基于 POT 方法的极值理论在基金净值预测中的应用

李晓康

(陕西理工学院数学系, 陕西 汉中 723000)

摘要：为对基金净值数据进行建模, 根据基金净值样本数据的尾部特点, 建立极大, 极小值分布的 GPD 模型, 运用 POT 方法确定临界值, 进而对参数进行估计, 并对模型进行检验. 最后, 运用建立的模型对一些极值点进行预测. 所得结果很好地描述了数据特点, 对极值点的预测符合实际.

关键词：极值分布; POT 方法; 基金净值; 估计

中图分类号：O212. 1 **文献标识码**：A **文章编号**：1008-5513(2010)05-0776-09

1 引言

极端事件打破自然界和人类社会经济活动的正常规律, 因而对自然界和人类社会经济产生巨大影响 (破坏). 研究极端事件的有力工具就是极值理论. 极值理论是统计学的重要分支. 极值理论分为一元, 多元极值理论 [1-4]. 一元极值理论模型主要由广义极值分布 (GEV) 和广义 Parto 分布 (GPD), 是基于不同样本选取方法而得到的渐近分布, 主要用于描述分布的尾部特征, 即超越阀值的分布. 极值理论已广泛应用于水文, 气象, 地震, 海洋, 交通, 工程设计等各个领域. 近年来, 极值理论及方法还被运用于银行, 保险, 股市等金融领域, 用于金融市场的风险管理及防范 [5-6]. 经典的统计极值理论是关于随机变量序列的最大 (小) 值渐进分布的理论, 基于这一理论的 POT 方法就是对超越一定门限的数据建模. 因此, 利用统计极值理论方法能够有效地对随机序列的最大 (小) 值的概率分布和数据序列的边际概率分布尾部进行建模, 基于相应的数学模型能够估计损失的风险, 从而为建立有效的风险防范和预警系统提供一定的理论依据.

2 一元极值理论

极值理论分为一元极值理论与二元极值理论. 一元极值理论主要对同一过程的单一指标进行研究; 二元极值理论主要对一些不同过程, 同一过程的几个不同指标, 几个地点的单一过程, 时间序列的一组延迟值进行研究. 目前, 极值理论与方法已广泛应用于银行, 股市, 保险等各领域.

收稿日期：2010-05-27.

基金项目：陕西理工学院科研基金 (SLG0919).

作者简介：李晓康 (1973-), 硕士, 讲师, 研究方向：应用概率统计.

2.1 GPD 分布

设 $x_i, i = 1, 2, \cdots, n$ 是来自分布函数为 $F(x)$ 的总体 X 的一组样本, 将其观测值由小到大排列: $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$, 称 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 为顺序统计量, 其中 $X_{(1)} = \min(X_1, X_2, \cdots, X_n)$ 称为样本极小值, $X_{(n)} = \max(X_1, X_2, \cdots, X_n)$ 称为样本极大值, 它们的极限分布称为极值分布. 极值理论涉及极小值与极大值 (统称为极值) 的极限分布问题. 广义 Pareto 分布 (GPD) 即双参数的广义极值分布, 是基于门限峰值法 (POT) 的样本极大值极限分布.

广义极值分布的分布函数 (GPD) 是基于以下定理:

定理 2.1 设 $(X_1, X_2, \cdots, X_n)$ 是独立同分布随机变量, 公共分布为 $F(x)$, 记

$$M_n^* = \max(X_1, X_2, \cdots, X_n)$$

则对足够大的 n, 有

$$P(M_n^* \le x) \approx G(x). \tag{2.1}$$

其中

$$G_{u,\xi,\sigma}(x) = \begin{cases} 1 - \Big[1 + \xi\Big(\dfrac{x-u}{\sigma}\Big)\Big]^{-1/\xi}, \xi \neq 0; \\ 1 - \exp\Big(-\dfrac{x-u}{\sigma}\Big), \text{其它}, \end{cases} \tag{2.2}$$

