doi:10. 16055/j. issn. 1672-058X. 2022. 0006. 008

高频数据下基于 LSTM 的协方差矩阵预测模型

包悦妍
(南京审计大学 统计与数据科学学院,南京 211815)

摘　要:协方差矩阵的建模与预测,对于金融风险管理、投资组合管理等至关重要。针对时间序列模型对高维变量预测精度较低的问题,利用长短记忆神经网络模型(LSTM),提出了基于深度学习的高频数据已实现协方差矩阵预测模型。利用金融高频数据得到已实现协方差矩阵,对其进行 DRD 分解,针对相关系数矩阵 $\boldsymbol{R}$ 进行向量化处理,利用向量异质自回归模型(HAR)预测已实现相关系数矩阵 $\boldsymbol{R}$;针对已实现波动率矩阵 $\boldsymbol{D}$,利用半协方差(semi covariance)思想,结合 LSTM 模型,得到已实现波动率矩阵 $\boldsymbol{D}$ 的深度学习预测模型,构建了 LSTM-SDRD-HAR 已实现协方差矩阵动态预测模型。LSTM 模型和 HAR 模型能捕捉实际数据的长期记忆性,半协方差有利于捕捉金融数据的杠杆性。实证分析表明:相较于传统向量 HAR 已实现协方差矩阵预测模型,LSTM-SDRD-HAR 预测已实现协方差矩阵更为准确,基于 LSTM-SDRD-HAR 预测已实现协方差矩阵构造的有效前沿组合投资效果更佳。

关键词:LSTM 模型;协方差矩阵预测;已实现半协方差;Markowitz 有效前沿
中图分类号:O212. 1　　**文献标志码**:A　　**文章编号**:1672-058X(2022)06-0065-06

0　引　言

协方差矩阵建模在风险管理、投资组合管理和资产定价方面有着重要的应用。同时,马科维茨[1]投资组合理论的提出进一步推动了协方差矩阵的一系列研究。新时代下大量数据带来了更多新的挑战,比如维数问题导致的估计的一致性、预测的精度等。目前,使用高频日内数据获得更可靠的低频收益率协方差矩阵是较为主流的方法。

Merton[2]最早提出基于日内收益平方和估计波动率的方法——已实现方差(Realized Variance, RV)。随后"已实现"方法也被运用到协方差估计上,得到了已实现协方差;Dong 等[3]基于 HAR 思想和 Engle(2002)的动态条件相关系数模型(Dynamic Conditional Correlation, DCC)方法,构造了 HAR-DRD 波动率矩阵预测模型;Callot 等[4]提出 VAR-LASSO 模型,他们对波动率矩阵做了对数化处理,然后使用向量自回归模型建模,结合 LASSO 方法对系数矩阵进行降维,估计出具有稀疏特性的系数矩阵;Bollerslev 等[5]在已实现协方差矩阵的基础上,将其分解成 3 部分:正部、负部和混合已实现半协方差矩阵,结果表明半协方差方法对经济信息反应更敏锐,有效提升了投资组合波动率的预测精度。深度学习因为其能处理更多种类的信息被广泛应用于波动率的预测上,如长短记忆神经网络模型(LSTM)、循环神经网络模型(RNN)、卷积神经网络(CNN)等方法。现有的研究一般将深度学习用来预测股票指数的波动率,如 Zhou 等[6]使用 CSI300 和百度每日 28 个搜索关键词作为 LSTM 模型的输

收稿日期:2021-10-13;**修回日期**:2021-11-14.
基金项目:国家社会科学基金(19BTJ035);江苏省研究生科研创新计划项目(KYCX20-1675).
作者简介:包悦妍(1997—),女,江苏无锡人,硕士研究生,从事高频金融研究.

