目录

摘要 ……1

Abstract ……1

第一章 引言 2

第二章 预备知识 3

2.1 盈余过程 3

2.2 随机数生成器 4

2.2.1 离散随机变量 4

2.2.2 连续随机变量 5

第三章 非参数估计理论介绍 6

3.1 核密度估计 6

3.1.1 介绍 6

3.1.2 统计性质 8

3.1.3 窗宽选择 12

3.2 离散关联核方法 14

3.2.1 介绍 14

3.2.2 统计性质 17

3.2.3 窗宽选择 19

3.3 数值模拟 21

3.3.1 连续随机变量 21

3.3.2 离散随机变量 23

第四章 模拟盈余过程 27

结论 33

参考文献 33

非参数方法在模拟保险公司破产概率中的应用

摘要：研究保险公司破产概率是风险论中的核心内容，而且计算破产概率有着重要现实意义。在以往的模型中，研究者们一直采用的都是参数方法。本文引入了两种非参数方法：核密度估计与离散关联核方法。核密度估计法是一种估计连续随机变量概率密度函数的方法。离散关联核方法是一种估计离散随机变量概率质量函数的方法。利用非参数方法可以不必对盈余过程中的随机变量在分布上进行假设，从而提高了适用性。随机模拟作为一种重要工具在科学研究中有着广泛的应用，利用随机模拟可以便捷地针对不同参数组合计算破产概率，计算结果可供以保险公司做理论与实际策略上的参考。

关键字：破产概率；核密度估计；离散关联核方法；随机模拟

Application of Non-parametric Method in Simulating Ruin Probabilities of Insurance Company.

Abstract: The research of ruin probabilities of insurance companies is one of the central topic in risk theory. Computing ruin probabilities also has crucial realistic meaning. In traditional models, researchers used parametric methods. Two non-parametric methods are introduced in this paper: kernel density estimation (KDE) and discrete associated kernel method (DAKE). KDE is a method designed for estimating probability density functions of continuous random variables. Its discrete counterpart, DAKE, is a method suitable for estimating probability mass functions of discrete random variables. Therefore, it's unnecessary to make assumptions about the distributions of random variables in the surplus process, which improves the applicability. Simulation, an important tool in scientific research, has a wide range of applications. Using simulation one can easily calculate ruin probabilities for different parameter combinations. The simulation results can serve the needs of theoretical and practical reference for insurance companies.

Keywords: Ruin Probabilities; Kernel Density Estimation; Discrete Associated Kernel Method; Simulation

第一章 引言

保险作为一种有效的风险管理方式已在全世界范围内被广泛接受。投保人通过向保险公司缴纳一定数量的保费来转移未来可能发生的风险。若在保期内发生事故造成财产生命损失，则保险公司需要向投保人支付合同商议的赔偿款。随着计算机的快速发展和商业需求的增加，保险运营中的一些重要环节需要精确的计算，其中破产概率就是一个重要方面。

经典的破产论是由瑞典精算师 Lundberg 于 1903 年首次提出[1]，在之后的百年中不断发展，相关理论结果可以参考 Asmussen 与 Albrecher 的论著 *Ruin probabilities*[2]。在计算和模拟破产概率时，传统的方法都会假定赔偿金额，索赔发生时间间隔以及保单到达时间间隔的具体分布，一般都假设为指数分布[3]。像这种假定数据服从某种给定函数形式的分布，而且该分布只由少数的参数决定的方法被称为*参数方法*。这种方法的一个极大的限制就是当样本的真实分布不符合假设分布时，会得到很差的结果。另一方面*非参数方法*并不限制分布的形式，更具有灵活性。本文的创新点就是将非参数方法引入到模拟破产概率这个经典问题中。

非参数核密度估计是一种用来估计连续型随机变量的概率密度函数的方法，最早由 Fix 和 Hodges 与 1951 年提出[4]，在之后的数十年有很多数学家提出了一些普适的算法和理论分析。在上世纪九十年代，Scott[5]和 Silverman[6]首先开始关注核密度估计在实际中的应用。在实际应用中，核密度估计中最重要的参数是窗宽(bandwidth)，它的确定一直是研究热点，因为它控制着核密度估计的效果。经典的经验准则由 Silverman 于 1986 年提出[7]，这种方法用正态分布作为真实分布来最小化积分均方误差(MISE)[8]。从而经验准则有一定的局限性。但是在近似正态分布的情况下，效果比较好。交叉验证法，最早由 Bowman 于 1984 年提出[9]，这种方法是选择将积分平方误差(ISE)最小化的窗宽，缺点是 ISE 方程常常有局部最小值，所以得到的窗宽难以保证是全局最优的。在 Jones 等人的综述中[10]，经验准则和交叉验证法被称为第一代窗宽选择方法，之后由 Sheather 和 Jones 在 1991 年提出的解方程嵌入法[11]被称为第二代方法，有着更好的表现。本文详细分析了核密度估计的统计性质以及上述三种窗宽选择方法并进行了数值实验。

离散关联核方法是一种用来估计离散型随机变量概率质量函数的方法。相较于核密度估计，离散关联核方法并不是一个热点研究课题，很多科学家都使用核密度估计中的连续核来进行概率质量函数估计，Marsh 和 Mukhopadhyay 于 1999 年提出用离散的泊松核[12]来估计概率质量函数但并没有产生太大的影响。之后在 2007 年 Kokonendji

等人提出了离散三角分布[13]，并在非参数概率质量函数估计中作为核函数。之后在2011 年[14]他们完整的阐述了离散关联核方法。同样的，离散关联核方法中最重要的问题也是窗宽选择。其中在 2011 年的论文中，受到核密度估计中交叉验证法的影响，Kokonendji 等人采用的是离散版本的交叉验证法。此方法一个严重的缺点是计算速度较慢，可从本文中的数值实验看出。之后在 2012 年，Zougab 等人提出了贝叶斯方法来计算窗宽[15]，此方法计算速度快，效果也好。本文中也对离散关联核方法的统计性质进行了分析并对比了上述两种窗宽选择方法。

本文首先介绍了破产模型与产生随机数的算法。接下来便对上面两种非参数方法进行详细讨论，并进行数值实验来对比窗宽选择方法。将上述两种非参数的方法引入后，本文将其应用在破产概率的模拟中以代替传统的参数方法，最后还计算了一个实例。

第二章 预备知识

2.1 盈余过程

对于任意的$t \geq 0$，令$U(t)$为保险公司在t时刻的盈余，$c(t)$为直到t时刻收取的保费，$S(t)$为直到t时刻的理赔量。假设在初始时刻，即$t = 0$时的盈余为u，则$U(t)$由下式给出：

$$U(t) = u + c(t) - S(t) \qquad t \geq 0. \tag{2.1}$$

因为理赔发生的时间和理赔的金额是不确定的，而且通常用随机变量来刻画，因此由随机过程的定义可知$\{U(t): t \geq 0\}$和$\{S(t): t \geq 0\}$是连续随机过程，分别称之为**盈余过程**和**总理赔过程**。

令$N(t)$表示截止到t时刻理赔发生的次数，假设当$t = 0$时，$N(t) = 0$，称$N(t): t \geq 0$为**理赔次数过程**。

令T_n表示第n次理赔发生的时间，则有：

$$N(t) = max\{n: T_n \leq t\}. \tag{2.2}$$

令$V_i = T_i - T_{i-1}$，表示第$i-1$与第i次理赔的时间间隔，则有：

$$T_i = V_1 + V_2 + \cdots + V_i, \tag{2.3}$$

其中$V_1, V_2, \cdots, V_i$是独立同分布的随机变量。由上述分析，可以把总理赔过程表达为：

$$S(t) = X_1 + X_2 + \cdots + X_{N_{(t)}}. \tag{2.4}$$

其中X_i表示第i次理赔产生的理赔额，且$X_1, X_2, \ldots, X_n$是独立同分布的随机变量。如果能够获取V_i和X_n的分布，我们就可以对总理赔过程过程进行模拟。

同样的，上述推理过程也适用于过程$c(t)$，只不过在通常的假设下，每笔保单的价格是固定的常数，若令$M(t)$表示截止到t时刻理赔发生的次数，并假设$M(0) = 0$。则有：

$$c(t) = a \times M(t), \tag{2.5}$$

上式中a是每笔保单的保费，$M(t)$可由下式表达：

$$M(t) = max\{m: W_1 + W_2 + \cdots + W_m \leq t\}, \tag{2.6}$$

其中W_i代表两次保单之间的时间间隔，是独立同分布的随机变量。

若在某一时刻t_0，出现了$U(t_0) < 0$，我们就称之为**破产**。

接下来令

$$T = min\{t: t \geq 0 \text{ and } U(t) < 0\}, \tag{2.7}$$

表示首次破产的时间。令

$$\psi(u) = \mathbb{P}(T < \infty), \tag{2.8}$$

表示破产概率，可以看出这是一个关于初始盈余的函数。进一步可以定义

$$\psi(u, t) = \mathbb{P}(T < t), \tag{2.9}$$

表示初始盈余为u时，在t之前发生破产的概率，显然这个量更加的具有现实意义。而且容易看出$\psi(u, t) \leq \psi(u)$.

