文章编号:1005-9679(2017)05-0020-08

指数成分股的高频波动相关性衡量与探究

王彦玮
(上海交通大学 安泰经济与管理学院,上海 200030)

摘 要: 以指数的高频波动率为研究对象,将其从一个崭新的角度进行了拆解:个股波动维度以及成分股波动相关性维度。按照不同指数走势的时间区间分类,对波动率的这两个维度与指数整体波动的相关性和一致性进行了考察及比较,以探究波动率变动的成因。最后,利用这样的拆解构建了高频波动率的预测模型,并将该模型的拟合和预测效果和其他模型进行了比较。

关键词: 高频波动;个股波动;波动相关性
中图分类号: F 832 **文献标志码:** A

The Correlation of High-frequency Volatility among Index Constituent Stocks

WANG Yanwei
(Antai College of Economics & Management, Shanghai Jiao Tong University, Shanghai 200030, China)

Abstract: This paper will study the high-frequency volatility of index from a brand-new perspective. The volatility of index will be decomposed into two separate dimensions-sum of volatility and correlation of volatility. The consistency between these two dimensions and the volatility of index will be checked and demonstrated separately under increasing stage, decreasing stage and fluctuating stage. Finally, this two variables will be applied to construct prediction equation for volatility which will be compared with other prediction models.

Key words: high-frequency volatility; sum of volatility; correlation of volatility

1 概述

对资产收益的波动率进行探究是目前金融学术研究的一个重要方向。资产收益的波动性之所以成为目前学界研究的热点,主要是因为波动率指标在资产定价、资产组合管理、投资风险管理等领域都有着不可取代的重要作用。对于我国资本市场来讲,由于金融衍生产品市场的逐步兴起、风险管理观念的成长以及数量化投资方式的发展,金融产品收益率的波动性研究自然日渐成为一个不可忽视的重要课题。

与此同时,随着计算机技术的普及以及数据存储技术的发展,很多学者在近年来对金融市场的研究中将研究数据从传统的日间数据逐步过渡到日内高频数据。由于采样更密集、包含的信息更多,相比较传统的低频数据时间序列,高频数据可以更好地发掘金融资产价格的运动规律以及微观结构,给我们提供更详尽的数据和信息。

首先,对于高频的价格数据,在波动率计算方法方面,由于原始数据采取了更密集的采样频率,与采用低频采样时有所不同。Andersen 和 Bollerslev (1998)[1] 率先提出了已实现波动率的概念,计算方

收稿日期:2016-12-20
作者简介:王彦玮(1992—),男,上海人,金融学硕士,主要研究方向:金融工程。E-mail:atswyw@126.com.

式为针对秒级别或者分钟级别的高频价格序列，求出对应的高频收益率并进行平方加和。

由于既往研究表明[2-4]，已实现波动率的测度方法具有无偏性以及较好的稳健性等优秀的统计特性，在本文中，对日内波动率的计算采用已实现波动率作为波动率测度。一般地，其定义式如式(1)所示：

$$RV_t = \sum_{i=1}^{n} r_i^2 \tag{1}$$

其中，RV_t 为 t 时间段内的高频已实现波动率；r_i 为 t 时间段内根据采样的价格序列得出的对应的收益率；n 为区间内收益率样本数。Andersen(2003)[5]证明随着采样频率的增加，已实现波动率会逐步趋近于对应区间的积分波动率，即：

$$\lim_{n\to\infty} \sum_{i=1}^{n} r_i^2 = \int_0 \sigma_s^2 \mathrm{d}s \tag{2}$$

需要指出的是，在计算收益率的时候，我们采取的是对数收益率，一方面因为其具备更好的统计特性，另一方面因为对数收益率是已实现波动率推导的理论基础。

2　本文创新点

目前，学术界对于高频波动率的测度和预测模型的研究已经较为成熟和充分，但是本文对于波动率的探究会从一个全新的角度进行。本文的研究对象为指数的高频收益波动情况，而对于计算得出的高频波动率，将其进一步拆解为两个维度：成分股波动行为的维度以及成分股之间波动行为相关性的维度。

定性地来看：成分股波动行为维度衡量的是个股各自的波动大小，而成分股之间波动行为相关性维度则是衡量构成指数的成分股的波动行为的一致性和相关性。

举例来讲，当成分股的波动较大、但是个股之间走势背离较大的时候，由于加权平均，指数层面的总体波动会出现一定的中和作用，往往会出现指数基本保持稳定、整体波动较小；反之，当个股之间的走势趋于一致的时候，也就是成分股波动相关性维度较大时，指数表现出来的波动也就相对较大。

本文做这样拆解的目的在于两点：第一，试图通过这样的拆解来探究不同的大盘趋势下指数波动的成因，即在指数出现较大幅度波动的情况下，分析其内在因素是个股波动还是个股波动相关性因素；第二，通过这样的拆解来从指数波动率数据中挖掘出更多的有效信息从而构建更好的波动率回归和预测模型。

