arXiv:2302.03172v1 [econ.EM] 7 Feb 2023

具有缺失数据的高维条件高斯状态空间模型*

Joshua C. C. Chan Aubrey Poon Dan Zhu¨普渡大学 厄勒布鲁大学 莫纳什大学

2023年1月

摘要

我们提出了一种高效的抽样方法，用于处理条件高斯状态空间模型中复杂的缺失数据模式和大量缺失观测值。两个重要示例是基于非平衡数据集的动态因子模型，以及包含多个频率变量的大型贝叶斯向量自回归模型。该方法的一个关键洞见是：给定观测数据后，缺失数据的联合分布是高斯分布。此外，该条件分布的逆协方差矩阵（即精度矩阵）是稀疏的，我们可以利用这一特殊结构大幅加快计算。我们通过两个实证应用展示了这一方法。第一个应用使用大型贝叶斯向量自回归模型，结合季度、月度和周度数据，生成每周国内生产总值估计值。第二个应用则通过带有随机波动率的动态因子模型，从包含一百多个月度变量的非平衡数据集中提取潜在因子。

JEL分类：C11、C32、C55

关键词：混频数据、非平衡面板、向量自回归、动态因子模型、随机波动率

*本文的早期版本曾以《状态空间混合频率向量自回归模型的有效估计：一种基于精度矩阵的方法》为题流传。感谢 Otilia Boldea、Francis Diebold、Gary Koop、James Mitchell 和 Michele Piffer 提出了许多建设性意见和宝贵建议。

1 引言

实证宏观经济学日益广泛地使用大规模时间序列模型，以充分利用大型数据集的广泛可用性。这一趋势有望实现更及时、更全面的分析，但也带来了新的挑战。首先，大型数据集通常由多个来源汇编而成，因此往往涉及复杂的缺失数据模式。一个突出的例子是混频数据：为纳入实时信息，人们采用混频数据，而非仅使用频率相同（且较低）的变量这一传统方法。例如，Aruoba、Diebold 和 Scotti（2009）结合每日、每周、每月和季度变量，构建商业状况指数以跟踪经济活动。Schorfheide 和 Song（2015）使用仅按季度提供的国内生产总值，以及其他季度和月度变量，获得月度频率的国内生产总值估计值。在这两种情况下，低频变量的高频观测值都被视为缺失数据。因此，缺失观测值的数量很大。

其次，从大型数据集中提取信息通常需要大规模时间序列模型。因子模型一直是实现这一目的的主力工具。得益于Ba ́nbura、Giannone 和 Reichlin（2010）以及 Koop（2013）的开创性研究，大型贝叶斯向量自回归如今已成为一种颇受欢迎的替代方案。此外，大量实证证据表明，允许模型具有异方差性、厚尾分布和异常值检测等灵活特征，对于改善样本内拟合和样本外预测表现至关重要（例如，Clark，2011；D’Agostino、Gambetti 和 Giannone，2013；Cross 和 Poon，2016；Stock 和 Watson，2016；Chan，2020）。因此，这些特征正日益被纳入动态因子模型和大型贝叶斯向量自回归模型中（近期的例子包括 Kastner 和 Huber，2020；Antolin-Diaz、Drechsel 和 Petrella，2021；Carriero、Clark、Marce

llino 和

2022年）。尽管近期在加快使用完整数据估计这些灵活的大型模型方面取得了许多进展，但相比之下，能够处理复杂缺失数据模式（其中缺失观测值数量众多）的高效算法仍不够成熟。

我们通过开发一种高效的抽样方法，在一步之内抽取所有缺失观测值，从而应对这些挑战。为使我们的方法具有广泛适用性，我们在条件高斯状态空间模型的一般框架下进行开发。因此，该方法适用于许多流行的大型模型，例如带有随机波动率的动态因子模型

或带有非高斯误差的混频向量自回归模型。此外，该设定可以轻松处理各种复杂的缺失数据模式，包括不平衡面板、混频设定，以及由于不同步数据发布而导致的样本末端“参差边缘”。得益于马尔可夫链蒙特卡罗（MCMC）方法的模块化特性，所提出的方法可以与任何针对完整数据条件高斯状态空间模型的高效抽样器直接结合使用。因此，本文补充了现有关于灵活大型模型快速估计的研究，并将其扩展到缺失数据设定。

所提出方法的核心洞见是：给定观测数据（以及其他模型参数和潜变量）后，所有缺失观测值的联合分布服从高斯分布。此外，该条件分布的精度矩阵（即协方差矩阵的逆矩阵）是稀疏的——事实上，对于许多常见的缺失数据模式，它是带状的，也就是说，它是稀疏的，且非零元素沿对角带排列。利用这些特殊结构可以大幅加快计算速度。具体而言，可以应用 Chan and Jeliazkov（2009年）的基于精度矩阵的抽样器，一步抽取所有缺失观测值。与标准的基于卡尔曼滤波的方法相比，这种方法的效率要高得多，尤其是在缺失观测值数量较多或状态向量维度较高时。[1]此外，这种方法很容易实现：只需定义适当的选择矩阵，将数据向量划分为观测数据和缺失数据即可。所提出的方法能够轻松处理许多复杂的缺失数据模式，例如包含多个频率变量的情形。

此外，我们的设定也能适用于存在额外信息、有助于提高对缺失观测值推断精度的情形。这一特性对于

[1]Chan and Jeliazkov（2009年）以及 McCausland、Miller 和 Pelletier（2011）的基于精度矩阵的抽样方法，旨在用于数据完整的线性高斯状态空间模型。该方法建立在早期关于高斯马尔可夫随机场（Rue，2001）和非参数回归（Chib 和 Jeliazkov，2006；Chib、Greenberg 和 Jeliazkov，2009）的研究基础上。由于易于实现且计算效率高，这种方法正越来越多地应用于广泛的实证研究中。近期实例包括对趋势通胀（Chan、Koop 和 Potter，2013；Chan，2017；Hou，2020）、时变菲利普斯曲线（Fu，2020）和股息增长（Pettenuzzo、Sabbatucci 和 Timmermann，2023）进行建模；估计产出缺口（Grant 和 Chan，2017a,b）；进行宏观经济预测（Cross 和 Poon，2016；Cross、Hou 和 Poon，2019）；以及拟合各种移动平均模型（Chan，2013；Chan、Eisenstat 和 Koop，2016；Dimitrakopoulos 和 Kolossiatis，2020；Zhang、Chan 和 Cross，2020）和动态因子模型（Kaufmann 和 Schumacher，2019；Beyeler 和 Kaufmann，2021）。