其中

$$x \in \begin{cases} (u, \infty), & \xi > 0, \\ (u, u - \sigma/\xi), & \xi < 0. \end{cases}$$

极值分布Ⅱ型, Ⅲ型分别对应于 $\xi > 0, \xi < 0$, Ⅰ型则是 $\xi \to 0$ 的极限. ξ 是形状参数, 当形状参数由小变大 (由负值变为正值) 时, 尾部逐渐变厚. u 是位置参数 (临界值), $\sigma > 0$ 是尺度参数. 当 $u = 0, \sigma = 1$ 时, 就是标准的 Pareto 分布. 极值分布模型在应用中主要用于拟合样本数据序列分布的尾部特征, 对随机序列的最大 (小) 的概率分布和数据序列的边际分布尾部进行建模, 以及各种分位数的估计, 如: 设某一极值的概率为 $G(x_p) = P(X < x_p) = 1 - p$, 则可得 GPD 分布的分位数

$$x_p = \begin{cases} u - \dfrac{\sigma}{\xi}(1 - (-\log(1-p))^{(-\xi)}), \xi \neq 0; \\ u - \sigma\log(-\log(1-p)), \text{其它} \end{cases} \tag{2.3}$$

在实际应用中, 根据其实际意义, 分位数 (x_p) 又称为回归期为 $(1/p)$ 的回报水平, 风险价值等. 将参数估计值带入上式, 即可得 GPD 分布分位数的估计:

$$\hat{x}_p = \begin{cases} u - \dfrac{\hat{\sigma}}{\hat{\xi}}(1 - (-\log(1-p))^{-\hat{\xi}}, \hat{\xi} \neq 0; \\ u - \hat{\sigma}\log(-\log(1-p)), \text{其它} \end{cases} \tag{2.4}$$

为此, 需要对模型的参数及门限值进行估计.

2.2 门限峰值 (POT) 法

门限峰值法, 简称 POT 方法. 门限峰值法较简单, 精度较高, 目前运用门限峰值法对极值分布进行统计推断, 可以得到很好的效果. POT 方法的前提条件是: (1) 超限发生的时间服从泊松分布; (2) 超限彼此相互独立, 且服从 GPD 分布; (3) 超限与超限发生的时间相互独立. 如果假设高于某一门限的事件发生的次数服从泊松分布, 且满足 POT 条件, 则可用 GPD 分布去拟合高于门限值的超限分布, 从而给出原分布序列的尾部估计. POT 方法是基于 PBdH 定理的, PBdH 定理 (Balkema De Haan 1974,Pickands 1975) 证明了在 MDA 条件下超额的分布弱收敛到广义 Pareto 分布. 此定理说明, 对充分大的临界值, 超额的分布

$$F_u(x-u) = P(X-u<x|X>u). \tag{2.5}$$

可以用 $G_{\xi,\sigma}(x)$ 近似, 其中参数 ξ,σ 未知. 等价地, 对 $x-\mu>0$, 超额的分布可用 (2.6) 式近似:

$$G_{\xi,\sigma}(x-\mu) = G_{\mu,\xi,\sigma}(x). \tag{2.6}$$

2.3 极值模型的临界值估计

从理论上讲, 极小值的临界值 u 应极小, 极大值的临界值 v 应极大, 但实际上, u 越小, v 越大, 用来估计尾部特征的样本观测值就越少, 所以, 需要确定适当的临界值, 确定临界值是极值分布的关键. 确定临界值的常用方法有 QQ 图和样本均超额函数图, 这里, 采取样本均超额函数图.

