区间数据模型在金融市场预测中的应用①

杨　威

（山西大学　管理与决策研究所　太原　030006）

摘　要：本文利用区间数据模型对金融市场进行区间预测分析，给出了有别于置信区间的新的区间预测方法。基于美国股票市场数据的实证分析结果表明，与传统的点值 Naïve 预测模型和 GARCH 类置信区间预测模型相比，本文提出的区间数据模型能够更好地利用数据信息，区间预测误差较小且具有统计显著性。此外，不同估计样本量、数据频度以及置信度的区间预测比较结果证实了区间数据模型的可靠性。因此，区间数据建模研究不仅为定量分析方法提供了新的研究视角，同时也能为市场交易和政策制定提供相关的决策依据。

关键词：区间数据；置信区间；预测

中图分类号：C931.1，F224.9　**文献标识码**：A　**文章编号**：（2016）01-0054-10

0　引言

金融市场预测长期以来都是金融研究中极富有挑战性的课题，学者们利用不同的计量经济和金融模型，基于可观测或可得样本数据建立统计模型对该问题进行了广泛的研究，为政府部门和相关行业、企业提供决策依据。然而，基于点值数据的计量模型仅能反映变量的水平或者波动单方面的信息，无法两者兼得，且无法直接给出市场预测的有效区间变化结构。例如：股票价格从开盘到收盘一直都在变化，但通常只用收盘价格来表示当天的价格情况，而该收盘价格其实只是该天特定时刻的一个特殊价格水平。在点值数据情形下的建模研究本身可能会导致部分信息的损失，如果对区间数据直接建模，即把区间数据作为一个整体来分析，能够包含更多的数据信息，减小信息损失。此外，区间数据广泛存在于经济、金融和社会生活中，可以通过下限和上限来刻画变量的变化范围，不仅具有丰富数据信息的优势，对经济决策制定而言，区间预测比点值预测能够提供更加全面的参考依据。因此，本文将针对金融市场中的区间数据建模，基于此给出金融市场价格的区间预测，并比较分析该区间建模方法的预测优势。

1　区间数据模型研究现状

1.1　区间数据模型理论研究现状

20 世纪 60 年代末，区间数据分析的理论和方法被提出，其应用逐步渗透到力学、数学、经济管理、工程等诸多领域。国内外早期区间数据的研究主要利用点值随机变量的区间取值来估计变量的分布函数[1-3]。然而，这些研究并没有考虑直接对区间数据整体进行建模。为了能够将区间数据模型与实际问题相融合，针对不同变量取值的特点，发展了因变量取值为区间情形[4-5]、自变量和因变量其中之一为区间且其他变量为点值情形[6]等系列成果。对于自变量和因变量均为区间数据的情形，一类基于不同的点值区间属性进行建模分析，另一类基于区间运算法则进行整体区间数据建模分析。

1.1.1　基于区间点值属性和点值运算的区间数据建模

由于区间数据可以唯一的由中点和极差（或者区间左、右边界）确定，一些学者利用这些点值数据

① 基金项目：国家自然科学基金青年项目（7150115）；教育部人文社科基金青年项目（14YJC630163）。

作者简介：杨威（1982—），男，山西大同人，博士，讲师，研究方向：区间时间序列分析、金融工程与风险管理，Email：yangwei@sxu.edu.cn。

来替代整个区间数据进行回归分析。代表性的区间数据建模方法是采用传统多元统计方法建立联立模型，其核心思想分为两大类[7]。第一，将区间数据的生成机制看成是中点和极差（或者区间左、右边界）两个独立的点过程，分别构建 MinMax 模型[8]和 CRM 模型[9]。但是，该方法不能保证模型本身暗含的假设条件，即预测区间左边界要小于等于区间右边界。Lima Neto and De Carvalho[10]在 CRM 模型中加入了极差约束条件，提出的 CCRM 模型能够确保预测区间结构的一致性。尽管如此，Gil 等[11]指出虽然利用区间点值属性研究线性区间模型的思路比较直观，但是模型参数估计不能依靠点值最小二乘方法，而需求解带有约束的最优化问题来保证参数估计的一致性，建议采用区间运算和区间最小二乘估计。第二，认为区间数据生成机制中的两个点过程相互影响，需在各自的模型方程中引入另一变量的滞后项，建立关于区间中点和极差（或者区间左、右边界）的向量自回归 VAR 模型[7]。然而，该方法依旧在传统的点值随机变量框架下进行研究，并没有将区间本身看作总体，估计过程中参数较多，估计效率和预测精度不理想[12]。

