基于宏观大数据的 GDP 即时预测

易艳萍 黄德金 王 熙*

摘 要：本文结合 EM 算法提出了基于 Lasso 方法的混频动态多因子模型；该模型适用于大数据环境下的 GDP 即时预测，并可对预测变动进行信息归因。本文采用该模型并使用大量月频宏观经济变量对我国季度 GDP 同比增速进行了即时预测。实证结果显示：（1）本文模型比传统动态多（单）因子模型和 MIDAS 模型具有更好的样本外预测表现；（2）在 2005—2022 年间，国家财政支出、社会消费品零售总额、工业增加值以及进口总值的同比增速对于 GDP 即时预测的影响最为突出。

关键词：即时预测；大数据分析；动态因子模型
DOI：10.13821/j.cnki.ceq.2024.03.10

一、引言与文献综述

党的十八大以来，以习近平同志为核心的党中央在宏观经济治理领域提出了一系列重大科学判断、政策方针，形成了一系列实践创新成果。2020 年 5 月，《关于新时代加快完善社会主义市场经济体制的意见》指出要“完善宏观经济治理体制”和“进一步提高宏观经济治理能力”。面对中国发展的战略机遇和风险挑战，党的二十大站在建设社会主义现代化强国和国家安全体系的战略高度，对完善社会经济治理体系作出新的部署。在这一背景下，及时、精准把脉宏观经济的整体运行情况，识别宏观经济在运行过程中的各类冲击，对于防范、化解宏观经济系统性风险和实现国家宏观经济治理体系现代化具有极为重要的意义。

自国民经济核算体系建立以来，国内生产总值（GDP）增速一直代表着一国宏观经济的整体发展情况，汇总了国家经济各个不同方面的发展状况，受到了决策监管层和社会经济各部门的持续关注。因此，及时、准确地预测 GDP 增速不但有利于决策监管部门及时发现经济运行中的风险点，提前做好精准性政策应对，从而有效地防范、化解宏观经济系统性风险，强化国家宏观经济治理的前瞻性；还有利于引导实体经济部门及时做好策略性应对，规避宏观风险。

虽然随着我国统计体系的不断完善，规范了各类宏观经济指标的发布流程，但是由于 GDP 是季度性统计指标，并且统计过程复杂，所以其更新频率不仅显著低于其他宏观指标，而且数据公布也存在 2、3 周左右的时滞。因此，为了实时监测 GDP 运行状

* 易艳萍，浙江大学经济学院和金融研究院；黄德金、王熙，北京大学经济学院。通信作者及地址：王熙，北京市海淀区颐和园路 5 号北京大学经济学院，100871；电话：（010）62753635；E-mail：wang.x@pku.edu.cn。本文受到国家自然科学基金面上项目（71973122、72373131）、国家社会科学基金项目（23BJL025）、北京市社会科学基金项目（21JCC082）的资助。本文作者感谢匿名审稿人的珍贵意见和建议，文责自负。

况，宏观经济学家们基于相对高频但有着碎尾[①]（ragged-edge）特性的其他宏观经济指标，通过各类统计工具，挖掘这些指标与 GDP 的内在联系，用以对 GDP 进行预测。随着相关宏观经济指标的不断发布，可以对 GDP 进行实时预测，即 GDP 的即时预测过程。

现有即时预测文献可以分为两大类。一类是基于混频回归类模型（mixed-data sampling（MIDAS）regressions）以及类似方法对 GDP 进行即时预测。譬如，Kuzin et al.（2011）结合了 MIDAS 和 VAR 分析框架即时预测了欧洲 GDP；刘汉和刘金全（2011）基于 MIDAS 模型即时预测了我国 GDP；类似的还有王维国和于杨（2016）以及张劲帆等（2018）。但是，混频类模型无法直接处理尾部不齐整的数据，需要对缺失数据进行填充，而且也无法囊括过多的解释变量（Schorfheide and Song，2015）。因此，在宏观大数据时代，混频类回归不但需要对所使用的宏观经济数据进行取舍，无法利用舍弃序列中新数据所提供的信息，也无法全面识别其他宏观经济变量发布对于 GDP 预测值变动的贡献。

另一类研究则基于混频动态因子类模型对 GDP 进行即时预测。Mariano and Murasawa（2003）通过将最大似然因子分析应用到季度-月度的混频模型中，首次在动态因子模型分析框架中引入了混频分析。Giannone et al.（2008）在动态因子框架下提出了一个解决碎尾数据的方法，他们采用了两步估计法：首先，仅采用可观测数据进行了主成分提取。然后，将卡尔曼平滑（Kalman Smoother）应用到带有缺失数据的数据集上重新进行因子和因子载荷估计，并在此基础上进行 GDP 即时预测。Bańbura and Modugno（2014）在卡尔曼滤波（Kalman Filter）分析框架下对缺失数据进行了遮掩处理，使用 EM 算法[②]基于所有可观测数据对潜因子进行了估计，并即时预测了欧洲 GDP。Bok et al.（2018）则采用同一方法对美国 GDP 进行了即时预测，做类似分析的还有 Bańbura et al.（2013）。王霞等（2021）使用了混频动态单因子模型（假设 1 个共同因子），选取了 6 个宏观经济变量，使用极大似然估计方法得到参数估计，对我国 GDP 进行了即时预测。

在我国，随着与 GDP 相关的宏观指标数量不断上升，对于 GDP 的即时预测也逐步进入大数据时代。虽然诸如 Giannone et al.（2008）和 Bok et al.（2018）等文献所使用的传统的 GDP 即时预测技术，仍然可以处理高维数据，但随着数据维度的提高，模型提取的潜在因子数目较多，存在过度拟合宏观经济数据中噪音的问题，导致模型泛化能力和样本外预测能力不足。与已有文献不同，本文基于高维宏观经济数据，在混频动态因子分析框架中，为了提高模型的样本外预测能力，我们结合存在缺失观测值的卡尔曼平滑[③]和 EM 算法，创新地提出了用 Lasso 方法（Tibshirani，2011 等）构建针对 GDP 增速的预测因子选择模块。本文所使用的 Lasso 方法，会使得 GDP 相对于部分因子的暴

① 在任一宏观经济指标公布时，我们均可以进行 GDP 即时预测。但由于宏观经济指标的公布日不一，因此在进行 GDP 即时预测时，存在部分高频宏观经济指标的尾部数据缺失，即碎尾数据。

② 当不可观测因子维度较高时，最大似然估计方法很难有效估计模型参数，为了解决这一问题，Watson and Engle（1983）在动态因子模型中引入了 EM 算法（expectation maximization algorithm）。

③ 我们在附录Ⅰ.3 详细解释了，当有缺失观测值时，如何计算卡尔曼滤波与卡尔曼平滑。本文的模型和数据均考虑有缺失值的情况，因此在不产生歧义的情况下，我们不再特别强调有缺失值这个前提条件。限于篇幅，附录未在正文列示，感兴趣的读者可在《经济学》（季刊）官网（https：//ceq. ccer. pku. edu. cn）下载。

露为零。这样一方面会降低在卡尔曼平滑中使用过多数据估计多个因子所产生的过拟合误差，提升对共同因子估计的准确度；另一方面，稀疏的预测模型也能提升预测部分的泛化能力，更有效地基于共同因子对 GDP 进行即时样本外预测。[①] 此外，本文还允许动态因子模型中的特质因子存在序列相关性，并对特质因子进行预测，从而在小样本或样本末端横截面不完整的应用中提高了对 GDP 即时预测的精度。更重要的是，在对 GDP 进行即时预测的基础上，我们的方法还可以通过信息分解将 GDP 预测值的时序变化分解到各个宏观经济指标的变动上（Bańbura and Modugno，2014）。不但能让政策制定者可以实时监测我国宏观经济运行的整体状况，还能精准定位当前宏观经济运行风险点，帮助决策者更加有前瞻性地及时针对经济形势作出政策响应，确保宏观经济的平稳运行。据我们所知，本文是首篇在动态混频多因子框架下对我国 GDP 进行即时预测，并使用信息分解定位我国宏观运行风险点的论文。