2.3.1 样本均超额函数

均超额函数的定义为:

$$e(u) = E(X-u|X>u). \tag{2.7}$$

即超过 u 的样本超额均值. 由于 X 的分布未知, 故经常采用如下的样本均超额函数. 样本均超额函数的定义为:

$$e(\mu) = \frac{1}{N_\mu}\sum_{i=1}^{N}(x_i-\mu)_+. \tag{2.8}$$

其中 $\sum_{i=1}^{N}(x_i-\mu)_+$ 表示阀值 μ 的条件样本余额观测值的总和, N_μ 表示超越阀值 μ 的条件样本余额观测值的个数. 样本均超额函数图为点 $(\mu,e(\mu))$ 构成的曲线.

2.3.2 临界值的确定

利用样本均超额函数图确定临界值, 首先将样本观测值排序, 作出样本均超额函数图, 看图形是否近似满足正斜率的线性关系, 如果满足, 再拟合线性关系, 取其与样本均超额函数图的首个交点为临界值.

3 实证分析

3.1 数据的选取与样本的确定

基金是近年来热销的金融产品, 但其单位净值每日波动, 具有一定的风险, 可视为随机变量. 在交易中, 人们关心的是一定阶段内的单位净值极大, 极小值, 故可以运用极值理论进行建模, 分析, 研究. 为了应用极值理论, 考虑到金融产品基金市值最大, 最小的变化, 选取易基 50 基金从 2006 年 6 月 16 日至 2008 年 4 月的 477 个交易日的每日净值为研究对象 (数据来源于中国基金网 (http://www.chinafund.cn)), 进一步从中选取极值样本, 对样本极大, 极小值进行分析, 建模, 分析其尾部特征, 进而对一些极值点进行预测.

3.2 样本数据的正态性检验

对选取的数据进行正态性检验, 利用 Matlab 软件中的 Toolbox//Statistics 工具箱中的 normplot 函数进一步对样本数据进行正态性检验, 结果如图 1:

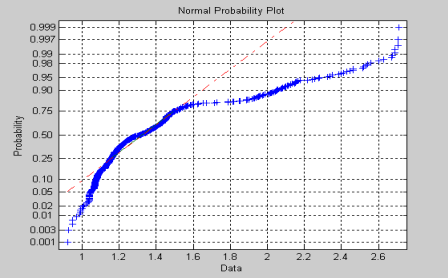

图 1 样本数据的正态性检验结果

从频率直方图, 基本统计结果及正态性检验结果可以看出, 样本数据为非正态分布, 偏峰, 厚尾, 符合极值分布的特征, 故可用极值分布进行建模.

3.3 极值模型的参数估计

3.3.1 样本极大, 极小值临界值的估计

首先做出样本极大值均超额曲线如图 2 和图 3:

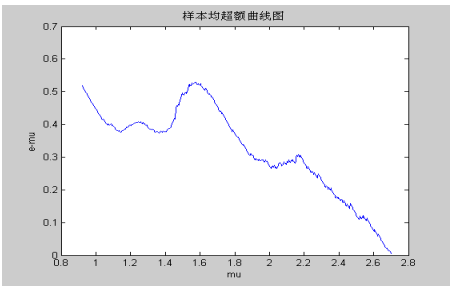

图 2 样本均超额曲线线图

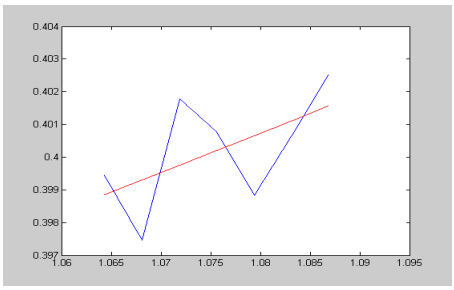

图 3 样本均超额曲线线性拟合图 ($(u = 1.06)$ 附近)

由上图可以看出, 在 1.2,1.4 及 2.0 附近的三个区间上样本均超额曲线呈线性增加, 故应在这 3 个点选择临界值, 由 2.3.2 所述的方法确定临界值为 1.06,1.28,2.05.

由图 4 和图 5 可以看出, 临界值在 1.06~2.05 之间均可近似满足线性关系. 对以上确定的样本极大值临界值, 可用 2.3.2 的方法进行参数估计.

