文章编号:0253-2778(2019)08-0635-10

基于分位数回归森林的 VaR 估计及风险因素分析

苟小菊,王 芊

(中国科学技术大学管理学院,安徽合肥 230026)

摘要:构建非参数、集成性的分位数回归森林算法,对上证综指和标普 500 指数的 VaR 进行了估计;同时构建了其他一些主流的方法,包括历史模拟、GARCH 族方法、弹性网络、门限分位数回归、CAViaR 等,进行检验和对比.通过对不同置信水平下的 VaR 估计进行多种检验,验证了该方法的有效性和稳健性.进一步,基于分位数回归森林模型定义了一种特征重要性度量方法,评估了各个因素对于风险值的影响权重大小,发现过去一日收益率对上证综指的风险值影响较大,而波动率对标普 500 指数的风险值影响较大,整体来看两国股市间的风险传导性较弱;并引入偏相依关系,动态地分析了各个因素在不同水平下对于风险值的作用方向,一定程度上弥补了机器学习算法在金融应用中一直存在的"黑箱性"问题.

关键词:分位数回归森林;在险价值;风险因素分析

中图分类号:F224;F832.51 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2019.08.007

引用格式:苟小菊,王芊. 基于分位数回归森林的 VaR 估计及风险因素分析[J]. 中国科学技术大学学报,2019,49(8):635-644.

GOU Xiaoju, WANG Qian. VaR estimation based on quantile regression forest and risk factors analysis[J]. Journal of University of Science and Technology of China, 2019,49(8):635-644.

VaR estimation based on quantile regression forest and risk factors analysis

GOU Xiaoju, WANG Qian

(*School of Management, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: Quantile regression forests as a nonparametric and ensemble method were built to estimate the VaR of Shanghai Composite Index and the S&P 500 Index at different confidence levels. Meanwhile, other methods were built for comparison, including historic simulation, GARCH, elastic net, threshold quantile regression model and CAViaR, and the superiority of the proposed method was verified. Further, a new measurement method of variable importance based on the quantile regression forest was defined to judge the importance of various factors on the risk value, and it was discovered that the past one day yield has the greatest influence on the risk value of the Shanghai Composite Index, and that the volatility has the greatest influence on S&P 500 Index risk value. At the same time, the risk conduction between China and US is weak. Further, by dynamically analyzing the partial dependence between the factors and risk value, the "black box" problem of machine learning used in financial applications has been remedied to some extent.

Key words: quantile regression forest; VaR; risk factor analysis

收稿日期:2018-06-13;**修回日期**:2018-11-02

基金项目:国家自然科学基金青年项目(71701191)资助.

作者简介:苟小菊(通讯作者),女,1962 年生,博士/副教授.研究方向:国际金融. E-mail: xjgou@ustc.edu.cn

0 引言

市场风险的管理是金融领域核心问题之一,巴塞尔协议提出的VaR方法是一种应用广泛的度量方法,目前对VaR的估计可以分为参数法和非参数法两种.

参数法通过对市场收益率的分布进行假定或通过特定形式的回归方法对于收益率的分位点进行估计. Morgan公司最早开发的经典工具RiskMetrics™通过GARCH类模型对VaR进行动态估计. Engle和Manganelli[1]从VaR的自回归角度对已有线性分位数模型进行了拓展,提出了CAViaR模型,下一日的VaR是之前VaR估计的函数. Meng和Taylor[2]则进一步对CAViaR模型进行了拓展,考虑了已实现波动率和日内价差、日间价差的因素,构建线性的分位数回归模型,并对多国股市进行了检验. 许启发和康宁[3]考虑了收益率的非对称异质效应,认为在极端的市场条件下收益率的自相关性较强,选取过去一日的收益率作为门限变量,通过门限分位数回归模型对于我国股市进行了研究. 罗克兵等[4]使用线性分位点回归模型,对于沪深两市高频连跌数据的VaR进行了估计,并根据回归模型的系数判断当连跌的收益率越大时未来的风险值也越大. Haugom等[5]则考虑了不同交易频率的投资者的影响,通过短期、中期、长期波动率指标,构建线性分位数回归对于一日VaR进行了估计. 同时由于金融市场收益率往往为厚尾的,极值模型通过捕捉收益率分布的厚尾特征,也得到了广泛的应用,如Liu等[6]构建了基于已实现波动率的异方差自回归HAR模型,并对残差项结合POT极值模型拟合,对沪深300指数的VaR进行估计,认为结合极值模型的估计最为稳健. 近年来,对于VaR估计的集成性方法也逐渐得到关注,以降低单一模型高估或者低估概率的发生: Halbleib和Pohemeier[7]首先提出了集成性的VaR预测方法,将多个单一模型的预测结果进行加权组合,实证检验发现相比于单一模型,集成性的模型更加稳健; Bayer[8]在此基础上引入了正则项约束,构建了集成性的弹性网络分位数回归模型.

而非参数方法不对模型进行过多假定,包括历史模拟法、蒙特卡洛模拟、收益率分布的核密度估计等[9].

随着机器学习方法的兴起,不少学者将机器学习应用到风险管理领域. Shim等[10]首先构建了分位数支持向量回归用于VaR的估计, Taylor[11]首先提出了分位数神经网络模型,并将其用于多期VaR的测度估计中;许启发等[12]在分位数神经网络模型的基础上,考虑到尾部数据的稀疏性,结合极值方法POT对于极端VaR进行了修正,弥补了尾部数据稀疏性的问题. 分位数回归森林[13]方法作为一种非参数的机器学习方法,其避免了对模型形式的具体设定,同时也是一种集成预测方法,其将各个决策树的结果进行组合,避免单棵决策树可能对风险的高估或者低估,因而在精度上相比于单一的预测方法会有提升,相对其他机器学习算法来说也不易产生过拟合现象. 目前分位数回归森林在金融领域的应用很少,方匡南[14]首先使用分位数回归森林方法对于下一日的VaR进行估计,发现该方法的估计结果很好.