所以，将指数波动率进行拆解，其中个股波动率和指标记做 sum_t，其计算公式定义为式(3)：

$$sum_t = \left(\sum_{i=1}^{n} \omega_{it}\sigma_{it}\right)^2 \tag{3}$$

其中，ω_{it} 为 t 时间段内成分股 i 在指数中的权重；σ_{it} 为在 t 时间段内个股 i 的已实现波动率的平方根；n 为指数成分股个数。从式(3)可以明显看出，该指标维度大小衡量的是成分股之间各自独立的波动行为。

类似地，将波动一致性指标记作 rho_t，并将其定义为：

$$rho_t = \frac{index_t}{sum_t} = \frac{index_t}{\left(\sum_{i=1}^{n} \omega_{it}\sigma_{it}\right)^2} \tag{4}$$

其中，$index_t$ 为指数在时间段 t 内的已实现波动率；sum_t 为时间段 t 内按照式(3)计算得出的个股波动率和指标。进一步展开式(4)可以得到：

$$rho_t = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{ij}\omega_{it}\omega_{jt}\sigma_{it}\sigma_{jt}}{\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_{it}\omega_{jt}\sigma_{it}\sigma_{jt}} \tag{5}$$

其中，ρ_{ij} 为成分股 i 和成分股 j 之间的波动率相关系数；σ_{it} 和 σ_{jt} 为在 t 时间段内个股的波动率；n 为指数成分股个数。从式(5)我们较为直观地做出判断，rho_t 指标本质是对成分股之间的波动率相关系数的一个加权平均。当个股之间波动相关系数 ρ_{ij} 全部等于 1 的时候，rho_t 会等于 1；反之，随着 ρ_{ij} 的下降，rho_t 也会随之发生下降。通过式(5)，也可以发现，该指标衡量的确实是指数成分股之间波动的相关性。

至此，文章完成了对两个波动维度的量化拆解和定义。下文的主要结构为：第三部分会对指数波动以及两个分维度进行描述性统计的初步分析；第四部分会使用 copula 模型对三者之间的关系进行更详尽的定性探究；第五部分会按照对波动率的拆解结果，尝试进行回归以及预测模型的构建；最后一部分简单总结全文，并进行后续工作的展望。

3　样本数据及其描述性统计

本文的研究对象为上证 50 指数及其成分股，所使用的高频数据均来源于新浪财经，高频数据的采样频率是每 3 秒一个价格数据。指数成分股的权重等其余数据信息均来源于东方财富 choice 数据终端，样本区间覆盖范围为 2013—2015 年的完整三年，区间内共包含 727 个交易日。其中，为了避免极端数值的影响，实证中将样本中“光大乌龙指”事件所发生的 2013 年 8 月 16 日对应的指数异常波动时间段去除。

对于所有有效样本数据，将五分钟，也就是 300 秒作为一个考察单位，对于每个考察区间，包含 100

个价格数据，利用包含在单个考察单位内的所有采样数据计算五分钟内的已实现波动率和其对应的两个波动维度的拆解数据。而对于每个交易日，我们拥有四个小时的交易时间，也就是说每个交易日可以获得 48 个波动率数据样本。

本文基于 Matlab2014b，eviews8.0 以及 excel 等平台开展数据处理和模型构建。

由于在不同的大盘走势情况下，指数的波动会呈现较大的差异，所以本部分的探讨将基于这三种大盘走势进行分段的探究。图 1 是样本区间内考察对象上证 50 指数每日收盘价走势图。众所周知，从 2014 年底开始到 2015 年年中，A 股迎来了一波快速上涨的牛市行情，但是接着出现了一波急剧下跌，三个月内，从接近 3 500 点的高位跌到 2 000 点以下。将这样两段时间定义为趋势性上涨和下跌阶段，剩下的定义为指数趋势性波动阶段。具体划分情况见图 1。

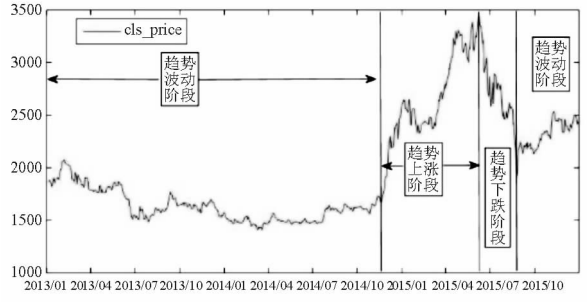

图 1　上证 50 收盘价走势及时间段划分

首先，按照前一部分给出的定义，对于波动相关性指标 rho_t 按照一般的认知进行分析。在整个指数呈现趋势上涨或者下跌的时间区间，由于成分股价格运动的趋势较为明显，成分股之间的波动相关性应该较为一致；而在盘整波动阶段，由于趋势性较弱，个股之间走势分歧较大，相应的波动相关性应该较弱。对于个股波动和维度 sum_t，在市场出现大幅波动的时间段内，个股的波动也可能显著提高。所以，我们合理预测，在指数趋势上涨、下跌以及波动阶段，波动性自身大小及其两个子维度的大小势必会呈现出不同的统计特性。

