基于EMD-QRF模型的QDII基金风险度量的研究

**摘要**：准确测度证券投资基金风险是提高 QDII 基金风险管理效果和促进证券投资基金市场健康发展的关键，契合深化金融业供给侧结构性改革的题中之意。本文提出了一种基于EMD和分位数回归森林相结合的的VaR估计方法，并进一步运用Kupiec检验将该方法与多种传统方法进行了实证对比，结果显示其具有较高的计量精度。因此本文的研究为EMD在风险评估领域中的运用提供了另外一种思路。

**关键词**：风险测度；分位数回归；随机森林；EMD分解

**1.引言**

随着我国经济实力日益增强和人民的不断富裕，国内越来越多的投资者开始进行全球化的资本配置。但由于我国资本市场尚未完全开放，绝大多数投资者只能通过QDII制度来参与到全球化的交易中去。QDII基金不仅为投资者提供多元的资产配置渠道，更是国内投资者规避我国经济转型阶段的经济风险的重要手段。准确测度证券投资基金风险是提高 QDII 基金风险管理效果和促进证券投资基金市场健康发展的关键，契合深化金融业供给侧结构性改革的题中之意。

金融资产的在险价值（Value at Risk，VaR）是一种重要的风险度量指标，它定义了在正常市场情况下，给定置信度金融资产可能遭受的最大损失。VaR实质为金融资产损失分布的一个特定分位数，Koenker等(1978)提出的线性分位数回归模型无需假定收益率序列的分布，就可以刻画解释变量在不同分位点对解释变量的影响，这为VaR风险测度提供了一种思路[1]。王新宇、赵绍娟（2008）利用3种不同类型解释变量的条件分位数回归模型对中国沪深股市的风险进行实证研究[2]。李邦国（2017）年利用分位数回归研究大豆期货市场风险，并与经典回归模型相比，VaR值准确率有所增加[3]。Bayer和Sebastian（2018）提出在分位数回归中引入正则项估计VaR[4]。刘演森（2018）提出贝叶斯Lasso分位数回归，并用于金融市场的VaR值估计[5]。随着对VaR准确度要求不断提高，人们纷纷选择将传统VaR方法与机器学习算法结合进行风险测度。徐启发等（2014）利用支持向量分位数回归给出多期VaR风险测度，并与传统VaR方法比较[6]。徐启发等（2014）利用神经网络分位数回归模型计算VaR值，发现无论在样本内还是样本外 ，都有很强的稳健性[7]。孙国强等（2018）提出将分位数回归森林应用于短期风电功率的概率密度预测，结果表明，与传统短期风电功率预测方法相比，分位数回归森林能够给出任意分位点下的预测结果，从而获得预测值的条件分布，对短期风电功率的概率密度预测结果更准确[8]。

金融市场受经济、政策、投资者心理等多种因素，各因素影响所引起的金融资产价格波动特征、时间范围也不尽相同，对金融时间序列数据去噪并提取出富有价值的信息对金融数据的预测有重要意义。经验模式分解(Empirical Mode Decomposition，EMD)是N. E. Huang等（1998）提出的一种基于数据的具有高时频分辨能力的信号分析方法[9]。与小波分析相比，它具有更高的分解效率和更好的时频特性，特别适用于各种非平稳非线性时间序列的处理和分析。近年来，关于EMD的理论和应用研究不断出现，已遍及医学，信号处理，机械诊断，气象和地震预测等领域，展现出良好的应用前景。李合龙、杨志（2015）将EMD将三峡债价格序列进行分解，结合分型理论对VaR估计，发现将EMD方法应用于风险计算拥有比传统计量方法更好的刻画和预测能力[10]。Awajan（2017）将EMD将MA模型结合用于每日股票市场时间序列数据的短期预测[11]。孟二浩(2018)在对径流的预测研究中发现，将EMD模型与SVM模型现结合，较SVM模型的预测效果有着显著提高[12]。

本文提出了一种基于EMD和分位数回归森林相结合的的VaR估计方法，并进一步运用Kupiec检验将该方法与多种传统方法进行了实证对比，结果显示其具有较高的计量精度。因此本文的研究为EMD在风险评估领域中的运用提供了另外一种思路。