混频应用至关重要，在这类应用中，需要将高频缺失观测值的线性组合映射到低频变量的观测值。我们的论文与 Eckert、Kronenberg、Mikosch 和 Neuwirth（2020年）以及 Hauber 和 Schumacher（2021年）的近期研究相关；他们也针对存在缺失观测值的情形，考虑了基于精度矩阵的抽样方法。不过，他们聚焦于动态因子模型，且后者未考虑混频设定。相比之下，我们的方法更为通用，适用于各种缺失数据模式下的任意条件高斯状态空间模型。

我们开展了一系列蒙特卡洛实验，以展示所提出的基于精度矩阵的方法在数值精度和计算速度方面的表现。具体而言，我们在多种设定下，使用所提出的抽样器和标准滤波方法估计混频向量自回归模型。结果表明，与传统的基于卡尔曼滤波的方法相比，所提出的基于精度矩阵的方法在计算上效率高得多。此外，该方法能够良好地扩展至高维情形，从而可以估计包含大量低频变量的向量自回归模型。

为展示所提出的基于精度矩阵的方法的灵活性，我们考虑了两个缺失数据模式迥异的宏观经济实证应用。在第一个应用中，我们使用一个包含随机波动率的大型混频贝叶斯 VAR，生成实际国内生产总值的每周估计值。这些高频 GDP 估计值可用于多种目的，例如监测当前经济状况，以及及时提供关键宏观经济变量的即时预测。为获得每周 GDP 估计值，我们使用 22 个变量拟合贝叶斯 VAR，这些变量具有 3 种不同频率：16 个周度变量、5 个月度变量以及季度实际国内生产总值。所有变量均按周频建模，而月度变量和季度变量的每周观测值则视为缺失数据。尽管缺失数据模式十分复杂——例如，不同月份和季度的周数各不相同——且缺失观测值数量众多，所提出的方法在计算上仍然高效且易于实现。

在第二个应用中，我们使用带随机波动率的动态因子模型，从 McCracken 和 Ng（2016）的 FRED-MD 数据集中实时提取潜在因子。FRED-MD 的每个数据版本都包含 128 个每月变量，但其中许多变量存在两类缺失值：样本开始时，部分近期构建的变量存在缺失观测值

；样本结束时，由于发布滞后而存在缺失值。我们采用所提出的方法，在动态因子模型下对缺失观测值进行抽样，并获得潜在因子。结果表明，即使在 COVID-19 疫情爆发初期显著下滑及随后反弹期间，第一因子仍能很好地反映整体经济状况。此外，我们的结果表明，仅使用不存在缺失值的变量可能会错误地反映潜在因子的动态变化，这凸显了纳入含有缺失值的变量所提供信息的重要性。

本文其余部分安排如下。第2节介绍所提出的基于精度矩阵的抽样方法，该方法用于在一般状态空间框架下抽取缺失观测值。第3节通过一系列蒙特卡洛实验，在多种混频设定下将所提出的抽样方法与标准的基于卡尔曼滤波的技术进行比较。第4节展示如何将所提出的抽样方法应用于两个常见的实证宏观经济应用。最后，第5节作出总结。

2 一般状态空间框架

本节介绍一种基于精度矩阵的方法，用于在一般状态空间框架下，根据多种信息集对缺失数据进行抽样。具体而言，我们首先推导给定观测数据和其他模型参数时缺失数据的联合条件分布，并证明该分布服从高斯分布。随后，我们讨论一种高效算法，用于从这个通常维度很高的高斯分布中生成样本。此外，由于在许多应用场景（例如混频设定）中，我们掌握有关缺失数据的额外信息，因此还将说明如何利用这些额外信息更新缺失数据的条件分布。

2.1 缺失数据的条件分布

我们的通用设定如下：针对一个由 $n \times 1$ 个变量构成的向量 $\mathbf{y}_t = (y_{1,\ t}, \ldots, y_{n,\ t})'$，其取值范围为 $t = 1, \ldots, T$，建立以下条件高斯状态空间模型：

$$\mathbf{y}_t = \mathbf{W}_t \boldsymbol{\alpha}_t + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma}_t), \tag{1}$$

$$\boldsymbol{\alpha}_t = \boldsymbol{\gamma} + \boldsymbol{\Phi}_1 \boldsymbol{\alpha}_{t-1} + \cdots + \boldsymbol{\Phi}_q \boldsymbol{\alpha}_{t-q} + \boldsymbol{\varepsilon}_t^{\boldsymbol{\alpha}}, \qquad \boldsymbol{\varepsilon}_t^{\boldsymbol{\alpha}} \sim \mathcal{N}(\mathbf{0}_k, \boldsymbol{\Omega}_t), \tag{2}$$

其中，$\mathbf{0}_m$ 表示一个 $m \times 1$ 零向量，$\boldsymbol{\beta}$ 是时不变参数向量，$\boldsymbol{\alpha}_t$ 是时变参数向量，$\boldsymbol{\Sigma}_t$ 和 $\boldsymbol{\Omega}_t$ 分别是观测方程和状态方程的协方差矩阵。协变量矩阵 $\mathbf{W}_t$ 和 $\mathbf{X}_t$ 可以包含 $\mathbf{y}_t$的滞后值。该框架涵盖了多种常用模型，包括动态因子模型和向量自回归模型。

请注意，该框架还将许多不同类型的误差过程纳为特例。例如，可以将 $\boldsymbol{\Sigma}_t$ 设定为 Cogley 和 Sargent（2005）、Primiceri（2005）、Carriero、Clark和Marcellino（2016）或 Kastner（2019）中的多元随机波动率过程。此外，还可以考虑各种类型的非高斯误差，例如通过设定 $\boldsymbol{\Sigma}_t = \lambda_t \mathbf{Q}$（其中 Q 是协方差矩阵，且 $\lambda_t \sim \mathcal{IG}(\nu/2, \nu/2)$）来指定 t 分布，或者通过设定 $\boldsymbol{\Sigma}_t = o_t^2 \mathbf{Q}$ 来指定 Stock 和 Watson（2016）中的异常值成分，其中 o_t 服从两部分分布，在 1 处有一个点质量，并在区间 (2, 10) 上服从均匀分布。当然，也可以采用上述多元随机波动率过程或非高斯误差的任意组合，例如 Chan（2020）和 Carriero、Clark、Marcellino 和 Mertens（2022）中的组合。