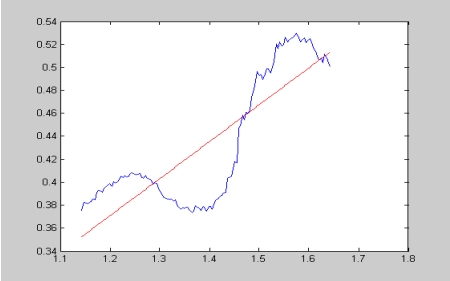

图 4 样本均超额曲线线性拟合图 ($u = 1.28$ 附近)

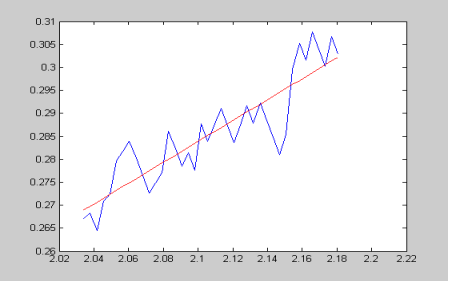

图 5 样本均超额曲线线性拟合图 ($u = 2.05$ 附近)

对样本极小值临界值的估计可按照上面的方法进行. 首先做出极小值样本超额曲线图 6.

从图 6 看, 样本极小值均超额曲线图基本成线性关系, 取其线性拟合的首个交点 1.20 为样本极小临界值的估计, 结果如图 7.

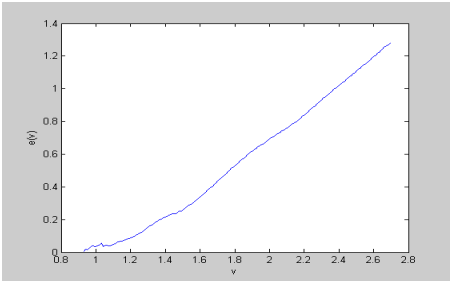

图 6 样本极小值均超额曲线图

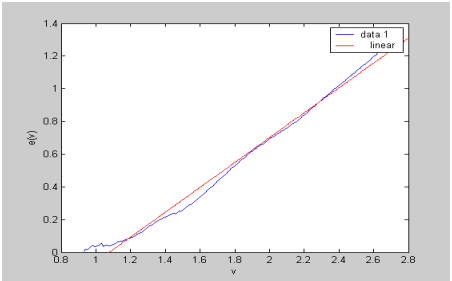

图 7 样本极小均超额曲线线性拟合图

3.3.2 参数的估计

对模型参数的估计采用 Dehann 矩估计法进行,Dehann 矩估计法的作法如下

设 n 为样本点个数, k 为超越门限 u 的观测值个数, 则门限 u 代表第 $n-k$ 大的数据, 最大, 第二, $\cdots$, 第 $n-k$ 大的变量, 记为 $X_n^*, X_{n-1}^*, X_{n-k}^* = u$, 则

$$\hat{\xi} = H_{k,n} + 1 - \frac{1}{2}\left(1 - \frac{H_{k,n}^2}{H_{k,n}^{(2)}}\right)^{-1}. \tag{3.1}$$

$$\hat{\sigma} = \rho H_{k,n}. \tag{3.2}$$

其中若 $\xi \hat{\geq} 0, \rho = 1$; 否则 $\rho = 1 - \hat{\xi}$

$$H_{k,n} = \frac{1}{k}\sum_{i=0}^{k-1}(\ln X_{n-i}^* - \ln X_{n-k}^*). \tag{3.3}$$

称为 Dehann 一阶矩, 表示从 $(n-k+1)$ 到 (n) 的后 (k) 个样本极大值与临界值的对数平均偏差.

$$H_{k,n}^{(2)} = \frac{1}{k}\sum_{i=0}^{k-1}(\ln X_{n-i}^* - \ln X_{n-k}^*)^2. \tag{3.4}$$

称为 Dehann 二阶矩, 表示从 $n-k+1$ 到 n 的后 k 个样本极大值与临界值的对数平方平均偏差.