**2 研究方法**

**2.1 EMD原理**

EMD方法从本质上讲是一种数据逐步分解技术，该技术可以将原信号中不同信号的波动或趋势分解出来，产生一系列不同尺度的分量，每一个分量称之为本征模函数（Instrinsic Mode Function，IMF）。每一个IMF分量都反映了原信号在不同尺度下的特征。

EMD算法基本流程如下：

（1）找出序列的所有极大值和极小值点。

（2）根据极大值和极小值进行三次样条插值构建的上包络线和下包络线。

（3）和均值 以及和的差值

（4）以代替原始信号,重复以上三步，直到与之间的某一设定值，即认为是一个IMF分量记，，

（5）重复以上四步，直至最后的分量符合筛选停止准则。一般直到小于某一设定值，或者变成一个单调函数时原始信号的EMD分解结束，则是剩余分量表示原序列的的均值或趋势项。

得到的分解形式如下：

上式满足恒等关系，即EMD技术可以完美地把原序列分解为多个分量，并且在此过程中没有出现信号的损失，保留了原序列的所有信息，这也是该技术较小波分析和傅里叶变换等其他处理方法大的优点。

**2.2分位数回归的的基本原理**

分位数回归是单因变量 Y 的条件分位数对自变量 X 进行回归，从而获得所有分位点下的回归预测模型。

在给定条件下，条件分布函数是的累计概率，即：

分位数为在给定条件下，的累计概率恰好为，即：

其中，为求最小值运算。

一般的条件分位数回归表示为：

可以是线性的，也可以是非线性的（如多项式函数，对数函数等），可以由下式估计

其中T为样本容量，为依赖分位数点的非对称损失函数，满足

VaR定义为在置信水平下，投资者持有特定的金融资产一段期间，其可能遭受的最大损失。记金融资产或资产组合的损失L的累积分布函数为 ,则置信水平时的VaR,就是收益率分布的第 分位数,也就是:

**2.3分位数回归森林**

分位数回归森林是随机森林算法的一种衍生， 通过结合分位数回归的特性，可提供因变量的全部条件分布信息。随机森林预测可以被看作是一个适应性近邻分类和回归过程，对于每一个对每一个， 可以得到原始n个观察值的一个权重集合。随机森林预测等价于因变量的加权均值，分位数回归森林的决策树是以标准随机森林算法生成，条件分布是通过观测到的因变量加权估计得到。一般而言，分位数回归森林算法包括如下步骤。

1. 生成k棵决策树，对于每棵决策树的每个节点，要考察所有的观测值，而不仅是平均值；，然后根据
2. 对于给定X，遍历所有的决策树，求出权重，然后对得到每个观测值得权重，其中
3. 利用得到的估计条件分布,其中
4. 条件分位数为

**2.4 Kupiec LR后验检验**

关于VaR模型估计结果最常用的统计检验方法是Kupiec（1995）提出的LR失败率检验。具体做法为：将统计观察序列中所有同时刻实际收益率超过模型所估计的VaR个数，设为。记观察序列长度为N，则失败频率为，Kupiec检验的零假设为，其似然比检验的统计量

**3.基于EMD-QRF的风险度量方法**

针对金融时间序列数据，随机因素多、信噪比大的特点。本文建立EMD-QRF模型对QDII基金的VaR风险测度进行估计。首先，采用EMD算法对原始序列进行处理，提取市场不同频率的趋势，并针对性的建立预测模型，从而有效的提高金融资产VaR的预测精度。为了减少建模任务量，对频率大小相似的分量进行重构为新的分量，并建立模型估计分量的VaR，将不同分量预测结果合并得到原始序列的VaR值。具体步骤如图1所示。

