_j^2 + \sum_{i=1}^{p-1}3^{p-i}\mu_i^4$$

$$= 3^p\sigma^4 + 6\sigma^4\sum_{i=1}^{p-1}3^{p-i}r^{2i} + 6\sigma^4\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{p-i}r^{2(i+j)} + \sigma^4\sum_{i=1}^{p-1}3^{p-i}r^{4i}$$

$$Kur(X_p) = \frac{E\{[X_p - E(X_p)]^4\}}{[Var(X_p)]^2} - 3$$

$$= \frac{3^p\sigma^4 + 6\sigma^4\sum_{i=1}^{p-1}3^{p-i}r^{2i} + 6\sigma^4\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{p-i}r^{2(i+j)} + \sigma^4\sum_{i=1}^{p-1}3^{p-i}r^{4i}}{[(\sum_{i=1}^{p-1}r^{2i} + 1)\sigma^2]^2} - 3$$

$$= \frac{3^p + 6\sum_{i=1}^{p-1}3^{p-i}r^{2i} + 6\sum_{i=1}^{p-1}\sum_{j=1}^{i-1}3^{p-i}r^{2(i+j)} + \sum_{i=1}^{p-1}3^{p-i}r^{4i}}{1 + 2\sum_{i=1}^{p-1}r^{2i} + \sum_{i=1}^{p-1}\sum_{j=1}^{p-1}r^{2(i+j)}} - 3$$

峰度随阶数变化趋势如图 3 所示.

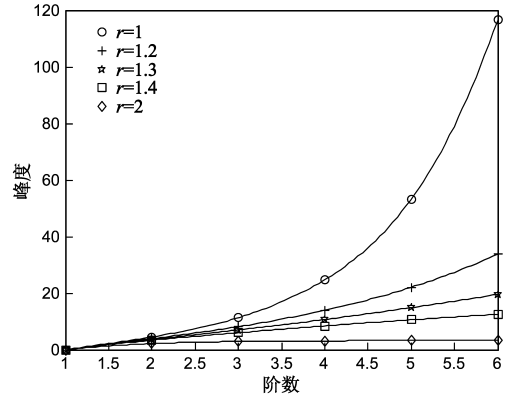

图3　p阶高斯分布迭代的峰度随阶数变化趋势曲线($\mu_i=r^i\sigma$)

从图 3 可以看出,在 $\mu_i = r^i\sigma(i = 1,\cdots,p)$的情况下,随着阶数的增加,高阶高斯分布迭代的峰度不断增加;当 $r\approx 1.3$ 时,峰度随阶数增长近似呈线性增长,当 $r < 1.3$ 时峰度随阶数增长趋势加剧,当 $r > 1.3$ 峰度随阶数增长趋势减缓.

上述工作表明,可以基于高斯分布构造具有尖峰肥尾特性的高阶高斯分布迭代,由于高斯分布具有良好的数学性质,很容易对高阶高斯分布迭代的数学性质进行推导.高阶高斯分布迭代可以用来刻画更多的不确定性现象,是一种从高斯分布向尖峰肥尾分布过渡的方法.

5　云模型和典型高斯分布迭代的样本数据统计分析实验

对于给定的具有"中间大、两头小"分布特征的 n 个数据样本 x_i,如何判定用几阶的高斯分布迭代来刻画比较合适,如何计算其参数,这其实也是逆向云模型发生器[8]研究的主要内容.二阶中心矩和四阶中心矩为

计算几种典型情况下的高阶高斯分布迭代的参数提供了手段.参数定义如下:Ex 表示样本期望,En_1 表示样本方差为随机变量时的期望,En_2 表示样本 2 阶方差(方差的方差)为随机变量时的期望,…,En_{p-1}表示样本 $p-1$ 阶方差为随机变量时的期望,He 为其方差. C_k 表示 k 阶中心矩(其中 C_2 为方差).

显然,Ex 具有直接的物理意义,很容易求得,

$$Ex = \frac{1}{n}x_i(i = 1, \cdots, n)$$

当 $p = 1$ 时,显然有 $c_2 = En_1^2$

当 $p = 2$ 时,$c_2 = En_1^2 + He^2$, $c_4 = 9He^4 + 18En_1^2He^2 + 3En_1^4$,可求得 En_1 和 He.