本文选取了 17 个与 GDP 增长密切相关的宏观经济指标（详情见第三部分的表 1），尽可能全面地涵盖了宏观经济运行中的各方面情况。本文的主要研究发现如下：第一，与直接使用 EM 算法估计的（传统）动态多因子模型相比，本文采用的基于 Lasso 方法的动态因子即时预测模型在预测我国季度 GDP 同比增速方面具有更高的精度。譬如，在对 GDP 同比增速的预测平均绝对误差（MAFE）上，我们的模型在整体区间内为 1.03 个百分点，而传统混频多因子模型平均绝对误差为 1.06 个百分点，单因子预测模型则为 1.40 个百分点。第二，在新冠疫情期间，我国 GDP 增速预测变得更加困难，本文考察的模型预测偏差均大于无疫情时期。但基于 Lasso 技术的即时预测动态因子模型依旧表现最佳，在疫情期间其平均绝对误差为 3.96 个百分点，传统混频多因子模型为 4.12 个百分点，而单因子模型则为 5.86 个百分点。第三，基于本文所提出的即时预测模型，在 2005 年第一季度至 2022 年第一季度期间，我们发现对于我国 GDP 即时预测的影响平均贡献最大的五个宏观经济变量分别是：国家财政支出、发电量、社会消费品零售总额、工业增加值以及进口总值的同比增长。

本文后文的安排如下：第二部分简要介绍了本文所提出的基于 Lasso 方法的即时预测动态因子模型；第三部分描述了数据收集以及预处理过程；第四部分讨论了实证研究发现；第五部分总结。由于篇幅限制，我们将相关技术细节和稳健性实证结果放进本文附录中。

二、研 究 方 法

（一）混频动态多因子模型

在对我国季度 GDP 同比增速进行即时预测时，本文使用了季度-月度混频动态模型。假设 $n=n_M+n_Q$ 个平稳的经济时间序列中，有 n_M 个月度同比增长率序列 y_t^M 和 n_Q 个季度同比增长率序列 $\overline{y}_t^Q$，记为 $y_t=[y_t^{M\prime},\ \overline{y}_t^{Q\prime}]'$，$t=1,\ \cdots,\ T$。与 Mariano and

① 在本文的估计中，Lasso 算法仅仅参与对 GDP 进行预测的因子选择，Lasso 算法会迫使 GDP 对于部分因子的暴露为零。比如当存在 6 个共同因子时，我们使用 6 个因子对其余 16 个宏观指标进行预测，但仅使用这 6 个因子中的某几个对 GDP 增速进行预测，因此，整体模型依旧为 6 因子模型。此外，由于 GDP 对于部分因子的暴露为零，在使用卡尔曼平滑时，会改变对于隐因子的相应估计。

Murasawa（2003）所提出的混频整合类似，我们假设模型是月频的，即时间下标 t 代表月度。我们将季度同比增长率 $\overline{y}_t^Q$ 分解为一系列潜在的月度增长率的加权平均，其具体表达式如下：

$$\overline{y}_t^Q=(1+L+L^2)\,y_t^Q=y_t^Q+y_{t-1}^Q+y_{t-2}^Q, \tag{1}$$

其中 y_t^Q 代表了在月度 t 未被直接观察到的潜变量①。

下文将所有月度序列记为 $y_t^*=[y_t^{M\prime},\ y_t^{Q\prime}]'$，$t=1,\ \cdots,\ T$，包括可直接观测的 y_t^M 和不可观测的 y_t^Q，其中上标 M 和 Q 分别代表这一时间序列指标的可观测频率为月频和季频。与 Bańbura and Modugno（2014）类似，我们假设 y_t^* 中的每个序列都由三个相互无关的部分组成：不可观测的共同因子部分、特质因子部分以及观测误差部分。因此，对于月度序列 y_t^*，存在如下动态多因子模型：

$$y_t^*=\Lambda f_t+\epsilon_t, \tag{2}$$

其中 y_t^* 是经过标准化后均值为零并具有单位方差的平稳 n 维向量过程，f_t 表示平稳的 r 维（隐）共同因子向量，$\epsilon_t=\left[\epsilon_{1,t},\ \epsilon_{2,t},\ \cdots,\ \epsilon_{n,t}\right]'=[\epsilon_t^{M\prime},\ \epsilon_t^{Q\prime}]'$ 表示特质因子与观测误差之和，其中特质因子可能存在时序自相关性。$\Lambda=\begin{bmatrix}\Lambda^M\\ \Lambda^Q\end{bmatrix}$ 为 $n\times r$ 维因子载荷阵，其中 Λ^M，Λ^Q 分别为 $n_M\times r$ 维和 $n_Q\times r$ 维矩阵。因此，$\Lambda f_t=\begin{bmatrix}\Lambda^M f_t\\ \Lambda^Q f_t\end{bmatrix}$ 是各共同因子对于月度序列 y_t^* 的影响的向量形式。此外，我们进一步假设（隐）共同因子向量 f_t 服从 VAR(1)过程：

$$f_t=Af_{t-1}+u_t,\quad u_t\sim\text{i.i.d.}N(0,Q), \tag{3}$$

其中 A 为 $r\times r$ 维的自回归系数，Q 为 u_t 的协方差矩阵。

同时，由于本文允许特质因子存在序列相关性，我们进一步假设各月度序列的特质因子服从 AR(1)过程，且不同序列的特质性因子不相关（如若相关则可进一步提取共同因子，并在状态空间估计过程中将其加入状态向量中）。我们将 ϵ_{it} 分解为特质性因子部分 $\tilde{\epsilon}_{i,t}$ 与观测误差部分 $\xi_{i,t}$，并将特质性因子 $\tilde{\epsilon}_{i,t}$ 的动态过程用如下方程描述：

$$\epsilon_{i,t}=\tilde{\epsilon}_{i,t}+\xi_{i,t},\ \xi_{i,t}\sim\text{i.i.d.}N(0,\kappa), \tag{4}$$

$$\tilde{\epsilon}_{i,t}=\alpha_i\tilde{\epsilon}_{i,t-1}+e_{i,t},\ e_{i,t}\sim\text{i.i.d.}N(0,\sigma_i^2), \tag{5}$$

其中 $\xi_t=[\xi_{1,t},\ \cdots\xi_{n,t}]'=[\xi_t^{M\prime},\ \xi_t^{Q\prime}]'$ 和 $\tilde{\epsilon}_t=[\tilde{\epsilon}_{1,t},\ \cdots,\ \tilde{\epsilon}_{n,t}]'=[\tilde{\epsilon}_t^{M\prime},\ \tilde{\epsilon}_t^{Q\prime}]'$ 不存在横截面相关性，但每个 $\tilde{\epsilon}_{i,t}$ 存在时序相关性。与 Bańbura and Modugno（2014）类似，本文假设观测误差的方差 κ 为一个 10^{-3}。

由于 $y_t^*=[y_t^{M\prime},\ y_t^{Q\prime}]'$ 中的 y_t^Q 不可直接观测，因此式（1）无法直接估计。根据式(1)至式(5)，我们可以得到如下关于 $y_t=[y_t^{M\prime},\ \overline{y}_t^{Q\prime}]'$ 的混频动态多因子模型②：

$$y_t=\tilde{\Lambda}\tilde{f}_t+\xi_t,\quad \xi_t\sim\text{i.i.d.}N(0,\tilde{R}), \tag{6}$$

① 举例说明：如果 t 代表 2019 年 4 月份，$\overline{y}_t^Q$ 则表示 2019 年 2、3 和 4 月份这三个月的“季度”同比增长率，而 y_t^Q，y_{t-1}^Q，y_{t-2}^Q 则分别代表 2019 年 4、3、2 月份的月度潜变量。此例中，$\overline{y}_t^Q$，y_t^Q，y_{t-1}^Q，y_{t-2}^Q 均不可直接观测。如果 t 代表 2019 年 6 月份，$\overline{y}_t^Q$，y_t^Q，y_{t-1}^Q，y_{t-2}^Q 的定义类似，但 $\overline{y}_t^Q$ 可直接观测。