我们关注 $\mathbf{y}_t$ 的某些元素缺失的情形。具体而言，将 $\mathbf{y}_t$ 划分为两个子向量 $\mathbf{y}_t^o$ 和 $\mathbf{y}_t^m$，其中 $\mathbf{y}_t^o$ 是由观测变量构成的 n_t^o 维向量，$\mathbf{y}_t^m$ 是由缺失变量构成的 n_t^m 维向量，且满足 $n_t^o + n_t^m = n$。请注意，此处 n_t^o 和 n_t^m 可以随时间变化，因此这一设定能够涵盖广泛的缺失数据模式，例如非平衡面板和参差边缘。此外，对于具有混合频率变量的情形，通常以最高频率表示时间指数，并将部分低频变量视为缺失。例如，在同时包含每月变量和季度变量的模型中，季度存量变量的每月值仅每 3 个月观测一次，其余值均视为缺失。[2]

[2]对于流量变量，其观测值可视为可映射的额外信息

最后，令 $N^o = \sum_{t=1}^{T} n_t^o$ 和 $N^m = \sum_{t=1}^{T} n_t^m$ 分别表示观测值和缺失值的总数，其中 $N^o + N^m = Tn$。为便于后文引用，将 $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')' \in \mathbb{R}^{Tn}$、$\mathbf{y}^o = (\mathbf{y}_1^{o\prime}, \ldots, \mathbf{y}_T^{o\prime})' \in \mathbb{R}^{N^o}$ 和 $\mathbf{y}^m = (\mathbf{y}_1^{m\prime}, \ldots, \mathbf{y}_T^{m\prime})' \in \mathbb{R}^{N^m}$ 向量堆叠起来。

处理缺失观测值 $\mathbf{y}^m$ 的一种常用方法，是将其视为待增广或抽样的潜变量。通常采用标准的卡尔曼滤波和平滑算法来实现这一方法。然而，这种方法的主要缺点是，当缺失观测值数量较多时，在高维情形下计算开销往往很大。尽管大规模动态因子模型和向量自回归模型日益受到青睐，这种显著的计算负担仍是实际应用含缺失数据的高维状态空间模型的一大障碍。此外，当缺失数据模式较为复杂时，基于卡尔曼滤波的算法也会变得更难实现。为解决这些计算和实现方面的问题，我们提出一种高效方法，在给定 $\mathbf{y}^o$以及其他模型参数和潜变量的条件下，对 $\mathbf{y}^m$ 进行联合抽样，记作 $\boldsymbol{\theta}$。所提出的方法概念简单，即使面对复杂的缺失数据模式也易于实现。

下文首先推导 $\mathbf{y}^m$的联合条件分布。为此，我们将 $\mathbf{y}$ 表示为 $\mathbf{y}^o$ 和 $\mathbf{y}^m$的函数：

$$\mathbf{y} = \mathbf{S}^o \mathbf{y}^o + \mathbf{S}^m \mathbf{y}^m, \tag{3}$$

其中，$\mathbf{S}^o$和 $\mathbf{S}^m$ 分别是 $Tn \times N^o$和 $Tn \times N^m$ 选择矩阵。特别地，$\mathbf{S}^o$ 和 $\mathbf{S}^m$ 的每一列都仅包含一个值为 1 的元素，其余元素均为 0——也就是说，$\mathbf{S}^o$和 $\mathbf{S}^m$ 分别总共包含 N^o 和 N^m 1个 1。此外，各列中的 1 位于不同的行，这意味着列向量线性无关。因此，矩阵 $\mathbf{S}^o$ 和 $\mathbf{S}^m$ 列满秩。

举例来说，假设 $T = 2$、$n = 3$ 和 $y_{3,\ 1}$、$y_{1,\ 2}$ 以及 $y_{3,\ 2}$ 缺失。那么，

对于缺失的高频值；此情形将在下一小节中进一步讨论。

$\mathbf{y}^o = (y_{1,\ 1}、\ y_{2,\ 1}、\ y_{2,\ 2})'$、$\mathbf{y}^m = (y_{3,\ 1}、\ y_{1,\ 2}、\ y_{3,\ 2})'$和

$$\begin{bmatrix} y_{1,1} \\ y_{2,1} \\ y_{3,1} \\ y_{1,2} \\ y_{2,2} \\ y_{3,2} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{\mathbf{S}^o} \begin{bmatrix} y_{1,1} \\ y_{2,1} \\ y_{2,2} \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{S}^m} \begin{bmatrix} y_{3,1} \\ y_{1,2} \\ y_{3,2} \end{bmatrix}.$$

作为第二个示例，假设 $y_{1,\ t}$ 仅在每隔 3 期的 $t = 3$、6、9、..... 时被观测到，而 $y_{2,\ t}$、.....、$y_{n,\ t}$ 则在 $t = 1$、.....、T 的每一期都被观测到。于是，$\mathbf{S}^o$ 是由对角块 $\mathbf{S}_1^o$、$\mathbf{S}_2^o$、.....、$\mathbf{S}_T^o$ 组成的分块对角矩阵，即 $\mathbf{S}^o = \text{diag}(\mathbf{S}_1^o, \mathbf{S}_2^o, \ldots, \mathbf{S}_T^o)$ 且 $\mathbf{S}^m = \text{diag}(\mathbf{s}_1^m, \mathbf{s}_2^m, \ldots, \mathbf{s}_T^m)$，其中当 t 可被 3 整除时，$\mathbf{S}_t^o = \mathbf{I}_n$ 且 $\mathbf{s}_t^m = \emptyset$；否则

$$\mathbf{S}_t^o = \begin{bmatrix} \mathbf{0}'_{n-1} \\ \mathbf{I}_{n-1} \end{bmatrix}, \quad \mathbf{s}_t^m = \begin{bmatrix} 1 \\ \mathbf{0}_{n-1} \end{bmatrix}.$$