在机器学习中一直存在"黑箱性"的特点,无法了解各变量在模型中的作用;以往的回归模型中,也较少分析各因素对于风险值的边际影响,或单纯依靠估计的回归系数进行分析,但这种相关性往往只是线性的相关.

本文在已有研究的基础上进行了进一步的研究,使用分位数回归森林方法,基于方匡南[14]的研究,进一步考虑了不同时刻的时变波动率对于风险值的影响,以取代静态波动率;同时在已有的分位数回归模型中,解释变量大多为本国股指前日收益率,这里进一步考虑了他国股指前日收益率,体现风险可能在国家间的传递.

本文还对模型中各个因素的作用进行了阐述. 通过定义一种基于分位数回归森林的变量重要性度量方法,判断各个因素在模型中的权重;引入了偏相依关系,并将其应用到分位数回归森林中,动态地探究各个特征在不同水平下、对风险值的作用方向,在一定程度上弥补了机器学习算法中一直存在的"黑箱性"问题.

1 相关VaR估计方法

VaR的定义为在给定的置信水平下,投资组合在未来特定的时间段内最大可能的损失. 本质上, VaR是确定随机变量分位点的值.

1.1 分位数回归森林方法

分位数回归森林(quantile regression forests)为一种机器学习算法,是建立在随机森林算法基础

上的一种衍生算法. 随机森林为一种集成预测方法，通过 Bootstrap 方法抽取多个样本集，对每个样本集建立一棵决策树，最终的预测结果为单棵决策树的组合.

具体来说，设训练样本为 (X_i,Y_i)，$i=1,\cdots,n$，其中特征变量 X 为 P 维，即 $X\in\mathbb{R}^p$. 决策树是一种树形结构，其中每个内部节点基于一个属性进行测试，根据测试结果的不同继续分配到不同的分支，每个叶节点代表最终输出的类别. 设 θ 为单棵决策树的参数，决定每个节点的分裂变量和树的深度等条件. 决策树的叶子节点记为 l，$l=1,\cdots,L$. 每个样本的特征取值 x 都属于叶子中一个区域，记为 $x\in\mathbb{R}_l$，同时将这个节点记为 $l(x,\theta)$. 对于训练样本 X_i，考虑其对应叶子节点上的权重向量 $w_i(x,\theta)$ 为

$$w_i(x,\theta)=\frac{I(X_i\in\mathbb{R}_{l(x,\theta)})}{\sum_{j=1}^{n}I(X_j\in\mathbb{R}_{l(x,\theta)})} \tag{1}$$

式中，$I(\cdot)$ 代表示性函数，显然式(1)的权重和为 1.

对于随机森林算法来说，新观测值的条件均值的估计 $E(Y\mid X=x)$ 即为 k 棵决策树的平均. 设每棵决策树的参数 θ_t 为独立同分布的，$t=1,\cdots,k$，则权重向量 $w_i(x)$ 为各个决策树的平均：

$$w_i(x)=\sum_{t=1}^{k}w_i(x,\theta_t)/k \tag{2}$$

则随机森林最终的预测为对因变量的加权平均

$$\hat{u}(x)=\sum_{i=1}^{n}w_i(x)\cdot Y_i \tag{3}$$

分位数回归森林建立在随机森林的基础上，对因变量全部的条件分布进行估计. 设给定条件 $X=x$ 时因变量 Y 的条件分布为

$$F(y\mid X=x)=P(Y\leqslant y\mid X=x)=E(I(Y\leqslant y)\mid X=x) \tag{4}$$

利用随机森林估计的权重 $w_i(x)$，条件分布的估计可以看成对于示性函数 $I(Y_i\leqslant y)$ 的加权平均. 因而给出对于条件分布的估计为

$$\hat{F}(y\mid X=x)=\sum_{i=1}^{n}w_i(x)\cdot I(Y_i\leqslant y) \tag{5}$$

同时可证明该方法对条件分布的估计是相合统计量[13]，当样本足够多时，式(5)对于条件分布的估计 $\hat{F}(y\mid X=x)$ 会依概率收敛到真实条件分布 $F(y\mid X=x)$.

进一步，可得到利用样本数据估计的 Y 的条件分位点 $Q_\tau(Y)$：

$$Q_\tau(Y)=\inf\{y:\hat{F}(y\mid X=x)\geqslant\tau\} \tag{6}$$

利用 VaR 与分位点的关系，则易得到通过分位数回归森林方法的 VaR 估计.

分位数回归森林中，模型的参数主要有两个，分别为树的数量和每个节点参与分裂变量的个数. 这里对于最优参数的选取方法为，在样本内数据集上进行训练，使得该参数建立的模型在样本内失败天数 N 最接近于理论失败天数.

1.2　其他分位数估计方法

作为对比，使用了其他的一些 VaR 的估计方法，包括门限分位数回归[15]、CAViaR-AS 方法[16]、GARCH 类方法、加权历史模拟法[17]方法. 具体的建模方法和步骤可参照相应的文献.