采用 Dehann 矩估计法进行参数估计时, Dehann 一, 二阶矩及点估计会随着 k 的不同而变化, 选取 k 的原则是各阶矩及点估计波动较小.

由以上讨论, 取极大临界值在 1.06~2.05 之间, 对应的超额样本个数在 51~465 之间, 故取 51~465 之间的超额样本数据, 用这些超额样本数据进行参数估计, 部分结果见表 1:

表 1　不同临界值下的极大值分布的 Dehann 矩及参数 ξ,σ 的点估计

k	u	Dehann 一阶矩	Dehann 二阶矩	$\hat{\xi}$	$\hat{\sigma}$
465	106	0.285 2	37.819 0	0.784 1	0.302 3
369	1.16	0.257 5	24.465 7	0.756 1	0.298 7
269	1.26	0.256 9	17.760 1	0.755 1	0.323 8
225	1.36	0.222 9	11.180 9	0.720 7	0.303 2
147	1.46	0.248 1	9.049 9	0.744 7	0.362 3
102	1.56	0.278 9	7.935 1	0.774 0	0.435 1
90	1.66	0.250 8	5.660 7	0.745 2	0.416 3
87	1.76	0.200 0	3.481 1	0.694 2	0.352 1
81	1.86	0.157 2	2.001 0	0.650 9	0.292 3
66	1.96	0.133 0	1.167 2	0.625 3	0.260 7

各个不同临界值下的 Dehann 一, 二阶矩及参数的点估计如图 8:

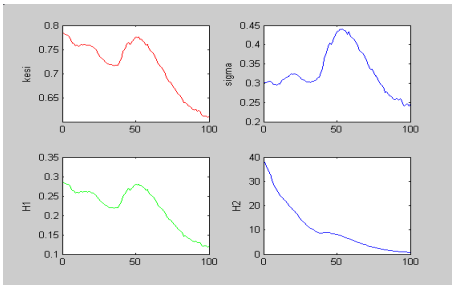

图 8　各个不同临界值下的 Dehann 一, 二阶矩及参数 (ξ,σ) 的 Dehann 点估计

从上表和图可以看出, 随着临界值的不同, Dehann 一, 二阶矩及参数 ξ,σ 的 Dehann 点估计都在发生波动. 从理论上讲, 极小值的临界值 u 应极小, 极大值的临界值 v 应极大, 但实际上, u 越小, v 越大, 用来估计尾部特征的样本观测值就越少, 所以, 需要确定适当的临界值. 从上表和图可以看出, 当极大值临界值 $u>1.46$ 时, 参数估计值波动较大, 而 $u<1.46$ 时参数估计值比较稳定, 综上所述, 取极大值临界值 $u=1.46$, 对应的极大样本个数为 147, 参数估计值 $\hat{\xi}=0.744\ 7, \hat{\sigma}=0.362\ 3$, 取极小值临界值为 1.20, 对应的超额样本数为 492 个, 用这些超额样本数据进行参数估计, 结果见表 2:

表 2　不同临界值下的极小值分布的 Dehann 阶矩及参数 ξ,σ 的点估计

k	v	Dehann 一阶矩	Dehann 二阶矩	$\hat{\xi}$	$\hat{\sigma}$
176	1.20	0.076 6	1.031 5	0.573 7	0.091 9

3.4 极大, 极小值尾部分布估计

将以上估计值带入 (2.1) 式, 可得样本极大值尾部分布的估计

$$G(x)=1-0.2970\left(1+0.744\ 7\left(\frac{x-1.46}{0.362\ 3}\right)\right)^{-1/0.744\ 7}. \tag{3.5}$$