一般而言，选择矩阵 $\mathbf{S}^o$ 和 $\mathbf{S}^m$ 是稀疏的，即使面对复杂的缺失数据模式，也可以轻松构造。

现在，将 (1) 在 $t = 1$、.....、T 上堆叠，并利用 (3) 中的表达式，可以将模型更紧凑地改写为

$$\mathbf{G}^o\mathbf{y}^o + \mathbf{G}^m\mathbf{y}^m = \mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_{Tn}, \boldsymbol{\Sigma}), \tag{4}$$

其中 $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_T')'$ 和 $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_T)$。[3] 我们假设 $\mathbf{G}^m$ 列满秩，这对于大多数常用模型都成立（且在实践中很容易验证）。下面给出两个示例，说明动态因子模型和 VAR(p)如何表示为式 (4) 的形式。

[3] 如果 (1) 的右端不包含 $\mathbf{y}_t$ 的任何滞后值，那么 $\mathbf{G}^m = \mathbf{S}^m$ 且 $\mathbf{G}^o = \mathbf{S}^o$。否则，$\mathbf{G}^o$ 和 $\mathbf{G}^m$ 将成为某些差分矩阵与选择矩阵的乘积，如示例 2 所示。

示例 1 考虑以下具有随机波动率的动态因子模型 :

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}_1\mathbf{f}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma}_t), \\ \mathbf{f}_t &= \boldsymbol{\Phi}_1\mathbf{f}_{t-1} + \cdots + \boldsymbol{\Phi}_q\mathbf{f}_{t-q} + \boldsymbol{\varepsilon}_t^{\mathbf{f}}, & \boldsymbol{\varepsilon}_t^{\mathbf{f}} &\sim \mathcal{N}(\mathbf{0}_k, \boldsymbol{\Omega}_t), \end{aligned}$$

其中 $\boldsymbol{\Sigma}_t = \text{diag}(e^{h_{1,t}}, \ldots, e^{h_{n,t}})$、$\Omega_t = \text{diag}(e^{h_{n+1,t}}, \ldots, e^{h_{n+k,t}})$ 是具有时变方差的对角矩阵，而 $\mathbf{y}_t$ 被划分为两个子向量 $\mathbf{y}_t^o$ 和 $\mathbf{y}_t^m$。利用式 (3) 中的恒等式，这个动态因子模型的观测方程可表示为式 (4) 的形式：

$$\mathbf{S}^o\mathbf{y}^o + \mathbf{S}^m\mathbf{y}^m = \mathbf{H}_{\mathbf{A}_1}\mathbf{f} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_{Tn}, \boldsymbol{\Sigma}),$$

其中 $\mathbf{H}_{\mathbf{A}_1} = (\mathbf{I}_T \otimes \mathbf{A}_1)$、$\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_T)$，且 $\otimes$ 表示克罗内克积。可以考虑一种更一般的动态因子模型，其中观测方程包含动态因子的滞后值，例如 $\mathbf{f}_{t-1}$、…、$\mathbf{f}_{t-p}$。在这种情况下，只需重新定义矩阵 $\mathbf{H}_{\mathbf{A}_1}$，将这些滞后值纳入观测方程即可。

示例 2 下一个示例是一个带有异常值成分的 VAR(p) :

$$\mathbf{y}_t = \mathbf{b}_0 + \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma}_t), \tag{5}$$

其中 $\boldsymbol{\Sigma}_t = o_t^2\mathbf{Q}$，Q 是协方差矩阵，且 o_t 服从两部分分布：在 1 处有一个点质量，并在区间 (2, 10) 上服从均匀分布。然后，将式 (5) 对 $t = 1, \ldots, T$ 逐项堆叠，可得

$$\mathbf{H}_{\mathbf{B}}\mathbf{y} = \mathbf{c}_{\mathbf{B}} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_{Tn}, \boldsymbol{\Sigma}), \tag{6}$$

其中

$$\mathbf{c_B} = \begin{bmatrix} \mathbf{b}_0 + \sum_{j=1}^{p} \mathbf{B}_j \mathbf{y}_{1-j} \\ \mathbf{b}_0 + \sum_{j=2}^{p} \mathbf{B}_j \mathbf{y}_{2-j} \\ \vdots \\ \mathbf{b}_0 + \mathbf{B}_p \mathbf{y}_0 \\ \mathbf{b}_0 \\ \vdots \\ \mathbf{b}_0 \end{bmatrix}, \ \mathbf{H_B} = \begin{bmatrix} \mathbf{I}_n & \mathbf{0}_{n\times n} & \cdots & \cdots & \cdots & \cdots & \cdots & \mathbf{0}_{n\times n} \\ -\mathbf{B}_1 & \mathbf{I}_n & \mathbf{0}_{n\times n} & \cdots & \cdots & \cdots & \cdots & \mathbf{0}_{n\times n} \\ -\mathbf{B}_2 & -\mathbf{B}_1 & \mathbf{I}_n & \mathbf{0}_{n\times n} & \cdots & & & \mathbf{0}_{n\times n} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ -\mathbf{B}_p & \cdots & & -\mathbf{B}_1 & \mathbf{I}_n & \mathbf{0}_{n\times n} & & \vdots \\ \mathbf{0}_{n\times n} & & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0}_{n\times n} & \cdots & \mathbf{0}_{n\times n} & -\mathbf{B}_p & \cdots & -\mathbf{B}_2 & -\mathbf{B}_1 & \mathbf{I}_n \end{bmatrix}. \tag{7}$$

再次利用（3）中的恒等式，我们得到

$$\mathbf{H_B S}^o \mathbf{y}^o + \mathbf{H_B S}^m \mathbf{y}^m = \mathbf{c_B} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_{Tn}, \boldsymbol{\Sigma}),$$

其形式与（4）相同，其中 $\mathbf{G}^o = \mathbf{H_B S}^o$、$\mathbf{G}^m = \mathbf{H_B S}^m$、$\mathbf{X} = \mathbf{I}_{Tn}$ 以及 $\boldsymbol{\beta} = \mathbf{c_B}$.[4]