本文将利用非参数的方法对V_i和X_n的概率分布进行估计，并模拟盈余过程，进而针对不同的初始值(u, t)计算破产概率。

2.2 随机数生成器

2.2.1 离散随机变量

设X是一离散随机变量，可能取的值为$\{0,1,2, \ldots, n\}$。$F_X(x)$是其分布函数，$p(x)$是其概率质量函数。下面的算法[16]可以产生服从$p(x)$的随机数。其中$uniform(0,1)$会产生一个服从$(0,1)$之间均匀分布的随机数。

Input : F_X
Output: A random number generated from given cdf F_X

$U \leftarrow uniform(0,1)$

$x \leftarrow 0$

while $F_X(x) < U$ **do**
| $x \leftarrow x + 1$
end
return x

算法 1 产生离散随机数

当算法终止时，$F_X(x) \geq U$且$F_X(x-1) < U$，因此$U \in (F_X(x-1), F_X(x)]$。由下式：

$$\mathbb{P}(X = x) = \mathbb{P}(U \in (F_X(x-1), F_X(x)]) = F_X(x) - F_X(x-1) = p(x), \tag{2.10}$$

可以证明当算法终止时，X的值就是一个服从给定概率质量函数的随机数。因此，对于**任意**给定的分布函数$F_X(x)$，利用上算法都可以产生服从该分布的随机数，而不仅限于常见的离散分布。

2.2.2 连续随机变量

设Y是一个连续随机变量，$F_Y(y)$是其分布函数，在文献[3]中，作者使用了反函数的方法来产生随机数，即：假设$F_Y^{-1}(y)$存在，则定义随机变量$Z = F_Y^{-1}(U)$，由下式：

$$F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(F_Y^{-1}(U) \leq y) = \mathbb{P}\big(U \leq F_Y(y)\big) = F_Y(y), \tag{2.11}$$

可知Z和Y服从相同的分布。但是这个方法有一定的局限性，即只有当分布函数的反函数解析表达式$F_Y^{-1}(y)$可以求出时，才可以使用这种办法。为了产生服从任意给定概率密度函数的随机数，下面引入一个新的方法：**拒绝方法(Rejection method)**。假设Y的概率密度函数为$f_Y(y)$，$f_Y(y)$仅在区间$[a,b]$上非零，且有 $f_Y(y) \leq k$，则可以按照如下算法[16]来产生随机数：

Input : fy, a, b, k
Output: A random number generated from given pdf f_y

while $TRUE$ **do**
  $y \leftarrow uniform(a, b)$
  $z \leftarrow uniform(0, k)$
  if $z < fy(y)$ **then**
    return y
  end
end

算法 2 产生连续随机数

假设点$P(Y', Z')$是一个二维的随机变量，在概率密度函数$f_Y(y)$与坐标横轴(Y')之间的区域内均匀分布。令R为在$f_Y(y)$下，且位于区间$[c, d]$内的区域，则有：

$$\begin{aligned}\mathbb{P}(c < Y' < d) &= \mathbb{P}(P\text{落在 R 内}) \\ &= \frac{Aera(R)}{Aera(f_Y(y))} \\ &= \frac{\int_c^d f_Y(y)dy}{1} \\ &= \int_c^d f_Y(y)dy,\end{aligned}$$

由概率密度函数的定义可知Y'和Y有相同的分布。这就证明了上述算法的正确性。

第三章 非参数估计理论介绍

3.1 核密度估计

3.1.1 介绍

核密度估计(Kernel Density Estimation，或 KDE)是用来估计连续随机变量的概率密度函数的一种方法，属于非参数估计理论。

假设$X_1, X_2, \cdots, X_n$是来自于某一总体的随机样本，且是独立同分布的。假设总体的概率密度函数为$f(x)$，则核密度估计的定义为：

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{3.1}$$

其中是h**窗宽**，$K(\cdot)$是**核函数**。

一个函数称之为核函数，如果满足如下三个条件：

$$K(x) \geq 0, \qquad -\infty < x < \infty, \tag{1}$$

$$K(-x) = K(x), \tag{2}$$

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{3}$$

根据公式(3.1)与上述核函数的定义，可以验证$\hat{f}(x)$的确是一个概率密度函数。首先根据核函数的性质(1)可得

$$\hat{f}(x) \geq 0, \qquad -\infty < x < \infty,$$

其次根据性质(3)可得：

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}(x)dx &= \int_{-\infty}^{\infty} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)dx \\ &= \sum_{i=1}^{n} \frac{1}{n}\int_{-\infty}^{\infty} K\left(\frac{x - X_i}{h}\right)d\frac{x}{h} \\ &= \sum_{i=1}^{n} \frac{1}{n} \\ &= 1.\end{aligned}$$

常见的核函数有下面几个：

$$\text{Epanechnikov: } \frac{3}{4}(1 - \mathrm{u}^2)\mathrm{I}(|\mathrm{u}| \leq 1),$$

$$\text{Gaussian: } \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\mathrm{u}^2\right),$$

$$\text{Triangle: } (1 - |\mathrm{u}|)\mathrm{I}(|\mathrm{u}| \leq 1),$$

$$\text{Uniform: } \frac{1}{2}\mathrm{I}(|\mathrm{u}| \leq 1).$$

下面是上述四个核函数的图像：

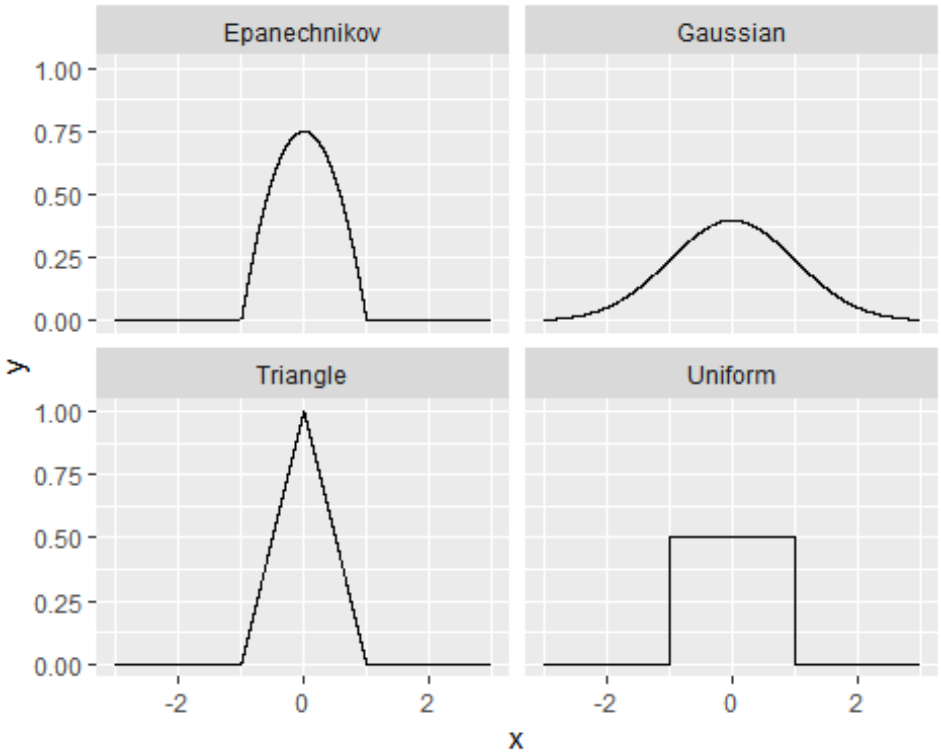

图 1 核函数图像

虽然核函数的种类有很多，但事实上选用不同的核对核密度估计的效果没有太大的影响，故本文不讨论核函数的选择，详细的讨论可以参看[17]6.2.3节。

3.1.2 统计性质

根据定义(3.1)可知，核密度估计是逐点计算的。比如给定一个点x_0，用

$$\hat{f}(x_0) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{h}\right),$$

来估计$f(x_0)$的值，也就是说这实际上在进行点估计。在下文，将会分析核密度估计的偏差，方差，均方误差等等。

下面首先考察核密度估计的**偏差(Bias)**：

$$\begin{aligned}\text{Bias}\big(\hat{f}(x_0)\big) &= \mathbb{E}\big[\hat{f}(x_0)\big] - f(x_0)\\ &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{x_0 - X_i}{h}\right)\right] - f(x_0)\\ &= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x_0 - x}{h}\right) f(x)dx - f(x_0) \qquad \left(\text{令} y = \frac{x_0 - x}{h}\right)\\ &= \int_{-\infty}^{\infty} K(y) f(x_0 - hy)dy - f(x_0),\end{aligned}$$

根据泰勒公式有：

$$f(x_0 - hy) = f(x_0) - f'(x_0)hy + \frac{1}{2}h^2y^2f''(x_0) + o(h^2), \tag{3.2}$$

将上式代入偏差计算公式可得：

$$\begin{aligned}\text{Bias}\big(\hat{f}(x_0)\big) &= \int_{-\infty}^{\infty} K(y) f(x_0 - hy)dy - f(x_0)\\ &= f(x_0)\int_{-\infty}^{\infty} K(y)dy - f'(x_0)h\int_{-\infty}^{\infty} K(y)ydy\\ &\quad + \frac{1}{2}h^2f''(x_0)\int_{-\infty}^{\infty} K(y)y^2dy - f(x_0) + o(h^2)\\ &= f(x_0) + \frac{1}{2}h^2f''(x_0)\int_{-\infty}^{\infty} K(y)y^2dy - f(x_0) + o(h^2)\\ &= \frac{1}{2}h^2f''(x_0)\mu(K) + o(h^2) \qquad \left(\text{令}\mu(K) = \int_{-\infty}^{\infty} K(y)y^2dy\right),\end{aligned}$$

故有：

$$\text{Bias}\left(\hat{f}(x_0)\right) = \frac{1}{2}h^2f''(x_0)\mu(K) + o(h^2), \tag{3.3}$$

根据上式可知核密度估计不是无偏估计，但h越小，则$\text{Bias}\big(\hat{f}(x_0)\big)$越小，而且当$h \to 0$时，$\text{Bias}\big(\hat{f}(x_0)\big) \to 0$。在(3.3)中，$\mu(K)$是一个只与核函数相关的值。另外一个事实是：在选定窗宽与核函数后，概率密度函数的曲率将会影响核密度估计的偏差。更具体的说，在曲率大（小）的点，使用核密度估计得到的偏差就更大（小）。

接下来计算核密度估计的**方差(Variance)**：

$$\begin{aligned}\mathbb{V}ar\left[\hat{f}(x_0)\right] &= \mathbb{V}ar\left[\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x_0-X_i}{h}\right)\right]\\ &= \frac{1}{nh^2}\mathbb{V}ar\left[K\left(\frac{x_0-X_i}{h}\right)\right]\\ &= \frac{1}{nh^2}\mathbb{E}\left[K^2\left(\frac{x_0-X_i}{h}\right)\right]-\frac{1}{n}\left(\frac{1}{h}\mathbb{E}\left[K\left(\frac{x_0-X_i}{h}\right)\right]\right)^2\\ &=: T_1+T_2,\end{aligned}$$

根据Bias$\left(\hat{f}(x_0)\right)$的推导，有:

$$\frac{1}{h}\mathbb{E}\left[K\left(\frac{x_0-X_i}{h}\right)\right]=f(x_0)+o(1),$$

因此可以得出：

$$T_2=O\left(\frac{1}{n}\right),$$

接下来计算T_1的值，首先计算：

$$\begin{aligned}\mathbb{E}\left[K^2\left(\frac{x_0-X_i}{h}\right)\right] &= \int_{-\infty}^{\infty}K^2\left(\frac{x_0-x}{h}\right)f(x)dx \qquad \left(令y=\frac{x_0-x}{h}\right)\\ &= h\int_{-\infty}^{\infty}K^2(y)f(x_0-hy)dy \qquad (由式(3.2))\\ &= hf(x_0)\int_{-\infty}^{\infty}K^2(y)dy+o(h^2)\\ &= hf(x_0)||K||_2^2+o(h^2)\\ &= hf(x_0)||K||_2^2+o(h),\end{aligned}$$