目前，使用集成方法对于 VaR 进行估计往往能取得较为理想的效果，分位数回归森林算法即为一种非线性的集成预测方法. 与之对比，弹性网络方法(elastic net)是一种线性的集成回归方法，集合了 LASSO 和岭回归的惩罚项. 设单一模型的分位数预测为 $q_{i,t}(\tau)$，$i=1,2,\cdots,M$. 则 M 个模型的集成预测为

$$q_t^c(\alpha)=\beta_0(\tau)+\beta_1(\tau)q_{1,t}(\tau)+\cdots+\beta_M(\tau)q_{M,t}(\tau)=q_t'(\tau)\beta(\tau) \tag{7}$$

$$\hat{\beta}(\tau,\lambda,\delta)=\arg\min_{\beta(\tau)}\frac{1}{T}\sum_{t=1}^{T}\rho_\tau(r_t-q_t'(\tau)\beta(\tau))+\lambda(\delta\|\beta(\tau)\|_1+(1-\delta)\|\beta(\tau)\|_2^2/2) \tag{8}$$

式中，

$$\rho_\tau(u)=u\cdot(\tau-I(u<0)) \tag{9}$$

λ 为正则化参数，δ 用于平衡 LASSO 和岭回归的比例. 使用的单一模型即为之前介绍的门限分位数回归、CAViaR、GARCH 类方法和历史模拟法. 参照 Hastie 等[18]的研究建议，这里选取 $\delta=0.5$. 对于正则化参数 λ 的选取，通过时间序列的交叉验证方法[8]确定.

2　实证分析

选取了上海证券综合指数(上证综指)和标准普尔 500 指数(标普 500 指数)作为样本指数，数据跨度从 2009 年 7 月至 2017 年 3 月. 数据来源于国泰安数据库，建模过程在 R 语言中完成. 为检验各模型的有效性，分别对 95%和 99%置信度的样本内外 VaR 进行估计和回测. 2009 年 7 月至 2015 年 12 月的数据作

为样本内数据,用于模型的训练和测试;2016年1月至2017年3月的数据作为样本外数据,用于进一步检验各模型的风险估计能力.

采用通常的对数收益率对股指序列的收益率进行计算.对于样本的描述性统计如表1所示.

表1 样本描述性统计

Tab.1 Sample descriptive statistics

	最小值	最大值	均值	标准差	偏度	峰度	1%分位点	5%分位点	LM统计量
上证综指	−8.87	5.60	0.00	1.54	−0.86	4.63	−5.39	−2.40	197.06*
标普500指数	−6.90	4.63	0.00	1.01	−0.42	3.91	−2.88	−1.63	432.65*

[注] LM统计量衡量收益率序列的ARCH效应,这里选取6阶滞后,*代表在1%的条件下显著,即具有ARCH效应.

2.1 变量选取

从三个方面考虑对于条件变量的选取:收益率在短期可能存在趋势效应,选取了过去三天的收益率作为变量;方匡南[14]在建模中使用前日收益率的均值方差作为条件变量,但未考虑收益率的异方差性.LM统计量可以检验是否存在ARHC效应,由表1可知,收益率序列存在ARCH效应,因而本文使用残差项服从t分布的GARCH(1,1)模型对收益率数据建模,将计算得到的条件波动率 σ_t 作为一个条件变量.同时大多数研究在选取条件变量时未考虑风险是否在不同国家之间存在传递,这里进一步考虑了两国之间可能存在的风险传导,即选取了前一日的他国股指收益率作为条件变量.

2.2 VaR估计的回测检验方法

通过Kupiec[19]检验和Christoffersen[20]提出的条件覆盖率检验对VaR的估计值进行检验.Kupiec检验用于判断实际失效次数与理论失效次数是否一致,其统计量服从自由度为1的卡方分布;条件覆盖率检验在Kupiec检验的基础上,考虑了失效事件之间是否相互独立,服从自由度为2的卡方分布.二者的原假设为模型是有效的,若拒绝原假设犯错的概率 P 小于显著性水平则说明模型无效;当模型同时通过了Kupiec检验和条件覆盖率检验时,说明模型对于风险值的估计较为准确.

2.3 95%置信度的VaR估计

分位数回归森林中的参数主要包括树的数量和每个节点参与分裂的变量数.这里对于最优参数的选取方法为:通过在样本内数据上进行模型构建,选取使得失败天数最接近于理论失败天数的参数组合.当估计置信水平为 $(1-\alpha)$ 的VaR时,设样本长为 N,树的数目为 k,节点参与分裂变量数为 m,模型预测为 $f(k,m)$,收益率的实际值为 y,$I(\cdot)$ 为示性函数,参数的选取方法为

$$\underset{k,m\in\mathbf{N}^+}{\operatorname{argmin}}\left|\sum_{i=1}^{n} I(y_i < f_i(k,m)) - \alpha N\right| \quad (10)$$

通过网格搜索,选取参数为500棵树,每个节点参与分裂的变量个数为3.在构建本文的分位数回归森林(QRF)的同时构建了门限分位数回归模型(TQR)、CAViaR-AS模型(CAViaR)、基于t分布和广义误差分布的GARCH(1,1)方法(GARCH-t,GARCH-ged)、历史模拟法(HS)和弹性网络方法(EN)以及使用前日收益率的均值方差作为条件变量的分位数回归森林模型(QRF-MS).对95%置信度下的VaR进行估计,各模型检验结果如表2所示.