② 推导过程见附录Ⅱ。

$$\widetilde{f}_t=\widetilde{A}\widetilde{f}_{t-1}+\widetilde{u}_t,\ \widetilde{u}_t\sim\text{i.i.d.}N(0,\widetilde{Q}),\tag{7}$$

其中

$$\widetilde{f}_t=\left[f_t',f_{t-1}',f_{t-2}',\widetilde{\epsilon}_t^{M\prime},\widetilde{\epsilon}_t^{Q\prime},\widetilde{\epsilon}_{t-1}^{Q\prime},\widetilde{\epsilon}_{t-2}^{Q\prime}\right]',\tag{8}$$

$$\widetilde{\Lambda}=\begin{bmatrix}\Lambda^M & 0_{nM\times r} & 0_{nM\times r} & I_{nM} & 0_{nM\times nQ} & 0_{nM\times nQ} & 0_{nM\times nQ}\\ \Lambda^Q & \Lambda^Q & \Lambda^Q & 0_{nQ\times nM} & I_{nQ} & I_{nQ} & I_{nQ}\end{bmatrix},\tag{9}$$

此处 $0_{nM\times r}$ 为 $n_M\times r$ 维的零矩阵，I_{nM} 为 $n_M\times n_M$ 维的单位矩阵（其他零矩阵和单位矩阵的定义类似）。$\widetilde{A}$，$\widetilde{u}_t$，以及 $\widetilde{R}$，$\widetilde{Q}$ 的表达式请参见附录Ⅱ。

我们将 GDP 增速作为观测向量 y_t 的最后一个元素，记为 $y_{n,t}$。由式（6）可知，季度 GDP 增速的观测方程可以表示为：

$$y_{n,t}=\widetilde{\Lambda}_n\widetilde{f}_t+\xi_{n,t},\tag{10}$$

其中 $\widetilde{f}_t$ 包含了共同因子和特质因子，$\widetilde{\Lambda}_n$ 是式（9）中 $\widetilde{\Lambda}$ 的最后一行，为 GDP 增速的因子载荷向量。值得再次强调的是，我们模型的频率是月频的，甚至是观测日频的，作为季频数据的 GDP 增速并不是在所有月度（日期）t 都可以直接观测的。接下来，我们简要介绍如何在 EM 算法框架内处理观测值 $y_{n,t}$ 出现周期性缺失的情况。

（二）EM 算法与压缩估计

由于不可观测因子 $\widetilde{f}_t$ 的维度较高，最大似然估计方法可行性低，因此本文采用了 EM 算法。y_t 和 $\widetilde{f}_t$ 的对数联合似然函数可以表达为 $l(Y,F;\theta)$，其中 $Y=[y_1,\cdots,y_T]$，$F=[\widetilde{f}_1,\cdots,\widetilde{f}_T]$，$\theta=\{\widetilde{\Lambda},\widetilde{A},\widetilde{R},\widetilde{Q}\}$ 为待估计参数。EM 算法的核心逻辑是通过递归迭代不断改进当前参数估计，最终找到最优参数估计。EM 算法的每一次迭代优化都分两步：在第一步（期望（E）步骤）中，已知上一次迭代中对参数 θ 的估计值，通过卡尔曼平滑计算 $l(Y,F;\theta)$ 的期望值（针对隐因子 F 求期望）；在第二步（最大化（M）步骤）中，基于 E 步骤所得到的 $l(Y,F;\theta)$ 期望值，寻找使其最大化的参数 θ，得到此次迭代的参数估计。EM 算法不断重复以上两个步骤，直至参数 θ 的估计值收敛（详情参见附录Ⅱ）。

在混频动态因子模型中，季度序列 $\overline{y_t^Q}$ 被视为具有周期性缺失值，且是月度隐变量 y_t^Q 的加权和，因此，在大部分观测方程中将会出现观测值缺失的情况。为了在存在缺失值情况下继续使用 EM 算法，本文引入 n 维对角矩阵 W_t 作为掩码矩阵[①]（mask matrix），其中当 $y_{i,t}$ 可观测时，W_t 的第 i 个对角元素等于 1；当 $y_{i,t}$ 缺失时，W_t 的第 i 个对角元素等于 0。我们将 y_t 分解为可观测部分和不可观测部分：

$$y_t=W_ty_t+(I-W_t)y_t.\tag{11}$$

由 Bańbura and Modugno（2014）与本文附录Ⅱ可知，EM 算法的第 $j+1$ 次迭代中，$\widetilde{\Lambda}$ 的估计量 $\widetilde{\Lambda}(j+1)$ 可以写作：

① 掩码矩阵这一术语源自自然语言处理（NLP）文献，但使用方式与我们并不完全相同，如在 Transformer 模型（Vaswani et al.，2017）中，掩码矩阵可以防止在翻译的过程中模型使用后续信息。

$$vec(\widetilde{\Lambda}(j+1)) = \Big(\sum_{t=1}^{T} \mathbb{E}_{\theta(j)}[\widetilde{f}_t \widetilde{f}_t' \mid \Omega_T] \otimes W_t\Big)^{-1} vec\Big(\sum_{t=1}^{T} W_t y_t \mathbb{E}_{\theta(j)}[\widetilde{f}_t' \mid \Omega_T]\Big), \tag{12}$$

此处 Ω_T 代表了在时间 T 能观察到的所有数据，$\theta(j)$表示在第 j 次迭代中对参数 θ 的估计值，$\otimes$ 代表克罗内克积（Kronecker product），$\mathbb{E}_{\theta(j)}[* \mid \Omega_T]$是基于参数估计值$\theta(j)$和信息集$\Omega_T$ 所计算的条件期望值（卡尔曼平滑）。对$\mathbb{E}_{\theta(j)}[\widetilde{f}_t \widetilde{f}_t' \mid \Omega_T]$、$\mathbb{E}_{\theta(j)}[\widetilde{f}_t' \mid \Omega_T]$等的计算（详情参见附录Ⅰ），在 EM 算法的 E 步骤中完成。记 $T^{obs} = \{t: y_{n,t}$可以直接观测$\}$，式（12）中对 GDP 增速的因子载荷向量$\widetilde{\Lambda}_n$ 的估计为：

$$\widetilde{\Lambda}_n'(j+1) = \Big(\sum\nolimits_{t\in T^{obs}} \mathbb{E}_{\theta(j)}[\widetilde{f}_t \widetilde{f}_t' \mid \Omega_T]\Big)^{-1}\Big(\sum\nolimits_{t\in T^{obs}} \mathbb{E}_{\theta(j)}[\widetilde{f}_t \mid \Omega_T] y_{n,t}\Big). \tag{13}$$

定理 1 GDP 增速作为观测向量的最后一个元素 $y_{n,t}$，式（12）和式（13）中对 GDP 增速的因子载荷向量$\widetilde{\Lambda}_n$ 的估计，等价于如下 OLS 回归所得到的参数估计：

$$y_{n,t} = \widetilde{\Lambda}_n\Big(\sum\nolimits_{t\in T^{obs}} \mathbb{E}_{\theta(j)}[\widetilde{f}_t \widetilde{f}_t' \mid \Omega_T]\Big)\Big(\sum\nolimits_{t\in T^{obs}} \mathbb{E}_{\theta(j)}[\widetilde{f}_t \mid \Omega_T]\mathbb{E}_{\theta(j)}[\widetilde{f}_t' \mid \Omega_T]\Big)^{-1}$$
$$\times \mathbb{E}_{\theta(j)}[\widetilde{f}_t \mid \Omega_T] + \eta_t, t \in T^{obs}. \tag{14}$$

由定理 1 可知，对于$\widetilde{\Lambda}_n$ 的估计，EM 算法每次迭代的 M 步骤等同于做 OLS 回归，这使得我们能将 Lasso 方法与 EM 算法结合起来，进一步提高 GDP 即时预测的精度。