其中$||K||_2$表示$K(x)$的L_2范数。有了上面的结果就可以计算T_1了：

$$\begin{aligned}T_1 &= \frac{1}{nh^2}\mathbb{E}\left[K^2\left(\frac{x_0-X_i}{h}\right)\right]\\ &= \frac{1}{nh}f(x_0)||K||_2^2+\frac{1}{nh^2}o(h)\\ &= \frac{1}{nh}f(x_0)||K||_2^2+o\left(\frac{1}{nh}\right),\end{aligned}$$

故有：

$$\mathbb{V}ar\left[\hat{f}(x_0)\right]=\frac{1}{nh}f(x_0)||K||_2^2+o\left(\frac{1}{nh}\right), \tag{3.4}$$

由上式可知，核密度估计的方差与nh成反比，为了使方差比较小，需要选择较大的h，或者增加样本个数n。另一方面，在给定窗宽和核函数后，在概率密度大的点，方差也会较大。

下面将偏差与方差综合起来考虑：**均方误差(Mean Square Error)**。

在定义式(3.1)中，有两个参数需要人为设定，分别是窗宽和核函数。根据前面对偏差和方差的分析可知：

$$\begin{aligned} h\uparrow \quad &\rightarrow \quad \text{Bias}\left(\hat{f}(x_0)\right)\uparrow, \mathbb{V}ar\left[\hat{f}(x_0)\right]\downarrow, \\ h\downarrow \quad &\rightarrow \quad \text{Bias}\left(\hat{f}(x_0)\right)\downarrow, \mathbb{V}ar\left[\hat{f}(x_0)\right]\uparrow, \end{aligned}$$

所以如何确定h的大小就是一个重要的问题。最小化均方误差就是一个有效的方法。其定义为：

$$\text{MSE}\left(\hat{f}(x_0)\right) = \mathbb{E}\left[\left(\hat{f}(x_0) - f(x_0)\right)^2\right] = \text{Bias}^2\left(\hat{f}(x_0)\right) + \mathbb{V}ar\left[\hat{f}(x_0)\right], \tag{3.5}$$

把(3.3)和(3.4)代入上式可得：

$$\text{MSE}\left(\hat{f}(x_0)\right) = \frac{1}{4}h^4[f''(x_0)]^2\mu^2(K) + \frac{1}{nh}f(x_0)||K||_2^2 + o(h^4) + o\left(\frac{1}{nh}\right), \tag{3.6}$$

上式中前两项称为**渐进均方误差**，可以求出使其值最小的窗宽：

$$h_{opt}(x_0) = \left(\frac{1}{n}\cdot\frac{f(x_0)||K||_2^2}{[f''(x_0)]^2\mu^2(K)}\right)^{\frac{1}{5}}, \tag{3.7}$$

最优窗宽的值依赖于未知概率密度函数及其二阶导数，因此在实践中，最优窗宽是无法计算的，除非能找到$f(x)$和$f''(x)$的合适的代替。此外，公式(3.7)计算得到的窗宽是依赖于x_0的，也就是说是局部最优窗宽，接下来计算全局的最优窗宽。

将(3.6)积分得到**积分均方误差(Mean Integrated Square Error)**：

$$\begin{aligned} \text{MISE}(\hat{f}) \quad &= \int_{-\infty}^{\infty}\text{MSE}\left(\hat{f}(x)\right)dx \\ &= \frac{1}{4}h^4\mu^2(K)\int_{-\infty}^{\infty}[f''(x)]^2\,dx + \frac{1}{nh}||K||_2^2\int_{-\infty}^{\infty}f(x)dx \\ &+ o(h^4) + o\left(\frac{1}{nh}\right) \\ &= \frac{1}{4}h^4\mu^2(K)\left||f''||_2^2 + \frac{1}{nh}\right||K||_2^2 + o(h^4) + o\left(\frac{1}{nh}\right), \end{aligned} \tag{3.8}$$

类似的，(3.8)中的前两项称为**渐近积分均方误差(AMISE)**，全局最优窗宽由最小化AMISE得到：

$$h_{opt} = \left(\frac{1}{n} \cdot \frac{||K||_2^2}{\mu^2(K)||f''||_2^2}\right)^{\frac{1}{5}}, \tag{3.9}$$

由上式可知全局最优窗宽仍然依赖于未知概率密度函数的二阶导数，所以在实践中仍然无法得到理论上的最优值。

3.1.3 窗宽选择

窗宽选择基本上分为两类，一种是**代入方法**，另一种是**交叉验证法**。代入方法的主要思想就是把$||f''||_2^2$的估计式代入式(3.9)中，然后计算出h_{opt}。

经验法则(Rule of Thumb)

Silverman 于 1986 年[7]提出了**经验法则(Rule of Thumb)**，用正态分布$\mathcal{N}(\mu, \sigma^2)$来代替(3.9)中的$f$，然后选择用高斯核，计算得到：

$$\hat{h}_{rot1} = \frac{4}{3}\hat{\sigma}n^{-\frac{1}{5}}, \tag{3.10}$$

其中，$\hat{\sigma}$是样本方差。上式容易受到极端值的影响，这是因为一个极端值的出现就会导致样本标准差的重大改变。一个更稳健的方法利用**四分位距(Interquartile Range(IQR))**来估计标准差，依然假设样本服从正态分布，而随机变量$Y \sim \mathcal{N}(0,1)$，则有：

$$\begin{aligned} \text{IQR} &= X_{0.75n} - X_{0.25n} \\ &= (\mu + \sigma Y_{0.75n}) - (\mu + \sigma Y_{0.25n}) \\ &= (Y_{0.75n} - Y_{0.25n})\sigma \\ &\approx 1.34898\sigma, \end{aligned}$$

故可得：

$$\hat{\sigma} = \frac{IQR}{1.34898}, \tag{3.11}$$

把上式代入(3.10)中可得

$$\hat{h}_{rot2} = \frac{4}{3}\frac{IQR}{1.34898}n^{-\frac{1}{5}}, \tag{3.12}$$

结合式(3.10)和(3.12)可得：

$$\hat{h}_{rot} = \min\left(\hat{h}_{rot1}, \hat{h}_{rot2}\right). \tag{3.13}$$

上面介绍的经验法则适用于真实分布是对称的，且与正态分布的差别不是很大的样本数据。如果差异过大，则会得到效果很差的窗宽。下面介绍另一种更为普适，而且效果在一般情况下更好的方法。

解方程嵌入法(Solve-the-Equation Plug-In Approach)。

解方程嵌入法也是代入方法的一种，该方法由 S.J.Sheather 和 M.C.Jones 于 1991 年[11]提出。其核心思想是计算出(3.9)中$||f''||_2^2$的估计值然后解方程即可。具体描述如下。

设 $\hat{S}(\hat{\alpha}_2(h))$ 是$||f''||_2^2||$的估计值，则求下列方程的根即可得到窗宽$\hat{h}_{pl}$：

$$h - n^{-\frac{1}{5}} \frac{||K||_2^2}{\mu^2(K)\hat{S}\big(\hat{\alpha}_2(h)\big)} = 0, \tag{3.14}$$

其中：

$$\hat{S}(\alpha) = [n(n-1)]^{-1}\alpha^{-5} \sum_{i=1}^{n}\sum_{j=1}^{n} \phi^{\mathrm{iv}}\left[\alpha^{-1}\big(X_i - X_j\big)\right], \tag{3.15}$$

$$\hat{\alpha}_2(h) = 1.357 \frac{\hat{S}(a)^{\frac{1}{7}}}{\hat{T}(b)}\, h^{\frac{5}{7}}, \tag{3.16}$$

上式中：

$$\hat{T}(b) = -[n(n-1)]^{-1}b^{-7} \sum_{i=1}^{n}\sum_{j=1}^{n} \phi^{\mathrm{vi}}\left[b^{-1}\big(X_i - X_j\big)\right], \tag{3.17}$$

上式中ϕ代表标准正态分布。式(3.16)中

$$a = 0.920\mathrm{IQR}n^{-\frac{1}{7}}, \qquad b = 0.912\mathrm{IQR}n^{-\frac{1}{9}}. \tag{3.18}$$

交叉验证法

积分平方误差(Integrated Square Error)是另一个衡量估计概率密度函数与真实概率密度函数间距离的值，可以通过最小化积分平方误差来计算h，这种方法由 Bowman[9]提出。

积分平方误差定义如下：

$$\mathrm{ISE}(h) = \int [\hat{f}(x) - f(x)]^2 dx, \tag{3.19}$$

将上式展开后可得：

$$\mathrm{ISE}(h) = \int \hat{f}^2(x)dx - 2\int \hat{f}(x)f(x)dx + \int f^2(x)dx, \tag{3.20}$$

上式右边第三项是与h无关的值，第一项是可以根据数据直接求出，对于第二项有：

$$\int \hat{f}(x)f(x)dx = \mathbb{E}\big[\hat{f}(X)\big], \tag{3.21}$$

可以看到上式右边是$\hat{f}(X)$的均值，故可以用平均值来估计，记估计的均值为$\mathbb{E}[\widehat{\hat{f}(X)}]$，则有：

$$\mathbb{E}\left[\widehat{\hat{f}(X)}\right] = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i), \tag{3.22}$$

其中

$$\hat{f}_{-i}(x) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} K\left(\frac{x - X_j}{h}\right), \tag{3.23}$$

上面这个式子把X_i排除在外，目的是在式(3.22)中使函数$\hat{f}_{-i}$自身与自变量X_i的取值不相关。式(3.20)中的第一项可由下式计算：

$$\int \hat{f}^2(x)dx = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n}(K * K)\left(\frac{X_j - X_i}{h}\right), \tag{3.24}$$

上式中$(K * K)(t) = \int K(\tau)K(t-\tau)d\tau$。 把$(3.20) - (3.24)$联立可得：

$$\begin{aligned} \mathrm{CV}(h) &= \mathrm{ISE}(h) - \int f^2(x)dx \\ &= \frac{1}{nh^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K * K)\left(\frac{X_j - X_i}{h}\right) \\ &\quad - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} K\left(\frac{x - X_j}{h}\right), \end{aligned} \tag{3.25}$$