本文构建的分位数回归森林模型对于样本内外的估计结果最好,若设定检验的显著性水平水平为0.1,则对于上证综指和标普500指数的样本内外的预测均通过了检验,准确估计了一日VaR值.其次为门限分位数回归模型、CAViaR模型和GARCH-ged模型以及基于过去收益率的均值方差的分位数回归森林方法,但是这几种方法对于样本外标普500指数的预测效果较差,未能通过检验.历史模拟法虽然在样本内数据上的拟合较好,但是在样本外的预测上效果较差.同时也发现基于弹性网络的集成预测并没有体现出绝对的优势,在0.1的显著性条件下多次被拒绝,且P值较小;可能源于其本质上还是一种线性的集成预测方法,其无法如分位数回归森林方法拟合数据间的非线性结构[8].而相较于过去收益率的均值方差作为输入变量的分位数回归森林模型,本文方法进一步提升了预测的可靠性,这可能是源于输入变量中包括了时变波动率和他国收益率的数据,包含了更多的信息.分位数回归森林模型作为集成性的方法,降低了出现过拟合的概率,更多的信息往往能进一步提高精度.

同时,为了说明分位数回归森林方法的估计效果,做出了该方法对于样本外VaR的估计与实际收

益率的对比图，如图 1 所示，收益率为负数即代表损失. 其中，浅色线条代表实际收益率，深色线条代表该方法估计的 VaR 值. 可见，分位数回归森林准确估计了样本外的风险值，尤其在 2016 年初熔断机制的试行导致我国股市剧烈动荡的时期，分位数回归森林算法可以较为准确地预估风险.

表 2　95%VaR 回测结果

Tab. 2　Results of 95% VaR backtesting

	模型	上证综指				标普 500 指数			
		LR_{UC}	P	LR_{CC}	P	LR_{UC}	P	LR_{CC}	P
样本内	QRF	0.00	0.98	1.24	0.54	0.15	0.70	0.73	0.70
	QRF-MS	0.00	0.98	1.24	0.54	0.15	0.70	0.73	0.70
	TQR	0.00	0.98	1.24	0.54	0.01	0.77	1.24	0.54
	CAViaR	0.31	0.58	0.37	0.83	0.03	0.87	0.77	0.65
	GARCH-t	16.75	0.00*	16.84	0.00*	3.79	0.05*	3.86	0.15
	GARCH-ged	0.19	0.66	0.26	0.88	0.89	0.35	1.12	0.57
	HS	0.02	0.89	0.27	0.87	0.03	0.86	0.03	0.98
	EN	5.10	0.02*	6.03	0.05*	3.79	0.05*	4.51	0.11
样本外	QRF	0.16	0.69	0.43	0.81	1.23	0.27	2.07	0.36
	QRF-MS	0.94	0.33	1.58	0.45	5.55	0.02*	7.71	0.02*
	TQR	0.02	0.90	0.16	0.92	4.11	0.04*	5.79	0.06*
	CAViaR	0.94	0.33	1.58	0.45	5.55	0.02*	7.71	0.02*
	GARCH-t	3.59	0.06*	4.05	0.13	11.85	0.00*	11.95	0.00*
	GARCH-ged	0.94	0.33	1.58	0.45	9.36	0.00*	9.53	0.01*
	HS	11.02	0.00*	11.13	0.00*	5.55	0.02*	7.71	0.02*
	EN	2.49	0.12	3.71	0.16	5.55	0.02*	7.71	0.02*

[注] * 代表在 10%显著性下拒绝原假设，模型失效. LR_{UC} 和 LR_{CC} 代表 Kupiec 检验和条件覆盖率检验的统计量，下同.

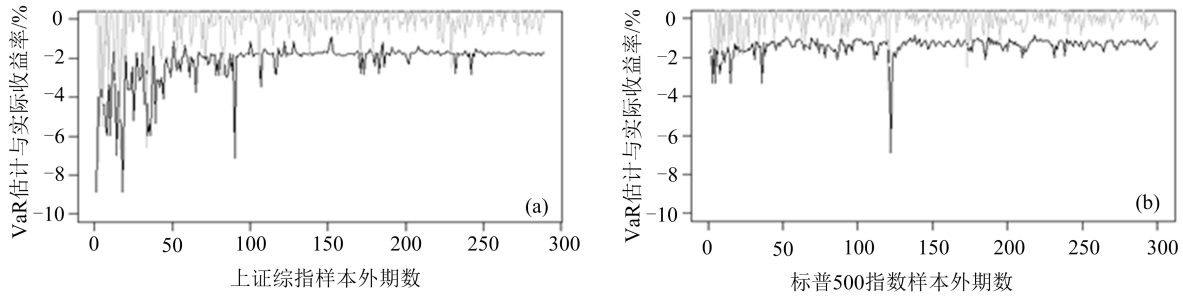

图 1　分位数回归森林 95%VaR 估计

Fig. 1　95% VaR estimation based on quantile regression forest

2.4　99%置信度的 VaR 估计

为进一步验证各方法极端情况下的预测能力和检验模型的稳健性，对于 99%置信度的样本内外的 VaR 进行了估计. 建模过程同上，对于分位数回归森林参数的选择与 95%置信度的估计一致，对于估计结果的检验如表 3 所示.