上式图形如图 9

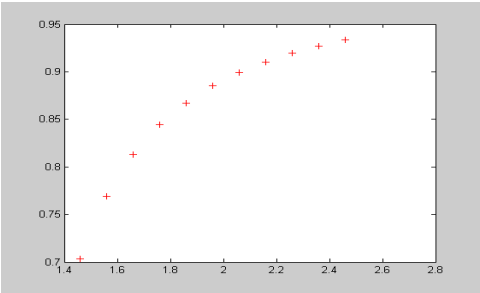

图 9 极大值的尾部分布的估计图

由上图可以看出, 当 $x=2.46$ 时, 其概率为 0.933 7, 随着 x 的增大, 其概率接近于 1. 利用上面得到的极大值尾部分布的估计对分位数进行估计, 由 (2.2) 式,

$$\hat{x}_p=\begin{cases}u-\dfrac{\hat{\sigma}}{\hat{\xi}}(1-(-\log(1-p))^{-\hat{\xi}},\hat{\xi}\neq 0;\\ u-\hat{\sigma}\log(-\log(1-p)),\text{其它}\end{cases} \tag{3.6}$$

可得极大值分布分位数的估计如表 3. 将极小值分布的各个参数的估计值带入, 可得极小值的

表 3 各不同置信度下的极大值分位数估计

p 值	极大值分位数
0.1	3.57
0.05	5.42

尾部分布的估计:

$$\hat{G}(x)=\left[1+0.573\ 7\left(\frac{1.20-x}{0.091\ 9}\right)\right]^{-1/0.573\ 7}. \tag{3.7}$$

上式图形如图 10.

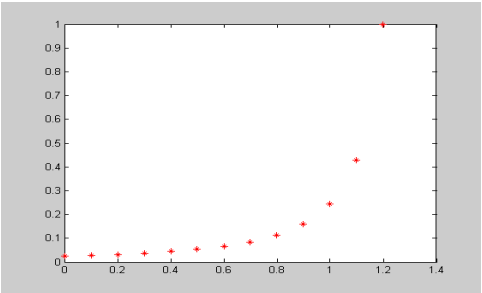

图 10 极小值的尾部分布的估计

同理, 可得极小值分位数的估计:

$$\hat{x}_p = \nu + \frac{\hat{\sigma}}{\hat{\xi}}((1-(1-p)^{-\hat{\xi}})). \tag{3.8}$$

各不同置信度下的极小值分位数估计如表 4:

表 4　各不同置信度下的极小值分位数估计

(p) 值	极小值分位数
0.1	1.80
0.05	2.09

3.5　模型的检验

对以上所得的模型及其结果进行检验, 采用 P-P 图检验和 Kolmogorov 检验.

3.5.1　P-P 图检验

首先对极大值样本 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)})$, 计算各 $p_i^{'}, p_i$ 和 $\hat{G}(x^{(i)})$, 然后做出点列 $(x^{(i)}, \hat{G}(x^{(i)})), i = 1, 2, \cdots, n$ 的图形, 观察其图形是否在直线附近, 结果如图 11 和图 12:

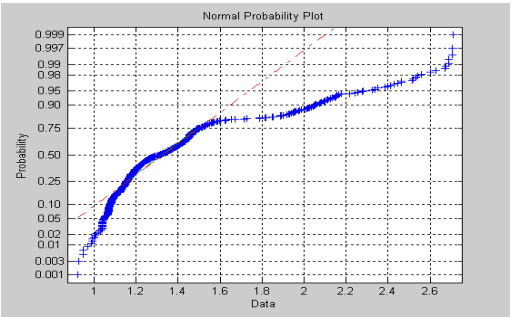

图 11　极大值分布 p-p 图

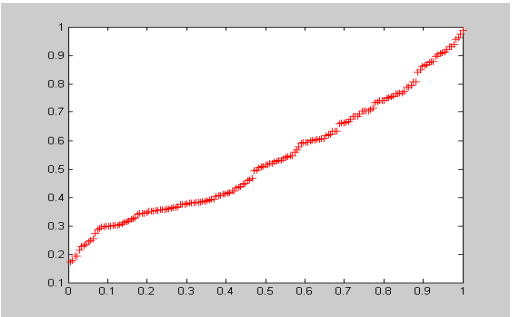

图 12　极小值分布 p-p 图

由图 11 和图 12 可看出, 两个图形均在直线附近, 即通过检验.