则由交叉验证法得到的窗宽由下式给出：

$$h_{cv} = arg\min_{h>0}\mathrm{CV}(h). \tag{3.26}$$

3.2 离散关联核方法

3.2.1 介绍

核密度估计是用来估计连续性随机变量的概率密度函数。为了估计离散型随机变量的概率质量函数，本节引入**离散关联核方法**。首先给出离散关联核的定义：

设概率质量函数f在集合$\mathbb{T}$上非零，x是$\mathbb{T}$中元素，h是窗宽。一个定义在支撑集$\mathbb{S}_x$上的概率质量函数$K_{x,h}(\cdot)$是离散关联核如果它满足下列条件：

$$x \in \mathbb{S}_x, \tag{1}$$

$$\lim_{h\to 0}\mathbb{E}\left(\mathcal{K}_{x,h}\right)=x, \tag{2}$$

$$\lim_{h\to 0}\mathbb{V}ar\left(\mathcal{K}_{x,h}\right)=0, \tag{3}$$

上面的条件中$\mathcal{K}_{x,h}$是对应于概率质量函数$K_{x,h}(\cdot)$的离散随机变量。

下面介绍几个离散关联核

例一：最简单的经验估计可以视为离散关联核的一个特例，定义：

$$K_{x,h}(y)=\mathbb{I}_{\{x\}}(y)=\begin{cases}1 & \text{if } y=x\\ 0 & \text{if } y\neq x\end{cases}\quad \forall y\in\mathbb{T}\quad \forall h\geq 0, \tag{3.27}$$

易知

$$x\in\mathbb{S}_x,\quad \mathbb{E}\left(\mathcal{K}_{x,h}\right)=x,\quad \mathbb{V}ar\left(\mathcal{K}_{x,h}\right)=0,$$

满足定义中的三个条件，故(3.27)是离散关联核。

例二(Discrete Triangular)：对称离散三角核由 Kokonendji 等[13]提出，其中$\mathbb{T}$可以是无界的，比如$\mathbb{N},\mathbb{Z}$，也可以是有界的，如$\{0,1\cdots,N\}$，对$\forall x\in\mathbb{T}$，定义随机变量$\mathcal{T}_{x,h}$，支撑集为$S_x=\{x,x\pm 1,x\pm 2,\cdots,x\pm a\}$，其中$a\in\mathbb{N}$是一个固定值。其概率密度函数由下式给出：

$$T_{x,h}(y)=\frac{(a+1)^h-|y-x|^h}{P(a,h)}\mathbb{I}_{\mathbb{S}_x}(y), \tag{3.28}$$

上式中$P(a,h)=(2a+1)(a+1)^h-2\sum_{k=0}^{a}k^h$是正则化系数。

例三(DiracDU)：Aitchison 和 Aitken[18]提出过一个适用于分类数据的离散分布，它可视为例一的拓展，其中$\mathbb{T}=\{0,1,\cdots,c-1\}$，而$c\in\mathbb{N}\setminus\{0,1\}$，定义随机变量$\mathcal{A}_{x,h}$，支撑集$\mathbb{S}_x=\mathbb{T}$，概率密度函数由下式给出：

$$A_{x,h}(y)=(1-h)\mathbb{I}_{\{x\}}(y)+\frac{h}{c-1}\mathbb{I}_{\mathbb{S}_x\setminus\{x\}}(y), \tag{3.29}$$

可以看出，当$h=0$时，就是例一介绍中的核。

若将离散关联核中的条件(3)换成

$$\lim_{h\to 0}\mathbb{V}ar\left(\mathcal{K}_{x,h}\right)=\mathcal{V}(0), \tag{3.30}$$

其中$\mathcal{V}(0)$是一个不依赖于x的 0 的邻域。加上条件(1),(2)，满足这三个条件的核称为**一阶**或**标准离散核**。这类核里面表现最好的是**二项核(Binomial Kernel)**。

例四(Binomial)：二项核[14]定义在$\mathbb{S}_x = \{0,1,\cdots,x+1\}$上，其中$x \in \mathbb{T}$，$\mathbb{T}$是自然数集，而$h \in (0,1]$，则有：

$$B_{x,h}(y) = \binom{x+1}{y}\left(\frac{x+h}{x+1}\right)^{y}\left(\frac{1-h}{x+1}\right)^{x+1-y}\mathbb{I}_{\mathbb{S}_x}(y), \tag{3.31}$$

记对应于上述概率质量函数的随机变量是$\mathcal{B}_{x,h}$，可得：

$$x \in \mathbb{S}_x, \quad \mathbb{E}\left(\mathcal{B}_{x,h}\right) = x+h, \quad \mathbb{V}ar\left(\mathcal{B}_{x,h}\right) = \frac{1-h}{x+1}(x+h).$$

下面是上述四个核函数当$x = 5, h = 0.1$时的图像

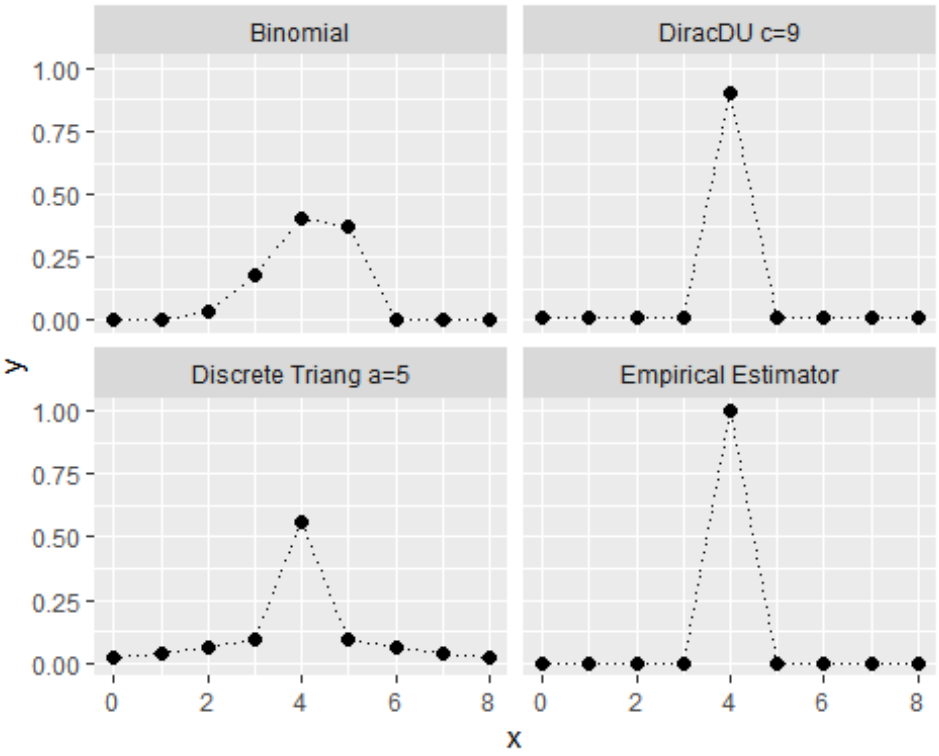

图 2 离散关联核图像

在了解离散关联核之后，下面给出离散关联核方法的定义。

设$X_1, X_2, \cdots, X_n$是来自于某一离散分布的随机样本，且是独立同分布的。假设总体的概率质量函数为$f(x)$，则离散关联核方法为：

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i), \tag{3.32}$$

其中$K_{x,h}(\cdot)$就是上面介绍的离散关联核。

需要注意的一点是式(3.32)并不是一个概率密度函数，因为没有正则化。但有如下结论

$$\sum_{x\in\mathbb{T}} \hat{f}\,(x) = C, \tag{3.33}$$

其中$C = C(n;h,K)$是一个正有限数，如果满足如下条件

$$\sum_{x\in\mathbb{T}} K_{x,h}\,(y) < \infty \quad \forall y \in \mathbb{T}, \tag{3.34}$$

这是因为把式(3.33)展开有：

$$C = \frac{1}{n}\sum_{i=1}^{n}\sum_{x\in\mathbb{T}} K_{x,h}\,(X_i),$$

而$K_{x,h}(y) > 0 \quad \forall y \in \mathbb{T} \cap \mathbb{S}_x$，故$C > 0$。又因为条件(3.34)可知$C < \infty$。

在一般的情况下，$C \neq 1$，但是可以计算出来，然后再用(3.32)除以这个常量就得到了正则化的概率质量函数了。故在下列讨论中，假设$C = 1$。

3.2.2 统计性质

设f的支撑集$\mathbb{T} = \mathbb{N}$，则其k阶有限差分$f^k(x)$有如下定义：

$$\begin{aligned} f^{(k)}(x) &= \left\{f^{(k-1)}(x)\right\}^{(1)} \\ f^{(1)}(x) &= \begin{cases} f(x+1) - f(x-1)/2 & \text{if } x \in \mathbb{N}\setminus\{0\} \\ f(1) - f(0) & \text{if } x = 0 \end{cases}, \end{aligned}$$

则有离散泰勒展开(见[19]第 351 页)：

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k + o((x-a)^k), \tag{3.35}$$

上式为f在任意一点$a \in \mathbb{T}$的展开。

类似的，接下来考察离散核密度方法的**逐点**统计性质。

首先计算**偏差**：

$$\begin{aligned}\text{Bias}\left(\hat{f}(x)\right) &= \mathbb{E}\left[\hat{f}(x)\right] - f(x)\\ &= \mathbb{E}\left[K_{x,h}(X_i)\right] - f(x)\\ &= \sum_{y\in\mathbb{T}\cap\mathbb{S}_x} f\,(y)K_{x,h}(y) - f(x)\\ &= \sum_{y\in\mathbb{T}\cap\mathbb{S}_x} f\,(y)\mathbb{P}\left(\mathcal{K}_{x,h}=y\right) - f(x)\\ &= \mathbb{E}\left[f\left(\mathcal{K}_{x,h}\right)\right] - f(x),\end{aligned} \tag{3.36}$$

接下来利用式(3.35)，把$f\left(\mathcal{K}_{x,h}\right)$在点$\mathbb{E}\left[\mathcal{K}_{x,h}\right]$展开可得：

$$\begin{aligned}f\left(\mathcal{K}_{x,h}\right) &= f\left(\mathbb{E}\left[\mathcal{K}_{x,h}\right]\right) + f^{(1)}\left(\mathbb{E}\left[\mathcal{K}_{x,h}\right]\right)\left(\mathcal{K}_{x,h} - \mathbb{E}\left[\mathcal{K}_{x,h}\right]\right)\\ &\quad + \frac{1}{2}f^{(2)}\left(\mathbb{E}\left[\mathcal{K}_{x,h}\right]\right)\left(\mathcal{K}_{x,h} - \mathbb{E}\left[\mathcal{K}_{x,h}\right]\right)^2 + o\left(\left(\mathcal{K}_{x,h} - \mathbb{E}\left[\mathcal{K}_{x,h}\right]\right)^2\right),\end{aligned} \tag{3.37}$$