入来预测指数波动率。另外,还有一些学者尝试将深度学习方法和传统时间序列方法结合来预测波动率,如 Psaradellis 等[7]提出一种将异质自回归模型(HAR)[8]和遗传算法支持向量机模型(GASVR)相结合的方法(HAR-GASVR)对波动率进行预测。许多研究表明深度学习模型在波动率预测与应用方面有着优秀的表现,但仅限于一维情况,对于预测多维协方差矩阵方面的研究几乎没有,深度学习模型的运行过程是个"黑匣子",模型的解释度差。与深度学习模型相反,时间序列模型虽然对高维数据的处理能力差,但它能够对波动率特有的性质,如长期记忆性、聚集性、杠杆性进行刻画,因此模型的可解释性强。

针对以上问题,提出了基于 LSTM 模型、DRD 分解、半协方差思想和 HAR 模型的协方差矩阵预测模型(LSTM-SDRD-HAR)。半协方差方法和 HAR 模型可以反映协方差矩阵存在的长期记忆性和杠杆性,让模型更好理解,LSTM 模型则可以提高预测精度,通过模型结合实现预测模型的强解释性和高预测准确度。本文先介绍了关于模型 LSTM-SDRD-HAR 的构建原理,然后对其进行统计评价和经济效益评价,结果表明协方差矩阵预测模型 LSTM-SDRD-HAR 的预测精度高,在投资组合中表现优秀。

1 模型与方法

1.1 已实现协方差矩阵与已实现半协方差矩阵

Andersen 等[9]于 1998 年提出了已实现波动率的概念,使得波动率变成"可观测值",就有了已实现协方差的定义:

$$\mathrm{RCOV}_{ij,t}=\sum_{k=1}^{M} r_{i,t,k/M} r_{j,t,k/M}$$

其中,$\mathrm{RCOV}_{ij,t}$ 是第 i 个资产与第 j 个资产在第 t 天的已实现协方差,$r_{i,t,k/M}$ 为第 i 个资产第 t 天的第 k 个对数收益率,则 n 个资产的已实现协方差矩阵为

$$\sum\nolimits_t=(\mathrm{RCOV}_{ij,t})_{n\times n}$$

Bollerslev 等在 2020 年进一步提出了已实现半协方差的概念。首先,定义函数 $p(x)\equiv\max\{x,0\}$,$n(x)\equiv\min\{x,0\}$,然后将已实现协方差矩阵分成 3 个部分,分别是正部、负部和混合部,假设考虑 n 个资产,定义分别如下:

$$\widehat{\boldsymbol{P}}_t\equiv\sum_{k=1}^{M} p(\boldsymbol{r}_{t,k/M})p(\boldsymbol{r}_{t,k/M})^{\mathrm{T}}$$

$$\widehat{\boldsymbol{N}}_t\equiv\sum_{k=1}^{M} \boldsymbol{n}(\boldsymbol{r}_{t,k/M})\boldsymbol{n}(\boldsymbol{r}_{t,k/M})^{\mathrm{T}}$$

$$\widehat{\boldsymbol{M}}_t\equiv\sum_{k=1}^{M} (\boldsymbol{p}(\boldsymbol{r}_{t,k/M})\boldsymbol{p}(\boldsymbol{r}_{t,k/M})^{\mathrm{T}}+\boldsymbol{n}(\boldsymbol{r}_{t,k/M})\boldsymbol{n}(\boldsymbol{r}_{t,k/M})^{\mathrm{T}})$$

其中,$\boldsymbol{r}_{t,k/M}$ 是一个 n 维向量,表示 n 个资产第 t 天第 k 个对数收益率序列,且 M 取任何数时都有 $\sum_t=\widehat{\boldsymbol{P}}_t+\widehat{\boldsymbol{N}}_t+\widehat{\boldsymbol{M}}_t$。同符号部(正部和负部)与混合部关于随机相关性和价格的跳的经济信息是有区别的,同符号部描述了动态杠杆效应,因此半协方差分解方法是对协方差矩阵的进一步刻画,它能捕捉到关于协方差矩阵更多的经济信息,可以显著提高预测投资组合回报方差的准确性。