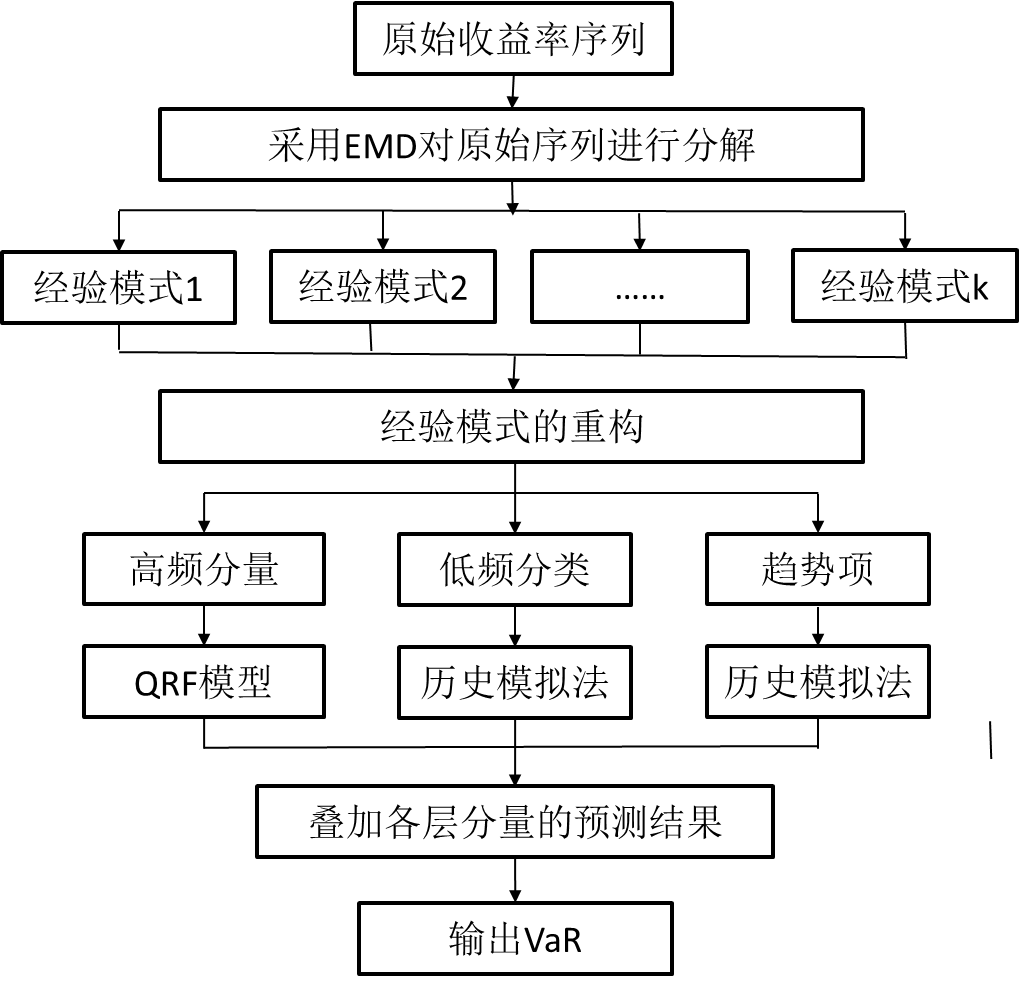


图1 基于EMD-QRF模型的

**3.1原始序列的分解**

采用EMD分解的方法将QDII基金的收益率序列分解为一系列低相关性、具有不同频率多个IMF分量和一个趋势项。每个IMF分量的序列曲线都是围绕零均值线的局部极大值和极小值基本对称的波形，代表着原序列在不同时间尺度的震荡变化。趋势项集中反映了序列的非平稳性，代表着原序列的总趋势。

**3.2经验模式分量的重构**

根据IMF分量的频率大小和统计特征将IMF分量分为高频项和低频项，然后对IMF分量进行重构。IMF分量的重构方法如下：

其中高频分量代表原序列的短期波动，低频分量代表原序列的中期波动，趋势分量代表着原序列的长期趋势。

**3.3相关统计量**

为了进一步挖掘IMF分量的的经济意义，我们引入以下指标:

1. 拟相关系数:

为经验模式i的模，为原序列的模。拟相关系数衡量IMF分量与原序列的相关性，拟相关系数越大，当前IMF与原序列相关性越大。

1. 方差贡献率

方差贡献率说明单个IMF分量的波动对整个序列波动的影响力的大小。

1. 平均周期

N表示原序列观测值个数，表示极大值个数。平均周期反映IMF分量波动的时间时间尺度。

**4 实证研究**

**4.1 数据来源与处理**

本文研究分析选取的数据是华夏全球精选QDII基金日净值数据，时间跨度为2010年1月4日至2019年4月29日，共1688个样本数据。为同时评价VaR风险度量在样本内和样本外表现的差异，将整个样本区间划分为两个部分。取2017年以前的1385个数据为样本内数据，2017年及以后的772个数据为样本外数据。数据来源于国泰安数据库。对QDII收益率数据进行对数收益率处理，其对数收益率处理公式为