定理 2（OLS post-Lasso estimator）对 GDP 增速的因子载荷向量$\widetilde{\Lambda}_n$ 的估计，即式（14）的回归，我们采用 Lasso 的方法对因子变量进行选择，并且在完成因子筛选后进行 OLS 回归估计因子载荷。如果式（6）至式（9）包含了真实的数据生成过程，则估计量是一致的，并可提高 GDP 即时预测的精度。

由于 Belloni and Chernozhukov（2013）指出 OLS post-Lasso 估计方法相较于 Lasso 回归，有更快的收敛速度以及更小的偏差，因此本文采用了前者。

由于本文对因子载荷矩阵 $\widetilde{\Lambda}$ 进行了参数约束，因此，本文采用 Bork（2009）和 Bork et al.（2009）在 EM 算法中施加参数限制的方式：$H_\Lambda vec(\widetilde{\Lambda}) = k_\Lambda$，其中 H_Λ 是 $q\times n\tilde{r}$ 维矩阵，$\tilde{r}$ 是$\widetilde{f}_t$的维度，k_Λ 是 $q\times 1$ 维向量。记$\widetilde{D} = \sum_{t=1}^{T} \mathbb{E}_{\theta(j)}[\widetilde{f}_t \widetilde{f}_t' \mid \Omega_T] \otimes [W_t \widetilde{R}(j)^{-1} W_t]$，$\widetilde{\Lambda}$ 的受约束估计量可以表达为：

$$vec(\widetilde{\Lambda}_r(j+1)) = vec(\widetilde{\Lambda}_u(j+1)) + \widetilde{D}^{-1} H_\Lambda' (H_\Lambda \widetilde{D}^{-1} H_\Lambda')^{-1}(k_\Lambda - H_\Lambda vec(\widetilde{\Lambda}_u(j+1))), \tag{15}$$

其中$\widetilde{\Lambda}_u(j+1)$（由式（12）给出）表示不受约束的参数估计值，$\widetilde{R}(j)$表示在第 j 次迭代中对协方差矩阵 $\widetilde{R}$ 的估计值。关于在 EM 算法中施加参数限制的技术细节（包括式（15）的推导），以及如何在受约束回归中做压缩估计，可参考附录Ⅱ.4 和附录Ⅱ.5。同时，EM 算法的第 $j+1$ 次迭代中，转移方程中的 $\widetilde{A}$ 的估计量 $\widetilde{A}(j+1)$可以写作：

$$\widetilde{A}(j+1) = \Big(\sum_{t=1}^{T} \mathbb{E}_{\theta(j)}[\widetilde{f}_t \widetilde{f}_{t-1}' \mid \Omega_T]\Big)\Big(\sum_{t=1}^{T} \mathbb{E}_{\theta(j)}[\widetilde{f}_{t-1} \widetilde{f}_{t-1}' \mid \Omega_T]\Big)^{-1}. \tag{16}$$

我们也可以对 $\widetilde{A}$进行参数约束，然后用类似式（15）的方法估计 $\widetilde{A}$。类似于定理 1，我

们可以证明 $\widetilde{A}(j+1)$、$\widetilde{A}(j+1)$都可以通过多元线性回归得到。因此，观测方程中的因子载荷矩阵和转移方程中的系数矩阵均可通过 OLS post-Lasso 方法估计。由于本文仅聚焦于 GDP 预测，我们采用了最为保守的因子筛选方式：仅对 GDP 观测方程使用变量选择。需要强调的是，本文分析仅考虑在共同因子个数 r 确定时，对 GDP 观测方程进行变量选择。在后文实证中我们会通过基于验证期样本，对因子个数进行实时选择。

（三）估计流程总结

以混频动态（$r=$）3 因子模型为例，我们简要总结针对本文模型的估计流程（当 $r=$任意正整数，可以类似操作）。这里我们假定除 GDP 增速外，其他所有宏观观测变量由 3 个共同因子驱动，因此，我们称这一模型为 3 因子模型。GDP 增速仅对其中部分因子存在非零因子暴露，我们通过 Lasso 方法对 GDP 观测方程进行变量选择，我们称此模型为 3 因子压缩模型。

估计流程总结如下：

步骤 1：使用样条法对缺失月频数据进行填充，并对月频数据使用主成分分析(PCA)，提取 3 个主成分因子，并基于这 3 个主成分，使用式（2）、式（3）分别进行回归，得到参数的初始值 $\theta(1)$。

步骤 2（EM 算法-E 步骤）：给定参数 $\theta(j)$，我们基于卡尔曼平滑计算隐因子时序特征估计：$\sum_{t=1}^{T}\mathbb{E}_{\theta(j)}[\widetilde{f}_t\,\widetilde{f}'_{t-1}\mid\Omega_T]$，$\sum_{t=1}^{T}\mathbb{E}_{\theta(j)}[\widetilde{f}_{t-1}\,\widetilde{f}'_{t-1}\mid\Omega_T]$，$\mathbb{E}_{\theta(j)}[\widetilde{f}_t\mid\Omega_T]$（参考附录Ⅰ式（Ⅰ7）—（Ⅰ10）和（Ⅰ13）—（Ⅰ16））。

步骤 3（EM 算法-M 步骤）：给定隐因子时序特征估计，我们估计参数 $\widetilde{A}$（式（16））、$\widetilde{\Lambda}$（式（12））、$\widetilde{R}$（附录Ⅱ式（Ⅱ16））以及 $\widetilde{Q}$（附录Ⅱ式（Ⅱ6））。对于 $\widetilde{\Lambda}$ 中 GDP 的因子载荷向量$\widetilde{\Lambda}_n$，我们则对式（14）进行 OLS post-Lasso 估计。如果考虑对 $\widetilde{\Lambda}$ 的线形约束，我们则采用式（15）估计。如果考虑对 GDP 因子载荷向量$\widetilde{\Lambda}_n$ 的线形约束，我们则对附录Ⅱ式（Ⅱ32）进行 OLS post-Lasso 估计。更新参数 $\theta(j+1)=\{\widetilde{\Lambda}(j+1),\widetilde{A}(j+1),\widetilde{R}(j+1),\widetilde{Q}(j+1)\}$。

步骤 4：重复步骤 2、步骤 3 直至参数估计收敛。

步骤 5：给定以上所估计的参数 θ，我们用卡尔曼平滑计算当季三个月的隐共同因子和特质因子的期望（附录Ⅰ式（Ⅰ7）—（Ⅰ10）和（Ⅰ13）—（Ⅰ16））。

步骤 6（预测）：预测当季 GDP 等于$\widetilde{\Lambda}_n$ 的估计值乘以当季三个月隐共同因子和特质因子的期望（式（10）或附录Ⅱ式（Ⅱ29））。

在 GDP 增速的即时预测过程中，每当新的宏观经济数据发布后，我们可以使用更新的信息集重新计算步骤 5 和步骤 6，可得到本季度 GDP 增速的最新预测。此外，我们可以对 GDP 即时预测变动进行信息分解，感兴趣的读者可参阅 Bańbura and Modugno (2014) 和本文附录Ⅲ。

三、数据收集以及预处理

本文一共选取了 17 个宏观经济指标，其中 GDP 和城镇居民可支配收入为季频指

标，其他宏观经济指标均为月频。为了兼顾数据可获得性以及数据涉及角度多样性，本文选取了价格型指标、内需型指标、外贸型指标以及政策型指标，力图数据能较为全面地反映我国整体经济形势。本文的数据样本时间跨度为 1995 年 1 月至 2022 年 4 月，主要数据来源为国家统计局官方网站。为了更符合我国的统计指标发布习惯，我们以各个宏观经济指标的时序同比增长率为研究标的。此外，由于春节因素及统计报表制度，国家统计局在部分年份将 1、2 月数据合并，并在之后仅公布在 2 月的累计值与累计同比增长率，因此，本文将部分年份的 1 月份数据作为缺失值处理，将每年前两个月的累计同比增长率作为 2 月的同比增长率。此外，由于国家统计局官方网站仅在 2013 年第一季度后开始公布城镇居民收入数据，我们借助中经网统计数据库完善了此变量在 2013 年以前的时序数据，并且对 2006 年以前月度公布的城镇居民可支配收入数据进行加权并计算同比增速。我们将所使用的宏观经济指标汇总至表 1 中，后文之中默认单位为百分点。