利用（4）中的表达式，接下来我们推导 $\mathbf{y}^m$ 在给定 $\mathbf{y}^o$ 以及其他模型参数和潜变量时的条件分布，并将这些量统称为 $\boldsymbol{\theta}$。直观地说，由于 $(\mathbf{y}^o, \mathbf{y}^m)$ 在给定 $\boldsymbol{\theta}$ 的条件下服从高斯联合分布，根据高斯分布的性质，$\mathbf{y}^m$ 在给定 $\mathbf{y}^o$ 和 $\boldsymbol{\theta}$ 的条件下也服从高斯分布。更确切地说，由（4）可知，$p(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta})$ 可表示为

$$\begin{aligned} p(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta}) &\propto \exp\left\{ -\frac{1}{2} \left(\mathbf{G}^m \mathbf{y}^m + \mathbf{G}^o \mathbf{y}^o - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} \right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{G}^m \mathbf{y}^m + \mathbf{G}^o \mathbf{y}^o - \mathbf{W}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta} \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[\mathbf{y}^{m\prime} \mathbf{G}^{m\prime} \boldsymbol{\Sigma}^{-1} \mathbf{G}^m \mathbf{y}^m - 2 \mathbf{y}^{m\prime} \mathbf{G}^{m\prime} \boldsymbol{\Sigma}^{-1} (\mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} - \mathbf{G}^o \mathbf{y}^o) \right] \right\}. \end{aligned}$$

接下来，令 $\mathbf{K}_{\mathbf{y}^m} = \mathbf{G}^{m\prime} \boldsymbol{\Sigma}^{-1} \mathbf{G}^m$，由于 $\mathbf{G}^m$ 列满秩，$\mathbf{K}_{\mathbf{y}^m} = \mathbf{G}^{m\prime} \boldsymbol{\Sigma}^{-1} \mathbf{G}^m$ 是一个 $N_m \times N_m$ 非奇异矩阵，因而可逆。此外，令 $\boldsymbol{\mu}_{\mathbf{y}^m} = \mathbf{K}_{\mathbf{y}^m}^{-1} \mathbf{G}^{m\prime} \boldsymbol{\Sigma}^{-1} (\mathbf{W}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} - \mathbf{G}^o \mathbf{y}^o)$。然后，对 $\mathbf{y}^m$ 进行配方，可将条件

[4]为方便记号，在推导过程中我们以初始条件 $\mathbf{y}_0$、……、$\mathbf{y}_{1-p}$ 为条件。这些初始条件可能包含缺失数据，但可以在单独的步骤中对其进行抽样。由于在大多数应用中，p 远小于 T，这一额外步骤的计算开销通常微不足道。或者，也可以通过重新定义 $\mathbf{y}$ 及其相关矩阵，将初始条件中的缺失数据与样本中的缺失数据联合抽样。

分布写为

$$
\begin{aligned}
p(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta}) &\propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}^{m\prime}\mathbf{K}_{\mathbf{y}^m}\mathbf{y}^m - 2\mathbf{y}^{m\prime}\mathbf{K}_{\mathbf{y}^m}\boldsymbol{\mu}_{\mathbf{y}^m}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\mathbf{y}^m - \boldsymbol{\mu}_{\mathbf{y}^m}\right)'\mathbf{K}_{\mathbf{y}^m}\left(\mathbf{y}^m - \boldsymbol{\mu}_{\mathbf{y}^m}\right)\right\}.
\end{aligned}
$$

因此，我们证明了给定观测数据时，缺失数据的联合条件分布为高斯分布，其均值向量为 $\boldsymbol{\mu}_{\mathbf{y}^m}$ ，精度矩阵为 $\mathbf{K}_{\mathbf{y}^m}$:

$$
(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}^m}, \mathbf{K}_{\mathbf{y}^m}^{-1}\right). \tag{8}
$$

由于 $\mathbf{G}^m$ 和 $\boldsymbol{\Sigma}$ 都是带状矩阵，因此精度矩阵 $\mathbf{K}_{\mathbf{y}^m}$ 也是带状矩阵。因此，我们可以使用 Chan and Jeliazkov（2009年）的基于精度矩阵的抽样器，高效地抽取 $\mathbf{y}^m$ 。我们在算法 1 中概述该抽样器。

算法 1 抽取 $(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}^m}, \mathbf{K}_{\mathbf{y}^m}^{-1}\right)$。

给定 $\boldsymbol{\mu}_{\mathbf{y}^m}$ 和 $\mathbf{K}_{\mathbf{y}^m}$，完成以下步骤。

1. 求出 $\mathbf{K}_{\mathbf{y}^m}$ 的 Cholesky 因子 $\mathbf{C}$，使得 $\mathbf{K}_{\mathbf{y}^m} = \mathbf{C}\mathbf{C}'$。
2. 通过回代求解 $\mathbf{C}'\mathbf{v} = \mathbf{x}$，得到 $\mathbf{v}$，其中 $\mathbf{x} \sim \mathcal{N}(\mathbf{0}_{N^m}, \mathbf{I}_{N^m})$。
3. 返回 $\mathbf{y}^m = \boldsymbol{\mu}_{\mathbf{y}^m} + \mathbf{v}$。

本文重点讨论使用马尔可夫链蒙特卡罗方法进行贝叶斯估计。但上述结果也适用于其他估计方法。例如，式 (8) 中的条件分布可与期望最大化算法结合使用，以获得 $\boldsymbol{\theta}$ 的最大似然估计。或者，也可以直接最大化观测数据似然，该似然可利用恒等式 $p(\mathbf{y}^o \mid \boldsymbol{\theta}) = p(\mathbf{y}^o, \mathbf{y}^m \mid \boldsymbol{\theta}) / p(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta})$ 进行计算。上述结果表明，右侧的两个密度均为高斯分布，且可快速计算。

目前为止的设定适用于存在缺失数据模式的情形，例如非平衡面板和不规则边缘。然而，在许多情况下，研究者掌握有关缺失数据的额外信息。一个典型例子是混频模型，其中低频流量变量的高频观测值被视为