1.2 HAR 模型

HAR 模型可用于描述波动率的长期记忆性,鉴于该模型结构简单且预测效果好,被广泛使用。Chiriac 等[10]将 HAR 模型从一维情形推广到高维情形,得到向量形式的 HAR。他们对已实现协方差矩阵 $\sum_t$ 进行拉直向量化,由于 $\sum_t$ 为对称矩阵,可只将其下三角部分进行向量化 $\boldsymbol{H}_t=vech(\sum_t)$,则 $\boldsymbol{H}_t$ 为 $n^*=n(n+1)/2$ 维向量,向量形式 HAR 模型为

$$\boldsymbol{H}_t=\boldsymbol{\theta}_0+\theta_1\boldsymbol{H}_{t-1}+\theta_2\boldsymbol{H}_{t-5|t-1}+\theta_3\boldsymbol{H}_{t-22|t-1}+\varepsilon_t$$

其中,$\boldsymbol{H}_{t-s|t-1}=\frac{1}{s}\sum_{i=1}^{s}\boldsymbol{H}_{t-i}$ 表示 s 天内 $\boldsymbol{H}_t$ 的平均值,$\boldsymbol{H}_{t-5|t-1}$ 表示 $\boldsymbol{H}_t$ 的周效应,$\boldsymbol{H}_{t-22|t-1}$ 对应于月效应,描述 $\boldsymbol{H}_t$ 的长期记忆性。为了模型简便和估计的有效性,假设常数项 $\boldsymbol{\theta}_0$ 是 n^* 维向量,回归系数参数 θ_1、θ_2 和 θ_3 都是标量。当 $\boldsymbol{H}_t$ 为一维已实现方差 RV_t 时,此时模型即为 Corsi 提出的(标量形式)HAR。

1.3 LSTM 模型

LSTM 网络是一种特殊的循环神经网络(RNN),它与 RNN 的区别在于中间状态的更新方式不同,这让 LSTM 能够解决梯度消失和梯度爆炸问题。不仅如此,在处理一些需要长期记忆问题,即当研究的序列较长时,LSTM 的表现优于 RNN。因此,LSTM 模型十分适合用来预测波动率,LSTM 模型构建如下:

$$f_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},x_t]+\boldsymbol{b}_f)$$

$$i_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},x_t]+\boldsymbol{b}_i)$$

$$\widetilde{C}_t=\tanh(\boldsymbol{W}_C\cdot[h_{t-1},x_t]+\boldsymbol{b}_C)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \widetilde{C}_t$$
$$O_t = \sigma(\boldsymbol{W_O} \cdot [h_{t-1}, x_t] + \boldsymbol{b_O})$$
$$h_t = O_t \circ \tanh(C_t)。$$

其中,σ 表示 sigmoid 函数,$\boldsymbol{W}$ 表示权重向量,$\boldsymbol{b}$ 表示偏置向量,h_{t-1} 为上一时刻的输出值,x_t 为当前时刻的输入值;基于激活函数 σ 的性质,输出值 f_t 在[0,1]之间,给 C_{t-1} 赋上权重,起到“遗忘”的作用;i_t 与 f_t 与之类似,起到权重的作用,i_t 会将重要的信息从单元状态 $\widetilde{C}_t$ 中挑选出来,更新原有单元状态;O_t 的计算方式与 f_t 以及 i_t 相同,最后根据新单元状态 C_t 和 O_t 得到当前时刻的输出 h_t。

1.4　LSTM-SDRD-HAR 模型的构建

LSTM-SDRD-HAR 模型就是通过整合 DRD 分解、半协方差、LSTM 模型和 HAR 模型来构建的,具体实施步骤:

步骤 1　根据 DRD 分解 $\sum_t = \boldsymbol{D_t R_t D_t}$ 计算得到 $\boldsymbol{D_t} = \boldsymbol{diag}(RD_{1,t}, \cdots, RD_{p,t})$,其中 $RD_{i,t} = \sqrt{RV_{i,t}} = \sqrt{\sum_{ii,t}}$。借鉴已实现半协方差的思想,计算第 i 只股票在 t 天的已实现波动率的正部和负部,记为 $RV^+_{i,t}, RV^-_{i,t}$:

$$RV^+_{i,t} = \sum_{k=1}^{M} \boldsymbol{p}^2(\boldsymbol{r}_{t,k/M})$$
$$RV^-_{i,t} = \sum_{k=1}^{M} \boldsymbol{n}^2(\boldsymbol{r}_{t,k/M})$$