由于以上宏观序列数据在每个月的发布时间并不相同，我们根据《国家统计局主要经济统计信息发布日程表》《中国海关统计数据公布时间表》，财政部官方网站，以及新闻搜索确定了以上宏观经济指标在 1995 年 1 月至 2022 年 3 月期间的实际发布日。对于每一个时间点，收集了在该时间点上所有可获得的实时数据。因此，在一个季度中，随着进行即时预测的时间点不同，我们所使用的信息集也会有所不同，同时样本尾部也会由于变量非同步发布而出现不平整的情况，即碎尾数据。我们所使用的卡尔曼平滑可以很好地处理这一由于数据发布日不一所导致的数据缺失问题。

表 1 宏观经济指标

指标名称	指标频率
居民消费价格指数	月度
工业生产者出厂价格指数	月度
工业增加值	月度
固定资产投资额累计	月度
社会消费品零售总额	月度
进口总值	月度
出口总值	月度
国家财政收入	月度
国家财政支出	月度
房地产投资	月度
商品房销售额	月度
M2 供应量	月度
M1 供应量	月度
M0 供应量	月度
发电量	月度
国内生产总值	季度
城镇居民人均可支配收入	季度

此外，我国部分指标的统计起始时间较晚。在我们的样本中，房地产与货币供应量相关指标统计起始时间最晚，从 2000 年起才开始统计。工业增加值、生产者出厂价格指数、进出口总值分别从 1998 年 7 月、1996 年 10 月、1994 年 8 月起开始统计。因此，为了保证模型中各变量都有充分的数据，我们从 2005 年 1 月开始对 GDP 季度同比增速进行样本外实时预测。随着时间的演进和新数据的发布，EM 和 Lasso 算法将会结合新数据重新估计混频多因子模型，并更新对 GDP 的样本外预测。因此我们在一个季度中的多个时间点会得到当季 GDP 不同的即时预测值，我们将 GDP 即时预测值的变化进行统计归因，分解至不同宏观指标的变动上。

四、实 证 结 果

在本部分中我们将从不同角度研究 GDP 即时预测模型的表现。我们采用了两类基准模型：基准模型 1，采用了上文所提及的动态因子模型，但是对于各类参数不施加任何变量选择结构，直接结合 EM 算法基于上文提及的 17 个宏观经济变量对 GDP 进行即时预测；基准模型 2，则是采用了 Bańbura and Modugno（2014）中对于状态转移方程的系数矩阵 $\widetilde{\boldsymbol{A}}$ 施加对角矩阵限制，并结合 EM 算法对 GDP 进行即时预测。最后，为了验证本文所提出的变量选择即时预测效果，本文仅对 GDP 的观测方程进行了变量选择。此外，我们采用距离预测时间点最近的一年数据作为验证数据集，对于 Lasso 的惩罚超参进行最优调参，在预测因子选择完成后我们进一步使用 OLS 回归进行参数估计。因此，在部分时间段，基于 Lasso 技术的动态混频因子模型可能会与基准模型 1 相同。最后，值得强调的是，在使用 Lasso 回归模型后 GDP 的部分因子载荷为 0，因此，给定 GDP 的因子载荷，我们在此后的 EM 迭代中使用压缩后的模型对隐因子进行估计，并对 GDP 进行即时预测。在下文中，我们将本文所提出的模型简称为压缩模型。

（一）模型预测整体表现

为了阐明模型的预测能力并使用更加贴合实时预测的流程，对于以上三类模型，我们分别基于 3、4、5 和 6 因子模型进行了即时最优模型选择[①]：我们使用样本中最近两个季度或三个季度作为验证集，用以选择最优因子模型。因为模型选择是实时变化的，我们称之为即时最优因子模型。换言之，对于压缩模型，我们会实时更新即时最优压缩模型；对于基准模型（1、2），我们也会实时选择即时最优基准模型（1、2）。此外，我们还估计了单因子模型[②]与混频预测 MIDAS 模型[③]。我们使用 2005 年以前的数据作为模型的训练数据，将 2005 年以后的数据作为样本外测试数据，采用延展数据窗口对我国 GDP 进行即时预测。为了考虑新冠疫情前后的宏观经济的结构性差异，我们将测试数据分为 2005—2019 年（新冠疫情前）和 2020—2022 年（新冠疫情期间）两个子样本，并

① 对于本部分结果的呈现，我们感谢匿名审稿人的宝贵意见。

② 单因子模型不存在因子选择的问题。

③ 此处，我们使用的 MIDAS 模型采用了本文所使用的 15 个月度宏观数据预测季度 GDP。对于缺失月度数据，我们使用上一期公布数据进行填充，最大滞后期为 2 个月，系数约束函数为 Beta 函数，对于具体估计过程，请参见 Ghysels et al.（2020）。

比较了不同模型在子区间以及在总体样本上的预测表现。最后我们基于 5 因子压缩模型对我国 GDP 即时预测进行了贡献分解，识别了对我国 GDP 预测信息贡献最大的几个宏观经济指标。

我们采用样本外均方预测误差（MSFE）来评价模型预测能力的优劣，其计算公式为：

$$MSFE=\frac{1}{T}\sum_{t=1}^{T}(y_t^f-y_t)^2, \tag{17}$$

其中 T 表示样本外预测样本中 GDP 同比增速发布的总次数，y_t 表示官方发布的真实值，y_t^f 表示模型对 GDP 的预测值（y_t＝所发布的 GDP 同比增速乘以 100）。这里样本外均方预测误差越小则代表模型预测精度越高。

我们将各类即时最优模型的样本外 MSFE 表现汇报在表 2 之中。整体而言，不论是整体预测结果还是新冠疫情前后的分区间预测结果，即时最优压缩模型均优于即时最优的基准模型 1 和基准模型 2。

从表 2 中我们可以发现，各类模型最优预测表现的优劣排序为：压缩模型、基准模型 2、基准模型 1 和单因子模型。以第二季度长度的验证集为例，在整体预测效果上压缩模型、基准模型 2 和基准模型 1 的 MSFE 误差对比单因子模型下降了 45.9%、41.1%和 40.4%。在疫情前预测效果上压缩模型、基准模型 2 和基准模型 1 的 MSFE 误差对比单因子模型下降了 22.3%、19.3%和 15.3%。在疫情后预测效果上压缩模型、基准模型 2 和基准模型 1 的 MSFE 误差对比单因子模型下降了 48.3%、44.2%和 42.6%。

表 2 GDP 即时预测的样本外 MSFE 表现（即时最优模型）

		最佳模型	模型类型	第二季度验证集	第三季度验证集
2005—2022 年	整体样本	因子压缩模型	压缩模型	4.529	4.514
			基准模型 1	4.986	4.974
			基准模型 2	4.899	4.907
			单因子模型	8.366	
			MIDAS	18.349	
2005—2019 年	新冠疫情前	因子压缩模型	压缩模型	0.706	0.689
			基准模型 1	0.732	0.719
			基准模型 2	0.769	0.790
			单因子模型	0.908	
			MIDAS	17.152	
2020—2022 年	新冠疫情后	因子压缩模型	压缩模型	30.019	30.019
			基准模型 1	33.345	33.345
			基准模型 2	32.432	32.351
			单因子模型	58.088	
			MIDAS	26.328	