1.1.2 基于区间总体和区间运算法则的区间数据建模

为了突破上述困境，学者们提出建立统一模型的思想和建模方法，实现区间模型研究的对象为区间取值的随机变量。基于区间运算法则对区间数据进行建模研究始于 Moore[13]提出的区间分析（interval analysis），其模型参数估计采用区间最小二乘（interval least squares，ILS）方法。然而，区间最小二乘估计过程依赖区间距离的定义形式。Diamond[14]利用仿射变换函数研究区间数据之间的线性关系，其参数估计最优化准则可视为是建立在 Hausdorff 区间距离 d_H 上普通最小二乘方法的区间推广。Gil 等[15-16]拓展了 Diamond[14]的研究视角，基于 d_W 区间距离给出了仿射变换函数系数估计最优解的判断准则，且参数的统计性质易于处理和解释[17]。进一步，Köer and Näther[18]利用 D_K 区间距离推广了 d_W 区间距离。González-Rodríguez[19]从凸紧集合视角将单变量和多变量线性区间模型结构形式进行了推广。Gil 等[11]基于 Hausdorff 区间距离 d_H 和区间距离 D_K 提出了检验线性区间模型系数显著与否的渐进方法，但参数估计理论仍然不完善。Hu 和 He[20]以及 He 和 Hu[21-22]提出用区间最小二乘法对正则化区间模型进行参数估计，并利用区间计算直接得到预测区间。但是，该方法仅使用了区间样本含有的水平信息，且缺少参数估计的大样本性质、极限分布以及参数假设检验等方法。为了保持区间样本信息的完整性，韩艾等提出了在参数估计中保持区间样本信息的方法，在一定程度上完善了 Hu 和 He[20]研究中的不足之处。Blanco-Fernández 等[23]提出了一种结构更为灵活的模型处理随机区间之间的线性关系，同时基于区间距离 d_θ 给出了参数估计方法，其参数 θ 是区间中点距离和区间极差距离的分配权重[24]。一般而言，基于 d_θ 区间距离的线性区间模型参数估计过程会依赖于距离参数 θ 的选择。Sinova 等[25]基于均方误差目标，通过理论分析和仿真实验的方法给出了最优距离参数 θ 的选择方法。Han 等[12]首次提出了针对区间时间序列数据建立线性自回归模型的分析框架，提出了线性时序区间模型一种最小 D_K 距离参数估计、极限分布和假设检验方法，搭建了区间时序计量模型的初步理论基础。

1.2 区间数据模型应用研究现状

关于区间数据模型的应用研究，已有文献多集中于比较分析区间预测的精度，也有少量文献将区间数据模型应用于经济和金融问题的解释。在区间数据模型预测应用方面，除了利用线性区间回归模型以外，学者们还提出了诸如平滑化方法、基于多层感知的人工神经网络智能方法以及 k-领域方法等[26, 7, 27]。Lima Neto and De Carvalho[9]利用仿真实验比较分析了基于不同区间点值属性的区间数据建模方法的预测表现。Lima Neto and De Carvalho[10]指出区间中点和区间极差的相关性会影响区间数据模型的预测精度。Hu 和 He[20]、He 等[28]、He 和 Hu[21-22]以及 Hu[29]利用股票市场和抵押贷款等数据的实证分析表明，基于区间运算得到区间预测要比传统基于点值运算和置信区间方法得到的区间预测精准。Yang 等[30]从区间时间序列模型出发，利用美国股票市场数据研究发现区间时序模型比传统点值 AR 模型在区间极差的短期和长期预测方面更具优势，并且统计检验表明该预测优势的稳健性。Maia 等[31]和 Maia 和 De Carvalho[27]将模型组合的思想引入到区间预测中，提出了基于区间点值属性和点值运算的（线性模型+神经网络）和（Holt 指数模型+神经网络）组合模型，通过仿真实验和实证分析表明模型组合方法能获得更加精准的区间预测。在区间数据模型经济和金融问题分析方面，基于 Han 等[12]所提出的区间时序模型框架，Yang 等[32]提出了区间虚拟变量，并且将虚拟变量与区间模型相结合，给出了能够度量危机事件对区间数据过

程影响的模型以及参数的经济解释。Yang 等[33]利用区间数据模型实证分析了美国次贷危机前后原油价格与美国股票市场相互作用的变化特征。

综上所述，现有的区间数据理论研究主要是从区间点值属性和区间数据整体出发，初步形成了区间数据建模的理论框架，而应用研究主要是比较不同区间数据模型的预测精度。Yang 等[30]研究表明，基于区间运算的区间时间序列模型比传统点值 AR 单方程模型能获得更加精准的区间预测。因此，本文进一步比较分析区间数据模型与基于点值模型的置信区间方法两者的区间预测精度。

2　区间数据预测模型

2.1　区间随机变量运算法则

设 $K_c(\mathbb{R})$ 为 $\mathbb{R}$ 中的非空紧区间构成的集合，且 $K_c(\mathbb{R})$ 中赋予加法和数乘运算，即对于任意的 A，$B\in K_c(\mathbb{R})$，我们有 $A+B$ 和 λA。在不同的区间运算假设下，由于缺少对称区间元素使得 $K_c(\mathbb{R})$ 不能够成为一个线性空间（Gil et al.，2007）。为此，学者们引入了 Hukuhara 差分（Hukuhara difference）概念，其形式定义为：对任意的 A，$B\in K_c(\mathbb{R})$，Hukuhara 差分 $C=A-_H B$ 满足