表 3 99%VaR 回测结果

Tab. 3 Results of 99% VaR backtesting

	模型	上证综指				标普 500 指数			
		LR_{UC}	P	LR_{CC}	P	LR_{UC}	P	LR_{CC}	P
样本内	QRF	0.00	0.95	0.31	0.86	0.01	0.94	0.32	0.85
	QRF-MS	0.20	0.65	0.44	0.80	0.11	0.74	4.02	0.13
	TQR	0.20	0.65	0.44	0.80	0.11	0.74	0.39	0.82
	CAViaR	3.76	0.05*	17.51	0.00*	1.81	0.18	6.31	0.04*
	GARCH-t	4.70	0.03*	4.77	0.09*	5.27	0.02*	5.34	0.07*
	GARCH-ged	9.02	0.00*	10.07	0.01*	3.20	0.07*	4.02	0.13
	HS	1.67	0.30	1.58	0.45	0.43	0.51	6.02	0.05*
	EN	0.52	0.47	0.72	0.70	6.83	0.01*	6.89	0.03*
样本外	QRF	0.00	0.95	0.07	0.97	0.38	0.54	0.41	0.82
	QRF-MS	0.00	0.95	0.07	0.97	0.38	0.54	0.41	0.82
	TQR	0.00	0.95	0.07	0.97	0.31	0.58	4.77	0.09*
	CAViaR	0.00	0.95	0.07	0.97	0.38	0.54	0.41	0.82
	GARCH-t	1.67	0.20	1.68	0.43	0.38	0.54	0.41	0.82
	GARCH-ged	1.28	0.26	1.45	0.48	0.38	0.54	0.41	0.82
	HS	5.81	0.02*	5.81	0.06*	0.38	0.54	0.41	0.82
	EN	1.67	0.20	1.68	0.43	0.38	0.54	0.41	0.82

由表 3 可见，两种分位数回归森林在极端情况下的预测能力较好，通过了全部的检验；而门限分位数回归方法的效果次之，但是在标普 500 指数的样本外预测上未能通过检验，各种失败事件的发生不是条件独立的，在市场动荡时期模型的效果会降低. 而弹性网络方法在 99%的条件下预测效果有了显著提升，除了在标普 500 指数样本内的预测高估了风险，其余的预测均通过了检验. 而相较而言，其余几种方法的预测效果相对较差，均多次出现未能通过检验的情况.

本文构建的分位数回归森林方法估计的风险值与实际收益率的走势如图 2 所示. 可见分位数回归森林估计的一日 VaR 值都在 10%以内，未出现如神经网络分位数回归[12]的估计值达到 30%的过拟合而导致在实际中失去作用的情况.

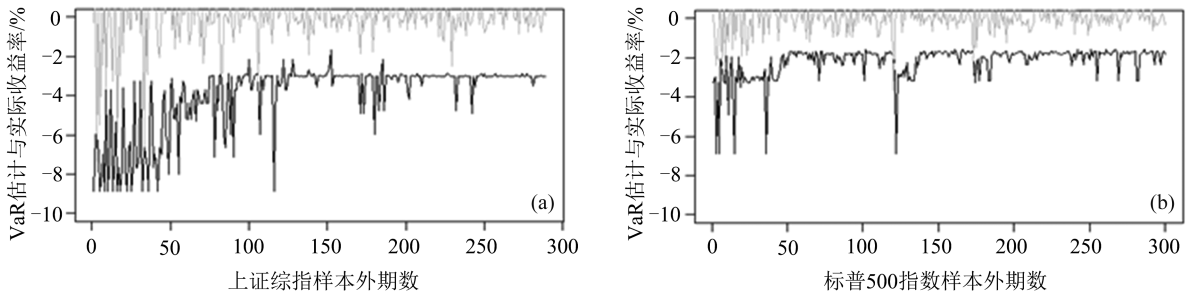

图 2 分位数回归森林 99%VaR 估计

Fig. 2 99% VaR estimation based on quantile regression forest

3 风险因素分析

上述检验发现分位数回归森林方法估计的 VaR 值在各种置信度下均较为准确，其非线性、集成性的优势得以体现. 但机器学习算法通常具有“黑箱性”的特点，以往的研究也往往忽视这一方面. 针

对本文构建的分位数回归森林模型中，各解释变量对于风险值有怎样的影响，本文提出了一种基于分位数回归森林的各因素重要性的度量方法，通过此方法衡量了各因素的重要性权重；并引入了偏相依关系，将其应用于构建的回归模型中，动态地度量各个因素对于 VaR 的作用方向.

3.1　变量重要性度量

在随机森林回归中，度量某特征重要性的方法主要有两种，一种方法为基于袋外数据 OOB 的方法，另一种则基于基尼系数的思想，计算使用该特征作为分裂变量的节点的均方误差减少的均值. 但是上述两种方法的重要性度量都是基于均方误差损失函数，因而度量的是该变量对于条件均值影响的重要性大小，无法完全体现对于分位点的影响权重. 本文定义一种基于分位数回归森林的变量重要性度量方法，并将其用于判断各因素对于分位点的影响程度大小. 基于训练数据，记其在给定置信水平 τ 下的预测值为 $\hat{F}_\tau(y \mid X_1, X_2, \cdots, X_j = x)$，对于某特征 j 随机扰动后的预测值记为 $\hat{F}_\tau(y \mid X_1, X_2, \cdots, X_j = \tilde{x})$. 因为 $\hat{F}_\tau(y \mid X = x)$ 会依概率收敛到真实分位点[13]，因而定义下式作为特征 j 的重要性度量：

$$\text{Importance}(X_j) = E_X[(\hat{F}_\tau(y \mid X_1, X_2, \cdots, X_j = x) - \hat{F}_\tau(y \mid X_1, X_2, \cdots, X_j = \tilde{x}))^2] \quad (11)$$

该值越大，说明特征扰动后，对于分位点的预测值影响越大，因而也就说明该特征在模型中的权重更大. 选择收益率的下 5%分位点作为因变量，为减少随机性的影响，计算 100 次式(11)的平均值作为重要性的度量，结果如表 4 所示.