为了对指数波动率 $index_t$、个股波动和指标 sum_t 以及波动相关性指标 rho_t 有一个总体的认识，将各个指标进行日内平均。定义如下：

$$index\_dayt=\frac{1}{48}\sum_{i=1}^{48} index_{ti} \tag{6}$$

$$rho\_dayt=\frac{1}{48}\sum_{i=1}^{48} rho_{ti} \tag{7}$$

$$sum\_dayt=\frac{1}{48}\sum_{i=1}^{48} sum_{ti} \tag{8}$$

图 2 和图 3 为指数波动情况和其拆解指标在样本区间内的变动情况。

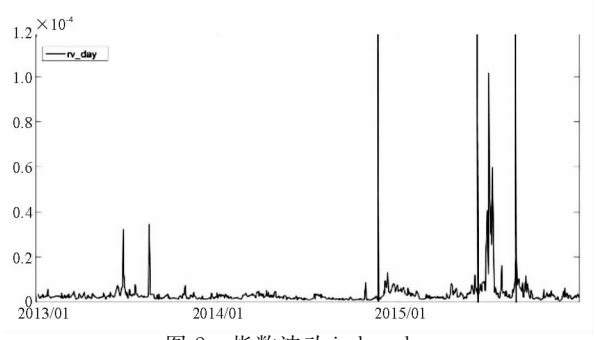

图 2　指数波动 index_day

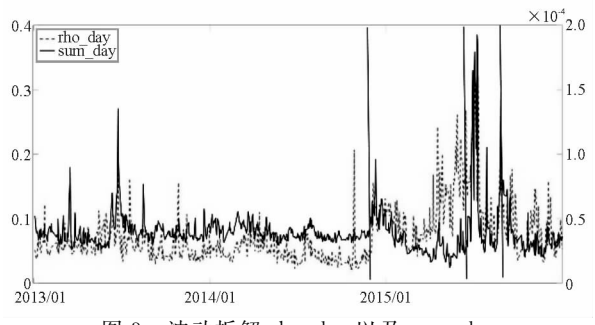

图 3　波动拆解 rho_day 以及 sum_day

从图 2 可以看出，趋势波动时间段内的指数波动，除去部分极端波动之外，相比较上涨和下跌阶段低，而下跌阶段的指数波动则远远大于其余时间段且极端波动数值较多。

接着观察拆解为两个具体维度的图 3。首先，对趋势上涨区间进行分析。实线所代表的个股波动维度 sum_dayt 和趋势波动区间相比没有显著差异，即在趋势上涨中，个股的价格波动行为并没有发生显著变化；与此不同的是，可以看到虚线所代表的波动一致性维度 rho_dayt 显著高于趋势波动阶段，这个结果与我们之前的预期是一致的。由于趋势性的存在，使得成分股之间的相关性呈现上升趋势，即趋势上涨时间段内指数波动升高的主要原因在于个股之间波动一致性的提高，而非个股波动行为的变动。

但是，趋势下跌阶段情况则不同，个股波动和波动一致性两个维度都远远高于其他两种阶段，两者的协同作用导致了图 2 中所示趋势下跌阶段指数波动的大幅增加，即趋势下跌阶段的指数波动增加是个股行为和成分股相关性提高带来的协同效应。

为了进一步进行定性的大小比较，表 1 列出波动率及其两个拆解维度的描述性统计数据。从偏度值和峰度值我们可以清楚看出分布的尖峰厚尾特性以及典型的右偏性。具体来看，首先，对于整体指数波动指标 $index_t$，和图 2 中反映的直观感受是一致

的，趋势下跌阶段的波动性显著较大，而趋势上涨阶段波动性又显著大于趋势波动阶段。对于相关性维度 rho_t，三种阶段的大小关系与指数波动的大小关系一致。最后，对于个股波动维度 sum_t，趋势下跌阶段依旧明显大于另外两者。

表 1　波动率数据描述性统计

item	period	mean	median	St. dev	skewness	kurtosis
index	fluct	2.66E-6	1.46E-6	4.83E-6	12.31	298.69
	rising	3.81E-6	2.35E-6	4.95E-6	7.26	125.71
	falling	1.09E-5	3.66E-6	3.34E-5	15.74	383.46
rho	fluct	0.059	0.042	0.059	4.21	29.12
	rising	0.10	0.084	0.085	2.79	13.51
	falling	0.14	0.11	0.11	2.45	9.30
sum	fluct	3.91E-5	3.57E-5	2.12E-5	18.75	842.24
	rising	3.37E-5	2.80E-5	2.28E-5	4.33	36.59
	falling	5.61E-5	3.93E-5	9.26E-5	12.04	226.18