$$A=B+C,\ C\in K_c(\mathbb{R}) \tag{1}$$

给定概率空间 (Ω, A, P)，映射 $X:\Omega\to K_c(\mathbb{R})$ 如果满足 $A\mid B_{d_*}$-可测性，则称为相应于 (Ω, A, P) 的区间随机变量，其中 d_* 是 $K_c(\mathbb{R})$ 中的距离测度，B_{d_*} 表示 $K_c(\mathbb{R})$ 上由 d_* 诱导的 σ-域。为了针对区间数据模型建立相应的估计方法，首先需要明确区间距离函数 $d_*(\cdot,\cdot)$ 的具体形式。比较直观的区间距离度量方式是 Hausdorff 距离，随后这种区间距离度量从不同角度得到了推广，例如：d_W 距离，d_θ 距离和 D_K 距离等[23, 25, 12]。其中，$D_K(\cdot,\cdot)$ 距离具有如下最为一般化的形式，且用 $\|\cdot\|_K$ 表示区间相应于核函数 K 的 L^2 距离。对任意的 A，$B\in K_c(\mathbb{R})$，有

$$D_K(A, B)=\sqrt{\int_{S^0}(s_A(u)-s_B(u))(s_A(v)-s_B(v))dK(u, v)} \tag{2}$$

其中，K 是一个对称半正定核函数，函数 s 是从空间 $K_c(\mathbb{R})$ 到 Hilbert 空间中闭凸锥 $[C(S^0), \|\cdot\|_K]$ 的一个等距映射，其定义形式为 $S_A(u)=\sup_{a\in A}\langle u, a\rangle$，$u\in R^1$，$S^0=\{u\in R^1,\ |u|=1\}=\{1, -1\}$。如果用 $\langle\cdot,\cdot\rangle_K$ 表示对应的内积运算结构，那么有

$$D_K^2(A, B)=\langle s_A-s_B, s_A-s_B\rangle_K \tag{3}$$

2.2　区间数据预测模型

基于区间数据运算法则，本文构建如下线性区间数据回归模型：

$$\Delta PS_t=\alpha_0+\beta_0 I_0+\sum_{i=1}^{p}\beta_i\Delta\hat{PS}_{t-i}+u_t \tag{4}$$

其中，α_0，β_0，β_i（$i=1, 2, \cdots, p$）均为待估参数；$\alpha_0+\beta_0 I_0=[\alpha_0-\beta_0/2, \alpha_0+\beta_0/2]$ 是区间截距项；$I_0=[-\frac{1}{2}, \frac{1}{2}]$ 是常值单位区间；$PS_t=[PSL_t, PSH_t]$ 和 $\hat{PS}_t=[PSH_t, PSL_t]$ 为金融资产区间价格过程其中 PSL_t 和 PSH_t 分别表示区间低价和区间高价；ΔPS_t 表示区间价格过程 $\{PS_t\}$ 的 Hukuhara 差分；$u_t=[u_{Lt}, u_{Ht}]$ 是相对信息集 I_{t-1} 的区间鞅差分序列过程并且满足 $E(u_t\mid I_{t-1})=[0, 0]$ 几乎处处成立。

上述区间数据模型的优势在于，其不仅能够充分利用数据信息进行更加有效的参数估计，而且可以从区间数据模型中得到传统点值（诸如高价、低价和极差）预测模型。

$$\Delta PSL_t=\alpha_0-\frac{1}{2}\beta_0+\sum_{i=1}^{p}\beta_i\Delta PSH_{t-i}+u_{Lt} \tag{5}$$

$$\Delta PSH_t=\alpha_0+\frac{1}{2}\beta_0+\sum_{i=1}^{p}\beta_i\Delta PSL_{t-i}+u_{Ht} \tag{6}$$

$$\Delta PSR_t=\beta_0-\sum_{i=1}^{p}\beta_i\Delta PSR_{t-i}+u_t^r \tag{7}$$

其中，PSR_t 表示区间过程的价格极差，ΔPSL_t，ΔPSH_t，ΔPSR_t 表示点值序列的一阶差分过程，u_t^r 是可加的信息项。注意到参数 α_0 在式(7)中不能识别，因为式(7)只用到了区间价格极差信息而不包含水平趋势信息。可以参照 Yang 等[32-33]，考虑区间数据协整关系的建模表达。基于此方法预测的区间高价和区间低价分别记为 Interval^H 和 Interval^L。

2.3　区间数据模型参数估计和检验

对于区间数据模型的参数估计，我们将采用 Han 等[12]所提出的两阶段最小 D_K 距离估计方法。假设区间数据回归模型的参数向量为 $\phi[\alpha_0, \beta_0, \beta_1, \cdots, \beta_p]$，那么最小 D_K 距离估计量 $\hat{\phi}$ 为

$$\hat{\phi} = \arg\min_{\phi \in \Phi} \hat{Q}_T(\phi) \tag{8}$$

其中 $\hat{Q}_T(\phi)$ 为区间数据模型的残差平方和，即

$$\hat{Q}_T(\phi) = \frac{1}{T}\sum_{t=1}^{T} q_t(\phi) \tag{9}$$

$$q_t(\phi) = \| \hat{u}_t(\phi) \|_K^2 = \| Y_t - Z'_t(\phi)\phi \|_K^2 = D_K^2[Y_t, Z'_t(\phi)\phi] \tag{10}$$