采用EMD方法对原始收益率序列进行分解，分解结果如图2所示。从图中可以看出，原序列经分解后得到8个IMF分量和一个趋势项，并且各个IMF分量呈现频率由高到低逐渐递减的趋势。各经验模式分量的相关统计量如表1所示。从表1中可以看出，高频项与原序列的相关性较高，且方差贡献率较大 。

为了减少建模任务量，将频率大小相似的分量进行重构为新的分量：高频分量，低频分量和趋势分量。

重构分量的走势如图3所示。从图中可以看出高频分量呈现高频波动特性，表现为近似随机游走的白噪声序列。而低频分量相对平滑，而趋势分量表现为一条平滑曲线。各重构分量的相关统计量如表2所示。从表2中可以看出，高频分量的方差贡献率为92.98%，平均周期为3天；低频分量房产贡献率仅占6.91%，平均周期为46天。

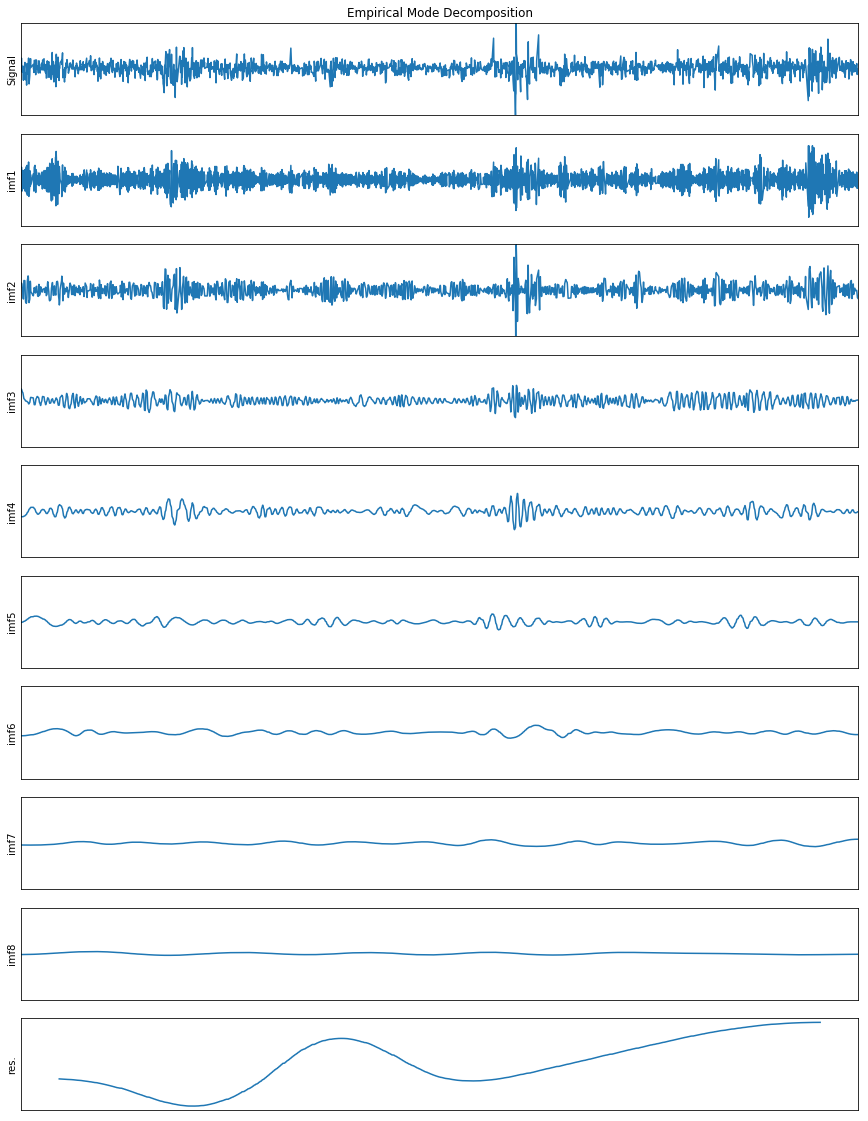


图2 收益率序列及各IMF分量

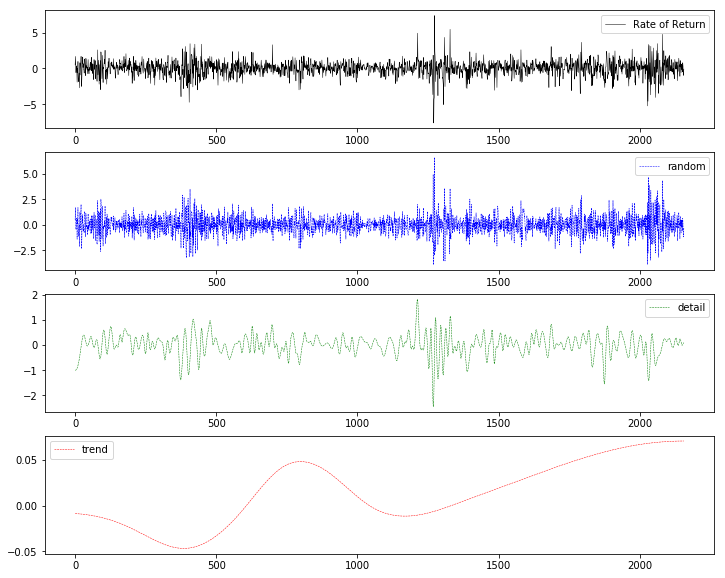


图3 重构分量的走势图

表1 IMF分量统计分析

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Std | Min | max | R | STDR | C |
| IMF1 | -0.012069 | 0.734368 | -3.207990 | 2.915314 | 0.663377 | 46.6591% | 2.9 |
| IMF2 | 0.001683 | 0.549000 | -3.940391 | 3.899327 | 0.507601 | 26.0768% | 6.1 |
| IMF3 | 0.001051 | 0.358642 | -1.432325 | 1.326255 | 0.337200 | 11.1284% | 12.3 |
| IMF4 | 0.004573 | 0.311082 | -1.561220 | 1.540034 | 0.253566 | 8.3726% | 21.6 |
| IMF5 | 0.004621 | 0.196065 | -0.677001 | 0.681461 | 0.150750 | 3.3259% | 42.3 |