表 4　变量重要性度量

Tab. 4　Variable importance measurement

变量名	上证综指		标普 500 指数	
	重要性	相对权重	重要性	相对权重
过去一日收益率	2.61	39.42%	0.29	28.17%
过去二日收益率	1.89	28.53%	0.12	11.20%
过去三日收益率	0.33	4.92%	0.13	12.61%
波动率	0.90	13.62%	0.35	34.19%
他国股指(标普 500 指数或上证综指)前日收益率	0.89	13.51%	0.14	13.83%

对于上证综指来说，对其风险值影响最大的因素为过去一日的收益率，重要性度量数值为 2.61，其次为过去二日的收益率、条件波动率等. 而对于标普 500 指数来说，影响最大的因素为条件波动率，其重要性度量数值为 0.35，其次依次为过去一日的收益率、上证综指前日收益率等. 过去一日、二日收益率对上证综指影响的权重较大，说明我国市场的风险延续性较强，当出现极端事件后隔日继续延续的可能性更大；而美国股市中波动率对其风险值影响相对较大，过去一日、二日收益率的影响相对较小，推测这可能源于其市场更为有效、市场弹性相对更强，因而当出现极端事件后的恢复能力更强.

由表 4 发现，上证综指各影响因素的重要性数值均明显大于标普 500 指数的，这里推测可能与中美两国股市成熟程度不同有关. 我国股市本质上属于新兴市场，管理机制尚未健全，股市波动性较强，受外界信息影响的程度较大；而美国股市较为成熟、监管机制相对完善，股市整体走势相对平稳，同时抵御外界因素影响的能力也更强.

从风险的相互传导上分析，标普 500 指数前日收益率对上证综指的重要性度量为 0.89，而反向的影响数值仅为 0.14，反映出影响的不对称性. 美股走向在一定程度上可以代表着全球经济的风向标，也会极大地影响着世界各国投资者对于未来经济的预期和投资的信心；标普 500 指数收益率对于上证综指的影响远大于二者的反向影响. 从相对权重上看，两国的他国股指前日收益率所占的影响权重较小，相对权重都在 13%左右，两国之间的风险传导并不显著. 这可能是由于我国目前特殊的资本市场结构造成的，国际资本尚不能自由流动，因而他国股指的收益率对于本国的影响力度较弱.

3.2　偏相依关系分析

在回归模型中，主要的度量自变量与因变量的相关方向的方法包括相关系数、回归模型的系数值等，但这些度量的是整体的相关性，同时线性分位数回归中的相关性为线性相关，而金融数据中存在大量的非线性相关. 偏相依关系可以度量各个特征在不同水平下对于因变量的影响方向，同时可以衡量变量之间的线性及非线性关系.

偏相依关系(partial dependence)由 Haste 等[21]提出，用于解释机器学习中一直存在的黑箱问题. 其定义如下：设所得的模型为 $f(\cdot)$，条件变量 X 为 p 维变量，即 $X = (X_1, X_2, \cdots, X_p)$，取其中两个不交子集 X_s 和 X_c，且有 $X_s \cup X_c = X$，则特征向量 X_s 取某一值 x 时与因变量的偏相依关系 P

定义为

$$P=E_{X_c}f(X_s=x,X_c)=\frac{\sum_{i=1}^{n}f(X_s=x,X_{c_i})}{n} \tag{12}$$

计算得到的 P 即为当特征向量 X_s 取特定值的时候模型对因变量的估计值，体现了该特征当前取值下因变量的期望数值. 这里取 X_s 为一维变量，计算时取各分位点的值代入式(12)，即可得到该特征在不同状态下与分位点的偏相依关系数值.

依据本文建立的分位数回归森林模型，由于篇幅所限同时结合之前的重要性度量，这里只列出过去一日收益率、条件波动率和他国股指(标普 500 指数或上证综指)前日收益率由小至大各 50 个分位点的值对收益率下 5%分位点的偏相依关系图，参见图 3 和图 4. 过去二日收益率和过去三日收益率与其的关系与过去一日收益率较为相似.