注：其中的峰度数据统一为超额峰度数值。

结合图 2、图 3 以及表 1 反映出来的信息，可以得出以下几点结论：

首先，在股市的趋势上涨和趋势下跌阶段其高频波动性会显著提升，而且在股市下跌中波动性增幅更大，这也反映出金融市场中普遍存在的不对称性。由于风险厌恶性的存在，相比较上涨，投资者对于下跌所做出的反应更剧烈，恐慌情绪对市场稳定性的破坏也更大。通俗来讲，也就是所谓的“杀跌”影响要大于“追涨”影响，从而造成下跌阶段中更多的极端波动时间段。

同时，通过对波动性两个成因的详细比较我们可以看出，在趋势上涨阶段，个股的波动行为并未出现提升，指数波动性提升的主要原因在于个股之间波动一致性的提升，但是在趋势下跌阶段，个股波动和波动一致性两个维度均出现了较大程度的提升。

4　copula 模型分析

上一部分，初步定性地对指数波动及其两个维度在不同走势情况下的表现进行了比较。为了进一步定量地衡量在不同的大盘走势下，拆解出的两个维度值和整体的指数波动的相关关系，将采用 copula 模型来进行定量分析。

copula 模型由统计学家 Sklar 在 1959 年首次提出，Nelson(2006)首次系统性地总结了 copula 理论的主要研究成果。

在对变量相关性的探讨方面，copula 函数可以将多个变量自身的边缘分布函数和已知的联合分布情况联系起来，进而可以通过拟合得出的连接函数计算得出变量之间相关关系的大小。同时，魏平等[6]的研究指出，copula 还具有一些优良的性质，比如 copula 函数在单调递增变换下可以保持函数以及秩相关性不变性，可以发掘变量之间的相关模式，等等。

张连增等[7]指出，copula 模型的构建可以分为三大类，参数法、半参数法以及非参数法。其中，参数法是对于变量的边际分布以及 copula 函数形式都做出分布假设并进行拟合；半参数法是不对边际分布做假设和拟合，直接使用经验分布，仅仅指定 copula 连接函数的形式进行拟合；非参数法则是对边际分布和 copula 联合函数都不做假设，直接进行估算。

由于我们的样本数据往往无法完全符合某种特定分布，所以为了避免在边缘分布拟合中带来误差累积，本文使用的估计方式是半参数估计方法，即不对变量的边缘分布进行模型假设和拟合，使用变量的经验分布函数代入。

连接函数方面，在金融领域中对 copula 函数的使用主要有两种：椭圆族函数以及阿基米德族函数[6]。其中，椭圆族函数主要包括正态 copula 以及 t-copula，共同点是描述的是对称关系，而相比较正态 copula 模型，t-copula 模型对于变量的尾部相关性更敏感。阿基米德族函数中主要使用的包括，Gumbel，Clyton，Frank 函数。其中，Gumbel 函数对于上尾部的权重更大，Clyton 对于下尾部权重更大，Frank 模型则呈现上下尾部对称分布。

基于 copula 模型，对变量之间相关程度的衡量参数主要包括 Kendall 秩相关系数、Spearman 秩相关系数、Gini 相关系数以及尾部相关系数等。

对于已知的 copula 函数 C(u,v)，令 τ 为 Kendall 秩相关系数，定义式如式(9)：

$$\tau=4\int_0^1\int_0^1 C(u,v)\mathrm{d}C(u,v)-1 \tag{9}$$

令 ρ 为 Spearman 秩相关系数，定义式如式(10)：

$$\rho=12\int_0^1\int_0^1 C(u,v)\mathrm{d}uv-3 \tag{10}$$

令 γ 为 Gini 秩相关系数，定义式如式(11)：

$$\gamma=2\int_0^1\int_0^1(|u+v-1|-|u-v|)\mathrm{d}C(u,v) \tag{11}$$

与 Spearman，Kendall 以及 Gini 这样的全局相关系数不同的是，尾部相关系数则更多集中关注变量分布的极端情况，上尾和下尾相关系数的定义分别如式(12)、(13)所示：

$$\lambda=\lim_{u\to 1^-}(X>G^-1(u)|Y>F^{-1}(u))$$

$$=\lim_{u\to 1^-}\frac{1-2u+C(u,u)}{1-u} \tag{12}$$

$$\lambda=\lim_{u\to 0^+}(X<G^{-}1(u)|Y<F^{-1}(u))$$

$$=\lim_{u\to 0^+}\frac{C(u,u)}{u} \tag{13}$$

以上定义的五种相关系数为正表示呈现正相关，而数值越接近 0，两者的相关度越小，越接近于 1，两者的相关度则越大。

对 $index_t$ 与 tho_t 以及 sum_t 分别进行 copula 拟合，并计算其相关性，试图找出在不同趋势下，这两个维度对整个指数波动的相关性质。可以预计的是，计算出的相关系数应该都是显著为正，而系数越接近于 1，表明该维度与指数波动的一致性越高。为了保证结果的有效性，将使用不同模型进行拟合，并得出不同的相关系数。所有结果汇总于表 2 至表 4中。