其中，Y_t 表示区间数据模型中的随机区间解释变量；$Z_t(\phi)$ 表示区间数据模型中参数向量 ϕ 的系数向量；$q_t(\phi)$ 为区间变量残差项；K 是半正定对称矩阵，$Z_t(\phi)$ 表示相应于参数向量 ϕ 的区间解释向量。基于参数估计量的一致性和渐进正态性，我们可以利用单值 Wald 方法在 H_0：$\boldsymbol{R}\phi^0 = r$ 原假设下对相关区间数据模型参数进行假设检验，其中矩阵 $\boldsymbol{R}$ 和 r 用来表述参数假设检验的结构形式。

3　数据样本选取及其统计分析

本文实证研究中选择美国股票市场中主要价格指数进行分析，其中包括美国股票市场标普 500 指数(Standard & Poor's 500 index，S&P500)、道琼斯工业指数(Dow Jones industrial average，DJIA)、纳斯达克指数(national association of securities dealers automated quotations，NASDAQ)。用 $PS_t = [PSL_t, PSH_t]$ 表示区间价格过程，PSR_t 表示区间价格的极差过程，所有 PSL_t 和 PSH_t 数据均采用日度股票指数最低价和最高价的对数价格形式。数据样本期为为 2003 年 1 月 3 日 ~ 2013 年 12 月 31 日，全部数据样本来自万得(Wind)数据库，数据样本基本统计量见表 1。

表 1　美国股票市场区间价格属性变量的基本统计分析

Table 1　Basic statistical analysis of interval price attribute variables in the U. S. stock market

股票指数	变量	均值	中值	最大值	最小值	标准差	偏度	峰度	JB 统计量	P 值
SP500	PSH	7.117 3	7.131 6	7.522 6	6.544 3	0.175 2	-0.332 5	2.968 0	51	0.000 0
	PSL	7.103 7	7.118 0	7.518 8	6.502 5	0.179 7	-0.385 5	3.060 1	69	0.000 0
	PSR	0.013 6	0.010 6	0.109 0	0.002 0	0.010 8	3.562 7	22.637 7	50 351	0.000 0
	ΔPSH	0.000 3	0.000 3	0.077 3	-0.071 0	0.009 4	-0.052 4	10.951 2	7295	0.000 0
	ΔPSL	0.000 3	0.000 8	0.087 2	-0.085 7	0.011 3	-0.465 6	16.222 3	20 271	0.000 0
DJIA	PSH	9.331 9	9.321 8	9.716 5	8.811 3	0.165 4	-0.161 6	2.784 2	17	0.000 2
	PSL	9.318 7	9.308 2	9.711 8	8.774 9	0.169 3	-0.211 5	2.866 9	23	0.000 0
	PSR	0.013 2	0.010 5	0.121 5	0.001 7	0.010 4	3.806 7	26.276 2	69 196	0.000 0
	ΔPSH	0.000 2	0.000 2	0.063 3	-0.059 6	0.008 6	-0.081 6	9.811 5	5356	0.000 0
	ΔPSL	0.000 3	0.000 7	0.083 9	-0.091 5	0.010 5	-0.425 5	17.428 1	24 101	0.000 0

续表

股票指数	变量	均值	中值	最大值	最小值	标准差	偏度	峰度	JB 统计量	P 值
NASDAQ	PSH	7.752 3	7.746 6	8.337 5	7.154 3	0.229 6	-0.052 0	3.067 4	2	0.412 8
	PSL	7.737 8	7.732 5	8.333 5	7.133 5	0.233 5	-0.076 7	3.089 6	4	0.162 1
	PSR	0.014 5	0.012 2	0.111 3	0.002 0	0.010 0	3.238 7	20.236 5	39 119	0.000 0
	ΔPSH	0.000 4	0.001 1	0.086 9	-0.074 2	0.010 9	-0.217 1	8.480 2	3487	0.000 0
	ΔPSL	0.000 4	0.000 8	0.106 5	-0.091 5	0.012 5	-0.088 4	11.756 4	8850	0.000 0

表 1 中呈现了美国股票市场中主要区间价格指数的各个点值属性变量的基本统计分析。数据结果表明，对于 S&P500 指数、DJIA 指数、NASDAQ 指数而言，各自价格过程的区间高价和区间低价在标准差、峰度、偏度方面有较为相似的特征；区间价格极差(日内波动性) 与区间高价差分和低价差分(日间波动性) 的特征有很大的不同；各个属性变量的 JB 统计量值很大且 P 值为 0(NASDAQ 股票指数高价、低价除外)，表明数据样本基本上不服从正态分布。虽然这些点值属性变量是从同一区间过程中提取的，但是不同变量所包含的信息是不同的。因此，本文尝试从区间数据整体建模的角度出发，研究区间数据模型在金融市场预测中的应用。