把上式代入(3.36)中可得

$$\text{Bias}\left(\hat{f}(x)\right) = f\left(\mathbb{E}\left[\mathcal{K}_{x,h}\right]\right) + \frac{1}{2}\mathbb{V}ar\left(\mathcal{K}_{x,h}\right)f^{(2)}(x) - f(x) + o(h). \tag{3.38}$$

接下来计算**方差**：

$$\begin{aligned}\mathbb{V}ar\left(\hat{f}(x)\right) &= \frac{1}{n}\mathbb{V}ar\left(K_{x,h}(X_i)\right)\\ &= \frac{1}{n}\left\{\mathbb{E}\left[K_{x,h}^2(X_i)\right] - \mathbb{E}^2\left[K_{x,h}(X_i)\right]\right\}\\ &= \frac{1}{n}\left\{\sum_{y\in\mathbb{T}\cap\mathbb{S}_x} f\,(y)\left[\mathbb{P}\left(\mathcal{K}_{x,h}=y\right)\right]^2 - \left[\sum_{y\in\mathbb{T}\cap\mathbb{S}_x} f\,(y)\mathbb{P}\left(\mathcal{K}_{x,h}=y\right)\right]^2\right\}\\ &= \frac{1}{n}f(x)\left[\mathbb{P}\left(\mathcal{K}_{x,h}=x\right)\right]^2 - \frac{1}{n}f^2(x) + R_{x,h},\end{aligned}$$

上式中：

$$\begin{aligned}R_{x,h} &= \frac{1}{n}\sum_{y\in\mathbb{T}\cap\mathbb{S}_x\setminus\{x\}} f\,(y)\left[\mathbb{P}\left(\mathcal{K}_{x,h}=y\right)\right]^2 + \frac{1}{n}f^2(x)\\ &\quad - \frac{1}{n}\left\{f(x) + \sum_{y\in\mathbb{T}\cap\mathbb{S}_x}\left[f(y) - f(x)\right]\mathbb{P}(\mathcal{K}_{x,h}=y)\right\}^2,\end{aligned} \tag{3.39}$$

下证$R_{x,h} = o\left(\frac{1}{n}\right)$: 对$\forall y \in \mathbb{S}_x \setminus \{x\}$，都可以找到一个关于$y$的常数$\eta(y)$使下式成立

$$
\begin{aligned}
0 \quad &\le \mathbb{P}(\mathcal{K}_{x,h} = y) \\
&\le \mathbb{P}\big(|\mathcal{K}_{x,h} - x| \ge \eta(y)\big) \\
&\le \frac{1}{\eta^2(y)}\mathbb{E}[(\mathcal{K}_{x,h} - x)^2] \\
&= \frac{1}{\eta^2(y)}\Big[\mathbb{V}ar(\mathcal{K}_{x,h}) + \big\{\mathbb{E}[\mathcal{K}_{x,h}] - x\big\}^2\Big] \\
&\to 0 \quad \text{当 } h \to 0,
\end{aligned}
$$

由上面的推导可知，当$h \to 0$时，$\mathbb{P}(\mathcal{K}_{x,h} = x) \to 1$。假设当$n \to \infty$时，$h = h(n) \to 0$，则有

$$
\begin{aligned}
\frac{R_{x,h}}{1/n} \quad &= \sum_{y \in \mathbb{T} \cap \mathbb{S}_x \setminus \{x\}} f\,(y)\big[\mathbb{P}(\mathcal{K}_{x,h} = y)\big]^2 + f^2(x) \\
&\quad - \left\{ f(x) + \sum_{y \in \mathbb{T} \cap \mathbb{S}_x} [f(y) - f(x)]\mathbb{P}(\mathcal{K}_{x,h} = y) \right\}^2 \\
&\to f^2(x) - f^2(x) = 0,
\end{aligned}
$$

因此$R_{x,h} = o\left(\frac{1}{n}\right)$，故有

$$
\mathbb{V}ar\big[\hat{f}(x)\big] = \frac{1}{n} f(x)\left\{\big[\mathbb{P}\big(\mathcal{K}_{x,h} = x\big)\big]^2 - f(x)\right\} + o\left(\frac{1}{n}\right). \tag{3.40}
$$

根据式(3.5)可计算离散核密度估计的均方误差(MSE)，这是一个局部的误差值，为了得到全局误差，可以计算积分均方误差：

$$
\begin{aligned}
\mathrm{MISE}\big(\hat{f}\big) \quad &= \sum_{x \in \mathbb{T}} \mathrm{MSE}\left(\hat{f}(x)\right) \\
&= \sum_{x \in \mathbb{T}} \left[f\big(\mathbb{E}\big[\mathcal{K}_{x,h}\big]\big) + \frac{1}{2}\mathbb{V}ar\big(\mathcal{K}_{x,h}\big) f^{(2)}(x) - f(x)\right]^2 \\
&\quad + \sum_{x \in \mathbb{T}} \frac{1}{n} f(x)\left\{\big[\mathbb{P}\big(\mathcal{K}_{x,h} = x\big)\big]^2 - f(x)\right\} + o\left(\frac{1}{n} + h^2\right),
\end{aligned} \tag{3.41}
$$

由前面关于偏差和方差的讨论可以得到$\mathrm{MISE} \to 0$，当$n \to \infty$时。

3.2.3 窗宽选择

交叉验证法

与连续情况相同，离散关联核方法也可以用交叉验证法来求最优窗宽，不同的是连续情况是积分，在离散情况下则是求和，具体描述如下。

离散情况的 ISE 定义如下：

$$\mathrm{ISE}(h) = \sum_{x\in\mathbb{T}} [\hat{f}(x) - f(x)]^2, \tag{3.42}$$

接下来的分析与连续情况的完全相同，只要把积分符号换成求和符号即可。离散版本的为：

$$\begin{aligned} \mathrm{CV}(h) \quad &= \mathrm{ISE}(h) - \sum_{x\in\mathbb{T}} f^2(x) \\ &= \sum_{x\in\mathbb{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} K_{x,h}(X_i) \right\}^2 - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} K_{X_i,h}(X_j), \end{aligned} \tag{3.43}$$

则由交叉验证法得到的窗宽由下式给出：

$$h_{cv} = arg\min_{h>0} \mathrm{CV}(h). \tag{3.44}$$

贝叶斯方法

之前介绍的各种选择窗宽的方法都视h为一固定的数，然而贝叶斯方法视h为一随机变量而且是与要估计的点x相关的随机变量，因此这个方法得到的窗宽不再是一个数，而是一个向量。

假设h的先验分布为$\pi(h)$，后验分布为$\pi(h|x)$。为了计算后验分布，需要求出**似然函数(Likelihood Function)**，因为x是根据分布$f(x)$产生的，但是h并不是$f(x)$的一个参数，故需要构造一个$f(x)$的近似函数，同时h又是该近似函数的参数：

$$f_h(x) = \sum_{y\in\mathbb{T}} f(y) K_{x,h}(y) = \mathbb{E}\left[K_{x,h}(Y)\right], \tag{3.45}$$

上式中Y是服从$f(x)$的随机变量。由前面关于偏差的推导可知，当$n \to \infty$时，$f_h(x)$是$f(x)$的一个很好的近似。如此便可得到似然函数：

$$L(x|h) = f_h(x), \tag{3.46}$$

根据贝叶斯公式可计算出后验分布：

$$\pi(h|x) = \frac{L(x|h)\pi(h)}{\int L(x|h)\pi(h)dh}, \tag{3.47}$$

但实际上$f_h(x)$的值也是计算不出来的，好在$f_h(x)$是一个均值，可以用平均值来估计。根据离散关联核方法的定义式(3.32)可知$\hat{f}(x)$就是$\mathbb{E}[K_{x,h}(Y)]$的平均值估计。在下式中为了强调h，把(3.32)中的$\hat{f}(x)$记为$\hat{f}_h(x)$。因此可以计算出后验分布的估计值：

$$\hat{\pi}(h|x)=\frac{\hat{f}_h(x)\pi(h)}{\int \hat{f}_h(x)\pi(h)dh}. \tag{3.48}$$

在平方误差下，最优窗宽由后验分布的均值给出：

$$\hat{h}(x)=\int h\hat{\pi}(h|x)dh, \tag{3.49}$$

详情可参见 PRML[20]的1.5.5小节。在特定的情况下，后验分布(3.48)和最优窗宽(3.49)可以算出解析式。Zougab 等人[15]计算出了使用二项核时的结果，二项核的定义可见(3.31)，先验分布$\pi(h)$选取 Beta 分布：

$$\pi(h)=\frac{1}{B(\alpha,\beta)}h^{\alpha-1}(1-h)^{\beta-1}, \tag{3.50}$$

其中：

$$B(\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}, \tag{3.51}$$

上式中$\Gamma(\cdot)$是 Gamma 函数。在此情况下可以得到最优窗宽：

$$\hat{h}(x)=\frac{\sum_{i=1}^{n}\sum_{k=0}^{X_i}\frac{x^k B(X_i+\alpha-k+1,x+\beta+1-X_i)}{(x+1-X_i)!\,k!\,(X_i-k)!}}{\sum_{i=1}^{n}\sum_{k=0}^{X_i}\frac{x^k B(X_i+\alpha-k,x+\beta+1-X_i)}{(x+1-X_i)!\,k!\,(X_i-k)!}}, \tag{3.52}$$

上式中，α,β是超参数，需要人为设定，不能从数据中计算出来。Zougab 等人提出的经验值为$\alpha=0.5,\beta=15$。

3.3 数值模拟

本节将应用前面所介绍的核密度估计与离散关联核方法来做数值模拟，以比较不同窗宽选择方法在实际情况中的表现。在下面两节中，将分别就连续的情况与离散情况进行数值模拟。

3.3.1 连续随机变量

选择指数分布为真实分布，并令$\lambda=1$，概率密度函数记为$f(x)$，则有：

$$f(x)=e^{-x}\qquad x>0. \tag{3.53}$$

记每次采样的样本个数为n，设N_{sim}为采样次数，这里定为 50 次。针对不同的n，将分别从上述指数分布中采样 50 次，在下面的实验中n将分别取$\{50,100,200,400,800,1000\}$。将经验法则，解方程嵌入法和交叉验证法求出来的窗宽分别记为h_{rot}，h_{pl}和h_{cv}。在下面的对比中，将采用渐进积分均方误差作为准则：

$$\text{AMISE} = \frac{1}{4}h^4\mu^2(K)\left||f''||_2^2 + \frac{1}{nh}\right||K||_2^2, \tag{3.54}$$