将 $RV^+_{i,t}, RV^-_{i,t}$ 做对数化处理后的序列$(\ln(RV^+_{1,t}), \ln(RV^-_{1,t}), \cdots, \ln(RV^+_{n,t}), \ln(RV^-_{n,t}))$作为 LSTM 模型的输入变量,对 $\boldsymbol{D_t}$ 每个元素 $RD_{i,t}$ 进行分别建模,在这里,取对数可以保证 $\boldsymbol{D_t}$ 的预测结果恒正。

步骤 2　对 $\boldsymbol{R_t}$ 作拉直向量化处理,由于 $\boldsymbol{R_t}$ 为对称矩阵且主对角线恒为 1,只需对其下三角矩阵进行拉直向量化,令 $\boldsymbol{y_t} = vech(\boldsymbol{R_t})$,利用 HAR 模型对 $\boldsymbol{y_t}$ 进行建模,再返回至矩阵形式。

步骤 3　将上述两步得到的预测值 $\widehat{\boldsymbol{D}}_t, \widehat{\boldsymbol{R}}_t$ 代入 $\widehat{\sum}_t = \widehat{\boldsymbol{D}}_t \widehat{\boldsymbol{R}}_t \widehat{\boldsymbol{D}}_t$,得到已实现协方差矩阵的预测。

LSTM-SDRD-HAR 模型通过 LSTM 模型和 HAR 模型刻画数据的长期记忆性,半协方差方法刻画数据的杠杆性,让模型更容易挖掘数据的特征,并且针对波动规律不同的部分 $\boldsymbol{D_t}$ 和 $\boldsymbol{R_t}$ 分别建模,具体问题具体分析,有利于找寻各自规律,提高模型的预测精度。

2　实证分析

实证采用上证 50 成分股中 10 只股票的 5 min 交易数据,这 10 只股票的选取方法是将成分股的股票代码从小到大排序,取排在前 10 的股票。5 min 高频收益率数据来源于锐思数据库,时间跨度为 2004-01-02—2019-12-31,将整个数据样本分为样本内数据和样本外数据,样本内数据为 2004-01-02—2016-02-05 期间数据,共 2 850 个交易日,样本外数据为 2016-02-06—2019-12-31 期间数据,共 943 个交易日。

2.1　数据处理

根据 10 只股票的 5 min 高频收益率计算已实现协方差矩阵。计算已实现协方差矩阵时,若某只股票缺失数据很多,将会出现已实现协方差矩阵不正定的情况,会给估计和预测带来较大误差,所以对交易日进行剔除处理,删去数据缺失率达到 25%的交易日,剔除后剩余 3 793 个交易日。对于数据缺失量少于 25%的交易日内数据,进行缺失值填补,填补规则为采用上一个时间段的收盘价价格,作为该时间段的收盘价格,计算出该 5 min 收益率。

表 1 展示了分解后 10 只股票已实现波动率 $\boldsymbol{D_t}$ 的均值与标准差,还有相关系数矩阵 $\boldsymbol{R_t}$ 的平均水平。从表 1 可知:10 只股票的已实现波动率均值在 0.02 附近,且不同股票的已实现波动率相对差异较大,体现了已实现波动率的相对独立特性。此外,这 10 只股票对应的已实现波动率标准差相近,一定程度上反应了 10 只股票的流动性和上证 50 市场的特性,即当市场报价集中在一个共同价格水平附近,市场的买方和卖方都愿意在当前价格附近,以较小的价差执行股票交易,波动率会较低且保持稳定。流动性强的市场会有更多的机遇,投资者通过合理制定投资策略就可以实现获利。在研究相关关系时,相关系数大于等于 0.3,就可以认为数据存在相关关系,从表 1 相关系数矩阵 $\boldsymbol{R_t}$ 的均值可以看出,同一市场内一些股票之间的相关性很小,但其他股票之间有相关关系,研究清楚其中的关系,有利于对已实现协方差矩阵更准确地预测并做出投资计划。在此,研究上证 50 的这 10 只股票是有意义和价值的。

