

高频数据与大数据方法 在通胀预测中的应用

Application of High Frequency Data and Big Data in Inflation Prediction

熊伟琪 张 鹏

摘 要：在近几年国际国内超预期事件频发的市场环境中，本文从通胀预测的现实需求出发，在建立稳定的单变量模型的基础上，加入高频指标，增加了模型预测的准确性和及时性。同时，在高频指标的选择部分，我们应用大数据降维技术，采用主成分分析方法选取更有价值的预测指标。最后，简要讨论了在疫情扰动、基期轮换、春节效应等特殊情形下通胀模型的预测表现。

关键词：通胀预测 大数据方法 特征缩减技术 主成分分析

Abstract: This paper seeks to address the demand for inflation forecasting. On the basis of establishing a stable univariate model, it adds high-frequency indicators, which increases the accuracy and timeliness of the model's forecasts. At the same time, it selects the dimension reduction technology of big data as high-frequency indicators and adopts the principal component analysis method to select more valuable prediction indicators. Lastly, it discusses the prediction performance of inflation models under special circumstances such as an epidemic disturbance, base period rotation and the Spring Festival effect.

Keywords: Inflation Forecast, Big Data Method, Feature Reduction Technology, Principal Component Analysis

熊伟琪、张鹏，中信银行总行金融市场部。本文系个人观点，与所在机构无关。



引言

在主要宏观经济指标中，国家统计局编制、统计并发布的消费者价格指数（CPI）一直受到社会各界的广泛关注。作为反映全国消费价格走势和通货膨胀（紧缩）程度的月度指标，CPI既是市场经济的基本变量，又是宏观调控的主要目标。CPI数值的高低变化，提供着丰富的、关于消费市场供求关系、货币政策调控方向以及金融市场利率变动的关键信息。因此，对于金融市场的参与者来说，对CPI变化趋势的预测和总体判断具有重要的实际意义。

本文结合单变量历史均值方法和高频数据建立通胀模型，并报告模型近三年的预测表现。实证结果再次佐证，在多元线性模型框架下，高频数据信息可以保证模型及时性，提高模型准确性。同时，本文采用主成分分析方法，跟踪一系列对CPI有较好预测效果的高频指标，在不同时期选取更有价值的预测指标。最后简要讨论在基期轮换、春节效应这类特殊情形下，对量化模型的调整情况。

常用通胀量化模型简介及预测表现回顾

通胀建模的基础理论与主要目标

宏观指标的建模方法，一般是基于经济金融、统计计量理论而建立的。但具体到模型的选择，则需要就量化建模的目标

来加以具体分析。举例来说，在宏观政策领域，动态随机一般均衡模型（DSGE）综合考虑了政府、企业、居民、银行等多部门的最优化行为，因其理论基础和政策分析上的优势，被经济合作与发展组织（OECD）、国际货币基金组织（IMF）等国际机构广泛使用；在资产定价领域，无套利模型衍生于经典金融理论，假定市场定价可以反映所有公开信息，已成为海内外金融市场资产定价的主流模型；通胀预测的理论基础源于约翰·泰勒根据美国货币政策实践提出的泰勒规则，但在通胀预测研究中发现，单变量模型由于结构简单、数据更新及时，比理论模型更具预测价值。所以，在通胀建模实际中首先要明确建模目标，才能更好地选择适当的模型。

在金融市场交易范畴内，通胀指标建模的意义主要体现在生成具有稳定性、及时性和准确性的预测结果。第一，稳定性指的是模型的预测表现具有可持续性，可以适应国内外超预期事件冲击下的复杂市场环境；第二，及时性指的是在确保指标预测质量的基础上，进一步加强预测时效性，抓住金融资产交易、市场风险应对的时间窗口；第三，准确性指的是在预期驱动下的金融资产价格博弈过程中，通过提高指标预测精度，形成策略以取得更佳资产配置表现。此外，在数据生成机制转换（如基期轮换等）时期，指标波动加大，

也需要预测分析人员进行应对和调整。

值得一提的是,在金融交易实务中,以CPI、采购经理指数(PMI)、社会融资规模等为代表的宏观经济金融指标,其数据的月度发布都会对市场造成一定冲击。原因是“预期差”的产生,即数据真实值发布与发布前的市场预期之间的差异。预期差是交易机会和市场风险的主要来源之一。因此,评价通胀指标预测表现,我们关注模型预测对CPI变动方向的判断能否平衡稳定性、及时性和准确性三个方面的要求。