当 $p \geqslant 3$ 时,变量增多,需要更多的方程来求解这些变量,可是高阶中心矩对数据的敏感性将增大,使得在实数内求解几乎不可能完成.

通过第 3 节数字特征的推导可以看出,He 在样本数据的分布特征中起着至关重要的作用,对于 Ex 和 He 中间的参数,我们不妨取 $En_1 = En_2 = \cdots = En_{p-1} = 0$ 为不同阶高斯分布迭代的典型参数,计算得到的 p 值表示目前的样本数据分布与典型高斯分布迭代之间的相似关系.

首先,利用正向云发生器[11]针对青年人这一概念,生成其年龄属性的样本云滴,算法如下:

```
Ex = 25;% 期望年龄
En = 5;% 青年人的年龄与期望年龄的方差为随机变量,期望是 5
He = 1;% 青年人的年龄与期望年龄的方差为随机变量,方差是 1
for i = 1:n    % 循环产生 n 个样本云滴
    En' = randn(1) * He + En;% 产生一个随机方差;
    X(i) = randn(1) * En' + Ex;% 产生一个样本云滴;
    end
```

利用第 4 节中 $\mu_i = 0(i = 1, \cdots, p), \sigma \neq 0$ 的情况下的推导结论有 $c_2 = He^2$, $c_4 = 3^pHe^4$,计算得到典型高斯分布迭代的参数:阶数 $p = 1.11$,样本期望 $Ex' = 25$,最后一阶的方差 $He' = 5.1$. 还可以对云模型中不同的 He 取值,计算典型高斯分布迭代的参数,如表 1 所示. 进一步地,给出通过云模型产生的云滴的分布与典型高斯分布迭代之间的对比,如图 4 所示.

可以看出对于一个 2 阶云模型产生的云滴分布,当 $\frac{He}{En} > 1$ 时,随着 He 的增大,云模型产生的云滴分布越来越接近典型 2 阶高斯分布迭代,可以通过相应的典型 2 阶高斯分布迭代来刻画,因此,当 $\frac{He}{En} > 1$ 时,逆向云发生器算法[11]可以通过求解典型高斯分布迭代的参数来获取,因为他们生成的云滴分布基本相同,这为雾化[12]后的逆向云发生器求解提供了新的手段,同时也为高阶云模型的逆向求解过程提供了思路.

表 1 云模型与典型高斯分布迭代之间的参数取值

云模型参数			典型高斯分布迭代参数			备注
Ex	En	He	P	Ex'	He'	
25	5	1	1.11	25	5.1	靠近 1 阶
25	5	3	1.54	25	5.8	处于 1 阶和 2 阶之间
25	5	5	1.8	25	7	靠近 2 阶
25	5	10	2	25	11	2 阶
25	5	100	2	25	101	2 阶

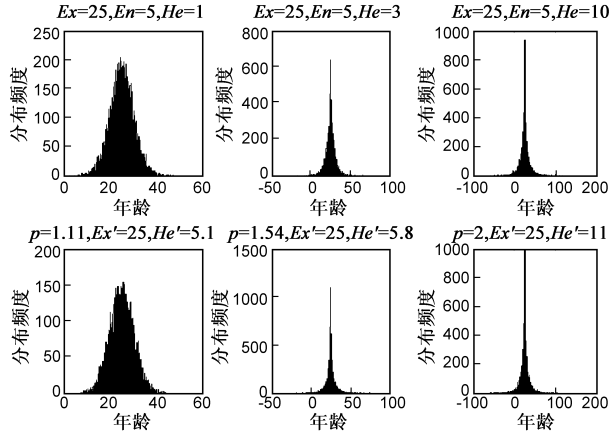

图4 云模型产生的云滴分布与典型高斯分布迭代之间的对比图

6 总结与展望

高阶高斯分布迭代及其期望、方差、四阶中心矩等数学性质是对高斯分布的扩展,建立了高斯分布与尖峰肥尾分布的桥梁. 本文的定义和研究方式是以一个方差 σ 作为初始的不确定性度量,层层迭代,每一层都是以上一层生成的随机样本作为方差,连同本层的期望生成一个随机样本,一方面作为本层的一个不确定性实现,另一方面作为下一层迭代的方差. 本文主要是从数学理论的角度对基于高阶高斯分布迭代的高阶云模型的数学特征进行了推导和分析,并针对产生的云滴样本进行了初步的统计分析,下一阶段将对高阶云模型及其应用研究展开进一步的工作.