为了进一步说明这一预测效果最终转化为预测 GDP 增速的误差，我们对以上模型计算了 MAFE 指标[①]：

$$MAFE=\frac{1}{T}\sum_{t=1}^{T}\mid y_t^f-y_t\mid. \tag{18}$$

以第二季度长度的验证集为例，在整体预测效果上压缩模型、基准模型 2、基准模型 1 和单因子模型的预测绝对误差分别为 1.037、1.063、1.103 和 1.405 个百分点，对比单因子模型，其他三个模型的预测误差分别下降了 26.2%、24.4%和 21.5%。在疫情前预测效果上压缩模型、基准模型 2、基准模型 1 和单因子模型的预测绝对误差分别为 0.599、0.603、0.637 和 0.736 个百分点。对比单因子模型，其他三个模型的预测误差分别下降了 18.7%、18.2%和 13.5%。在疫情后预测效果上压缩模型、基准模型 2、基准模型 1 和单因子模型的预测绝对误差分别为 3.960、4.128、4.211 和 5.862 个百分点。对比单因子模型，其他三个模型的预测误差分别下降了 32.4%、29.6%和 28.2%。

以第三季度长度的验证集为例，在整体预测效果上压缩模型、基准模型 2、基准模型 1 和单因子模型的预测绝对误差分别为 1.036、1.060、1.112 和 1.405 个百分点，对比单因子模型，其他三个模型的预测误差分别下降了 26.3%、24.6%和 20.9%。在疫情前预测效果上压缩模型、基准模型 2、基准模型 1 和单因子模型的预测绝对误差分别为 0.597、0.600、0.648 和 0.736 个百分点，对比单因子模型，其他三个模型的预测误差分别下降了 18.9%、18.6%和 12.0%。在疫情后预测效果上压缩模型、基准模型 2、基准模型 1 和单因子模型的预测绝对误差分别为 3.960、4.128、4.211 和 5.862 个百分点。对比单因子模型，其他三个模型的预测误差分别下降了 32.4%、29.6%和 28.2%。

为了进一步探究压缩模型的优越性来源，我们将固定因子个数的基准模型和压缩类模型对于 GDP 即时预测的样本外 MSFE 表现汇报在表 3 之中。从表 3 可以看出，不论是整体样本还是新冠疫情前后的两个子样本内，压缩模型一直是最佳 GDP 即时预测模型。从整体预测效果来看，GDP 即时预测效果最好的五个模型依次是：6 因子压缩模型、6 因子基准模型 1、5 因子基准模型 1、5 因子压缩模型以及 6 因子基准模型 2。而相对于最佳的 6 因子压缩模型，次优的四个模型其预测误差上升了 0.73%、2.77%、5.01%和 6.82%。我们可以发现本文所考察的多因子模型在整体样本的预测误差均显著小于单因子模型。此外，我们发现不论是在新冠疫情发生前还是新冠疫情发生后的样本内，压缩模型均是最佳预测模型。

表 3　各类 GDP 即时预测的样本外 MSFE 表现

		最佳模型	模型类型	3 因子	4 因子	5 因子	6 因子
2005—2022 年	整体样本	6 因子压缩模型	压缩模型	5.721	5.687	4.586	4.367
			基准模型 1	5.753	5.748	4.488	4.399
			基准模型 2	5.409	4.880	4.910	4.665
			单因子模型		8.366		

① 由于篇幅限制，我们未将 GDP 即时预测的样本外 MAFE 表现单独制表，仅对实证结果进行了梳理。

（续表）

		最佳模型	模型类型	3 因子	4 因子	5 因子	6 因子
2005—2019 年	新冠疫情前	5 因子压缩模型	压缩模型	0.946	0.905	0.675	0.797
			基准模型 1	0.946	0.885	0.697	0.749
			基准模型 2	0.861	0.764	0.816	0.890
			单因子模型		0.908		
2020—2022 年	新冠疫情后	6 因子压缩模型	压缩模型	37.553	37.570	30.658	28.169
			基准模型 1	37.804	38.167	29.763	28.735
			基准模型 2	35.728	32.325	32.204	29.831
			单因子模型		58.088		

从表 3 中还可以看出，新冠疫情后 GDP 即时预测误差显著大于新冠疫情前的即时预测误差。此外，我们会发现因子数目较高的即时预测模型有着更佳的表现：以压缩模型为例，6 因子模型预测表现优于 5 因子模型，5 因子模型预测表现优于 4 因子模型。可以发现，不论是压缩类模型，还是基准模型 1，其预测误差随着因子数目的上升而下降。整体而言，新冠疫情期间较大因子数目的 GDP 即时预测模型表现更加优秀。此外，虽然基准模型 2 对模型结构施加了限制，但是不论是整体样本还是各子样本，基准模型 2 并没有列位前三模型之中，这反映出，Bańbura and Modugno（2014）施加的人为限制存在较大局限性，无法获得实证证据支撑。

其次，在新冠疫情前区间内，最佳的五个 GDP 即时预测模型为 5 因子压缩模型、5 因子基准模型 1、6 因子基准模型 1、4 因子基准模型 2 以及 6 因子压缩模型。相对于 5 因子压缩模型，其他四个次优模型的预测误差上升了 3.26%、10.96%、13.19%和 18.07%，并且在我们所考察的所有模型中，除了两个模型，其他模型预测效果均优于单因子模型。

最后，在新冠疫情后区间内，最佳的五个 GDP 即时预测模型为 6 因子压缩模型、6 因子基准模型 1、5 因子基准模型 1、6 因子基准模型 2 以及 5 因子压缩模型。而次优的四个模型相对于 6 因子压缩模型，其预测误差上升 2.01%、5.66%、6.26%和 8.84%，并且单因子模型在新冠疫情后其预测误差相较以上多因子模型上升了约 100%。

我们在表 4 中汇报了基于不同模型对 GDP 同比增速即时预测与实际发布 GDP 同比增速的样本外相关系数。

表 4　各类 GDP 即时预测的样本外相关系数表现

		最佳模型	模型类型	3 因子	4 因子	5 因子	6 因子
2005—2022 年	整体样本	6 因子压缩模型	压缩模型	0.720	0.721	0.781	0.793
			基准模型 1	0.719	0.721	0.786	0.791
			基准模型 2	0.738	0.766	0.764	0.777
			单因子模型		0.609		

（续表）

		最佳模型	模型类型	3 因子	4 因子	5 因子	6 因子
2005—2019 年	新冠疫情前	5 因子压缩模型	压缩模型	0.935	0.940	0.952	0.940
			基准模型 1	0.936	0.943	0.949	0.943
			基准模型 2	0.937	0.941	0.939	0.936
			单因子模型		0.931		
2020—2022 年	新冠疫情后	6 因子压缩模型	压缩模型	0.413	0.406	0.490	0.518
			基准模型 1	0.411	0.421	0.481	0.496
			基准模型 2	0.440	0.464	0.462	0.486
			单因子模型		0.220		

从表 4 可以看出，就相关系数而言，不论是整体样本还是新冠疫情前后的两个子样本内，压缩模型也一直是最佳 GDP 即时预测模型。从整体预测效果来看，GDP 即时预测效果最好的五个模型依次是：6 因子压缩模型、6 因子基准模型 1、5 因子基准模型 1、5 因子压缩模型以及 4 因子压缩模型。在新冠疫情前，GDP 即时预测效果最好的五个模型依次是：5 因子压缩模型、5 因子基准模型 1、4 因子基准模型 1、6 因子基准模型 1、2 因子基准模型 2。在新冠疫情期间，GDP 即时预测效果最好的五个模型依次是：6 因子压缩模型、6 因子基准模型 1、5 因子压缩模型、6 因子基准模型 2 和 5 因子基准模型 1。此外，本文还在附录Ⅳ中将训练窗口固定为 10 年，并使用与预测期最近的两/三个季度作为 Lasso 超参数调试的验证集重新估计并测试各类模型的样本外表现，得到了与上文类似的结果。

（二）模型预测时序表现

由于不同宏观经济指标在每个月的发布时间并不相同，因此在每个数据公布日，我们均可以更新 GDP 即时预测结果。我们在图 1 中汇报了 5 因子压缩模型对于 GDP 即时样本外预测的时序走势。可以发现，因子压缩模型可以较好地预测 GDP 的实际增速，在经济整体增速的上升期和下降期均能提前预测 GDP 增速上升和下跌的趋势，有助于决策者及时判断当前经济整体运行状况。