表 1 描述性统计

Table 1 Descriptive statistics

股票名称	D_t		R_t 均值									
	均值	标准差										
浦发银行	0.018 5	0.010 7	1.000 0									
东风汽车	0.024 8	0.017 5	0.226 9	1.000 0								
上海机场	0.020 2	0.010 8	0.246 6	0.186 7	1.000 0							
五矿发展	0.026 9	0.015 1	0.303 3	0.265 9	0.230 9	1.000 0						
葛洲坝	0.023 5	0.013 6	0.283 2	0.249 2	0.225 0	0.306 7	1.000 0					
特变电工	0.021 8	0.012 4	0.308 5	0.260 0	0.235 2	0.328 7	0.315 9	1.000 0				
广州发展	0.021 8	0.012 5	0.230 8	0.225 4	0.187 2	0.257 4	0.235 8	0.253 2	1.000 0			
同方股份	0.024 7	0.013 4	0.319 5	0.258 8	0.243 9	0.337 5	0.301 8	0.330 1	0.248 1	1.000 0		
上汽集团	0.022 3	0.012 0	0.313 2	0.248 1	0.236 2	0.283 6	0.266 6	0.286 8	0.216 6	0.294 7	1.000 0	
国金证券	0.027 4	0.016 0	0.338 8	0.261 6	0.237 5	0.339 0	0.318 2	0.344 2	0.245 3	0.344 4	0.290 7	1.000 0

经过比较每一只股票的已实现波动率 $RD_{i,t}$ 和已实现相关系数 $R_t(i,j)$ 时间序列图,发现已实现波动率 $RD_{i,t}$ 和已实现相关系数 $R_t(i,j)$ 的变化规律不同,限于篇幅,下文仅以浦发银行为例。图 1 描绘的是第一只股票浦发银行的已实现波动率 $RD_{1,t}$,大致在 2008 年和 2016 年,$RD_{1,t}$ 波动得比较剧烈,整体上波动幅度不大,较为稳定;图 2 描绘的是浦发银行与第二只股票东风汽车的已实现相关系数 $R_t(1,2)$,其波动幅度大,说明已实现相关系数 $R_t(1,2)$ 反映了一定时间内市场情况的变化。比较图 1 和图 2,可以直观地看到:已实现波动率 $\boldsymbol{D}_t$ 和相关系数 $\boldsymbol{R}_t$ 两者的形态是不同的,已实现波动率 $\boldsymbol{D}_t$ 的时间序列图有明显的波峰,变化值小且稳定;而相关系数 $\boldsymbol{R}_t$ 波峰不明显,波动大且频繁,因此有必要对 $\boldsymbol{D}_t$ 和 $\boldsymbol{R}_t$ 分别进行研究。

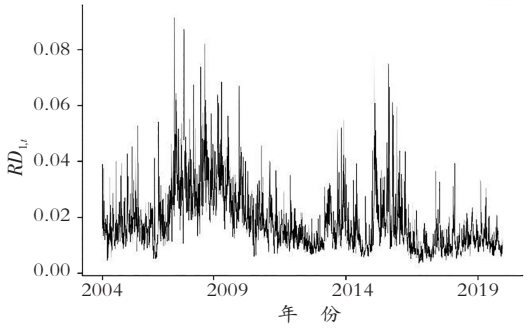

图 1 已实现波动率 $RD_{1,t}$ 时间序列图

Fig. 1 Time series of realized volatility $RD_{1,t}$

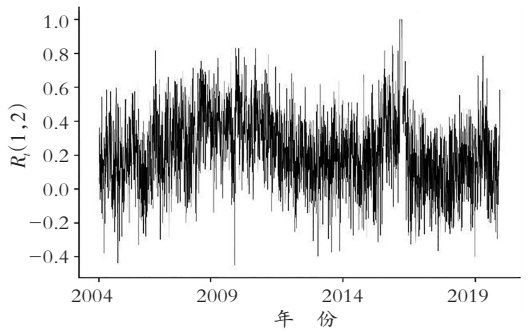

图 2 相关系数 $R_t(1,2)$ 时间序列图

Fig. 2 Time series of correlation coefficient $R_t(1,2)$

2.2 预测评价

选取均方根误差,F_{RMSE} 和均方误差 F_{MAE} 这两个指标来评价已实现协方差矩阵预测模型的预测能力,指标定义如下:

$$F_{\mathrm{RMSE}} = \frac{1}{T}\sum_{t=1}^{T} F_{\mathrm{RMSE}t} = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{n^2}tr\left(\left[\boldsymbol{H}_t - \sum\nolimits_t\right]\left[\boldsymbol{H}_t - \sum\nolimits_t\right]^{\mathrm{T}}\right)}$$

$$F_{\mathrm{MAE}} = \frac{1}{T}\sum_{t=1}^{T} F_{\mathrm{MAE}t} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n^2}\left\|\boldsymbol{H}_t - \sum\nolimits_t\right\|_1$$

在实证研究中,$n=10$,$T=943$。

为了展示构建的已实现协方差矩阵的预测效果,在此将其与 HAR、EWMAQ、HAR-DRD、LASSO-VAR、LASSO-DRD、LSTM-DRD-HAR 模型进行比较。

表 2 展示了 7 个模型在样本外的预测误差,可以看到 HAR 模型存在的明显问题就是预测效果差。EWMAQ 模型是在指数移动平均模型(EWMA)的基础上将估计误差考虑进模型的调参过程,结果显示这个做法确实带来了预测效果的提升,结果与 HAR-DRD 模型相差无几,但 EWMAQ 在计算过程中会涉及四阶矩,计算较为复杂且耗时久。与 EWMAQ 相比,剩下 6 个模型的计算过程简便且耗时短。表 2 结论主要如下:对比 HAR 和 HAR-DRD、LASSO-VAR 和 LASSO-DRD 模型发现,经过 DRD 分解再建模比不分解直接建模预测结果更准确,这说明事先进行 DRD 分解是有必要的,它将不同结构的组成部分分离开来,对不同部分建模找寻各自的规律可以提高准确性,所以考虑的模型中大部分都采用了 DRD 分解;表中预测效果最好的是 LSTM 类模型,LSTM 算法比 LASSO 算法的模型精度提升至少 8%,LSTM-SDRD-HAR 的样本外预测结果仅次于 LSTM-DRD-HAR。总体来说,DRD 分解和 LSTM 模型能有效改进模型预测精度,并且不会增加计算复杂度。

表 2　已实现协方差矩阵样本外预测结果

Table 2　Out-of-sample prediction results of the realized covariance matrix

模　型	F_{RMSE} (10^{-4})	F_{MAE} (10^{-4})	基于 EWMAQ 提升百分比 (F_{RMSE})%	基于 EWMAQ 提升百分比 (F_{MAE})%
HAR	24.570 3	24.435 0		
EWMAQ	2.991 8	1.889 9		
HAR-DRD	2.280 5	1.178 0	20.7	37.7
LASSO-VAR	1.687 2	0.984 7	43.6	47.9
LASSO-DRD	1.444 6	0.909 9	51.7	51.9
LSTM-DRD-HAR	0.877 4	0.639 7	70.7	66.2
LSTM-SDRD-HAR	1.203 3	0.820 4	59.8	56.6

2.3　经济评估

为了评估波动率矩阵预测的经济价值,考虑马科维茨有效前沿。假定投资者是风险厌恶型的,则在相同的年化预期收益率 μ_p 下,他们会选择风险更小的资产;同样地,如果风险水平相同,投资者们就会选择高收益资产。在这里,最优投资组合就是下面这个问题的解:

$$\min_{w_{t+s|t}} \boldsymbol{w}'_{t+s|t}\widehat{H}_{t:t+s}\boldsymbol{w}_{t+s|t}$$

$$\text{s.t.}\quad \boldsymbol{w}'_{t+s|t}E_t[r_{t:t+s}]=\frac{s\mu_p}{252},\boldsymbol{w}'_{t+s|t}\iota=1$$