3.5.2 Kolmogorov 检验

下面利用 Kolmogorov 检验来检验经验分布与理论分布的一致性. 对极大值分布, $x \in R_1, \| G_n - G \| = \sup \mid G(x) - G_n(x) \mid = 0.079\ 6$. 对极小值分布, $x \in R_1, \| G_n - G \| = \sup \mid G(x) - G_n(x) \mid = 0.075\ 0$. 从表 5 和表 6 可以看出, 对极大, 极小值分布, 在显著性水平为 0.05 和 0.01 下均通过 Kolmogorov 检验, 说明所给出的理论分布估计与实际分布是一致的.

表 5　极大值分布的 Kolmogorov 检验

α 值	临界值 c_α	n	$c_\alpha/\sqrt{n}$	结论
0.05	1.358	147	0.112	接受 H
0.01	1.628	147	0.134 3	接受 H

表 6 极小值分布的 Kolmogorov 检验

α 值	临界值 c_α	n	$c_\alpha/\sqrt{n}$	结论
0.05	1.358	176	0.102	接受 H
0.01	1.628	176	0.123	接受 H

4 结论与评价

本文对给出的基金净值历史数据, 在对数据进行直观分析的基础上, 注意到数据的偏态, 厚尾的特征, 运用极值理论进行建模, 对临界值运用 POT 方法进行估计, 对模型的参数采用 Dehann 矩估计法进行估计, 并对模型进行了检验. 以上结果表明, 所得模型与实际数据符合较好, 故可以将模型的结果运用于实际, 对该基金净值的最大 (小) 值在一定概率意义下做出预测, 由表 3, 表 4 的结果可得: 在 90 以预测, 该基金净值的最大 (小) 值为 3.57(1.80); 在 99 此结果可以作为在实际交易 (卖出, 买进) 的参考.

由于采用的是历史数据建立的模型, 模型的参数也是根据历史数据做出的, 故在实际应用中, 应及时补充和更新数据对模型参数进行修正, 以便得到尽可能准确的估计.

参 考 文 献

[1] Dodd E L. The greatest and least variate under general laws of error[J]. Trans. Amer. Math. Soc., 1923,25:525-539.

[2] Frechet M. Sur la loi de probalilite de ecart maximum[J]. Ann. Soc. Polonaise Math. Cracow, 1927,06:93-95.

[3] Fisher R, Tippett L. Limiting forms of the frequency distributions of the largest of smallest member of a sample[J]. Proc. Camb. Phil. Soc., 1928,24:180-190.

[4] Hann L D. On regular variation and its application to the weak convergence of sample extremes[J]. Mathematical Centre racts, 1970,1:30-33.

[5] 高丽君, 李建平, 徐伟宣, 等. 基于 POT 方法的商业银行操作风险极端值估计 [J]. 运筹与管理, 2007,1(16):112-117.

[6] 高洪忠. 用 POT 方法估计损失分布尾部的效应分析 [J]. 数理统计与管理, 2003,4(23):64-69.

Application of extreme value theroy in predicting fund based on POT method

LI Xiao-kang

(Mathematics Department, Shaanxi University of Technology, Hanzhong 723000,China)

Abstract: For moding the data of net value of fund, according the tail feature of sample data of net value of fund, the paper estaiblishes the GPD model of maximum and minimum distrbution, using the POT method to determine the critical value, then estimates the parameters, and the models were tested. Finally, using the model to predict extreme points. The results well describes the feature of data, the forecast of the extreme points accordes the reality.

Keywords: extreme value distribution, POT method, net of fund, estimation

2000MSC: 62G07