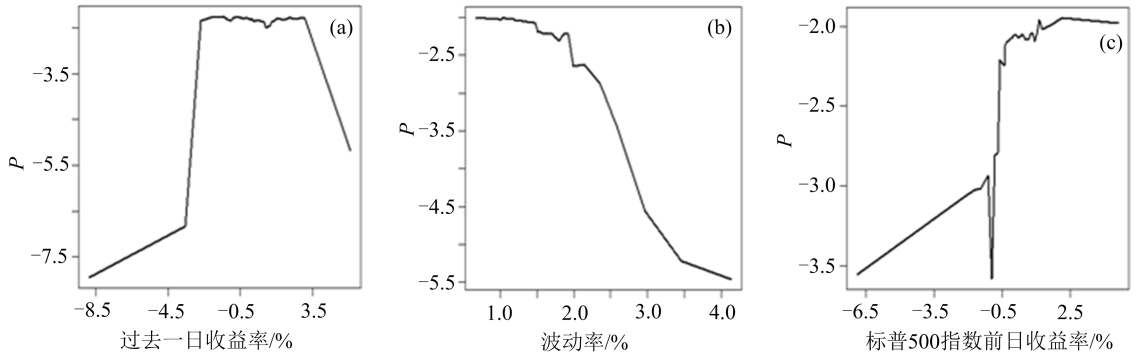

图 3 上证综指与各因素的偏相依关系

Fig. 3 Partial dependence plot for Shanghai Composite Index

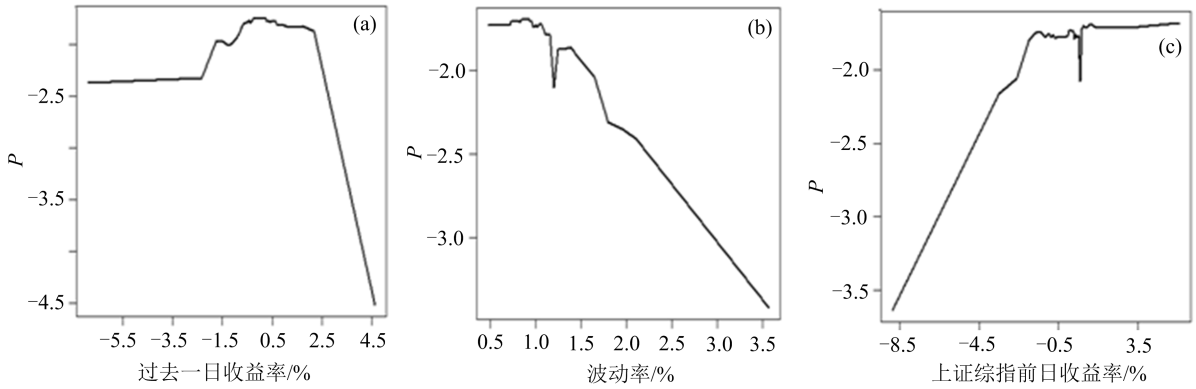

图 4 标普 500 指数与各因素的偏相依关系

Fig. 4 Partial dependence plot for S&P 500 Index

研究发现，对于上证综指和标普 500 指数，各因素对于风险值的影响方向较为相似.

具体来说，过去一日收益率对未来一日风险存在一种非线性的影响，当过去一日收益率很低时，未来一日股市继续走低的可能性增大，指数走势对于风险值的影响存在一种短期的趋势效应. 但是，当一日收益率过高时，说明市场过热，未来很可能出现反转，风险值反而会加大.

预期的条件波动率对于当日的风险值的影响较为线性，二者存在一种正向相关的关系. 波动率越大，股市震荡的幅度越大，这时发生风险的可能性越大. 投资者可以关注市场的相关波动率指标，当市场预期会出现大幅震荡、较高的换手率等指标时，需要防范风险事件的发生.

他国股指前日收益率对两国的风险值在影响方向上也较为相似. 虽然之前的重要性度量发现中美两国的风险传导较弱，但是在方向上依然呈现了一定的相似性. 当他国股指前一日发生极端事件时，下一日本国股指收益率出现极端事件的可能性会相对加大. 美国作为金融中心，其代表了全球经济的走向，同时其股市的走向很大程度上也会影响着世界各国投资者的信心和对经济的预期，投资者中往往

存在羊群效应，若美国股市发生大跌，则我国投资者的情绪可能会受到影响、对经济的前景持悲观预期进而跟随抛售股票，导致我国股市走势进而下跌，即出现风险的溢出效应. 而随着中国在世界经济中地位的增强，沪港通、深港通的开放，A 股纳入明晟指数等进程，中国股市与世界各国的联系也在加强，更多境外合格投资者参与其中，更多的北上资金一定程度上也会加大风险的传导，与美国股市呈现一定的同步性特征. 我国投资者需要关注美国主要股市前一日走向，当国际市场出现较大的利空消息、美股前一日暴跌时，未来一日我国股市走低是大概率事件，投资者应考虑降低风险敞口，减少潜在损失.

4　结论

本文引入了非参数、集成性的分位数回归森林方法，对于上证综指和标普 500 指数不同置信水平的 VaR 进行估计. 输入特征中考虑了过去三日的收益率、时变的波动率和他国股指收益率的影响. 通过 Kupiec 检验和条件覆盖率检验，对于样本内外的估计值分别进行了检验，验证了该方法的有效性.

同时，针对机器学习中的黑箱问题，本文对模型中各因素的作用进行了阐述：定义一种变量重要性度量方法，度量了各个因素对于风险值的影响程度大小；发现过去一日收益率对于上证综指的影响相对较大，而波动率对于标普 500 指数的影响相对较大；引入了机器学习中的偏相依关系分析，研究了各因素在不同水平下对于 VaR 的作用方向，发现过去一日收益率在门限范围内有风险的延续效应，波动率与风险呈正相关，而中美两国之间存在一定程度的风险的传递效应.

本文目前只考虑了该方法在中美两国市场的应用，未来可进一步在更多的市场中对其稳健性进行检验. 机器学习算法在市场风险的管理中有广阔的应用空间，未来可进一步尝试计算尾部的 ES 值、对多日 VaR 进行估计等.

参考文献(References)

[1] ENGLE R F, MANGANELLI S. CAViaR[J]. Journal of Business & Economic Statistics, 2004, 22(4): 367-381.

[2] MENG X, TAYLOR J W. An approximate long-memory range-based approach for value at risk estimation[J]. International Journal of Forecasting, 2018, 34(3): 377-388.

[3] 许启发，康宁. 门限分位数自回归模型及在股市收益自相关分析中应用[J]. 系统工程理论与实践，2015，35(12)：2993-3007.
XU Qifa, KANG Ning. Threshold quantile autoregressive model with application to auto-correlation analysis of stock returns[J]. Systems Engineering: Theory & Practice, 2015, 35(12): 2993-3007.