通胀建模的实证研究与预测表现

通胀预测的建模思路说明

在对中国通胀数据的实证预测研究中,我们采用公开可得的数据完成建模,形成可复制、可验证、可追溯的预测与回测结果。宏观指标预测工作的流程为:数据挖掘(Data)→建模设计(Model)→参数估计(Estimate)→预测表现(Performance)。

首先,从通胀指标定义上来看,我国CPI物价的调查内容包括食品衣着、居住、交通通信、教育文娱、医疗保健等八大类,262个基本分项的商品和服务价

格。其中,食品类基本分项在其中占比约两成,非食品类占八成。因此,一个天然的建模思路是先分别预测CPI分项增速,之后按权重加总,得到CPI总体增速:

$$\hat{Y}_{t+h} = w_t(\text{食品}) \times \hat{Y}_{t+h,W}(\text{食品}) + w_t(\text{非食品}) \times \hat{Y}_{t+h,Y}(\text{非食品}) + \epsilon_t \quad (1)$$

$$1 = w_t(\text{食品}) + w_t(\text{非食品}) \quad (2)$$

式(1)中 $w_t(\text{食品})$ 表示食品项权重, $w_t(\text{非食品})$ 表示非食品项权重。权重的计算方法根据每月官方发布的分项增速确定。以2022年8月的国家统计局发布^①为例,“从同比看,食品价格上涨6.1%……影响CPI上涨约1.38个百分点。”反推得到CPI食品项的权重 $w_t(\text{食品})$ 约为17.8%。非食品项权重 $w_t(\text{非食品})$ 约为82.2%^②。

通胀预测的数理模型基础

在通胀预测的实证研究中发现,理论模型假设严格,不适宜描述实际市场环境。而单变量模型结构简明,往往具有更好的预测表现。历史通胀数据显示,以一年12个月为一个完整周期,月度CPI数据具有明显季节性特征。数据驱动的历史均值模型^③对这类季节性时间序列具有较好预测效果,近月预测中误差更小,模型更

① 国家统计局城市司首席统计师董莉娟解读2022年8月份CPI和PPI数据[EB/OL].(2022-09-09)[2022-11-04].http://www.stats.gov.cn/tjsj/sjjd/202209/t20220909_1888073.html.

② 食品项权重计算公式: $w_t(\text{食品}) = \text{食品涨幅对CPI拉动} / \text{食品涨幅}$ 。非食品项权重: $w_t(\text{非食品}) = 1 - w_t(\text{食品})$ 。

③ 历史均值实质上就是CPI单变量模型的一种特殊情况。



稳定。因此，我们选择单变量模型作为通胀建模的基准模型，其核心公式如下：

$$\hat{Y}_{t+h,Y} = F(Y_{t|t}) + \epsilon_{Y,t} \quad (3)$$

上述等式右边的 $Y_{t|t}$ 表示截至第 t 期的CPI环比数据， $F(\cdot)$ 表示以历史CPI环比数据为自变量的函数。在历史均值模型中，该函数简化为历史同期的数据均值： $F(Y_{t|t}) = Y_t + Y_{t-1 \times 12} + \dots + Y_{t-(L-1) \times 12} + Y_{t-L \times 12}$ ，其中 $Y_{t-L \times 12}$ 表示向历史回溯 L 年的同期环比。等式左边的 $\hat{Y}_{t+h,Y}$ 表示单变量模型对未来第 h 期的指标预测。 $\epsilon_{Y,t}$ 为模型残差项。

$$\hat{Y}_{t+h,W} = c_t + W_t \beta_w + \epsilon_{W,t} \quad (4)$$

式（3）中单变量模型的优势在于保证了模型中枢水平的稳定，对通胀月度变化的方向判断较为准确。但是，物价水平（特别是食品物价）常常受到短期供需错配等因素的冲击。因此我们采用式（4）所示的多元线性模型，在其中纳入体现短期食品价格冲击的高频指标 W_t 。公式（4）中， β_w 为模型参数，表示变量集 W_t 对CPI的边际影响。 c_t 、 $\epsilon_{W,t}$ 分别为模型的常数项和残差项。