| IMF6 | 0.003625 | 0.162722 | -0.483119 | 0.619004 | 0.101343 | 2.2909% | 93.7 |
| IMF7 | -0.010109 | 0.130959 | -0.303061 | 0.330883 | 0.084620 | 1.4838% | 179.7 |
| IMF8 | 0.002316 | 0.080383 | -0.142367 | 0.178201 | 0.056266 | 0.5590% | 431.2 |
| Trend | 0.013525 | 0.034596 | -0.046627 | 0.070403 | 0.026721 | 0.1036% | — |

表2 重构分量统计分析

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Std | Min | max | R | STDR | C |
| Random | -0.004762 | 1.002178 | -6.738452 | 8.051202 | 0.963616 | 92.9765% | 3 |
| Detail | 0.000453 | 0.273263 | -0.924154 | 1.362698 | 0.225787 | 6.9127% | 46 |
| Trend | 0.013525 | 0.034596 | -0.046627 | 0.070403 | 0.026721 | 0.1108% | — |

**4.2 模型的构建**

4.2.1 基于分位数回归的VaR估计

运用分位数回归技术计算QDII基金VaR，借鉴王新宇、赵绍娟（2008）的方法引入滞后5项的收益率序列作为解释变量，同时考虑到波动率对收益率的影响在上述模型的基础上加入新的解释变量波动率，波动率的估计参照已实现的方差的做法。

记为滞后5期的平均收益率，波动率的估计值为

以下建立分位数回归模型，分别得到不同置信水平下的VaR估计，并将分位点分别为0.01,0.05,0.1的VaR值与QDII基金收益率序列如图3所示，青色虚线左边为样本内数据的拟合情况，青色虚线右边为样本外数据的拟合情况。