4　区间预测比较方案及判别准则

为了便于实证分析，我们在这里简要介绍一下本文中其他不同的区间预测方法以及预测结果的精度判别准则。

4.1　基于点值数据模型的置信区间预测方法

传统的区间预测方法是指预测值以一定的概率落于某个区间，这个区间是在某个置信水平下得到的，即以置信区间作为未来点值过程的区间预测。本文将比较研究基于点值 GARCH(1，1) 模型(置信) 区间预测与区间时间序列模型区间预测的精度。

$$\begin{cases}\Delta PC_t = c + \gamma \Delta PC_{t-1} + \varepsilon_t \\ \sigma_t^2 = \delta_0 + \delta_1 \varepsilon_{t-1}^2 + \delta_2 \sigma_{t-1}^2 \\ (\varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots) \sim N(0, \sigma_t^2)\end{cases} \tag{11}$$

其中，$\{PC_t\}$ 表示金融资产的收盘价格序列；c，γ，δ_0，δ_1，δ_2 均为 GARCH 模型中的待估参数。注意到式(11) 描述了点值过程的水平和波动特征，相应的置信区间反映了未来取值的不确定性，但是该种区间预测方法与我们所提出的区间数据建模及预测方法完全不同。因此，本文将比较分析区间数据方法与置信区间方法的区间预测优势，其中置信区间的构建将考虑不同的置信水平。基于此方法预测的区间高价和区间低价分别记为 $GARCH^H$ 和 $GARCH^L$。

此外，我们将考虑区间数据的 Naïve 预测方法，并将该方法的预测结果作为其他各个预测模型表现的比较基准。假设随机区间过程服从于鞅序列或者随机游走过程，那么在时刻 t 可以获得的所有信息 I_t 条件下，区间过程在时刻 $t+1$ 的区间预测值等于时刻 t 的真实区间值，即

$\hat{PS}_{t+1} = [\hat{PSL}_{t+1}, \hat{PSH}_{t+1}] = [PSL_t, PSH_t]$。基于此方法预测的区间高价和区间低价分别记为 $NAIVE^H$ 和 $NAIVE^L$。

4.2　区间预测精度判别准则

基于区间数据模型和点值 GARCH 类置信区间方法的区间预测精度比较可以通过以下两种判别准则来进行。

首先，我们可以利用平均绝对误差(mean absolute deviation，MAD) 和平方误差(mean squared error，MSE) 这两个常用的误差判断准则。令 $\{PA_t\}$ 为观测到的区间属性变量(例如：PSR_t、PSL_t 和 PSH_t)，而 $\{\hat{PA}_t\}$ 为相应的预测值，$t = 1, 2, \cdots, n$，n 为观测样品数量。那么 MAD 和 MSE 定义如下：

$$MAD \triangleq \frac{1}{n}\sum_{t=1}^{n}|PA_t - \hat{PA}_t|\ ,\ MSE \triangleq \frac{1}{n}\sum_{t=1}^{n}(PA_t - \hat{PA}_t)^2 \tag{12}$$

一般而言，较小的 MAD 和 MSE 表示相应模型的区间预测精度较高。

其次，为了从统计显著性角度验证不同模型的区间预测优势，我们可以利用修正的 Diebold-Mariano 检验方法（modified Diebold-Mariano，MDM）[34-35]。令 e_{it} 和 e_{jt} 分别表示由模型 i 和模型 j 所得到的预测误差，平方预测误差定义为 $L(\hat{PA}_{it}) = e_{it}^2$ 和 $L(\hat{PA}_{jt}) = e_{jt}^2$，将损失差分序列设定为 $d_t = L(\hat{PA}_{it}) - L(\hat{PA}_{jt})$，那么检验模型 i 与模型 j 预测表现的差异性等价于检验损失差分序列 d_t 的总体均值是否为 0。假设序列 d_t 是协方差平稳且有短记忆特征，检验原假设：两个模型具有相同的预测表现。Harvey 等[35]给出了如下修正的 Dieold-Mariano 检验统计量：

$$\left[\frac{T+1-2h+T^{-1}h(h-1)}{T}\right]^{1/2} \bar{d}/V_h(\bar{d})^{1/2} \tag{13}$$

其中，$V_h(\bar{d}) = T^{-1}\left[\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k\right]$，$\gamma_k$ 是序列 d_t 的 k-阶自协方差，h 为预测阶数。该检验统计量具有渐进 t_{T-1} 分布。该检验可以利用两种预测误差测度 MAD 和 MSE 分别进行，对于 Model 1/Model 2 而言，显著负的 MDM 检验统计量表示模型 1 产生的预测误差 MAD 或者 MSE 要显著的比模型 2 产生的预测误差要小。

5 实证结果分析

为了说明本文提出的区间数据模型的预测优势，我们将基于美国股票市场数据对比分析不同模型的区间预测表现，同时在预测误差 MAD 和 MSE 比较分析基础上，利用 MDM 方法进行统计检验，为验证区间数据模型的预测优势提供有力的支撑。