上式中除了h以外所有的值都可计算出来，包括$||f''||_2^2$：

$$||f''||_2^2 = \frac{1}{2},$$

因为此时真实分布是已知的。表 1 是数值模拟所得到的数据，AMISE_{rot}，AMISE_{pl}和AMISE_{cv}分别代表使用经验法则，解方程嵌入法和交叉验证法得到的误差，为了提高准确度。表格中的结果是计算 50 次后求的均值。可以看出就误差而言经验法则和解方程嵌入法明显比交叉验证法效果好，其中经验法则又比解方程嵌入法要好，但是区别并不明显。

n	AMISE_{rot}	AMISE_{pl}	AMISE_{cv}
50	0.0194154	0.0232440	0.0413755
100	0.0111105	0.0142996	0.0276860
200	0.0063974	0.0084600	0.0187823
400	0.0036073	0.0053141	0.0130824
800	0.0020696	0.0033057	0.0088048
1000	0.0017233	0.0028408	0.0076841

表格 1 渐进积分误差对比

表 2 是窗宽h的方差对比。因为同一分布进行 50 次采样，所以计算出的h是有变化的，这种变化程度大小可视为计算窗宽方法的稳定性。从中可以看出稳定性最好的是解方程嵌入法而且其方差只是经验法则的五分之一左右，稳定性最差的是交叉验证法。

n	$Var(h_{rot})$	$Var(h_{pl})$	$Var(h_{cv})$	$\frac{Var(h_{rot})}{Var(h_{pl})}$
50	0.0076279	0.0017618	0.0068728	4.329633
100	0.0036151	0.0007495	0.0021738	4.823182
200	0.0014681	0.0002494	0.0008468	5.886168
400	0.0004448	0.0000876	0.0003500	5.077086
800	0.0002085	0.0000273	0.0001049	7.636338
1000	0.0001192	0.0000174	0.0000744	6.871620

表格 2 方差对比

3.3.2 离散随机变量

选择泊松分布为真实分布，令$\lambda = 8$，并记概率质量函数为$f(x)$，则有：

$$f(x) = \frac{8^x e^8}{x!} \qquad x \in \mathbb{N}. \tag{3.55}$$

记每次采样的样本个数为n，设N_{sim}为采样次数，这里定为 50 次。针对不同的n，将分别从上述泊松分布中采样 50 次，在下面的实验中n将分别取$\{25,50,100,200,400,800\}$。将交叉验证法和贝叶斯方法求出来的窗宽分别记为$h_{cv}$和$h_{bayes}$。在下面的对比中，将采用积分均方误差的估计值作为标准：

$$\widehat{\mathrm{MISE}} = \frac{1}{N_{sim}} \sum_{t=1}^{N_{sim}} \sum_{x \in \mathbb{T}} [\hat{f}_h^{[t]}(x) - f(x)]^2, \tag{3.56}$$

上式中$\hat{f}_h^{[t]}(x)$表示以第t次采样得到的结果为数据，以h为窗宽，以二项核为核函数计算出来的密度估计值。

表 3 是数值模拟所得到的数据，$\widehat{MISE}_{bayes}$与$\widehat{MISE}_{cv}$分别代表使用贝叶斯方法得到的误差和使用交叉验证法得到的误差。可以发现，当样本数量相对较小时$(n = 25,50,100)$，贝叶斯方法的误差更小，当样本数量相对较大时$n = 200,400,800$，交叉验证法的误差更小，但是差距并不明显，可以从表格第四列可以看出来。其中 RED 的定义如下：

$$\mathrm{RED} = \frac{\widehat{MISE}_{CV} - \widehat{MISE}_{bayes}}{\widehat{MISE}_{bayes}} \times 100\%. \tag{3.57}$$

n	$\widehat{MISE}_{bayes}$	$\widehat{MISE}_{CV}$	RED
25	0.0079629	0.0096125	20.7%
50	0.0038683	0.0040914	5.8%
100	0.0024465	0.0024513	0.2%
200	0.0011867	0.0011819	-0.4%
400	0.0007080	0.0006977	-1.5%
800	0.0004948	0.0004826	-2.5%

表格 3 积分均方误差对比

表 4 是贝叶斯方法与交叉验证法求窗宽所需时间的对比，可以看到，当样本量比较小的时候($n = 25,50$)，贝叶斯方法比交叉验证法要快几百倍，当样本量继续增加，贝叶斯方法比交叉验证法快了几千倍。由于两种方法精度相差不大，但时间效率相差极大，故在后面实例分析中将采用贝叶斯方法来求窗宽。

n	t_{bayes}	t_{CV}	$\frac{t_{CV}}{t_{bayes}}$
25	0.0035007s	1.359394s	388.3198
50	0.0081762s	5.206485s	636.7823
100	0.0164803s	17.482640s	1060.8199
200	0.0419147s	55.361410s	1320.8129
400	0.0808404s	206.322200s	2552.2168
800	0.1895512s	742.719400s	3918.3049

表格 4 运行时间对比

下面三张图直观的展示了运行时间[1]的差别。图 3 是贝叶斯方法运行时间图，可以看出运行时间随样本量线性增加。图 4 是交叉验证法运行时间图，运行时间则是随样本量平方增长。图 5 将两种方法运行时间放在一起比较，可以看出两种方法在运行时间上不是同一量级。