[4] 罗克兵，叶五一，董筱雯. 高频连涨连跌收益率的分位点 Granger 因果检验与条件 VaR 估计[J]. 中国科学技术大学学报，2016，46(11)：919-927.
LUO Kebing, YE Wuyi, DONG Xiaowen. Granger causality test in quantiles and conditional VaR estimation of continuously rising and falling returns[J]. Journal of University of Science and Technology of China, 2016, 46(11): 919-927.

[5] HAUGOM E, RAY R, ULLRICH C J, et al. A parsimonious quantile regression model to forecast day-ahead value-at-risk[J]. Finance Research Letters, 2016, 16: 196-207.

[6] LIU G, WEI Y, CHEN Y, et al. Forecasting the value-at-risk of Chinese stock market using the HARQ model and extreme value theory[J]. Physica A: Statistical Mechanics & Its Applications, 2018, 499, 288-297.

[7] HALBLEIB R, POHLMEIER W. Improving the value at risk forecasts: Theory and evidence from the financial crisis[J]. Journal of Economic Dynamics & Control, 2012, 36(8): 1212-1228.

[8] BAYER S. Combining value-at-risk forecasts using penalized quantile regressions[J]. Econometrics & Statistics, 2018, 8: 56-77.

[9] NIETO M R, RUIZ E. Frontiers in VaR forecasting and backtesting[J]. International Journal of Forecasting, 2016, 32(2): 475-501.

[10] SHIM J, KIM Y, LEE J, et al. Estimating value at risk with semiparametric support vector quantile regression[J]. Computational Statistics, 2012, 27(4): 685-700.

[11] TAYLOR J W. A quantile regression neural network approach to estimating the conditional density of multiperiod returns[J]. Journal of Forecasting, 2000, 19(4): 299-311.

[12] 许启发，陈士俊，蒋翠侠，等. 极端 VaR 风险测度的新方法：QRNN+POT[J]. 系统工程学报，2016，31(1)：33-44.
XU Qifa, CHEN Shijun, JIANG Cuixia, et al. A new method for extreme value at risk measure: QRNN+POT[J]. Journal of Systems Engineering, 2016, 31(1): 33-44.

[13] MEINSHAUSEN N, RIDGEWAY G. Quantile

regression forests[J]. Journal of Machine Learning Research, 2006, 7(6): 983-999.
[14] 方匡南. 随机森林组合预测理论及其在金融中的应用[M]. 厦门:厦门大学出版社, 2012:187-205.
[15] 缪柏其, 叶五一. 高等计量经济学基础[M]. 北京:高等教育出版社, 2013:163-164.
[16] 张颖, 张富祥. 分位数回归的金融风险度量理论及实证[J]. 数量经济技术经济研究, 2012(4): 95-109.
ZHANG YIN, ZHANG Xiangfu. The theory and empirical research of quantile regression in financial risk measurement[J]. The Journal of Quantitative & Technical Economics, 2012(4): 95-109.
[17] ŽIKOVIĆ S, AKTAN B. Decay factor optimization in time weighted simulation: Evaluating VaR performance[J]. International Journal of Forecasting, 2011, 27(4): 1147-1159.
[18] HASTIE T, TIBSHIRANI R, WAINWRIGHT M. Statistical Learning with Sparsity: The Lasso and Generalizations[M]. Boca Raton, FL: Chapman & Hall/ CRC, 2015:57.
[19] KUPIEC P H. Techniques for verifying the accuracy of risk measurement models[J]. Social Science Electronic Publishing, 1995, 3(2):73-84.
[20] CHRISTOFFERSEN P F. Evaluating interval forecasts[J]. International Economic Review, 1998, 39(4): 841-862.
[21] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical Learning[M]. New York: Springer, 2009: 369-370.

(上接第624页)

[9] FAN Wenping, JIANG Xiaoyun. Parameters estimation for a one-dimensional time fractional thermal wave equation with fractional heat flux conditions[J]. Acta Phys Sin, 2014,63(14): 140202. (in Chinese)
[10] DE JAGER E M, JIANG F. The Theory of Singular Perturbation [M]. Amsterdam: North-Holland Publishing Co, 1996.
[11] BARBU L, MOROSANU G. Singularly Perturbed Boundary-Value Problem [M]. Basel: Birkhauserm Verlag AG, 2007.
[12] FAYE L, FRENOD E, SECK D. Singularly perturbed degenerated parabolic equations and application to seabed morphodynamics in tided environment [J]. Discrete Contin Dyn Syst, 2011, 29(3): 1001-1030.
[12] SAMUSENKO P F. Asymptotic integration of degenerate singularly perturbed systems of parabolic partial differential equations[J]. J Math Sci, 2013, 189 (5): 834-847.
[13] GE H X, CHENG R J. A meshless method based on moving kriging interpolation for a two-dimensional time-fractional diffusion equation[J]. Chin Phys B, 2014,23(4): 040203.
[15] MO Jiaqi. Homotopiv mapping solving method for gain fluency of a laser pulse amplifier[J]. Science in China, Ser G, 2009, 39(7): 1007-1010.
[16] MO Jiaqi, LIN Wantao, LIN Yihua. Asymptotic solution for the El Nino time delay sea-air oscillator model[J]. Chin Phys B, 2011, 20(7): 070205.
[17] MO Jiaqi, LIN Yihua, LIN Wantao, et al. Perturbed solving method for interdecadal sea-air oscillator model [J]. Chin Geographical Sci, 2012, 22(1): 42-47.
[18] XU Jianzhong, ZHOU Zongfu. Existence and uniqueness of anti-periodic solutions to an nth-order nonlinear differential equation with multiple deviating arguments[J]. Ann Diff Eqs, 2012, 28(1): 105-114.