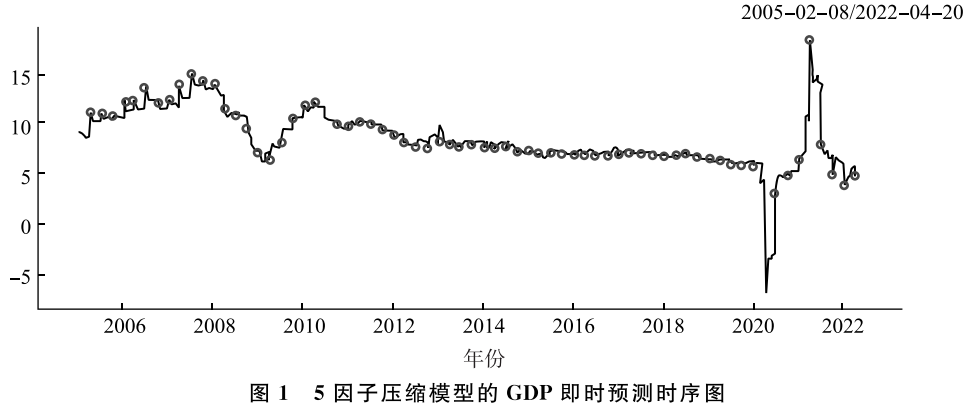

图 1　5 因子压缩模型的 GDP 即时预测时序图

注：其中实线为 GDP 同比增速预测值，圆圈为公布的 GDP 同比增速。

在图 2 中，我们汇报了 5 因子压缩模型、基准模型 1、基准模型 2 在 2008 年第三季度至 2009 年第一季度期间（典型的宏观经济下行）的 GDP 即时预测走势。我们可以发现随着逐渐靠近 GDP 增速发布日期，三个模型的预测值逐步逼近真实的 GDP 增速，而压缩因子模型的收敛速度最快，对于 2009 年第一季度经济增速下滑的判断，我们的模型可以较数据发布日提前 61 日预知。对于经济上行期的预测演化，本文附录Ⅳ汇报了一个典型的宏观经济上行期中的时序预测情况。

（三）信息增益分解

如前文所述，在高维数据背景下，相较于直接使用 MIDAS 回归，混频动态因子不但有着更高的预测精确度，其另一优势是可以将 GDP 增速即时预测值的变化分解至不同宏观指标的公布信息上。信息分解可以理解为：随着时间不断演进，新宏观数据指标的发布[①]会修正我们对潜在共同因子的估计（预期），进而改变我们对当季 GDP 增速的预测。因此，我们可以将任一时间段内 GDP 增速预测值的变化分解至期间所有新发布的宏观经济序列之上。具体计算方法可参考 Bańbura and Modugno（2014）和本文附录Ⅲ。

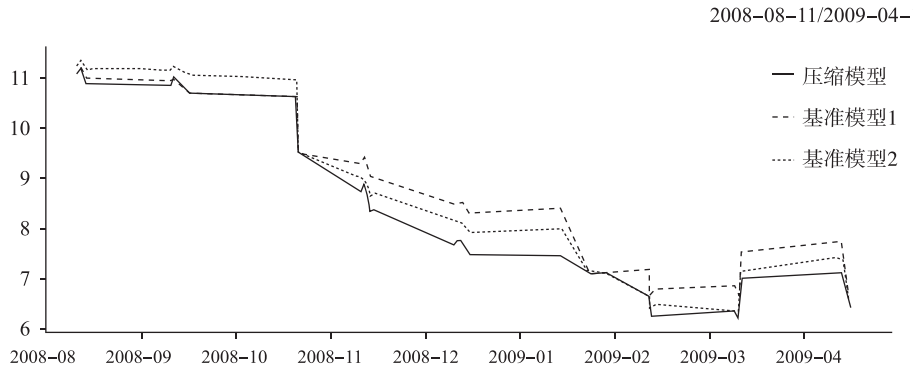

图 2 5 因子模型的 GDP 即时预测时序图

注：2008 年第三季度至 2009 年第一季度，圆圈代表真实值。

在表 5 中，我们基于 5 因子压缩模型，汇报了 2005 年第一季度至 2022 年第一季度 GDP 即时预测的变动分解。我们从各个宏观变量引起 GDP 即时预测的平均变动（以变动绝对值的平均值衡量）和总变动（以变动平方的总计衡量）两个角度来分析我国不同宏观经济变量可以为 GDP 即时预测所带来的信息增益，下文中我们简称这两者为平均信息增益和总信息增益。

表 5 基于 5 因子压缩模型的 GDP 即时预测变动分解

指标名称	平均信息增益	总信息增益
国家财政支出	0.057	0.872
发电量	0.046	0.804

① 更确切地说，是公众未预期到的宏观数据指标的变化部分导致了我们对潜在因子估计的改变，因此，这部分冲击又被称为信息冲击。

（续表）

指标名称	平均信息增益	总信息增益
社会消费品零售总额	0.045	3.470
工业增加值	0.042	1.979
进口总值	0.040	1.186
工业生产者出厂价格指数	0.038	1.033
出口总值	0.026	0.678
国家财政收入	0.021	0.695
居民消费价格指数	0.021	0.252
M1 供应量	0.019	0.287
M0 供应量	0.018	0.205
房地产投资	0.018	0.129
M2 供应量	0.016	0.148
固定资产投资额	0.014	0.272
商品房销售额	0.007	0.039

从表 5 中可以发现，在平均信息增益方面，能给 GDP 即时预测带来最多平均信息增益的五个宏观变量分别为国家财政支出、发电量、社会消费品零售总额、工业增加值以及进口总值的同比增长。在总信息增益方面，能给 GDP 即时预测带来最多信息增益的五个宏观变量分别为社会消费品零售总额、工业增加值、进口总值、工业生产者出厂价格指数以及国家财政支出的同比增长。在两种衡量信息增益的测算中，带来信息增益最多的前五个宏观变量中有四个是一致的。这些发现从侧面验证了因子压缩模型在预测信息增益分解方面的稳定性，也反映出个体消费、工业生产、进出口以及国家财政是在对我国 GDP 进行即时预测时应重点关注的宏观经济指标。

为了进一步阐示本文所使用的信息增益分解，我们在图 3 中汇报了基于 5 因子压缩模型对 2008 年第四季度 GDP 即时预测的信息增益分解。为了更好地展示 GDP 即时预测的信息增益分解过程，我们将宏观变量划分为消费、投资、进出口、货币供应量、工业生产及国家财政收支共六组（见附录Ⅳ表Ⅳ4）。从图 3 可以看出，在 2008 年 11 月 10 日，我们对于 2008 年第四季度的 GDP 增速初始预测值为 8.72%。随后，11 月 11 日 CPI 及进出口数据公布使得预测值略微上升 0.16%，11 月 12 日货币供应量公布使得预测值略微下降 0.14%。11 月 13 日工业增加值及发电量数据发布带来了负面冲击，预测估计值下降 0.41%。2018 年 12 月 10 日，进出口数据及 PPI 数据发布，进一步降低预测值约 0.71 个百分点。12 月 11 日和 12 日公布的 CPI 数据和社会消费品零售总额数据影响不大。直到 12 月 15 日，工业增加值、发电量数据以及货币供应量数据发布，其中工业生产类数据进一步压低预测值 0.29%。此后 GDP 同比预测基本稳定在 7.45%上下，直到 2009 年 1 月 22 日官方发布 2008 年第四季度 GDP 同比增速为 7.1%。本文附录Ⅳ也展示了基于固定窗口训练的因子模型的信息分解结果，结果与前文类似。