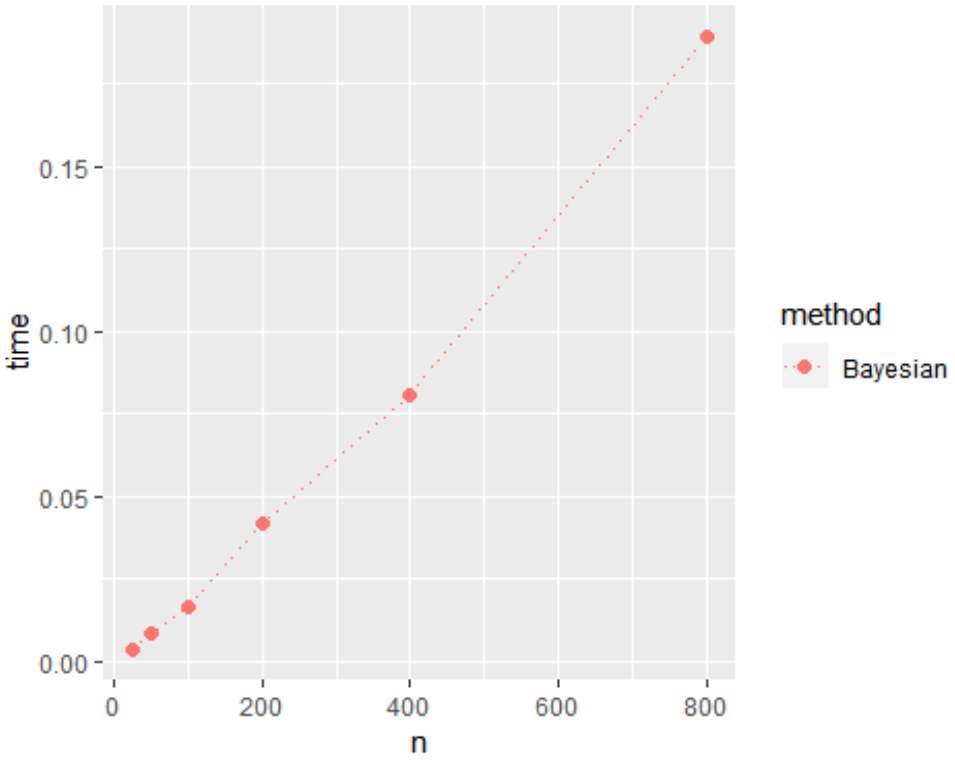

图 3 贝叶斯方法运行时间图

[1] 实验所用机器：CPU 最大频率 3.2GHz，四核四线程，RAM: 12G；R 语言版本 3.5.1；操作系统 Win10 x64。

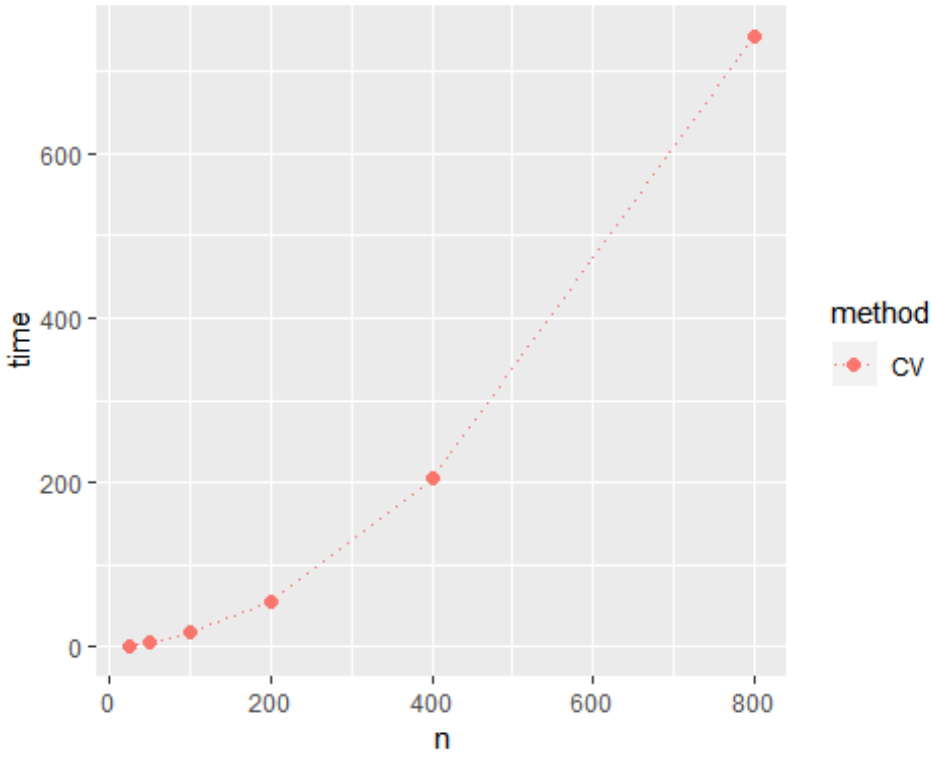

图 4 交叉验证法运行时间图

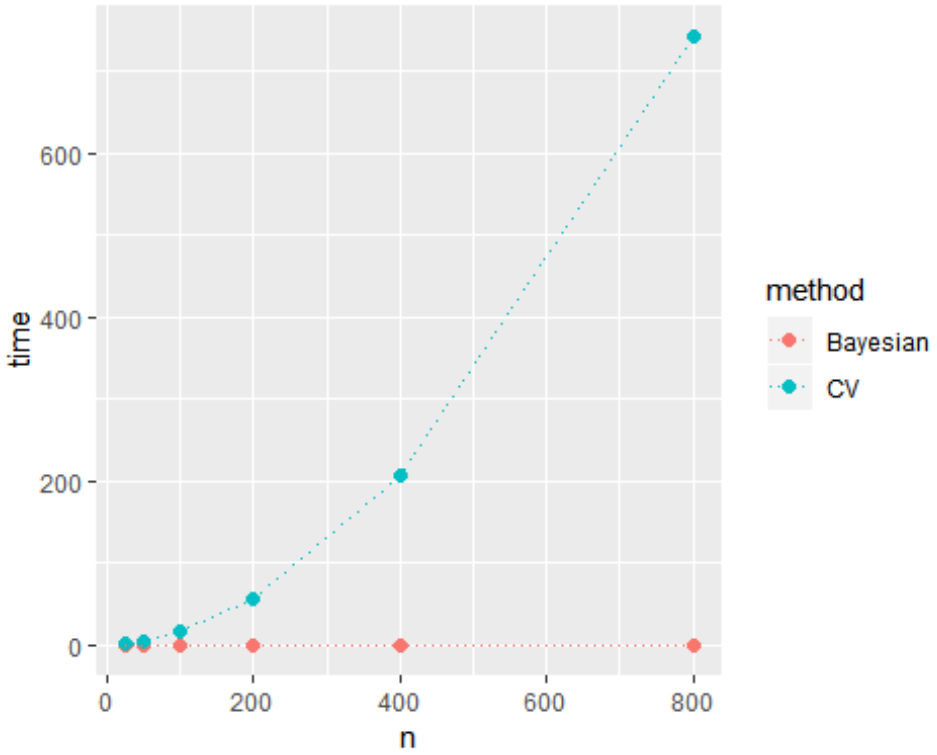

图 5 贝叶斯方法与交叉验证法运行时间对比图

第四章 模拟盈余过程

本节将应用非参数方法对盈余过程进行模拟。

盈余过程的定义式为：

$$U(t) = u + c(t) - S(t),$$

根据盈余过程一节的分析，并假设每天收取的保费为固定值c，则上式可以写为：

$$U(t) = u + ct - \sum_{j=1}^{N(t)} X_j\,, \tag{4.1}$$

在上式中，由盈余过程一节的分析可知，需要模拟两个总体的随机数：理赔发生时间间隔V，还有理赔金额X。

对于理赔金额X，利用历史数据，可以用核密度估计的方法来计算概率密度函数，并记为$claim_amount_pdf$，然后根据随机数生成器一节中介绍的拒绝方法(算法 2)来产生服从概率密度函数$claim_amount_pdf$的随机数，即理赔金额。记上述产生随机理赔额的函数为$continuous_rng$，该函数就是算法 2 的一个实现。

另一方面，对于理赔时间间隔V，采用关联核密度估计的方法，可以估计出T的概率质量函数pmf，进一步求出它的累积分布函数cdf并记之为$claim_interval_cdf$，然后利用随机数生成器一节中介绍的算法 1，来产生随机数。记该随机数发生器为$discrete_rng$。

假设初始资金为u，每日收取保费为c，停止模拟盈余过程的时间为t_stop，模拟盈余过程的次数为num，K为其中破产的次数。在模拟开始的时候，记录下保单到达的时间与索赔到达的时间，分别为t_policy, t_claim，其中索赔到达的时间由理赔时间间隔的随机数发生器$discrete_rng$产生。若保单先到达，则公司资金将增加c，并记录下次保单到达的时间。如索赔先到达，则公司将损失一定的财产，具体的金额将由产生随机理赔额的函数$continuous_rng$算出，此时需要检测公司是否处于负资产状态，如果处于负资产状态，则意味着破产，那么破产次数K将增加 1，如果没有处于负资产状态，则继续模拟此次盈余过程。最终的破产概率由K/num计算得出。完成的算法描述见算法 3。

进一步假设该公司设定了一个红利边界 $barrier$，当盈余 S 超出 $barrier$ 时，超出的那部分盈余作为红利分给股东。在这种情况下，只需在赛算法 3 中增加一个 if 语句以在每次收取保费后判断此时盈余是否已经达到红利边界 $barrier$，若达到，则将盈余设为 $barrier$。详情可见算法 4。

Input : $num, c, t\_stop, u, continuous\_rng, claim\_amount\_pdf$
$, a, b, k, discrete\_rng, claim\_interval\_cdf$
Output: Ruin probability

$K \leftarrow 0$
for *m in 1:num* **do**
  $t\_policy \leftarrow 1$
  $t\_claim \leftarrow discrete\_rng(claim\_interval\_cdf)$
  $S \leftarrow u$
  $t \leftarrow t\_stop$
  while $min(t\_policy, t\_claim) \leq t$ **do**
    $t\_first \leftarrow min(t\_policy, t\_claim)$
    if $t\_policy \leq t\_claim$ **then**
      $S \leftarrow S + c$
      $t\_policy \leftarrow t\_first + 1$
    else
      $S \leftarrow S - continuous\_rng(claim\_amount\_pdf, a, b, k)$
      $t\_claim \leftarrow t\_first + discrete\_rng(claim\_interval\_cdf)$
    end
    if $S < 0$ **then**
      $K \leftarrow K + 1$
      break
    end
  end
end
return K/num

算法 3 模拟破产概率

Input : $num, c, t\_stop, u, continuous\_rng, claim\_amount\_pdf$
$, a, b, k, discrete\_rng, claim\_interval\_cdf, barrier$
Output: Ruin probability with surplus bound

$K \leftarrow 0$
for *m in 1:num* **do**
  $t\_policy \leftarrow 1$
  $t\_claim \leftarrow discrete\_rng(claim\_interval\_cdf)$
  $S \leftarrow u$
  $t \leftarrow t\_stop$
  while $min(t\_policy, t\_claim) \leq t$ **do**
    $t\_first \leftarrow min(t\_policy, t\_claim)$
    if $t\_policy \leq t\_claim$ **then**
      $S \leftarrow S + c$
      if $S > barrier$ **then**
        $S \leftarrow barrier$
      end
      $t\_policy \leftarrow t\_first + 1$
    else
      $S \leftarrow S - continuous\_rng(claim\_amount\_pdf, a, b, k)$
      $t\_claim \leftarrow t\_first + discrete\_rng(claim\_interval\_cdf)$
    end
    if $S < 0$ **then**
      $K \leftarrow K + 1$
      break
    end
  end
end
return K/num

算法 4 有红利上界时模拟破产概率

下面是一个具体的算例。

现有某保险公司在某地区的车险记录，其中包括了819个理赔金额数据，和818个理赔时间间隔数据，每天收取的保费为 7.8 万元。记理赔金额数据为$claim$，记理赔时间间隔为$time_interval$。前六个数据见表 5。从表中可以看出，索赔金额是连续型随机变量，而理赔时间间隔应该是离散型随机变量。因此将分别应用核密度估计与离散关联核方法来估计它们的分布。

理赔金额	52.2882	74.3657	48.5505	38.9804	51.2097	50.6000
理赔间隔	15	38	3	20	5	4

表格 5： 理赔金额与时间间隔前六个数据

为了获得数据的大体结构，可以画出理赔金额与理赔时间间隔的直方图：

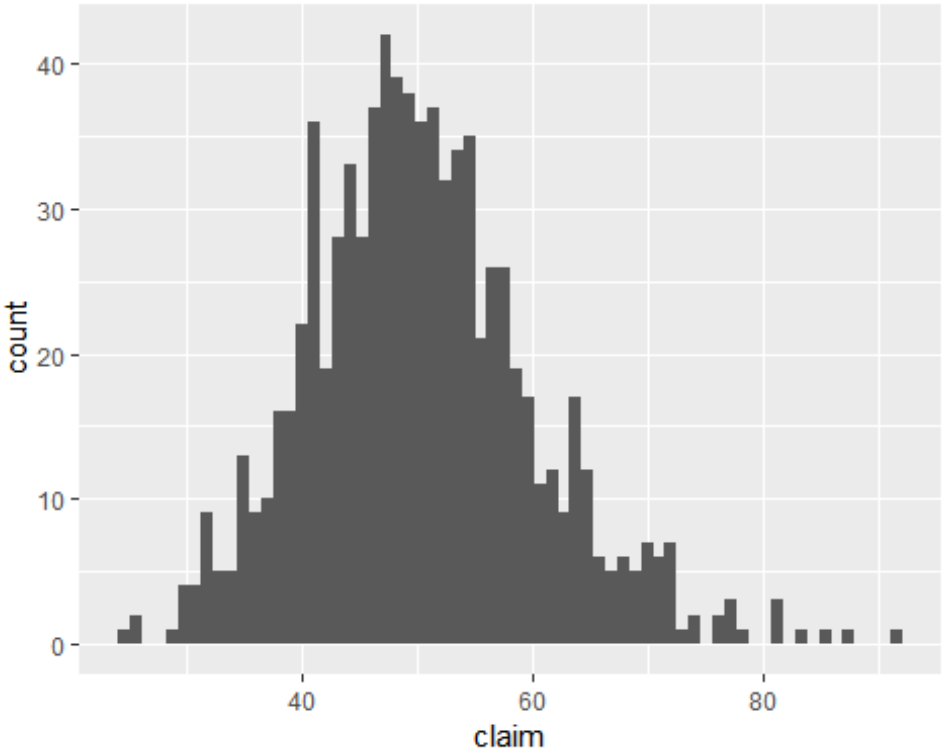

图 6 理赔金额直方图

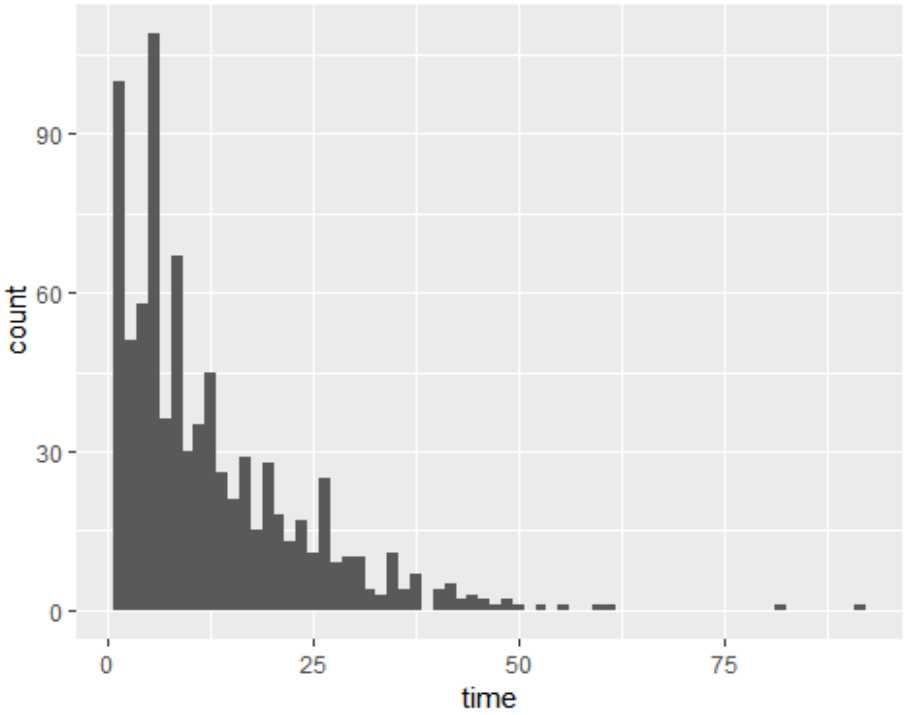

图 7 理赔时间间隔直方图

从第一张图可以看出，理赔金额的数据分布像是一个正态分布，因此将选用高斯核为核函数，并用经验法则选取窗宽。第二张图像像是一个指数分布，但是数据是正整数，故采用离散关联核方法，以二项核为核函数，用贝叶斯方法计算窗宽。下面是估计得到的概率密度函数和概率质量函数的图像。