在高频指标变量的选择方面，肖争艳等（2022）应用机器学习方法识别我国通胀的主要影响因素，发现食品价格上涨和通胀预期是我国大多数轮次通胀的重要原

因。表1列示在通胀建模预测中跟踪和应用的14个主要高频指标。具体来看，我们在食品和非食品两大主要分项下，各平均选择了7个与CPI序列之间相关性较好，并具有一定预测作用的高频指标。在关注猪肉、蔬菜等高波动指标对CPI食品项影响的同时，兼顾国内、国际能源价格对CPI非食品项的影响。

加入高频数据模型可以提供兼具及时性和准确性的通胀预测。以2022年8月的实际预测情境为例，国家统计局一般在每月上中旬发布上月的CPI数值，在统计局在9月9日公布CPI真实值前，我们可以采用截至8月底的高频数据描述短期物价的变动因素，完成模型的参数估计和数值预测。在保证通胀预测及时性的前提下，改进模型预测的准确性。

评价通胀预测模型的表现

金融市场交易实战范畴内，我们首先要关注的是指标预测在样本外检验（out-of-sample test）中的表现。首先厘清两个基本概念，第一，模型的拟合指的是使用一段历史数据（“样本内数据”）完成模型估计，得到模型参数。第二，模型的预测指的是根据模型估计参数和对解释变量的预测，完成指标预测。模型拟合的数值与样本内数值之差，称为拟合误差^①。在

① 样本内参数优化流程说明：首先设定目标函数，常用目标函数（也称损失函数）一般为拟合误差的一次或二次函数。之后问题转化为求解目标函数最小值，求解在样本内平均拟合误差最小的模型参数。

表 1 CPI 通胀预测模型的参考高频数据集

	指标名称	频率	起始日期	发布时间	数据来源
食品	生猪全国平均出场价	周	2016-06-30	2022-08-29	发改委
	白条猪批发价	周	2005-01-31	2022-08-24	商务部
	猪肉平均批发价	日	2013-11-18	2022-08-26	农业部
	山东蔬菜批发价格指数	日	2013-01-01	2022-08-29	山东省物价局
	带鱼大宗价	周	2007-04-30	2022-08-26	农业部
	草鱼大宗价	周	2006-05-31	2022-08-26	农业部
	黄花鱼大宗价	周	2005-07-31	2022-08-26	农业部
	鸡蛋平均价	周	2013-01-01	2022-08-29	农业部
	面粉平均价	周	2008-01-04	2022-08-26	农业部
非食品	螺纹钢期货价格	日	2009-03-27	2022-08-29	上海期货交易所
	动力煤期货价格	日	2013-09-26	2022-08-29	郑州商品交易所
	焦煤期货价格	日	2013-03-22	2022-08-29	大连商品交易所
	生产资料价格指数	周	2007-08-17	2022-08-26	商务部
	南华工业品指数	日	2004-06-01	2022-08-29	南华期货
	南方八省电厂日耗	周	2022-01-18	2022-08-22	Wind新闻整理
	唐山高炉开工率	周	2009-01-09	2022-08-26	Wind新闻整理
	汽车轮胎开工率	周	2013-07-18	2022-08-25	中国橡胶信息贸易网
	江浙织机开工率	日	2013-09-04	2022-08-25	隆众化工

官方数据公布后，模型预测的数值与实际数值之差，称为预测误差。从预测的流程看，包括历史样本选择、模型估计、参数优化、模型预测等部分。样本外检验就是基于预测误差（而非拟合误差）的检验过程，是以最小化样本外预测误差为目标的模型参数优化。

模型的样本内与样本外检验，对于宏观指标建模都十分关键。样本内检验显示

模型对历史一段时间内的数值拟合效果，模型参数可以用于解释变量与指标之间的经济学联系；样本外检验模拟模型在未来一段时间内，指标预测的表现，筛选出简单灵活的模型。对通胀预测的实证研究发现，具有更好样本内拟合表现的模型，不一定在样本外预测中表现更佳^①。金融从业人员需要结合当前经济周期和指标数据特征建模，并在真实环境下回溯检验模型

① 更多讨论请参考Swanson & Xiong (2018)。



预测表现。

我们考虑CPI季节性特征，建立了单变量模型，并应用于2020年8月至2021年10月的通胀预测。同时，通过逐月更新高频数据集，迭代多元线性模型参数，建立了单变量加高频数据的多元预测模型，完

成2021年11月至2022年8月的通胀预测。

我们比较量化模型预测值与官方公布的CPI真实值之间的预测误差和预测方向准确性，形成样本外的预测检验结果（表2）。通胀模型的实际预测表现（表3）指向以下几点主要结论。

