非参数方法在沪深股市收益率分布的应用

陈 娟

(浙江工商大学数量经济系，浙江杭州 310035)

摘 要：沪深大盘指数的收益率分布函数并不服从通常人们所认为的正态分布。本文采用一种新的方法——非参数核密度估计，对大盘指数的收益率分布函数进行研究。这种新方法不仅很好地刻画了收益率分布的尖峰和肥尾特征，而且比一般的正态分布更能捕捉市场的风险特征，结论也更加准确。

关键词：收益率；非参数估计；核密度函数；窗宽

中图分类号：F830; O212 **文献标识码**：A **文章编号**：1008-309(2005)03-0022-05

一、引 言

在描述股价行为的经典计量模型中，股市的收益率通常被假定是服从正态分布。但是许多计量金融学家对这一经典假设做了大量的研究并发现，收益率的分布并不服从正态分布这一假设。事实上，大多数价格的变化存在很明显的尖峰现象，也就是说相对正态分布而言，在均值附近的数据点特别多。许多学者认为这只不过是由一些“异常值”所引起，从而在统计分析中将这些“异常值”去掉。例如，国内学者陶亚民[1]认为，上海股市收益率分布是服从正态分布的，但这却是在剔除了“异常点”的基础上得到的结论。然而 Mandelbrot[2]认为将这些“异常值”值从数据中去掉是不可取的。因为“异常值”的出现并不是一种偶然现象，尖峰和肥尾现象几乎是所有股票收益率数据所共有的。这说明“异常值”本身反映了股票收益率并不服从正态分布这一假定。陈启欢[3]也通过实证研究的方法得到我国股市收益率分布曲线并不服从正态分布。封建强、王福新[4]利用几种不同的分布函数来刻画收益率分布，并且利用稳定的 Pareton 分布和 t 分布来拟合了沪深股市的收益率分布。但是正如作者在文章最后所讲到的一样，从精确的意义上来讲 Pareton 分布和 t 分布都不能很好的描述收益率数据。同样李亚静[5]也检验得到收益率分布的非正态性。在收益率分布非正态的情况下，我们又该如何对收益率分布进行估计呢？对于这个问题，本文从一个新的角度来进行了说明，即利用非参数核密度估计的方法来对沪深指数的收益率分布进行研究，从而得到一些与以往不同的结论。

二、收益率分布的正态性检验

考虑到数据的代表性与完整性，本文选取了上证综合指数和深证成分指数作为沪深股市的代表，以每日的收盘价为分析对象。考虑到数据的平稳性，样本取值范围为：2001 年 1 月 4 日至 2003 年 12 月 31 日，上证和深成分别共 710 个有效数据(数据来源于 http://quote.stockstar.com/stock/

收稿日期：2004-11-22

作者简介：陈娟(1978-)，女，山西太原人，硕士研究生，研究方向：金融风险管理

external_history.asp?code=shzs000001），分别计算它们的日收益率 $R_{t+1}=\frac{P_{t+1}-P_t}{P_t}$，$P_t$ 是第 t 日的收盘指数，P_{t+1} 是第 t+1 日的收盘指数。计算得到沪深大盘指数收益率数据的统计特征见表 1，收益率数据的直方图见图 1。

表 1　沪深股市收益率的统计特征

	均值	中值	最大值	最小值	标准差	偏度	峰度	JB 值
上海	-0.000455	-0.000208	0.094014	-0.06543	0.013617	0.871323	10.9938	2002.56
深圳	-0.000439	-0.000334	0.095299	-0.069302	0.014374	0.801475	10.32585	1682.44