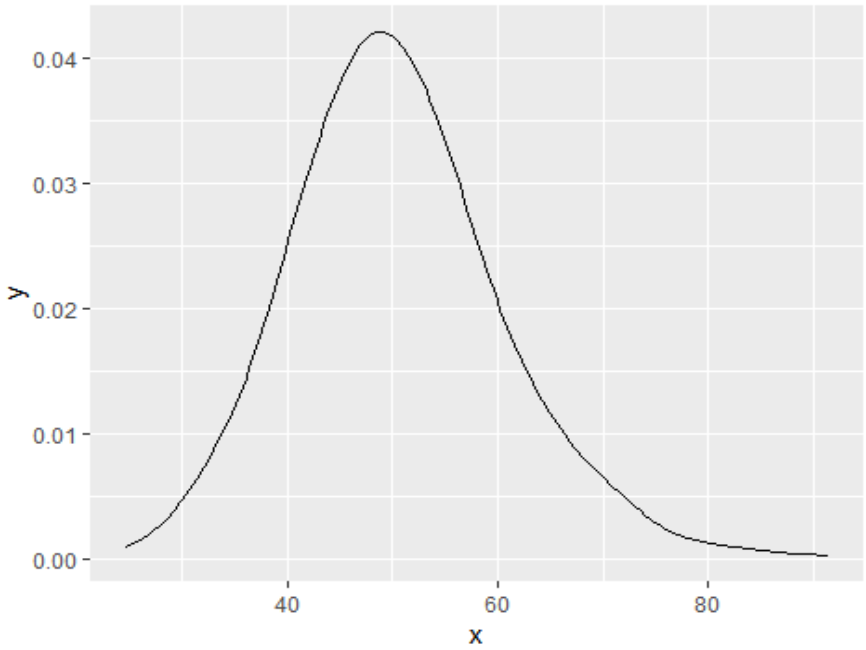

图 8 概率密度函数图

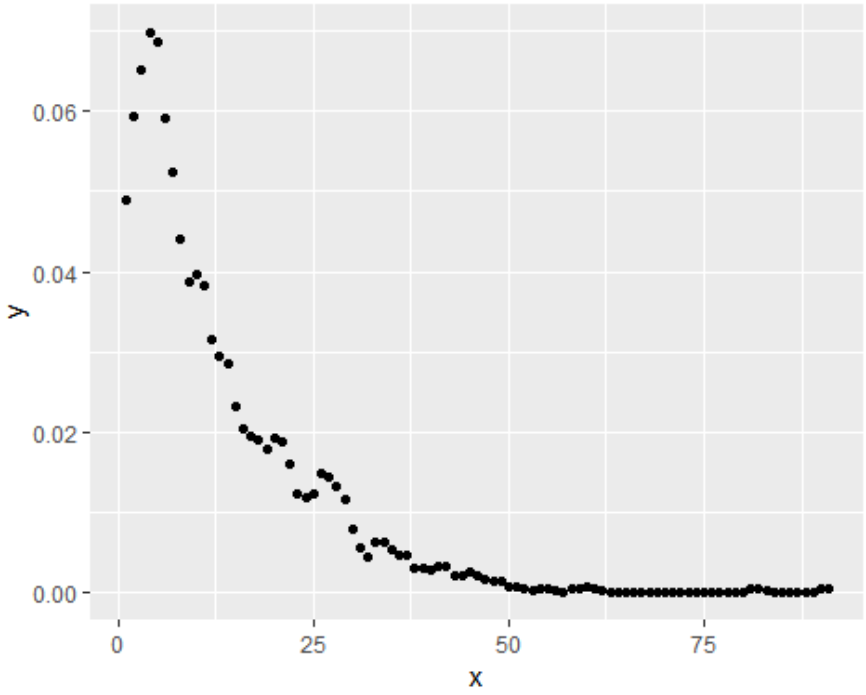

图 9 概率质量函数图

在得到估计的分布后，即获得了$claim_amount_pdf$与$claim_interval_cdf$便可以用算法 3 模拟盈余过程以得到破产概率，下面的表 6 是不同的初始资产u和t组合得到的破产概率值。从表中可以看出，在相同的运营时间下，初始资金越多，破产的概率越低；在相同的初始资金下，运营时间越久，破产概率就会越大。因此有充足的启动资金将会降低保险公司的破产概率。

(u,t)	$t=50$	$t=60$	$t=70$	$t=80$	$t=90$	$t=100$
$u=50$	0.1260	0.1272	0.1280	0.1284	0.1284	0.1284
$u=75$	0.0608	0.0624	0.0636	0.0638	0.0642	0.0644
$u=100$	0.0244	0.0252	0.0254	0.0256	0.0258	0.0258
$u=125$	0.0136	0.0148	0.0150	0.0152	0.0152	0.0154

表格 6 模拟破产概率值

下面的表 7 是当初始资金$u=75$时，有红利边界的模拟破产概率值，为简洁起见，表中将 *barrier* 记为 b。

(b,t)	$t=50$	$t=60$	$t=70$	$t=80$	$t=90$	$t=100$
$b=100$	0.2066	0.2456	0.2876	0.3212	0.3576	0.3884
$b=150$	0.0744	0.0834	0.0942	0.1036	0.1122	0.1194
$b=200$	0.0616	0.0634	0.0650	0.0666	0.0676	0.0692

表格 7 有红利边界的模拟破产概率值

由上表可知，当红利边界不变时，随着时间的增加，破产概率总体上呈现增加趋势；当时间不变时，随着红利边界的增加，破产概率呈现下降趋势。与表 6 第二行的数据对比可知，有红利边界的破产概率要比没有红利边界的破产概率大。

结论

本文研究了非参数核估计与离散关联核方法在模拟保险公司破产概率中的应用。非参数方法不事先假定理赔额，理赔时间间隔的概率分布，而是根据数据本身来计算出估计的概率分布，有更大的灵活性和实用性。本文还详细的讨论了核密度估计与离散关联核方法的偏差，方差以及均方误差，还讨论了多种窗宽选择方法。另外本文还给出了产生连续和离散随机数的算法，以及盈余过程随机模拟的算法。

参考文献

[1] LUNDBERG F. (1) Approximerad framställning af Sannolikhetsfunktionen:(2) Återförsäkring af Kollekrivrisker[M]. Almquist, 1903.

[2] ASMUSSEN S, ALBRECHER H. Ruin probabilities[M]. World scientific Singapore, 2010, 14.

[3] 孙立娟, 顾岚. 保险公司破产概率的估计及随机模拟[J].系统工程理论与实践,2000(07):63-68.

[4] FIX E, HODGES JR J L. Discriminatory analysis-nonparametric discrimination: consistency properties[R]. California Univ Berkeley, 1951.

[5] SCOTT D W. On optimal and data-based histograms[J]. Biometrika, Oxford University Press, 1979, 66(3): 605－610.

[6] SILVERMAN B W. Choosing the window width when estimating a density[J]. Biometrika, 1978, 65(1): 1－11.

[7] SILVERMAN B W. Density Estimation for Statistics and Data Analysis[M]. Taylor & Francis, 1986.

[8] GRUND B, HALL P, MARRON J. Loss and risk in smoothing parameter selection[J]. Journal of Nonparametric Statistics, Taylor & Francis, 1994, 4(2): 133－147.

[9] BOWMAN A W. An alternative method of cross-validation for the smoothing of density estimates[J]. Biometrika, Oxford University Press, 1984, 71(2): 353－360.

[10] JONES M C, MARRON J S, SHEATHER S J. A brief survey of bandwidth selection for density estimation[J]. Journal of the american statistical association, Taylor & Francis, 1996, 91(433): 401－407.

[11] SHEATHER S J, JONES M C. A reliable data-based bandwidth selection method for kernel density estimation[J]. Journal of the Royal Statistical Society: Series B (Methodological), Wiley Online Library, 1991, 53(3): 683－690.

[12] MARSH L C, MUKHOPADHYAY K. Discrete Poisson kernel density estimation-with an application to wildcat coal strikes[J]. Applied Economics Letters, Routledge, 1999, 6(6): 393－396.

[13] KOKONENDJI C, SENGA KIESSÉ T, ZOCCHI S S. Discrete triangular distributions and non-parametric estimation for probability mass function[J]. Journal of Nonparametric Statistics, Taylor & Francis, 2007, 19(6-8): 241－254.

[14] KOKONENDJI C C, KIESSE T S. Discrete associated kernels method and extensions[J]. Statistical Methodology, Elsevier, 2011, 8(6): 497－516.

[15] ZOUGAB N, ADJABI S, KOKONENDJI C. Binomial kernel and Bayes local bandwidth in discrete function estimation[J]. Journal of Nonparametric Statistics, Taylor & Francis, 2012, 24(3): 783－795.

[16] JONES O, ROBINSON A P, MAILLARDET R. Introduction to scientific programming and simulation using R[M]. Chapman; Hall/CRC, 2009.

[17] SCOTT D W. Multivariate density estimation: theory, practice, and visualization[M]. John Wiley & Sons, 2015.

[18] AITCHISON J, AITKEN C G. Multivariate binary discrimination by the kernel method[J]. Biometrika, Oxford University Press, 1976, 63(3): 413－420.

[19] SCHUMAKER L. Spline functions: basic theory[M]. Cambridge University Press, 2007.

[20] BISHOP C M. Pattern recognition and machine learning[M]. springer, 2006.