表2 量化模型预测表现（2020年8月—2022年8月）

日期		量化预测	官方公布	预测表现	
年	月			方向判断	预测误差
2020	8	2.41	2.40	—	0.01
	9	2.10	1.70	✓	0.40
	10	0.90	0.50	✓	0.40
	11	(0.10)	(0.50)	✓	0.40
	12	(0.40)	0.20	低估	(0.60)
2021	1	(0.29)	(0.30)	✓	0.01
	2	(1.04)	(0.20)	低估	(0.84)
	3	0.51	0.40	✓	0.11
	4	1.24	0.90	✓	0.34
	5	1.53	1.30	✓	0.23
	6	1.11	1.10	✓	0.01
	7	0.79	1.00	✓	(0.21)
	8	1.09	0.80	高估	0.29
	9	0.84	0.70	高估	0.14
	10	1.50	1.50	✓	0.00
	11	2.67	2.30	✓	0.37
	12	1.30	1.50	✓	(0.20)
2022	1	1.10	0.90	✓	0.20
	2	1.00	0.90	高估	0.10
	3	1.00	1.50	✓	(0.50)
	4	1.70	2.10	✓	(0.40)
	5	2.52	2.10	高估	0.42
	6	2.36	2.50	✓	(0.14)
	7	2.80	2.70	✓	0.10
	8	2.70	2.50	✓	0.20

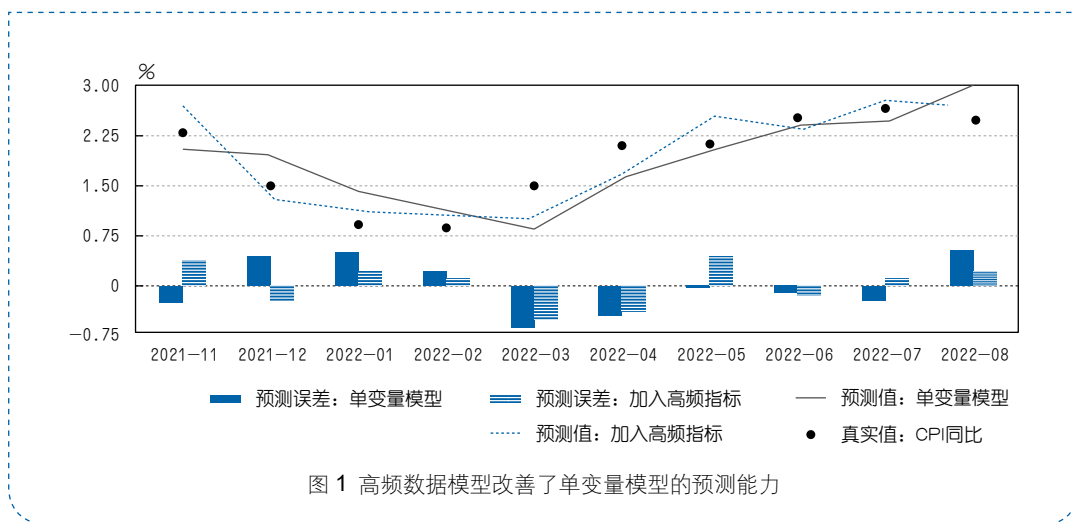
一是在后疫情环境下，数据驱动的量化模型仍具较好预测表现。我们采用季节性模型的思想，基于CPI环比历史均值的预测方法在2020年8月至2021年10月的15次预测中，11次准确预测了CPI同比的变动方向，准确率达到73%；平均预测误差（绝对值）仅为0.27，预测精度达到70.4%^①。

二是在新一轮猪周期启动时期，在量

化预测模型中纳入高频数据，可以进一步改善模型的预测效果。具体来看，我们在CPI季节性特征的基础上，综合考虑具有较好预测效果的预测指标，包括猪肉、鲜菜、粮油、PMI、高炉开工等高频数据并形成预测。在2021年11月至2022年8月的10次预测中，8次准确预测了CPI同比的变动方向，准确率提高至80%；平均预测误差（绝对值）为0.26，预测精度提高至86.2%。