缺失数据，而这些缺失观测值与多个低频观测值相关联。接下来，我们将说明如何利用额外信息更新缺失数据的条件分布。

2.2 基于额外信息进行条件化

上一节讨论了如何在给定观测数据 $\mathbf{y}^o$ 和模型参数 $\boldsymbol{\theta}$的条件下，高效抽样缺失数据向量 $\mathbf{y}^m$ 。许多应用中都能获得额外信息，而且通常需要或有必要将新信息纳入分析。例如，在同时包含每月变量和季度变量的混频设定中，一种常见做法是将季度流量变量的每月观测值视为缺失值，再通过某些跨期约束将这些缺失值映射到观测值。另一个例子是参差边缘设定：某些变量的最新值尚未发布，但已有高质量的即时预测（例如来自专业预测者调查的预测）。下文将说明如何调整所提出的抽样方法，以处理各种包含额外信息的设定。

为使所提出的框架具有通用性，能够处理各种信息设定，假设有一个额外的 $M \times 1$ 可观测变量向量（记为 $\mathbf{z}$），可用于改进对 $\mathbf{y}^m$的推断。我们考虑两种将 $\mathbf{z}$ 与 $\mathbf{y}^m$联系起来的映射。第一种情况是，假设 $\mathbf{z}$ 可以通过以下线性方程组精确映射到缺失数据：

$$\mathbf{z} = \mathbf{M}\mathbf{y}^m, \tag{9}$$

其中，$\mathbf{M}$ 是一个 $M \times N^m$矩阵，用于指定 M精确线性关系。我们将这类额外信息称为硬约束。一个重要的例子是基于 Mariano 和 Murasawa（2003，2010）提出的对数线性近似、适用于混频设定的跨期约束。更具体地，假设 $y_{i,\ t}^m$ 是第 i 个变量在第 t 月缺失的每月值。令 $z_{i,\ t/3}$ 表示相应的已观测季度值（注意，$z_{i,\ t/3}$ 仅在每三个月中的一个月被观测到）。那么，对该季度变量的算术平均值进行标准对数线性近似，可表示为

如下：

$$z_{i,t/3} = \frac{1}{3}y^m_{i,t} + \frac{2}{3}y^m_{i,t-1} + y^m_{i,t-2} + \frac{2}{3}y^m_{i,t-3} + \frac{1}{3}y^m_{i,t-4} \tag{10}$$

对于 $t = 3$、6、9、.....。通过将 (10) 叠加并适当定义 M，可将 (10) 中的精确线性约束写成 (9) 的形式。对于平衡的月度和季度变量，$M = N^m/3$。

尽管 (10) 中考虑的映射在技术上基于对数线性近似，但在大多数应用研究中，它被视为精确的线性关系。更恰当的做法或许是明确允许存在测量误差或近似误差。此外，在其他一些情形下，允许存在测量误差也是合适的（例如，将专业预测人员的即时预测映射到基础内生变量）。因此，我们考虑一种包含测量误差的替代映射，其形式如下：

$$\mathbf{z} = \mathbf{M}\mathbf{y}^m + \boldsymbol{\varepsilon}^z, \quad \boldsymbol{\varepsilon}^z \sim \mathcal{N}(\mathbf{0}_M, \mathbf{O}), \tag{11}$$

其中，O 是固定的对角协方差矩阵，用于表示测量误差的大小。我们将这类额外信息称为软约束。

在介绍了纳入额外信息的一般框架后，接下来讨论如何根据这一新的信息集调整缺失数据的抽样。首先考虑硬约束的情况。回顾一下，仅以观测数据和模型参数为条件时，缺失数据服从 (8) 所指定的高斯分布。因此，以 (9) 中的精确线性约束为条件对缺失数据进行抽样，相当于从退化高斯分布 $\mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}^m}、\mathrm{K}^{-1}_{\mathbf{y}^m}\right)$ 1(My^m =z) 中抽样，其中 1(·) 是指示函数。可以使用高效算法从 $\mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}^m}、\mathrm{K}^{-1}_{\mathbf{y}^m}\right)$，使得 My^m =z 中抽样，例如 Rue and Held (2005) 中的算法 2.6，以及 Cong、Chen 和 Zhou (2017) 中的算法 2。具体而言，我们可以先使用算法 1 从 u $\sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}^m}、\mathrm{K}^{-1}_{\mathbf{y}^m}\right)$ 中抽样。然后，通过计算以下内容，更新加入 z =My^m 后的条件集合

$$\mathbf{y}^m = \mathbf{u} + \mathbf{K}^{-1}_{\mathbf{y}^m}\mathbf{M}'(\mathbf{M}\mathbf{K}^{-1}_{\mathbf{y}^m}\mathbf{M}')^{-1}(\mathbf{z} - \mathbf{M}\mathbf{u}).$$

可以证明，y^m 服从分布 (y^m | y^o, $\boldsymbol{\theta}$, My^m =z)。Rue and Held（2005）中的算法 2 描述了一种高效的实现方法，可避免显式计-

算 $\mathbf{K}_{\mathbf{y}^m}$ 或 $\mathbf{M}\mathbf{K}_{\mathbf{y}^m}^{-1}\mathbf{M}'$的逆。使用该实现，对 $\mathbf{z} = \mathbf{M}\mathbf{y}^m$ 进行条件化所需的额外计算成本对于 $M \ll N^m$而言相对较低。对于较大的M，该算法会涉及几个大型稠密矩阵，计算量可能会更大。

算法 2 在硬约束下抽样 $(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta}, \mathbf{z} = \mathbf{M}\mathbf{y}^m)$，其中 $(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}^m}, \mathbf{K}_{\mathbf{y}^m}^{-1}\right)$。

给定参数 $\boldsymbol{\mu}_{\mathbf{y}^m}$ 和 $\mathbf{K}_{\mathbf{y}^m}$，完成以下步骤。

1. 求得 $\mathbf{K}_{\mathbf{y}^m}$ 的 Cholesky 因子 $\mathbf{C}$，使得 $\mathbf{K}_{\mathbf{y}^m} = \mathbf{C}\mathbf{C}'$。2. 使用算法 1 抽样 $\mathbf{u} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{y}^m}, \mathbf{K}_{\mathbf{y}^m}^{-1}\right)$。3. 通过前向和后向代入求解 $\mathbf{C}\mathbf{C}'\mathbf{U} = \mathbf{M}'$，得到 $\mathbf{U}$。4. 求解 $\mathbf{M}\mathbf{U}\mathbf{V} = \mathbf{U}'$，得到 $\mathbf{V}$。5. 返回 $\mathbf{y}^m = \mathbf{u} + \mathbf{V}'(\mathbf{z} - \mathbf{M}\mathbf{u})$。