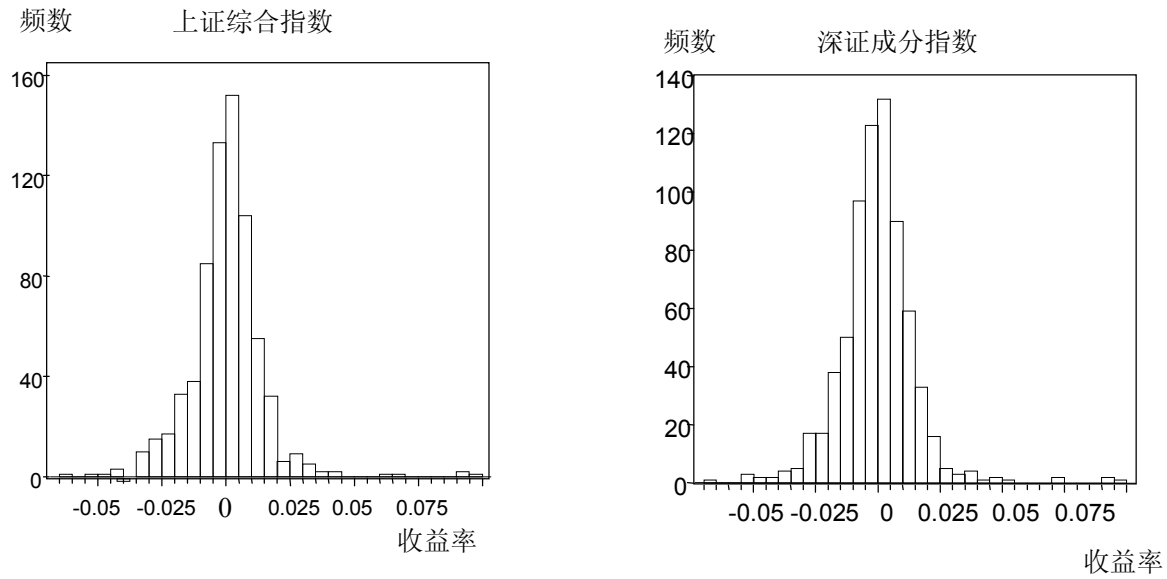

图 1　沪深股市收益率分布的直方图

从表 1 可以看出，沪深股市收益率的均值都很小，但其变化范围较大，而且具有明显的偏度、较大的峰度以及比正态分布更厚的尾部（正态分布的偏度是 0，峰度是 3）。这些数据都表明沪深股市收益率分布是非正态的。此外，通过 JB 检验值也可以得到拒绝收益率分布服从正态分布的假定。

三、收益率分布的非参数估计

在明确了沪深股市收益率分布非正态的情况下，传统的参数估计收益率分布的方法就是行不通的。因此，本文考虑从非参数估计的角度出发来研究收益率的分布特征。非参数估计不事先假定变量之间的结构关系，而是通过已知数据直接估计这种结构关系。相反，参数估计则要事先假定好变量之间的某种结构关系，然后利用数据对未知的参数进行估计，从而来确定这种结构关系。对于变量之间结构关系的某种假定在通常情况下是很难做到的，因此这使得参数估计结果的准确性受到一定的质疑。非参数估计则恰好弥补了这种局限性。它直接对结构关系的估计给我们的研究带来很多方便。本文就利用了非参数估计的这个特点对沪深股市收益率进行了研究。

由于股票收益率的波动较大，其分布也就较不均，所以采用灵活性较大的非参数密度估计是

比较恰当的。其中，核估计方法又是非参数密度估计中有关单样本模型较为典型的估计方法。因此，本文就采用核密度函数估计，借鉴谭英平[6]的方法，即利用皮尔逊 x^2 拟合优度检验，对沪深收益率分布进行研究。

具体思想方法如下：

假定窗宽 h 有一系列可能的合理取值，针对具体的样本数据我们可以通过一些“预运算”给出窗宽取值的某个限定区间。在这个限定区间中，我们对每个选定的 h 值，计算样本数据的核密度估计值，理论上选择目标函数值最小的函数形式就作为最优的核估计密度函数。这里我们可以将拟合优度检验的统计量看成目标函数的一种，那么它就可以作为确定最佳窗宽的衡量标准。值得注意的是，在选择一个合适的窗宽时，我们要将估计结果的拟合度和平滑性来综合考虑，而不是简单的以目标函数值最小为标准。只有这样，选择出的最佳窗宽才充分考虑了拟合度与平滑性这两个方面的内容。

具体方法如下：

1．将沪深股市收益率数据进行分组，确定在各组中收益率数据的实际频数。考虑到收益率数据分布的大概范围，本文将其分为八组，具体分组情况如下：

表 2　沪深股市收益率的分组频数

区 间 范围	[0.05, 0.1]	[0.01, 0.05]	[0.005, 0.01]	[0, 0.005]	[0, -0.005]	[-0.005, -0.01]	[-0.01, -0.05]	[-0.05, -0.1]
上海	152	104	111	5	133	85	118	2
深圳	132	90	123	6	123	97	135	4

2．假设收益率数据所包含的密度函数是 $f(x)$，非参数核估计下的密度函数是

$$\hat{f}(x)=\frac{1}{\sqrt{2\pi}nh}\sum_{i=1}^{n}\exp\left[-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2\right] \qquad (1)$$

那么，原假设和备择假设分别是：

$$H_0: f(x)=\hat{f}(x); \qquad H_1: f(x)\neq\hat{f}(x)$$

3．根据 x^2 拟合优度检验的统计量的定义，我们得到样本数据的统计量值为：

$$\chi^2=\sum_{i=1}^{n}\frac{(f_i-np_i)^2}{np_i} \qquad (2)$$

其中，f_i 为第 i 组的收益率数据的实际频数，np_i 是按照核估计的密度函数计算得到的收益率数据的理论频数。因为，

$$np_i=n\Pr(C_{i-1}<X\leq C_i)=n\left[\hat{F}(C_i)-\hat{F}(C_{i-1})\right] \qquad i=1,\Lambda\ ,m \qquad (3)$$