表3 量化模型预测表现的数值分析（2020年8月—2022年8月）

	平均预测误差	CPI同比平均值	预测偏离度	预测精度
模型：历史均值（2020-08至2021-10）	0.27	0.90	29.6%	70.4%
模型：高频数据（2021-11至2022-08）	0.26	1.90	13.8%	86.2%
全区间（2020-08至2022-08）	0.26	1.30	20.4%	79.6%



① 预测精度为：1 - （平均预测误差绝对值/平均CPI同比绝对值），衡量模型在一段时间内预测值的准确/偏离程度。



CPI预测指标集的跟踪与分析：特征缩减技术的应用

综上所述，数据驱动的量化模型预测表现依赖于对解释指标的选择。在实践中，分析人员需要根据预测表现判断指标数据集的有效性，纳入对CPI具有较好指示意义的指标，剔除不再适用的指标。随着大数据技术的发展，研究人员可以使用更新频率更高的日度、周度、旬度发布的数据完成对宏观指标的预测。对于大数据资源的处理和加工也变得更加重要。机器学习中的特征缩减技术（如PCA、Lasso、Elastic Net等）就是一类经典的大数据方法。特征缩减技术可以实现从预测指标的大数据集中去粗取精，提取对预测最有帮助的信息，剔除对预测相关程度低或重复的信息。

在多元统计分析应用中，主成分分析有几方面优势：一是它实现了对解释变量 W_t 数据集的降维，一定程度上避免模型过拟合风险；二是从原始数据集中提取的主成分 $F_t = \{F_{1,t}, F_{2,t}, \dots, F_{14,t}\}$ 能最大程度反映原始数据集 W_t 的变动特征信息^①。

主成分分析通过线性映射 Λ 将原始变量集 W_t 的信息转换为线性不相关的新变

量，即主成分 F_t 。公式（5）中， $\Lambda_{(14 \times 14)}$ 为线性组合的系数（loadings）。式（6）描述主成分分析的线性求解过程，即求解样本数据^②协方差矩阵的最大特征值和特征向量。按照系数加权的线性组合 $W_t \Lambda$ 包含最多的包含原始数据集波动的信息，并按对原始数据集总体方差的贡献从大到小排列^③。

$$W_{t,(N \times 14)} = F_{t,(N \times 14)} \Lambda'_{(14 \times 14)} + e_{t,(N \times 14)} \quad (5)$$

$$\max \text{tr}(\Lambda' W' W \Lambda) \quad (6)$$

$$\text{s.t. } \frac{\Lambda' \Lambda}{N} = I$$

我们应用主成分分析方法，检验2020年8月—2022年8月14个主要高频数据对指标预测主成分的边际贡献，据此选出更有价值的预测指标。主要研究结论分为两部分，一是在后疫情时期，该方法实现了数据信息集由十四维到三维的降维过程（图3右下）。前三个主成分指标可以分别解释原始数据集36.5%、18.2%和16.1%的总体变动，累计解释70.8%的总体变动（表4）。

$$W_t(14 \times 1) \xrightarrow{\text{主成分分析 (PCA)}} F_t(3 \times 1)$$

① 此外，主成分分析通过正交变换，解决由于指标间存在相关性，导致模型估计出现多重共线性的问题。

② 在应用中，首先将样本数据标准化，排除原始变量量纲对主成分分析的影响。

③ 式（5）中， $\text{tr}(\cdot)$ 代表矩阵的迹，也是矩阵对角线上的元素之和。关于主成分分析方法的更多技术细节可参考何晓群（2019）。

表 4 前三个主成分对原始数据方差的解释比例
(2020 年 8 月—2022 年 8 月)