接下来，我们考虑软约束的情形。实质上，我们根据 (11) 中指定的新信息更新缺失数据 $\mathbf{y}^m$ 的条件分布。因此，可以将 (8) 中 $\mathbf{y}^m$ 的原始高斯分布视为“先验分布”，将 (11) 中的新信息视为“似然”。然后，通过标准的贝叶斯更新，我们得到

$$(\mathbf{y}^m \mid \mathbf{y}^o, \boldsymbol{\theta}, \mathbf{z}) \sim \mathcal{N}\left(\overline{\boldsymbol{\mu}}_{\mathbf{y}^m}, \overline{\mathbf{K}}_{\mathbf{y}^m}^{-1}\right), \tag{12}$$

其中

$$\overline{\mathbf{K}}_{\mathbf{y}^m} = \mathbf{M}'\mathbf{O}^{-1}\mathbf{M} + \mathbf{K}_{\mathbf{y}^m}, \quad \overline{\boldsymbol{\mu}}_{\mathbf{y}^m} = \overline{\mathbf{K}}_{\mathbf{y}^m}^{-1}\left(\mathbf{M}'\mathbf{O}^{-1}\mathbf{z} + \mathbf{K}_{\mathbf{y}^m}\boldsymbol{\mu}_{\mathbf{y}^m}\right).$$

由于在大多数应用中，矩阵 $\mathbf{M}$、$\mathbf{O}$ 和 $\mathbf{K}_{\mathbf{y}^m}$ 均为带状矩阵，因此精度矩阵 $\mathbf{K}_{\mathbf{y}^m}$也为带状矩阵。因此，可以直接应用算法 1 中的基于精度矩阵的抽样器，高效地抽取 $\mathbf{y}^m$；我们只需分别用 $\mathbf{K}_{\mathbf{y}^m}$ 和 $\overline{\boldsymbol{\mu}}_{\mathbf{y}^m}$替换 $\mathbf{K}_{\mathbf{y}^m}$ 和 $\boldsymbol{\mu}_{\mathbf{y}^m}$。

与针对硬约束情形的算法 2 相比，从 (12) 抽样的速度快得多，而且在高维情形下表现良好。对于 Mariano 和 Murasawa（2003，2010）这类近似跨期约束，后者的抽样器自然更可取。对于其他精确的跨期约束，实证上可以通过将 $\mathbf{O}$ 的对角元素设得非常小（例如 10^{-8}），来近似这些硬约束。

3 蒙特卡洛研究

本节开展一系列蒙特卡洛实验，以评估所提出的基于精度矩阵的方法在抽取潜在缺失观测值方面相较于基于卡尔曼滤波的方法的速度和准确性。在第一小节中，我们考察混频设定，其中缺失数据是低频变量的高频观测值。随后，我们在下一小节中考察非平衡面板。

所有数据集均由以下向量自回归模型生成：

$$\mathbf{y}_t = \mathbf{b}_0 + \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma}),$$

其中，$\mathrm{y}_t = (\mathrm{y}_t^{o\prime}, \mathrm{y}_t^{m\prime})'$是一个混合频率数据向量，$\mathrm{y}_t^o$是一个高频变量（已观测）向量，$\mathrm{y}_t^m$是一个低频变量的高频观测值（缺失）向量。此外，低频变量 $z_{i,\ t/3}, i = 1, \ldots, n^m$在 $t = 3, 6, 9, \ldots$ 时被观测到，可通过式（9）或式（11）为缺失的 y_t^m 提供信息。对于基准情形，我们设定 $p = 5$。此外，我们按如下方式生成模型参数。令 $\mathbf{b}_0 = 0.01 \times \mathbf{1}_n$。第一个向量自回归系数矩阵的对角元素独立同分布于均匀分布 $\mathcal{U}(0, 0.5)$，而非对角元素服从$\mathcal{U}(-0.2, 0.2)$。较高阶向量自回归系数矩阵的所有元素均独立同分布于 $\mathcal{N}(0, 0.05^2/l^2)$，其中 l 为滞后阶数。误差协方差矩阵 $\boldsymbol{\Sigma}$ 由逆 Wishart 分布 $\mathcal{IW}(n + 10, 0.07\mathrm{I}_n + 0.03\mathbf{1}_n\mathbf{1}_n')$生成。

对于每个模拟数据集 $r = 1, \ldots, R$，我们使用 4 种方法估计缺失观测值 y_t^m：采用式（9）中硬性跨期约束的基于精度矩阵的抽样器、采用式（11）中软约束的基于精度矩阵的抽样器、Carter 和 Kohn（1994）的模拟平滑器（由 Schorfheide 和 Song（2015）提供的代码实现），以及 Durbin 和 Koopman（2002）的模拟平滑器。[5] 对于 Carter 和 Kohn（1994）和 Durbin 和 Koopman（2002）的模拟平滑器，我们

[5] Schorfheide 和 Song（2015）的实现采用紧凑状态空间表示来抽取缺失观测值。该表示将每月观测值从测量方程中的状态向量中移除。因此，它降低了状态向量的维度，通常也更有效率。相比之下，Durbin 和 Koopman（2002）的模拟平滑器要求模型采用标准的伴随形式（详情见附录B），这会导致状态向量的维度更高。

施加硬约束。对于采用软约束的基于精度矩阵的抽样器，我们将测量误差协方差矩阵 $\mathbf{O}$ 的对角元素设为 10^{-8}。因此，我们将其视为硬约束的一种近似，以便比较这 4 种方法的结果。

最后，我们为 VAR 系数采用标准正态先验，为误差协方差矩阵采用逆 Wishart 先验。具体而言，设 $\boldsymbol{\beta} = \text{vec}\left([\mathbf{b}_0, \mathbf{B}_1, \ldots, \mathbf{B}_p]'\right)$ 表示所有 VAR 系数构成的 $k \times 1$ 向量，其 $k = n(np+1)$。那么，先验分布为$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}_k, \mathbf{I}_k)$ 和 $\boldsymbol{\Sigma} \sim \mathcal{IW}(5, \mathbf{I}_n)$。