其中，$\hat{F}(x)$ 是根据核估计密度函数 $\hat{f}(x)$ 计算得到的分布函数。由于本文采用的是标准正态核函数，所以可以推导出核估计条件下的收益率分布函数是：

$$\hat{F}(x)=\Pr(X\leq x)=\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{x-x_i}{h}\right) \qquad (4)$$

将（4）式代入（3）式可得：

$$np_i = \sum_{i=1}^{n}\left[\Phi\left(\frac{C_i - x_i}{h}\right) - \Phi\left(\frac{C_{i-1} - x_i}{h}\right)\right] \qquad (5)$$

这样，我们就可以计算在不同窗宽下的核估计密度函数的统计量值。

选择一个最佳窗宽，使得估计的密度函数即通过检验，又具有较好的平滑度。换句话说就是，让通过检验的窗宽尽可能的大。

当原假设成立时，x^2 统计量近似服从自由度为 m-k-1 的 x^2 分布，其中 k 代表总体分布模型中需要估计的参数个数，而本文的模型并不涉及任何具体的参数模型，因此不存在显形参数。但从核估计密度函数的表达式来看，窗宽 h 是需要估计的唯一变量，那么我们有理由把它看成该检验问题中唯一的一个需要确定的参数。这样就可以说，x^2 统计量是服从 m=8，k=1,自由度是 6 的 x^2 分布。这里我们认为显著水平 a=0.05。通过计算得到窗宽一系列可能的取值区间，0.0012、0.0015、…、0.0023、0.0025，以及与此相对应的 x^2 检验统计量值，如下表 3 和表 4：

表 3　上证综合指数收益率的窗宽和相应 x^2 统计量

窗宽 h(%)	0.12	0.15	0.17	0.18	0.19	0.20	0.21	0.23
统计量 x^2	3.7376	6.3736	8.6823	10.0099	11.4544	13.0157	14.6929	18.3804

表 4　深圳成份指数收益率的窗宽和相应 x^2 统计量

窗宽 h(%)	0.15	0.17	0.18	0.19	0.20	0.21	0.22	0.23
统计量 x^2	6.1044	8.0020	9.0789	10.2419	11.4902	12.8225	14.2371	15.7320

在显著水平 a=0.05 时，自由度是 6 的 x^2 分布的临界值是 $\chi^2_{0.95} = 12.592$ 。因此，我们可以发现，使检验统计量比临界值小的最大窗宽，上证 $h_{sh} = 0.0019$, 深成 $h_{sz} = 0.0020$ 。这样我们就得到了上证和深成收益率分布核密度估计的最佳窗宽。

四、非参数估计下的收益率密度函数及实际应用

在核估计的窗宽确定后，我们就可以得到收益率的核估计密度函数的确定形式：

$$\hat{f}_{sh}(x) = \frac{1}{710 * 0.0019 * \sqrt{2\pi}} \sum_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{x - x_i}{0.0019}\right)^2\right]$$

$$\hat{f}_{sz}(x) = \frac{1}{710 * 0.0020 * \sqrt{2\pi}} \sum_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{x - x_i}{0.0020}\right)^2\right]$$

此时，我们就可以画出上证和深成在非参数核密度估计下的收益率分布图：

考虑在非参数核密度估计的情况下，收益率的期望和方差：

$$EX = \int_{-\infty}^{\infty} t f_n(t)dt = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h} \int_{-\infty}^{\infty} t \exp\left(-\frac{(t - x_i)^2}{2h^2}\right)dt$$