其中,$\boldsymbol{w}_{t+s|t}$ 是 $n\times1$ 维的基于第 t 天得到的第 $t+s$ 天的投资组合权重向量,ι 是元素全为 1 的 $n\times1$ 维向量,$\frac{s\mu_p}{252}$是标准化到日收益率的目标收益率。为了评估不同模型的预测效果,对他们的"事后"条件投资组合的均值和标准差进行比较:当给定权重 $\boldsymbol{w}_{t+s|t}$ 时,就能计算出均值 $r^p_{t:t+s}=\boldsymbol{w}'_{t+s|t}\boldsymbol{r}_{t:t+s}$ 和标准差 $\sigma^p_{t:t+s}=\sqrt{\boldsymbol{w}'_{t+s|t}\boldsymbol{Y}_{t:t+s}\boldsymbol{w}_{t+s|t}}$。本文还考虑理想情况下的有效前沿,即 $\widehat{\boldsymbol{H}}_{t:t+s}=\sum_{t:t+s}$。结果如图 3 所示($s=1$)。

图 3 中"Oracle"代表的是理想情况,在所有有效前沿的上方,很明显它是最优的。星标的位置表示全局最小方差组合,从图 3 可以看出:HAR 模型的全局最小方差组合风险最大,收益仅高于 EWMAQ 模型;而 EWMAQ 模型的全局最小方差组合的收益最低。在经济评价中也有着和统计评价相一致的结果:LSTM 类模型的全局最小方差组合风险最小,因为模型 LSTM-DRD-HAR 精准的预测,所以其有效前沿曲线与理想情况相近,而模型 LSTM-SDRD-HAR 因为对重大事件敏感,所以能有效规避风险,在同等的风险水平下获得更高的收益;LSTM 类优于 LASSO 类模型,LASSO 类又优于 HAR、EWMAQ 和 HAR-DRD 类;考虑 DRD 分解的模型,它们的有效前沿曲线在没有考虑 DRD 分解模型的上方,说明 DRD 分解也有利于改善经济评价。综上,LSTM 算法和 DRD 分解不仅可以提高预测精度,在投资组合优化方面也起到了积极作用,其中模型 LSTM-SDRD-HAR 的综合评价最高。

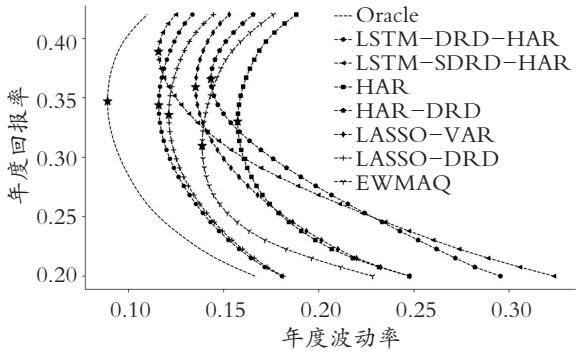

图 3　马科维茨有效前沿图

Fig. 3　Markowitz's effective frontier diagram

3　结　论

本文主要利用 DRD 分解、LSTM 和半协方差方法,构造了已实现协方差矩阵预测模型 LSTM-SDRD-HAR,并通过上证 50 中 10 只股票的实证分析,对模型的经济效益进行评价。首先,利用 DRD 分解,对分解后已实现波动率矩阵 $\boldsymbol{D}_t$ 和相关系数矩阵 $\boldsymbol{R}_t$ 分别建模,有利于捕捉各自动态规律;其次,为了让预测模型能对重大事件的发生有更敏锐的洞察力,结合了半协方差的思想,引入 $RV^{+}_{i,t}$ 和 $RV^{-}_{i,t}$;最后对相关系数矩阵 $\boldsymbol{R}$ 进行拉直向量化,采用 HAR 模型对其建模,更好地刻画波动率长期记忆性的特征。实证结果表明:利用 LSTM 模型和 DRD 分解的方法,可以有效提高模型预测精度;LSTM-SDRD-HAR 模型不仅有较高的预测准确度,在投资组合中的应用表现更为突出,对重大事件的反应更灵敏,能为投资者决策提供理论支撑,从而降低损失,增加收益。

近年来,深度学习在预测低维时间序列时表现出很好的预测效果,如何利用深度学习预测优势,结合传统时间序列模型的特征优势,从而提升协方差矩阵的预测能力,值得深入分析与研究。

参考文献(References):

[1] MARKOWITZ H. Portfolio selection [J]. Journal of Finance, 1952(7): 77—91.