	总体方差的解释比例	累计方差解释比例
第一个主成分	36.5%	36.5%
第二个主成分	18.2%	54.7%
第三个主成分	16.1%	70.8%

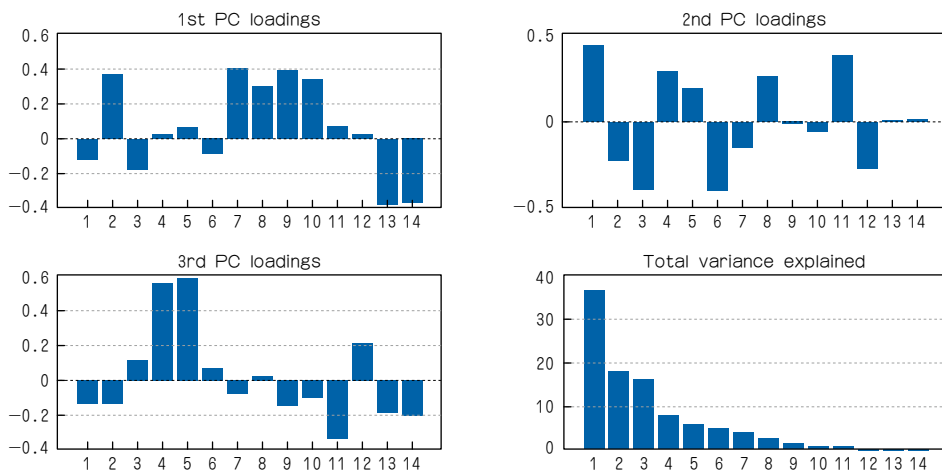
二是跟踪分析对三个主成分边际贡献较大的高频指标，我们总结几点主要特征如下。以权重系数的绝对数值计，在第一个主成分指标中，有明显拉动作用的指标有动力煤价格、生产资料价格、白条猪和生猪价格，集中反映的是维持居民基本生产生活的必选消费成本。在第二个主成分指标中，有明显拉

动作用的指标有涤纶库存、大带鱼和蔬菜价格，主要体现的是居民的可选消费成本。在第三个主成分指标中，拉动效果显著的指标有汽车半钢胎、汽车全钢胎开工率，主要与居民生活消费中的交通运输和服务项目成本相关。

特殊情况下的调整

2022年以来，伴随疫情冲击、俄乌冲突等超预期事件冲击，通胀量化模型预测表现的稳定性也面临着更严峻的考验。以下我们简要回顾CPI指标预测在特殊月份的模型预测值调整情况。

疫情冲击：对通胀预测的实证研究



注：左上、右上、左下的三张图显示 14 个高频数据指标的第一、二、三个主成分的权重系数。权重系数的绝对值越高，该因素的变动对主成分会产生越大的影响。右下图显示 14 个主成分指标对大数据集总体方差的解释比例。

图 2 对通胀预测高频数据集的主成分分析结果



发现，量化模型在2022年3、4月有明显低估，但5月却出现了高估。原因在于，2022年3月在深圳、上海、北京等多地出现散发疫情，运输物流受到一定限制，造成消费端价格阶段性上行。进入5月后，新一轮猪周期启动、鲜菜大量上市，缓解了前期的涨价压力。

基期轮换：被预测数据的生成机制变化，往往会影响量化模型的预测表现。我们对2021年的最近的一次基期轮换的实证研究结果显示，由于统计篮子内指标品类型的增删和权重的调整，通胀基期轮换时点前后，模型的预测误差将出现一定波动，同时高频数据的预测作用也有所减弱。

春节效应：历史经验显示，受居民送礼囤货等需求拉动，每年春节前CPI环比趋于上涨并在节后回落，需进行相应调整。此外，由于不同年份春节的公历时间错位，若两年间的春节日期落在不同月份，对应月份的同比数值波动将有所加大。

文章主要结论

本文报告常用通胀量化模型在疫情后三年（2020—2022年）时期的预测表现。实证研究结果再次佐证，在多元线性模型中纳入高频预测指标，可以改进模型对于通胀变动方向的判断，进一步提高预测精度。此外，应用特征缩减技术中的主成分分析方法，在通胀预测中辅助定位预测效果较好的指标。最后，讨论在疫情扰动、基期轮换、春节效应等特殊情况下对量化模型相应的调整情况。[\[N\]](#)

学术编辑：曾一已



参考文献：

- [1]何晓群.多元统计分析(第五版)[M].北京:中国人民大学出版社,2019.
- [2]肖争艳,陈衍,陈小亮,陈彦斌.通货膨胀影响因素识别——基于机器学习方法的再检验[J].统计研究,2022,39(06):132-147.
- [3]熊伟琪,张鹏.穿越周期的通胀预测:季节性模型及高频数据的适用性[J].金融市场研究,2021(06):90-98.
- [4]Norman R Swanson,Wei qi Xiong,Xi ye Yang.Predicting interest rates using shrinkage methods, real-time diffusion indexes, and model combinations[J].Journal of Applied Econometrics,2020,vol. 35(5):587-613.
- [5]Norman R Swanson,Wei qi Xiong.Big data analytics in economics: What have we learned so far, and where should we go from here?[J].Canadian Journal of Economics,2018,vol. 51(3):695-746.