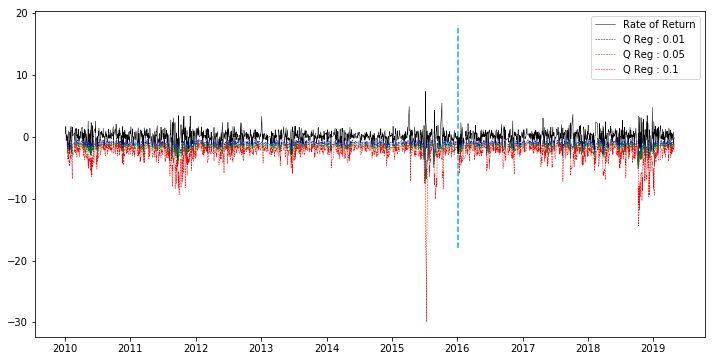


图4 基于分位数回归的VaR估计

4.2.2 基于分位数回归森林的VaR值估计

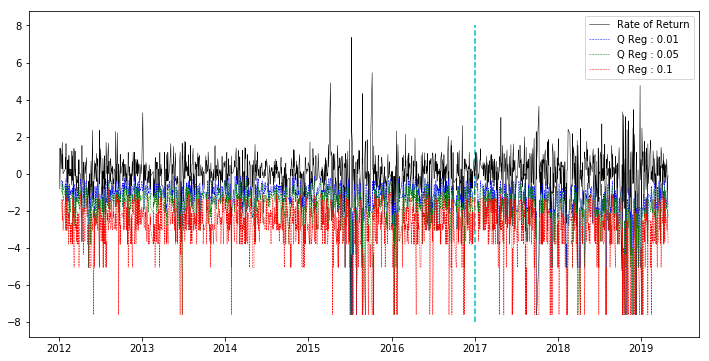
运用分位数回归森林计算QDII基金VaR，为了避免特征输入对模型估计结果的影响，选取上述分位数回归的的解释变量作为输入特征估计QDII基金的VaR风险测度。分位数回归森林的每个叶子都记录着所有观测的响应值。利用分位数回归森林计算QDII得VaR值，只需完全展开树，建立收益率的分布，取相应的分位数点就得到VaR值。分位点取0.01，0.05，0.1的VaR值与QDII基金收益率序列如图3所示，青色虚线左边为样本内数据的拟合情况，青色虚线右边为样本外数据的拟合情况。

图5 基于分位数回归森林的VaR估计

4.2.3 基于EMD-QRF模型的VaR估计

运用EMD-QRF模型估计QDII基金的VaR值，首先采用EMD算法对原始序列进行处理，提取市场不同频率的趋势，对频率大小相似的分量进行重构为新的分量。然后对高频分量建立QRF预测VaR值，对趋势分量和低频分量利用历史模拟法计算VaR值。最后将三个分量的VaR值叠加，得到原序列的VaR值。分位点取0.01，0.05，0.1的VaR值与QDII基金收益率序列如图3所示，青色虚线左边为样本内数据的拟合情况，青色虚线右边为样本外数据的拟合情况。

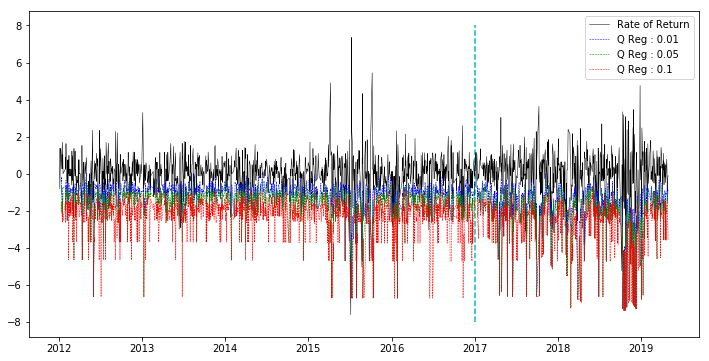


图5 基于分位数回归森林的VaR估计

4.3 对比分析结果

以下利用Kupiec似然比检验方法，将VaR预测数值与实际数据的差距进行量化，从而对分位数回归模型、分位数回归森林模型和EMD-QRF模型进行对比，判别哪个模型能更准确的预测上证综合指QDII基金收益率数据的VaR。样本内数据的样本容量为1142，样本外数据的样本容量为534。首先将得到的VaR序列与实际损失序列进行对比，得到不同模型不同置信水平下的LR统计量以及统计量所对应的p值。得到的结果如表3所示。

在99%置信水平下，对于样本内数据 ，分位数回归森林和EMD-QRF模型都能适应样本内VaR风险测度。从LR值上看，EMD-QRF的LR值与分位数回归森林相比较大，说明EMD-QRF的在样本内的预测精度相对较低。对于样本内数据，分位数回归和分位数回归森林的LR值相对较大，而EMD-QRF模型依然能适应样本内VaR风险测度。