5.1 样本外预测比较分析

对于不同模型的样本外区间预测比较，我们采用样本外一阶滚动预测，分别计算区间高价和区间低价的预测误差 MAD 和 MSE，误差计算结果见表 2。

表 2 美国股票市场指数的样本外区间预测误差结果

Table 2 Out-of-Sample interval forecasting error results of the U. S. stock market indices

Variables	S&P500		DJIA		NASDAQ	
	MAD	MSE (E^{-04})	MAD	MSE (E^{-04})	MAD	MSE (E^{-04})
$NAIVE^H$	0.0061	0.8627	0.0056	0.7218	0.0075	0.0115
$Interval^H$	0.0059	0.7980	0.0054	0.6646	0.0073	0.0110
$GARCH^H$	0.0068	0.8652	0.0061	0.6736	0.0083	0.0127
$NAIVE^L$	0.0071	0.0127	0.0066	0.0109	0.0086	0.0155
$Interval^L$	0.0068	0.0120	0.0062	0.0100	0.0084	0.0151
$GARCH^L$	0.0070	0.0104	0.0063	0.0083	0.0085	0.0140
$NAIVE^R$	0.0058	0.0073	0.0056	0.0071	0.0057	0.0069
$Interval^R$	0.0050	0.0058	0.0049	0.0056	0.0051	0.0057
$GARCH^R$	0.0089	0.0156	0.0078	0.0121	0.0109	0.0197

从表 2 的区间预测误差结果可以看出，针对美国股票市场 S&P500 指数、DJIA 指数、NASDAQ 指数的区间高价、区间低价和区间极差的预测，本文提出的区间数据模型较 Naïve 预测方法和 GARCH 类置信区间预测方法有较小的预测误差 MAD。在预测稳定性方面，除了区间低价预测表现外，区间数据模型的区间高价预测和区间极差预测表现较 Naïve 预测方法和 GARCH 类置信区间预测方法有较小的预测误差 MSE。

进一步，为了检验不同模型的样本外预测优势，本文基于预测误差 MAD 和 MSE 进行 MDM 统计检验，具体结果见表 3 ~ 表 5。

表 3　美国股票市场 S&P500 指数的样本外区间预测比较检验结果

Table 3　Out-of-Sample interval forecasting comparison test results of S&P500 index in the U. S. stock market

检验标准	Model 1/Model 2	高价		低价		极差	
		统计量	*P* 值	统计量	*P* 值	统计量	*P* 值
基于 MAD 的 MDM 检验	GARCH/NAÏVE	4. 0276	0. 0000	-0. 8300	0. 4081	4. 9430	0. 0000
	Interval/NAÏVE	-2. 8375	0. 0053	-3. 5167	0. 0000	-7. 2754	0. 0000
	Interval/GARCH	-4. 4325	0. 0000	-0. 3122	0. 7554	-5. 4339	0. 0000
基于 MSE 的 MDM 检验	GARCH/NAÏVE	0. 9172	0. 3608	-1. 3208	0. 1890	2. 3668	0. 0195
	Interval/NAÏVE	-0. 3177	0. 7513	-1. 2585	0. 2106	-3. 0030	0. 0032
	Interval/GARCH	-1. 1753	0. 2421	1. 1800	0. 2403	-2. 4635	0. 0151

表 4　美国股票市场 DJIA 指数的样本外区间预测比较检验结果

Table 4　Out-of-Sample interval forecasting comparison test results of DJIA index in the U. S. stock market

检验标准	Model 1/Model 2	高价		低价		极差	
		统计量	*P* 值	统计量	*P* 值	统计量	*P* 值
基于 MAD 的 MDM 检验	GARCH/NAÏVE	3. 3046	0. 0012	-1. 4257	0. 1564	4. 3207	0. 0000
	Interval/NAÏVE	-2. 8049	0. 0058	-4. 0146	0. 0000	-6. 5968	0. 0000
	Interval/GARCH	-4. 2052	0. 0000	0. 4072	0. 6845	-5. 0890	0. 0000
基于 MSE 的 MDM 检验	GARCH/NAÏVE	-1. 3376	0. 1835	-1. 3569	0. 1773	2. 0750	0. 0400
	Interval/NAÏVE	0. 4672	0. 6412	-2. 3297	0. 0214	-2. 0955	0. 0381
	Interval/GARCH	0. 8400	0. 4025	1. 2739	0. 2051	-2. 2894	0. 0237

表 5　美国股票市场 NASDAQ 指数的样本外区间预测比较检验结果

Table 5　Out-of-Sample interval forecasting comparison test results of NASDAQ index in the U. S. stock market