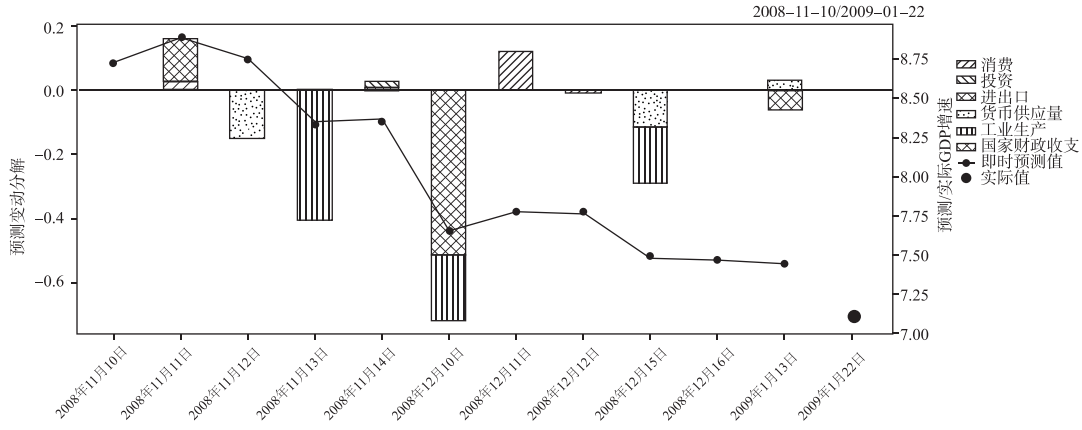

图 3 基于 5 因子压缩模型 2008 年第四季度 GDP 即时预测的贡献分解

五、结论与政策建议

党的二十大报告指出，“党中央团结带领全党全军全国各族人民有效应对严峻复杂的国际形势和接踵而至的巨大风险挑战，以奋发有为的精神把新时代中国特色社会主义不断推向前进”，并提出要“增强干部推动高质量发展本领、服务群众本领、防范化解风险本领。加强干部斗争精神和斗争本领养成，着力增强防风险、迎挑战、抗打压能力”。纵观党的二十大报告，共计 16 次提及风险。提高“防风险”“化解风险”的能力是党的二十大对宏观经济治理部门的一项重要要求。

在这一时代背景下，精确的 GDP 即时预测不但可以助力实时把脉宏观经济，准确及时地监测我国宏观经济运行的整体状况，精准定位宏观经济运行的风险点，还有利于我国实施前瞻性、针对性的政策调控，有利于宏观经济治理相关部门防范、化解宏观经济系统性风险，确保经济安全，完善宏观经济治理体系，从而助力经济高质量发展。

与已有 GDP 即时预测的文献相比，首先，本文首次在动态多因子混频框架中进行了模型压缩的研究，这一创新在已有文献的基础上进一步提高了 GDP 即时预测的样本外精度。其次，本文使用了更为广泛的宏观经济数据指标，大大拓宽了我国宏观经济预测模型可以使用的有效信息范畴，并提高了 GDP 的可预测频率。最后，在研究我国 GDP 即时预测的文献中，本文也是首篇将信息增益分解用于识别对 GDP 预测信息贡献度的论文。

展望未来研究，本文尽管已经考虑了广泛的宏观经济变量，并开创性地在 GDP 即时预测模型中融合了机器学习分析技术，但由于篇幅原因，本文并未在如何使用非线性方法针对 GDP 进行即时预测方面进行展开。若是使用粒子滤波分析框架，并加入基于神经网络拟合方法的非线性关系，可能会进一步提高多因子混频即时预测的准确性。未来，我们也将考虑基于宏观经济学与资产定价学理论，从金融市场和大宗商品市场中提取市场对于宏观经济整体运行情况的情绪指标，并引入电信、交通数据以完善我国宏观经济实时数据库，进一步提高对于 GDP 预测的频率，从“发布日频”提升至真正的日

频，助力监管部门、学界与业界更好地把握和理解我国经济现实问题，推进学科融合交叉发展。

参 考 文 献

[1] Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin, "Now-Casting and the Real-Time Data Flow", In: Elliott, G., C. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*. Amsterdam: Elsevier, 2013, 2, 195-237.

[2] Bańbura, M., and M. Modugno, "Maximum Likelihood Estimation of Factor Models on Datasets with Arbitrary Pattern of Missing Data", *Journal of Applied Econometrics*, 2014, 29 (1), 133-160.

[3] Belloni, A., and V. Chernozhukov, "Least Squares after Model Selection in High-Dimensional Sparse Models", *Bernoulli*, 2013, 19 (2), 521-5.

[4] Bok, B., D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti, "Macroeconomic Nowcasting and Forecasting with Big Data", *Annual Review of Economics*, 2018, 10, 615-643.

[5] Bork, L., "Estimating US Monetary Policy Shocks Using a Factor-Augmented Vector Autoregression: An EM Algorithm Approach", SSRN 1348552, 2009.

[6] Bork, L., H. Dewachter, and R. Houssa, "Identification of Macroeconomic Factors in Large Panels", CREATES Research Paper, 2009 (2009-43).

[7] Ghysels, E., V. Kvedaras, and V. Zemlys-Balevičius, "Mixed Data Sampling (MIDAS) Regression Models", In: Vinod, H. D. and C. R. Rao (eds.), *Handbook of Statistics*. Amsterdam: Elsevier, 2020, 117-153.

[8] Giannone, D., L. Reichlin, and D. Small, "Nowcasting: The Real-Time Informational Content of Macroeconomic Data", *Journal of Monetary Economics*, 2008, 55 (4), 665-676.

[9] Kuzin, V., M. Marcellino, and C. Schumacher, "MIDAS vs. Mixed-Frequency VAR: Nowcasting GDP in the Euro Area", *International Journal of Forecasting*, 2011, 27 (2), 529-542.

[10] 刘汉、刘金全，"中国宏观经济总量的实时预报与短期预测——基于混频数据预测模型的实证研究"，《经济研究》，2011 年第 3 期，第 4—17 页。

[11] Mariano, R. S., and Y. Murasawa, "A New Coincident Index of Business Cycles Based on Monthly and Quarterly Series", *Journal of Applied Econometrics*, 2003, 18 (4), 427-443.

[12] Schorfheide, F., and D. Song, "Real-Time Forecasting with a Mixed-Frequency VAR", *Journal of Business & Economic Statistics*, 2015, 33 (3), 366-380.

[13] Tibshirani, R., "Regression Shrinkage and Selection via the Lasso: A Retrospective", *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2011, 73 (3), 273-282.

[14] Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need", *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 6000-6010.

[15] 王维国、于扬，"基于混频回归类模型对中国季度 GDP 的预报方法研究"，《数量经济技术经济研究》，2016 年第 4 期，第 108—125 页。

[16] 王霞、司诺、宋涛，"中国季度 GDP 的即时预测与混频分析"，《金融研究》，2021 年第 8 期，第 22—41 页。

[17] Watson, M. W., and R. F. Engle, "Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models", *Journal of Econometrics*, 1983, 23 (3), 385-400.

[18] 张劲帆、刚健华、钱宗鑫、张龄琰，"基于混频向量自回归模型的宏观经济预测"，《金融研究》，2018 年第 7 期，第 34—48 页。

Nowcasting GDP with Big Macroeconomic Data

YI Yanping
(Zhejiang University)
HUANG Dejin WANG Xi*
(Peking University)

Abstract: Combining expectation maximization (EM) algorithm and Lasso method, we propose a methodology to estimate a mixed-frequency dynamic factor model on large macroeconomic panels with ragged edges. We apply this approach to nowcast China's gross domestic product (GDP) growth and decompose the resulting forecast revision into contributions from the news. We find that, (1) Our method improves the out-of-sample forecast accuracy in nowcasting Chinese GDP growth, compared to the existing methods; (2) From 2005Q1 to 2022Q1, growth of fiscal expenditure, growth in retail sales, growth of industry's value-added and growth of total imports affected the real-time forecast of China's GDP growth the most.

Keywords: nowcasting; big data analysis; dynamic factor model

JEL Classification: C53, C61, E27

* Corresponding Author: WANG Xi, School of Economics, Peking University, Haidian District, Beijing 100871, China; Tel: 86-10-62753635; E-mail: wang.x@pku.edu.cn.