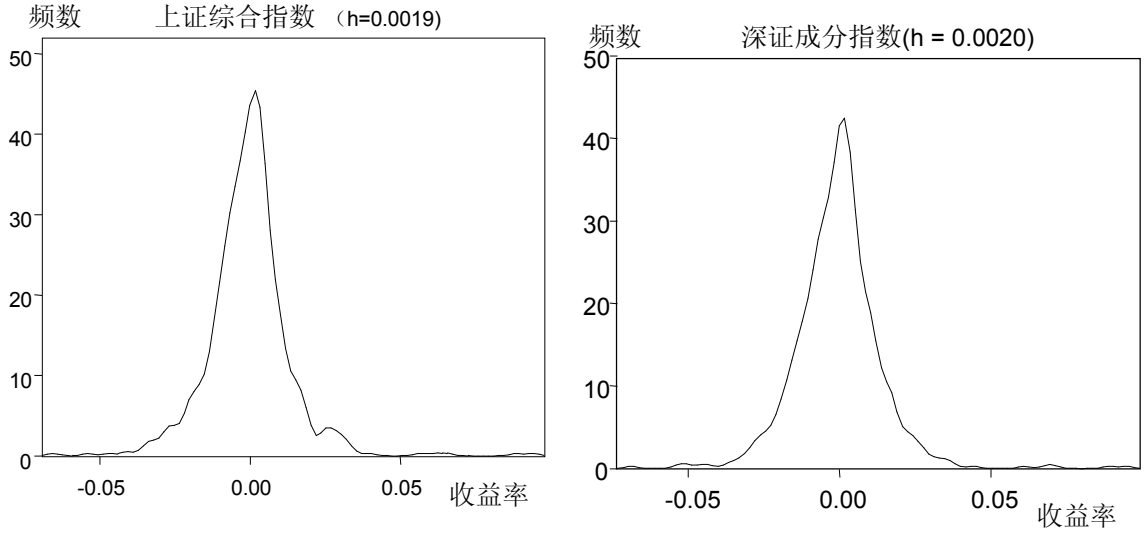

图 2　沪深股市收益率的密度函数图

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(hy + x_i)\exp\left(-\frac{y^2}{2}\right)dy$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i \qquad (6)$$

$$E(X^2) = \int_{-\infty}^{\infty}t^2\hat{f}(t)dx = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2\pi}h}\int_{-\infty}^{\infty}t^2\exp\left(-\frac{(t-x_i)^2}{2h^2}\right)dt = h^2 + \frac{1}{n}\sum_{i=1}^{n}x_i^2 \qquad (7)$$

$$Var(X) = E(X^2) - [E(X)]^2 \qquad (8)$$

通过公式（6）（7）（8），计算出核估计密度函数的期望和方差，见表 5：

表 5　沪深股市非参数估计与实际的收益率的统计特征比较

	核估计的期望	核估计的方差	实际的均值	实际的方差
上证指数	0.0000539	0.0002126	0.0000539	0.0002092
深成指数	-0.0002156	0.0002424	-0.0002156	0.0002386

从表 5 可以看出，核估计收益率的期望与原来数据的均值是相等的，但是方差却不同，比实际数据的方差偏大。

由公式（4）我们知道在核估计密度函数下的收益率分布函数形式，此时我们就可以计算收益率落在不同区间时概率值的大小，计算结果见表 6：

表 6　沪深股市收益率的区间概率值

区间	<-0.05	[-0.05,0]	[-0.01,0]	[-0.005,0]	[0,0.005]	[0,0.01]	[0,0.05]	>0.05
上证	0.00264	0.49726	0.32406	0.20267	0.18505	0.33448	0.49295	0.00704
深成	0.00498	0.49502	0.30015	0.16925	0.18650	0.31967	0.49146	0.00854

表 6 的计算结果表明：上海和深圳股票市场的收益率下跌大于 0.05 的可能性分别是 0.264％和 0.498％，而上涨大于 0.05 的可能性分别是 0.704％和 0.854％。这样的结果说明，我国沪深两市出现高收益的可能性比较大。与一些西方成熟的股票市场相比，我国股市仍然具有高收益，高风险的特征。

五、总结

本文通过对沪深股票市场的收益率进行非参数核密度估计研究发现，非参数估计方法对收益率分布给出一个比较合适的拟合效果。在此基础上，本文计算还得到在非参数估计下的收益率的期望和方差，以及收益率落在各个区间的概率值。这为我国股票市场收益率分布的研究提供了一个新的视角。

参考文献

[1] 陶亚民, 蔡明超, 杨朝军. 上海股票市场收益率分布特征的研究[J]. 预测, 1999, (2): 57-58

[2] Mandelbrot B B. New Method in Statistical Economics [J]. Journal of Political Economy, 1963, 71: 421-440

[3] 陈启欢. 中国股票市场收益率分布曲线的实证[J]. 数理统计与管理, 2002, (5): 9-11

[4] 封建强, 王福新. 中国股市收益率分布函数研究[J]. 中国管理科学, 2003, (1): 14-21

[5] 李亚静, 朱宏泉. 沪深股市收益率分布的时变性[J]. 数学的实践与认识, 2002, (2): 228-233

[6] 谭英平. 非参数密度估计在个体损失分布中的应用[J]. 统计研究, 2003, (8): 40-44

Application of Nonparametric Method in the Returns Distributions in Shanghai and Shenzhen Stock Markets

CHEN Juan

(Zhejiang Gongshang University, Hangzhou, China 310035)

Abstract: This paper has studied the returns distributions in Shanghai and Shenzhen stock markets. The results indicate that the returns distributions are not normal distributions. Therefore, the author uses a new way--nonparametric kernel density estimators, to estimate the returns distributions of stocks. By the new estimators, we research the returns distributions and get the new results. The way is very useful at depicting the characteristics of sharp peaks and fat tails of returns distributions.

Key words: Returns; Nonparametric estimator; Kernel density function; Bandwidth