[2] MERTON R C. On estimating the expected return on the market: an exploratory investigation [J]. Journal of Financial Economics, 1980, 8(4): 323—361.

[3] DONG H O, PATTON A J. High-dimensional copula - based distributions with mixed frequency data[J]. Journal of Econometrics, 2016, 193(2): 349—366.

[4] CALLOT L, KOCK A B, MEDEIROS M C. Modeling and forecasting large realized covariance matrices and portfolio choice[J]. Journal of Applied Econometrics, 2017, 32(1): 140—158.

[5] BOLLERSLEV T, LI J, PATTON A J, et al. Realized semicovariances[J]. Econometrica, 2020(88): 1515—1551.

[6] ZHOU Y L, HAN R J, XU Q, et al. Long short-term memory networks for CSI300 volatility prediction with Baidu search volume[J]. Concurrency and Computation: Practice and Experience, 2019(31): 4721—4725.

[7] PSARADELLIS I, SERMPINIS G. Modelling and trading the U. S. implied volatility indices: evidence from the VIX, VXN and VXD indices[J]. International Journal of Forecasting, 2016, 32(4): 1268—1283.

[8] CORSI F. A simple approximate long-memory model of realized volatility[J]. Journal of Financial Econometrics, 2009(7): 174—196.

[9] ANDERSEN T G, BOLLERSLEV T. Answering the skeptics: yes, standard volatility models do provide accurate forecasts [J]. International Economic Review, 1998, 39(4): 885—905.

[10] CHIRIAC R, VOEV V. Modelling and forecasting multivariate realized volatility [J]. Journal of Applied Econometrics, 2011, 26(6): 922—947.

Covariance Matrix Prediction Model Based on LSTM Using High Frequency Data

BAO Yue-yan

(School of Statistics and Data Science, Nanjing Audit University, Nanjing 211815, China)

Abstract: The modeling and prediction of the covariance matrix is very important for financial risk management and investment portfolio management. To solve the problem of low prediction accuracy of time series models for high-dimensional variables, long short memory neural network model (LSTM) is used to propose a covariance matrix prediction model using high-frequency data based on deep learning. The model uses financial high-frequency data to obtain the realized covariance matrix, performs DRD decomposition on the realized covariance matrix, vectorizes the correlation coefficient matrix $\boldsymbol{R}$, and uses the vector heterogeneous autoregressive model (HAR) to predict the realized correlation coefficient matrix $\boldsymbol{R}$. Based on the realized volatility matrix $\boldsymbol{D}$, this paper uses the idea of semi covariance, combined with the LSTM model, obtains the deep learning prediction model of the realized volatility matrix $\boldsymbol{D}$, and constructs the dynamic prediction model of realized covariance matrix, LSTM-SDRD-HAR. The LSTM and HAR model can capture the long-term memory of actual data, and the semi-covariance is conducive to capturing the leverage of financial data. The empirical analysis shows that compared with the traditional vector HAR prediction model, LSTM-SDRD-HAR has a more accurate prediction of the realized covariance matrix. The effective frontier portfolio investment structured by LSTM-SDRD-HAR prediction is better.

Key words: LSTM model; prediction of covariance matrix; realized semi-covariance; Markowitz effective frontier

责任编辑:李翠薇

引用本文/Cite this paper:

包悦妍. 高频数据下基于 LSTM 的协方差矩阵预测模型[J]. 重庆工商大学学报(自然科学版),2022,39(6):65—70.

BAO Yue-yan. Covariance matrix prediction model based on LSTM using high frequency data[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(6):65—70.