检验标准	Model 1/Model 2	高价		低价		极差	
		统计量	*P* 值	统计量	*P* 值	统计量	*P* 值
基于 MAD 的 MDM 检验	GARCH/NAÏVE	4. 3144	0. 0000	-0. 5710	0. 5691	6. 4524	0. 0000
	Interval/NAÏVE	-2. 2752	0. 0246	-2. 9711	0. 0036	-8. 1462	0. 0000
	Interval/GARCH	-4. 4036	0. 0000	-0. 2690	0. 7884	-6. 7407	0. 0000
基于 MSE 的 MDM 检验	GARCH/NAÏVE	2. 6206	0. 0099	-1. 1783	0. 2409	2. 9029	0. 0044
	Interval/NAÏVE	-0. 3928	0. 6951	-1. 3376	0. 1835	-3. 1734	0. 0019
	Interval/GARCH	-2. 4953	0. 0139	1. 0688	0. 2872	-2. 9273	0. 0041

表 3 ~ 表 5 的统计检验结果表明：首先，基于 MAD 的统计检验表明区间数据模型的区间高价和区间极差预测均要显著优于 Naïve 预测模型和 GARCH 类置信区间预测模型，而 GARCH 类置信区间预测方法表现最差。其次，虽然区间数据模型的区间低价预测的统计检验显著性不强，但是统计量多数是负值，因此区间数据模型在区间低价预测有一定优势。第三，基于 MSE 的统计检验有类似的结论，表明区间数据模型的预测优势具有稳定性。

5.2 稳健性检验

为了检验区间数据模型预测优势的稳健性，我们考虑了多种不同参数对模型区间预测表现的影响。在利用美国股票市场区间数据的实证研究中，除了基于统计误差 MAD 和 MSE 的比较分析外，还利用 MDM 统计检验方法验证区间数据模型预测优势的显著性。此外，文中还对不同的样本估计窗宽、不同数据频度以及不同的置信水平进行了区间预测比较分析。各种参数条件下均得到类似的结论，进一步证实了区间数据模型的可靠性。

6 结论及展望

传统点值数据模型已经在计量经济金融分析领域得到了广泛的应用，但是点值数据模型往往损失了部分数据信息，因此本文提出了区间数据建模方法并实证分析其预测能力。实证结果表明区间数据模型比传统的 Naïve 预测模型和 GARCH 类置信区间预测模型在区间高价、区间低价和区间极差预测方面均有明显优势，并且该优势在统计上是显著的。因此，区间数据建模理论和方法可以为金融市场预测分析提供更有价值的参考信息，具有更加广泛的应用前景，值得进一步深入研究。

参考文献：

[1] Wang, Z., J. C. Cardiner. A class of estimation of the survival function from interval-censored data [J]. *The Annals of Statistics*, 1996, 24: 647-658.

[2] 郑祖康，丁邦俊. 关于区间数据的分布函数估计问题 [J]. 应用概率统计，2004，20 (2)：119-125.
Zheng, Z., B. Ding. Problems of estimating a distribution function with interval censored data [J]. *Chinese Journal of Applied Probablility and Statistics*, 2004, 20 (2): 119-125. (in Chinese)

[3] 李坟华，于珊珊，郭均鹏. 基于区间分析的投资组 VaR 计算新方法 [J]. 数理统计与管理，2013，32 (3)：564-570.
Li, W., S. Yu, J. Guo. Mew Method of VaR about the investment portfolio based on interval analysis [J]. *Journal of Applied Statistics and Management*, 2013, 32 (3): 564-570. (in Chinese)

[4] Hong, H., E. Tamer. Inference in censored models with endogenous regressors [J]. *Econometrica*, 2003, 71 (3): 905-932.

[5] Hashimoto, E. M., E. M. M. Ortega, G. A. Paula, M. L. Barreto. Regression models for grouped survival data: Estimation and sensitivity analysis [J]. *Computational Statistics and Data Analysis*, 2011, 55: 993-1007.

[6] Manski, C. F., E. Tamer. Inference on regressions with interval data on a regressor or outcome [J]. *Econometrica*, 2002, 70 (2): 519-546.

[7] Arroyo, J., R. Espínola, C. Maté. Different approaches to forecast interval time series: A comparison in finance [J]. *Computational Economics*, 2011, 37: 169-191.

[8] Billard, L., E. Diday. From the statistics of data to the statistics of knowledge: symbolic data analysis [J]. *Journal of the American Statistical Association*, 2003, 98 (462): 470-487.

[9] Lima Neto, E. A., F. A. T. De Carvalho. Center and range method for fitting a linear regression model to symbolic interval data [J]. *Computational Statistics and Data Analysis*, 2008, 52: 1500-1515.

[10] Lima Neto, E. A., F. A. T. De Carvalho. Constrained linear regression models for symbolic interval-valued variables [J]. *Computational Statistics and Data Analysis*, 2010, 54: 333-347.

[11] Gil, M. A., G. González-Rodríguez, A. Colubi, et al. Testing linear independence in linear models with interval-valued data [J]. *Computational Statistical and Data Analysis*, 2007, 51: 3002-3015.

[12] Han, A., Y. Hong, S. Wang. Autoregressive conditional models for interval-valued time series data. Available at: http://economics.yale.edu/sites/default/files/hong-120926.pdf, 2012 [2012-9-1].