表 2　趋势波动阶段 copula 模型拟合情况

	model	ρ	τ	γ	λ_u	λ_t
	gaussian	0.990	0.913	0.934	0.855	0.999
rho_t	student	0.759	0.568	0.624	0.413	0.994
&	gumbel	0.586	0.419	0.470	0.507	0.995
$index_t$	clayton	0.582	0.414	0.470	0.024	0.990
	frank	0.892	0.706	0.770	0.105	0.991
	model	ρ	τ	γ	λ_u	λ_t
	gaussian	**0.992**	**0.923**	**0.942**	**0.872**	**0.999**
sum_t	student	**0.855**	**0.677**	**0.731**	**0.598**	**0.996**
&	gumbel	**0.839**	**0.656**	**0.712**	**0.732**	**0.997**
$index_t$	clayton	**0.621**	**0.447**	**0.506**	**0.026**	**0.990**
	frank	**0.99**	**0.913**	**0.934**	**0.855**	**0.999**

表 3　趋势上涨阶段 copula 模型拟合情况

	model	ρ	τ	γ	λ_u	λ_t
	gaussian	**1.000**	**0.982**	**0.997**	**0.971**	**1.000**
rho_t	student	**0.983**	**0.893**	**0.920**	**0.853**	**0.999**
&	gumbel	**0.976**	**0.874**	**0.906**	**0.909**	**0.999**
$index_t$	clayton	**0.930**	**0.781**	**0.834**	**0.076**	**0.991**
	frank	**0.993**	**0.924**	**0.943**	**0.342**	**0.993**
	model	ρ	τ	γ	λ_u	λ_t
	gaussian	0.977	0.869	0.898	0.783	0.998
sum_t	student	0.646	0.471	0.525	0.373	0.994
&	gumbel	0.686	0.503	0.559	0.592	0.996
$index_t$	clayton	0.362	0.248	0.283	0.016	0.99
	frank	0.831	0.632	0.703	0.081	0.991

表 4　趋势下跌阶段 copula 模型拟合情况

	model	ρ	τ	γ	λ_u	λ_t
	gaussian	0.996	0.943	0.958	0.905	**0.999**
rho_t	student	**0.952**	**0.811**	**0.850**	0.691	0.997
&	gumbel	0.872	0.695	0.749	0.766	0.998
$index_t$	clayton	0.767	0.580	0.645	0.037	0.990
	frank	0.980	0.873	0.905	0.230	0.992
	model	ρ	τ	γ	λ_u	λ_t
	gaussian	**0.997**	**0.950**	**0.964**	**0.917**	0.999
sum_t	student	0.949	0.807	0.847	**0.706**	**0.997**
&	gumbel	**0.891**	**0.720**	**0.772**	**0.787**	**0.998**
$index_t$	clayton	**0.781**	**0.594**	**0.660**	**0.038**	0.990
	frank	**0.995**	**0.936**	**0.951**	**0.384**	**0.994**

注：为了便于比较，两个维度中较大的指标加粗标注。

结合表中的数据，进行分析。首先，从上下尾部的极端相关性来看，可以看到下尾相关性很大，但是上尾相关性较小。这是因为波动率的右端极端值与其他数值落差较大且数目较小，导致上尾相关性的降低。但是，在波动率较低的区域，数值分布较为集中，所以造成了较大的上下尾不对称。

对于不同趋势下的模型数据进行对比分析，整体来讲，可以看出，趋势下跌区间的相关性整体高于其余两种波动阶段，表明存在较为显著的行情下跌时，两个维度的解释性都会得到提高。

接着对三个阶段进行内部比较。首先，对于趋势波动阶段，各模型的结论较为一致，即个股波动维度与整个指数波动情况的一致性更高。对于趋势上涨阶段，结论则相反，波动相关性与指数波动性的一致性更好，可以认为波动一致性的变动是指数波动更好的解释因素，这与在前一部分描述性统计中的观察结论是一致的。最后，对于趋势下跌阶段，各模型的结论出现了不一致，且两个因素与整体波动的相关系数差距较小，我们认为这表明两个因素均存在较大的一致性，即在指数趋势下跌阶段，指数波动的两个维度均随着指数波动自身的增长出现了显著增长，并共同导致了指数波动的大幅增加。

在该部分，文章使用了 copula 模型对指数波动的两个解释维度进行了相关性探究。根据拟合得出的相关性的大小，得出结论：

在指数呈现趋势波动的时候，指数波动大小变动的主要成因为个股拉动；在指数波动上涨阶段，指数波动大小变动的主要因素则是个股波动的相关性变动导致；在指数波动下跌阶段，个股行为和成分股波动相关性维度都是重要的成因。