在95%和90%置信水平下，对于样本内数据，虽然三个模型在VaR风险估计都差强人意，但整体来看，分位数回归森林和的LR都比分位数回归森林LR值小，说明分位数回归森林和EMD-QRF模型的表现相对较好，而对于样本外数据，EMD-QRF的LR值显著小于分位数回归模型和分位数回归森林。

综上，EMD-QRF模型和分位数模型在VaR的估计上，较传统的分位数回归模型更有优势，且相较于分位数回归森林模型，EMD-QRF模型有更好的泛化能力。

表3 Kupiec LR检验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | 置信度 | 失败天数 | 失败率 | LR统计量 |
| 训练集 | 分位数回归 | 90% | 142 | 0.102973 | 0.134269 |
| 95% | 89 | 0.064540 | 5.643786 |
| 99% | 26 | 0.018854 | 8.665477 |
| 分位数回归森林 | 90% | 145 | 0.105147 | 0.400133 |
| 95% | 75 | 0.052212 | 0.140077 |
| 99% | 14 | 0.010153 | 0.003214 |
| EMD-QRF模型 | 90% | 161 | 0.116751 | 4.102579 |
| 95% | 87 | 0.063089 | 4.609458 |
| 99% | 17 | 0.012328 | 0.702722 |
| 测试集 | 分位数回归 | 90% | 92 | 0.120104 | 3.253836 |
| 95% | 54 | 0.070496 | 6.042876 |
| 99% | 22 | 0.028721 | 18.014232 |
| 分位数回归森林 | 90% | 53 | 0.400133 | 8.958479 |
| 95% | 28 | 0.036554 | 3.203363 |
| 99% | 15 | 0.019582 | 5.552421 |
| EMD-QRF模型 | 90% | 68 | 0.088773 | 1.110684 |
| 95% | 39 | 0.050914 | 0.013390 |
| 99% | 10 | 0.013055 | 0.658601 |

**5.结论**

本文通过分位数回归，分位数回归森林和EMD-QRF三个模型对华夏全球精选QDII基金的历史收益率进行VaR风险度量，并且进行了对比研究。实证结果表明，EMD-QDII模型能有效给出VaR风险测度估计，且具有很好的泛化能力。基于EMD-QDII的VaR模型可以较好、稳健的反映QDII基金的市场风险。 “ 数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已 ”。本文为了说明模型的有效性，没有在特征工程上做文章。在模型的实际应用中，应根据研究的问题和数据选择更好的特征，从而使得在金融资产风险度量上表现出更出色的性能。

**参考文献**

1. Koenker R,Bassett G.Regression quantiles[J].Econometrica,1978,46(1):33-50
2. 王新宇, 赵绍娟. 基于分位数回归模型的沪深股市风险测量研究[J]. 中国矿业大学学报, 2008, 37(3).
3. 李邦国. 基于分位数回归的大豆期货市场的风险分析[J]. 中国商论, 2017(33):9-11.
4. Bayer, and Sebastian. "Combining Value-at-Risk Forecasts Using Penalized Quantile Regressions." Econometrics and Statistics (2017):S2452306217300680.
5. 刘演森. 贝叶斯分位数回归模型及其在金融风险管理中的应用[D]. 山东大学, 2017.
6. 许启发, 张金秀, 蒋翠侠. 基于支持向量分位数回归多期VaR测度[J]. 系统工程学报, 2014, 29(2):202-214.
7. 许启发, 徐金菊, 蒋翠侠, et al. 基于神经网络分位数回归的VaR金融风险测度[J]. 合肥工业大学学报(自然科学版), 2014(12):1518-1522.
8. 孙国强,梁智,俞娜燕, et al..基于EWT和分位数回归森林的短期风电功率概率密度预测[J]. 电力自动化设备, 2018, v.38；No.292(08):165-172.
9. Huang, Norden E., et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis." Proceedings Mathematical Physical & Engineering Sciences 454.1971(1998):903-995.
10. 李合龙, 杨志. 基于EMD的VaR估计方法及实证研究[J]. 统计与决策, 2015(1):163-166.
11. Awajan, Mohd Tahir Ismail Ahmad M. , and S. A. Wadi . "A HYBRID APPROACH EMD-MA FOR SHORT-TERM FORECASTING OF DAILY STOCK MARKET TIME SERIES DATA." Journal of Internet Banking and Commerce 22.1(2017):1-10.
12. 孟二浩.基于改进EMD混合模型月径流预测研究[D]. 西安理工大学, 2018.