[13] Moore, R. E. Interval Analysis [M]. Englewood Cliffs, N. J.: *Prentice-Hall*, 1966.

[14] Diamond, P. Least squares fitting of compact set-valued data [J]. *Journal of Mathematical Analysis and Applications*, 1990, 14: 531-544.

[15] Gil, M. A., M. T. López-García, M. A. Lubiano, et al. Regression and correlation analyses of a linear relation between randomintervals [J]. *Test*, 2001, 10: 183-201.

[16] Gil, M. A., M. A. Lubiano, M. Montenegro, et al. Least squares fitting of an affine function and strength of association for

interval-valued data [J]. *Metrika*, 2002, 56: 97-111.

[17] Bertoluzza, C., N. Corral, A. Salas. On a new class of distances between fuzzy numbers [J]. *Mathware & Soft Computing*, 1995, 2: 71-84.

[18] Körner, R., W. Näther. On the variance of random fuzzy variables. In: Bertoluzza, C., MA. Gil, DA Ralescu (eds) Statistical Modeling, Analysis and Management of Fuzzy Data [M]. *Physica-Verlag*, 2002, 87: 22-39.

[19] González-Rodríguez, G., A. Blanco, N. Corral, et al. Least squares estimation of linear regression models for convex compact randomsets [J]. *Advances in Data Analysis and Classification*, 2007, 1: 67-81.

[20] Hu, C., L. He. An Application of Interval Methods to Stock Market Forecasting [J]. *Journal of Reliable Computing*, 2007, 13 (5): 423-434.

[21] He, L., C. Hu. Impacts of interval computing on stock market variability forecasting [J]. *Computational Economics*, 2009, 33: 263-276.

[22] He, L., C. Hu. Midpoint method and accuracy of variability forecasting [J]. *Empirical Economics*, 2010, 38: 705-715.

[23] Blanco-Fernández, A., N. Corral, G. González-Rodríguez. Estimation of a flexible simple linear model for interval data based on setarithmetic [J]. *Computational Statistical and Data Analysis*, 2011, 55: 2568-2578.

[24] Trutschnig, W., G. González-Rodríguez, A. Colubi, et al. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid andspread [J]. *Information Science*, 2009, 179: 3964-3972.

[25] Sinova, B., A. Colubi, M. A. Gil, G. González-Rodríguez. Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric [J]. *Information Sciences*, 2012, 199: 109-124.

[26] Muñoz San Roque, A., C. Maté, J. Arroyo, et al. iMLP: Applying multi-layer perceptrons to interval-valued data [J]. *Neural Processing Letters*, 2007, 25 (2): 157-169.

[27] Maia, A. L. S., F. A. T. De Carvalho. Holt's exponential smoothing and neural network models for forecasting interval-valued time series [J]. *International Journal of Forecasting*, 2011, 27: 740-759.

[28] He, L., C. Hu, M. Casey. Prediction of variability in mortgage rates: interval computing solutions [J]. *Journal of Risk Finance*, 2009, 10 (2): 142-154.

[29] Hu, C. A note on probabilistic confidence of the stock market ILS interval forecasts [J]. *Journal of Risk Finance*, 2010, 11 (4): 410-415.

[30] Yang, W., A. Han, S. Wang. Forecasting financial volatility with interval-valued time Series data [J]. *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, 2014, 7: 1224-1233.

[31] Maia, A. L. S., F. A. T. De Carvalho, T. B. Ludermir. Forecasting models for interval-valued time series [J]. *Neurocomputing*, 2008, 71: 3344-3352.

[32] Yang, W., A. Han, K. Cai, S. Wang. ACIX model with interval dummy variables and its application [J]. *Procedia Computer Science*, 2012, 9: 1273-1282.

[33] Yang, W., A. Han, S. Wang. Analysis of the interaction between crude oil price and US stock market based on interval data [J]. *International Journal of Energy and Statistics*, 2013, 1 (2): 85-98.

[34] Diebold, F. X., R. S. Mariano. Comparing predictive accuracy [J]. *Journal of Business & Economic Statistics*, 1995, 13: 253-263.

[35] Harvey D., S. Leybourne, P. Newbold. Testing the equality of prediction mean squared errors [J]. *International Journal of Forecasting*, 1997, 13: 281-291.

The Application of Interval Data Models in the Financial Market Forecasts

Yang wei

Institute of Management and Decision, Shanxi University, Taiyuan 030006, China

Abstract: This paper used interval data model to give interval prediction of financial markets, and proposed a new interval forecasting method which was different from confidence interval method. The empirical analysis results based on the U. S. stock market data showed that, compared with the traditional point-valued Naïve forecasting model and GARCH confidence interval forecasting model, the interval model proposed in this paper could better

utilize the data information and had a small interval prediction error which was statistically significant. In addition, some other comparison results of interval forecasts based on different estimating samples, data frequency and confidence levels proved the reliability of this interval data methodology. Therefore, the interval data modelling not only provided a new research perspective for quantitative analysis, but also provided related decision reference for market trading and policy making.

Key words: Interval data; Confidence interval; Forecast