5　波动率拟合和预测模型

在对拆解出的变量进行了性质分析之后，文章在本部分将试图将其应用在构建指数波动预测领域。在传统的波动率预测模型中，最广为使用的是 GARCH 模型以及 SV 模型，但是在低频领域得到大家广泛认可的这两个模型在高频数据建模领域存在一定的局限性。Andersen 等人(2003)[5] 发现对已实现波动率取对数后，近似符合正态分布，结合其长记忆性特征提出 ARFIMA-Ln(RV)模型；Crosi(2009)[8] 基于异质市场理论，提出 HAR-RV 模型，通过不同时间跨度的已实现波动率的线性拟合来刻画市场波动信息。目前来讲，对于高频波动率的探究大部分都是以 ARFIMA 以及 HAR-RV 两个模型为基础进行改进和拓展。

由于相比较 ARFIMA 模型，HAR-RV 模型有更好的经济学意义，所以本文选取 HAR-RV 作为波动率预测模型。该模型主要的经济学意义是通过将不同时间内的波动率来互相叠加。

但是过往的经验显示，金融资产的价格并不是一直保持连续的波动，而是存在着跳跃性的价格变动。Merton[9] 的研究指出，金融资产的价格在一般情况下会遵循一个连续的路径运动，但是在异常波动或者重大事件的影响下，会出现不连续的跳跃现象。所以，在传统的价格随机变动过程方程的基础上，需要增加离散跳跃项：

$$dp(t)=\mu(t)dt+\sigma(t)dw+\kappa(t)dq(t) \tag{14}$$

其中，$\mu(t)$ 为漂移项，反映价格的均值过程；$\sigma(t)$ 为价格的瞬间波动，存在左极限且右连续；$w(t)$ 为标准布朗过程；$q(t)$ 是时变强度为 $\lambda(t)$ 的计数过程，满足 $\lambda(t)dt=P(dq(t)=1)$；$\kappa(t)$ 为价格对数序列的跳跃成份，$\kappa(t)=p(t)-p(t^-)$ 等式右侧前两项为传统的价格波动描述。对式(14)做二次变差可以求出实际价格波动率[10]：

$$QV_t=\int_{t-1}^{t}\sigma^2(s)ds+\sum_{i\in(t-1,t)}\kappa_{t,i}^2 \tag{15}$$

其中，$\sum_{i\in(t-1,t)}\kappa_{t,i}^2$ 为时间区间内 $q(t)$ 次的跳跃平方和。本文所使用的已实现波动率 RV 即 QV_t 的一致估计量。

Andersen 和 Bollerslev 等人通过将已实现波动率分解为连续样本路径方差和跳跃方差这两个维度，将原先的 HAR-RV 模型进一步改进为 HAR-RV-CJ 模型[10-11]，通过构建已实现双幂次变差来实现跳跃波动的分离，其中连续路径方差 C 为 $\int_{t-1}^{t}\sigma^2(s)ds$，离散路径方差 J 为 $\sum_{i\in(t-1,t)}\kappa_{t,i}^2$。为了进一步提高估计量稳定性，Adersen(2009)[12] 对估计方法提出改进，构造出收敛性和稳定性更好的估计算法，本文也将采用改进方法进行计算和拟合。

同时，根据过往的研究，由于波动率本身极端值落差较大，所以进行取对数操作之后可以获得更好的正态分布，从而 LNHAR-RV-CJ 模型可以获得更好的预测和拟合特性[13-14]。

于是，仿照常规的 LNHAR-RV-CJ 模型，尝试建立高频波动预测模型，将原模型中的日内波动考察单位改为 5 分钟内的波动率，原先的周波动率以日波动率代替，月波动率以周波动率代替。得到的回归模型如下：

$$\begin{aligned}Ln(RV_t^m)=&\alpha+\beta_{m1}\ln(J_{t-1}^m)+\beta_{w1}\ln(J_{t-1}^w)+\\&\beta_{m2}\ln(C_{t-1}^m)+\beta_{d2}\ln(C_{t-1}^d)+\\&\beta_{w2}\ln(C_{t-1}^w)+\varepsilon_t\end{aligned} \tag{16}$$

其中：

RV_t^m 为五分钟已实现波动率；

J_{t-1}^m 为五分钟跳跃路径方差，即前一个五分钟内的跳跃路径方差指标；

$J_{t-1}^d=\frac{1}{48}\sum_{i=1}^{48}J_i^m$ 为日跳跃路径方差，即前一个交易日所有的 48 个五分钟样本区间计算得出的跳跃路径方差指标的均值；

$J_{t-1}^w=\frac{1}{5}\sum_{i=1}^{5}J_{t-i}^d$ 为周跳跃路径方差，即过去五个交易日日跳跃路径方差的均值。

续路径方差各变量定义可类比上述定义。

同时，为了下文的对比工作，同样对原始的 LN-HAR-RV 模型进行回归拟合：

$$\begin{aligned}Ln(RV_t^m)=&\alpha+\beta_{m1}\ln(RV_{t-1}^m)+\beta_{d1}\ln(RV_{t-1}^d)+\\&\beta_{w1}\ln(RV_{t-1}^w)+\varepsilon_t\end{aligned} \tag{17}$$