致谢 感谢模糊集合专家王立新博士对本文研究内容尤其是数学性质证明的指导和帮助.

参考文献

[1] 王梓坤. 概率论基础及其应用[M]. 北京:北京师范大学出版社,1996.

[2] Embrechts P, Klupperberg C, Mikosch T. Modelling Extremal Events for Insurance and Finance [M]. Berlin: Springer-Verlag, 1997.

[3] Mandjes M. Overflow behavior in queues with many long-tailed inputs [J]. Journal of Applied Probability, 2000, 32(4): 1150 – 1167.

[4] Baltrunas A. Some asymptotic results for transient random walks with applications to insurance risk [J]. Journal of Applied Probability, 2001, 38(1): 108 – 121.

[5] Werner T, Upper C. Time variation in the tail behavior of bund futures returns [J]. Journal of Futures Markets, 2004, 24(4): 387 – 398.

[6] 李德毅,孟海军,史雪梅.隶属云和隶属云发生器[J].计算机研究与发展,1995,32(6):16 – 21.
Li deyi, Meng haijun, Shi Xuemei. Membership clouds and membership cloud generators [J]. Journal of Computer Research and Development, 1995, 32(6): 16 – 21. (in Chinese)

[7] 宋远俊,李德毅,杨孝宗,崔东华.电子产品可靠性的云模型评价方法[J].电子学报,2000,28(12):75 – 77.
SONG Yuan jun, LI De-yi, YANG Xiao-zong, CUI Dong-hua. Reliability evaluation of electronic products based on cloud models [J]. Acta Electronica Sinica, 2000, 28(12): 75 – 77. (in Chinese)

[8] 陈晖,李德毅,沈程智,张飞舟.云模型在倒立摆控制中的应用[J].计算机研究与发展,1999,36(10):1180 – 1187.
CHEN Hui, LI De Yi, SHEN Cheng Zhi, ZHANG Fei Zhou. A clouds model applied to controlling inverted pendulum [J]. Journal of Computer Research and Development, 1999, 36 (10): 1180 – 1187. (in Chinese)

[9] 杜▮,李德毅.基于云的概念划分及其在关联采掘上的应用[J].软件学报,2001,12(2):196-203.
DU Yi, LI De yi. Concept partition based on cloud and its application to mining association rules [J]. Journal of Software, 2001, 12(2): 196 – 203. (in Chinese)

[10] 王立新.正态云的基本数学性质及云滤波器[Z].个人通信,2011.5.30.

[11] 李德毅,杜▮.不确定性人工智能[M].北京:国防工业出版社,2005.

[12] 刘禹,李德毅,张光卫.云模型雾化特性及在进化算法中的应用[J].电子学报,2009,37(8):1651 – 1658.
Liu Yu, Li Deyi, Zhang Guangwei. Atomized feature in cloud based evolutionary algorithm [J]. Acta Electronica Sinica, 2009, 37(8): 1651 – 1658. (in Chinese)

作者简介

刘玉超　男.1980 年 12 月出生,山东烟台人.2005 年获解放军理工大学计算机专业硕士学位,中国电子系统工程研究所工程师,现为清华大学计算机科学与技术系博士研究生,从事云模型、粒计算、数据挖掘等方面的研究.
E-mail: yuchao_liu@163.com

马于涛　男.1980 年 3 月出生,湖北武汉人.2007 年获得武汉大学计算机专业博士学位,现为武汉大学软件工程国家重点实验室副教授,主要从事软件工程、云计算、复杂网络等方面的研究工作.

张海粟　男.1983 年生,安徽人.解放军理工大学计算机专业博士研究生,国防信息学院讲师,主要研究领域为数据挖掘、云计算、复杂网络.

陈桂生　男.1961 年生,湖南人.2007 年获得清华大学计算机专业博士学位,现为中国电子系统工程研究所高级工程师,主要研究领域为云模型、人工智能、复杂网络.