其中，RV_{t-1}^m 为前一个五分钟已实现波动率；RV_{t-1}^d 为日已实现波动均值；RV_{t-1}^w 为周已实现波动率均值，定义类比上文所述。

如前文所述，HAR-RV-CJ 模型的本质是在原先 HAR-RV 模型的基础上，将波动率拆解为连续路径部分和跳跃路径部分，而本文之前则是将指数波动率拆解为个股行为维度和相关性维度。受此启发，将得出的 rho_t 与 sum_t 指标类比 C_t^m 与 J_t^m，并按照上面的方法计算出日度均值和周度均值，由此得出模型表达式如下：

$$\begin{aligned}Ln(RV_t)=&\alpha+\beta_{m1}\ln(rho_{t-1}^m)+\beta_{d1}\ln(rho_{t-1}^d)+\\&\beta_{w1}\ln(rho_{t-1}^w)+\beta_{m2}\ln(sum_{t-1}^m)+\\&\beta_{d2}\ln(sum_{t-1}^d)+\beta_{w2}\ln(sum_{t-1}^w)\varepsilon_t\end{aligned} \tag{18}$$

其中，rho_{t-1}^m，sum_{t-1}^m 为分钟波动指标；rho_{t-1}^d，sum_{t-1}^d 为日度波动指标均值；rho_{t-1}^w，sum_{t-1}^w 为周度波动指

标均值。

同时，为了更好地阐述模型的解释能力，对模型的样本外预测能力进行考察，而为了测算预测模型的估计效果，引入多个常见的预测效果评估指标：误差均方根（RMSE）、绝对误差平均（MAE）、相对误差绝对值平均（MAPE）以及希尔不等系数（TIC）。分别定义为：

$$RMSE=\sqrt{\frac{1}{n}\sum_{t=T+1}^{T+n}(\hat{y}t-yt)^2} \tag{19}$$

$$MAE=\frac{1}{n}\sum_{t=T+1}^{T+n}|\hat{y}t-yt|^2 \tag{20}$$

$$MAPE=\frac{100}{n}\sum_{t=T+1}^{T+n}\left|\frac{\hat{y}t-yt}{yt}\right| \tag{21}$$

$$TIC=\sqrt{\frac{1}{n}\sum_{t=T+1}^{T+n}(\hat{y}t-yt)^2}\Big/\left(\sqrt{\frac{1}{n}\sum_{t=T+1}^{T+n}(\hat{y}t)^2}+\sqrt{\frac{1}{n}\sum_{t=T+1}^{T+n}(yt)^2}\right) \tag{22}$$

其中 T 为样本容量，n 为样本外预测数目，$\hat{y}$ 为预测值，yt 为真实值。由表达式可以看出，以上四个误差评判指标越小，说明预测效果越好。

表 5　趋势波动阶段模型拟合与预测效果对比

	LNHAR-RV	LNHAR-RV-CJ	LNHAR-RV-RS
α	−2.541***	−1.865***	−0.888***
β_{m1}	0.493***	0.020***	0.421***
β_{d1}	0.177***	0.003	0.173***
β_{w1}	0.151***	0.030***	0.210***
β_{m2}	/	0.488***	0.736***
β_{d2}	/	0.162***	0.169***
β_{w2}	/	0.135***	0.081**
Adjusted R-squared	0.352	0.355	**0.372**
RMSE	0.735	0.720	**0.705**
MAE	0.551	0.532	**0.523**
MAPE	4.133	4.067	**4.002**
TIC	0.028	0.027	**0.027**

由于在上文的探究中，发现在不同的大盘走势情况下，各指标与整体波动的相关程度存在显著差异，所以本部分的拟合和预测会分大盘走势阶段进行。对比模型为 LNHAR-RV 与 LNHAR-RV-CJ，对于趋势波动、趋势上涨和趋势下跌三种阶段都采取将样本数据的 90%用作拟合，将剩下的 10%用作样本外预测。为了便于表达，将构造的式(18)模型简称为 LNHAR-RV-RS 模型。

表 6　趋势上涨阶段模型拟合与预测效果对比

	LNHAR-RV	LNHAR-RV-CJ	LNHAR-RV-RS
α	−1.271***	−0.775***	−0.198
β_{m1}	0.492***	0.020***	0.350***
β_{d1}	0.277***	0.010	0.282***
β_{w1}	0.147***	0.007	0.204***
β_{m2}	/	0.452***	0.726***
β_{d2}	/	0.247***	0.123**
β_{w2}	/	0.184***	0.184***
Adjusted R-squared	0.421	0.422	**0.440**
RMSE	0.822	0.825	**0.806**
MAE	0.651	0.658	**0.630**
MAPE	5.233	5.297	**5.067**
TIC	0.032	0.033	**0.031**

表 7　趋势下跌阶段模型拟合与预测效果对比

	LNHAR-RV	LNHAR-RV-CJ	LNHAR-RV-RS
α	−0.873***	−0.521**	−0.312
β_{m1}	0.829***	0.034***	0.354***
β_{d1}	0.121***	0.028	0.154***
β_{w1}	-0.012	0.042	0.351***
β_{m2}	/	0.673***	0.992***
β_{d2}	/	0.187***	0.226***
β_{w2}	/	−0.057	−0.204***
Adjusted R-squared	0.725	0.758	**0.760**
RMSE	0.728	0.698	**0.668**
MAE	0.566	0.550	**0.534**
MAPE	4.472	4.337	**4.222**
TIC	0.028	0.027	**0.026**

注：表 5-7 中上半部分为拟合结果，下半部分为预测误差度量，其中***，**，* 分别代表拟合系数在 1%、5%以及 10%的水平下显著。对于预测模块，三个模型中表现最好的模型都加粗标注。

结合表 5-7 中的数据，可以看出 LNHAR-RV-RS 模型在三个阶段预测和拟合效果都是最好的。与预期有落差的是，在高频领域，LNHAR-RV-CJ 模型表现不够稳定，在趋势上涨阶段的预测中表现甚至不及传统模型。通过对比各个模型拟合系数的显著性看出，对于 LNHAR-RV-CJ 模型，跳跃路径方差部分除了分钟数据均较为显著之外，日度和周度均值都较为不显著，这也直接影响了该模型的解释和预测能力。LNHAR-RV-RS 模型中两个维度的拆解在回归中均显现出较为理想的显著性和解释能力。

对这一差异是这样理解的:对于波动率数据,其跳跃分量一般出现在较短时间内。长期来讲,冲击效应会很不显著,对于回归以及预测的作用也就相对较小。但是对于该拆解方法,由于个股行为维度和相关性维度为两个对称的维度,所以这样的拆解就避免了其中某个变量长期不显著的问题,构建的模型从而获得了更好的解释和预测能力。

6 结论和展望

本文尝试对指数波动进行拆解。首先,对拆解的两个维度和波动率自身进行了初步的描述性统计分析,进一步使用 copula 模型进行了定量的一致性分析,最后以此为基础进行了波动率解释和预测模型的构建。相关结论表明,这样的拆解不仅可以向我们很好地解释和阐述不同大盘走势下指数波动增加的主要成因,也可以给波动率解释和预测模型的构建带来更多的有效信息,达到更好的拟合与预测效果,这样的结果也给对波动率的研究和预测工作带来了一种崭新的思路。

以本文现有的工作为基础,在后续的探究中,还可以设想依据这样的拆解,对指数的极端波动时间段进行预警或者构建对应的指数买卖策略实现套利。

参考文献:

[1] 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.

[2] BARNDORFF-NIELSEN O E, GRAVERSEN S E, SHEPHARD N. Power variation and stochastic volatility: a review and some new results[J]. Journal of Applied Probability, 2004, 41(1):133-143.

[3] KIM C, MARK P. Realized range-based estimation of integrated variance[J]. Journal of Econometrics, 2007, 141(2):323-349.

[4] 徐正国,张世英. 多维高频数据的"已实现"波动建模研究[J]. 系统工程学报, 2006, 21(1):6-11.

[5] ANDERSENT G, BOLLERSLEV T, DIEBOLD F X, et al. Modeling and forecasting realized volatility [J]. Econometrica, 2003,71(2):579-625.

[6] 魏平,刘海生. Copula 模型在沪深股市相关性研究中的应用[J]. 数理统计与管理, 2010, 29(5):890-898.

[7] 张连增,胡祥. Copula 的参数与半参数估计方法的比较[J]. 统计研究,2014(2):91-95.

[8] CORSI F. A simple approximate long-memory model of realized volatility[J]. Social Science Electronic Publishing, 2009, 7(2):174-196.

[9] MERTON R C. Option pricing when underlying stock returns are discontinuous[J]. Working Papers, 1975, 3(1/2):125-144.

[10] ANDERSEN T G, BOLLERSLEV T, DIEBOLD F X. Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility[J]. Review of Economics & Statistics, 2007, 89(4):701-720.

[11] HUANG X, TAUCHEN G. The relative contribution of jumps to total price variance[J]. Journal of Financial Econometrics, 2005, 3(4):456-499.

[12] ANDERSEN T G, DOBREV D, SCHAUMBURG E. Jump-robust volatility estimation using nearest neighbor truncation[J]. Journal of Econometrics, 2010, 169(15533):75-93.

[13] ANDERSEN T G, BOLLERSLEV T, DIEBOLD F X, et al. The distribution of realized stock return volatility[J]. Journal of Financial Economics, 2001, 61 (1):43-76.

[14] DEO R, HURVICH C, LU Y. Forecasting realized volatility using a long-memory stochastic volatility model: estimation, prediction and seasonal adjustment[J]. Journal of Econometrics, 2003, 131(1/2): 29-58.