3.1 低频变量的缺失观测值

我们考虑维度各异的数据生成过程，其中 $T = 300$：小型（$n = 6, n^o = 5, n^m = 1$）、中型（$n = 11, n^o = 10, n^m = 1$）和大型（$n = 16, n^o = 15, n^m = 1$）。我们还研究了未观测变量数量较多且 $n^m = 5$ 的情形。对于每种设计，我们按照上述方法从混频 VAR 中生成 $R = 100$个数据集。随后，我们使用吉布斯采样器对每个数据集进行拟合，该采样器依次抽取缺失观测值和模型参数。具体而言，我们采用 4 种不同方法对缺失观测值进行抽样。为评估这些方法的准确性，我们计算估计的缺失观测值与实际值之间的均方误差（MSE）。具体来说，对于每个含有缺失观测值 $\mathbf{y}_1^{m(i)}, \ldots, \mathbf{y}_T^{m(i)}, i = 1, \ldots, R$ 的数据集，以及每种后验均值向量为 $\widehat{\mathbf{y}}^{m(i,j)}, j = 1, \ldots, 4$ 的方法，我们计算 $\text{MSE}_i(\widehat{\mathbf{y}}^{m(i,j)}) = \sum_{t=1}^{T} ||\mathbf{y}_t^{m(i)} - \widehat{\mathbf{y}}_t^{m(i,j)}||^2/T$，其中 $||\cdot||$ 为 ℓ^2-范数。我们进一步对每种设计的 R 个均方误差取平均，以概括准确性，结果见表1。

我们还报告了基于 15,000 次马尔可夫链蒙特卡洛抽样（预热期为 5,000 次抽样）的计算时间。[6] 由于四种方法的目标都是从同一分布中抽样——即给定观测数据和模型参数时缺失观测值的条件分布——原则上它们应得到相同的均方误差。事实上，在我们的模拟中，它们得到的均方误差非常相近；细微差异主要源于数值误差和模拟误差。就运行时间而言，显然，在多种设定下，所提出的基于精度矩阵的方法

[6]计算时间基于一台标准台式机，该电脑配备 Intel Xeon W-2223@ 3.6GHz 处理器和 16 GB 内存，代码使用 MATLAB 实现。

比基于卡尔曼滤波的方法具有更高的计算效率。

表1：使用四种方法估计缺失观测值的均方误差和计算时间：所提出的带有跨期硬约束的基于精度矩阵的方法（P-hard）、带有跨期软限制的基于精度矩阵的方法（P-soft）、Schorfheide 和 Song（2015）中实现的 Carter 和 Kohn（1994）模拟平滑器（CK），以及 Durbin 和 Koopman（2002）模拟平滑器（DK）。

		MSE				计算时间（分钟）			
n^m	n^o	P-hard	P-soft	DK	CK	P-hard	P-soft	DK	CK
1	5	0.004	0.004	0.004	0.005	0.7	0.6	7	5
1	10	0.004	0.004	0.004	0.004	3	3	31	9
1	15	0.004	0.004	0.004	0.004	13	13	61	16
5	5	0.005	0.005	0.005	0.005	8	4	23	24
5	10	0.005	0.005	0.005	0.005	16	12	51	35
5	15	0.004	0.004	0.004	0.005	38	35	106	51

表1报告了完整的MCMC估计的运行时间。随着向量自回归模型的维度增加，模拟VAR系数的计算复杂度占据主导地位，这可能会给人一种四种方法的运行时间趋于一致的印象。为了更好地了解所提出的方法在更广泛设置下的表现，下面我们仅比较抽样缺失观测值的运行时间。

首先，图1报告了在一系列 n^o 和 n^m 下，使用四种方法对缺失观测值进行十次抽样的运行时间。显然，两种基于精度矩阵的方法相比基于卡尔曼滤波的方法表现良好，并且都能很好地扩展到高维设置。此外，当存在大量变量具有缺失观测值时，带有软限制的变体尤其高效。

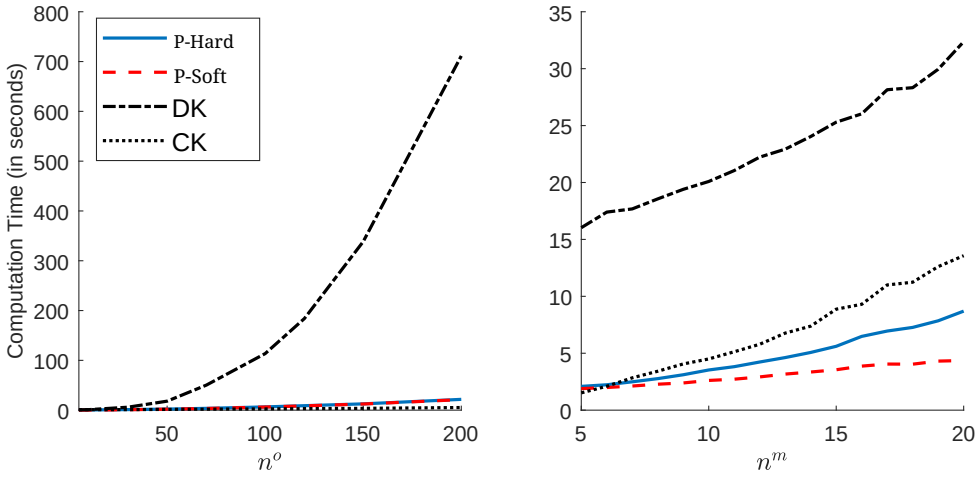


图1：抽取10个样本的计算时间与 n^o 和 n^m的关系，其中二者分别表示观测变量和部分未观测变量的数量； $T = 300$ 和 $p = 5$。四种方法分别为：采用硬性跨期约束的基于精度矩阵的抽样器（P-hard）、采用软约束的基于精度矩阵的抽样器（P-soft）、Schorfheide 和 Song（2015）实现的 Carter 和 Kohn（1994）模拟平滑器（CK），以及 Durbin 和 Koopman（2002）的模拟平滑器。

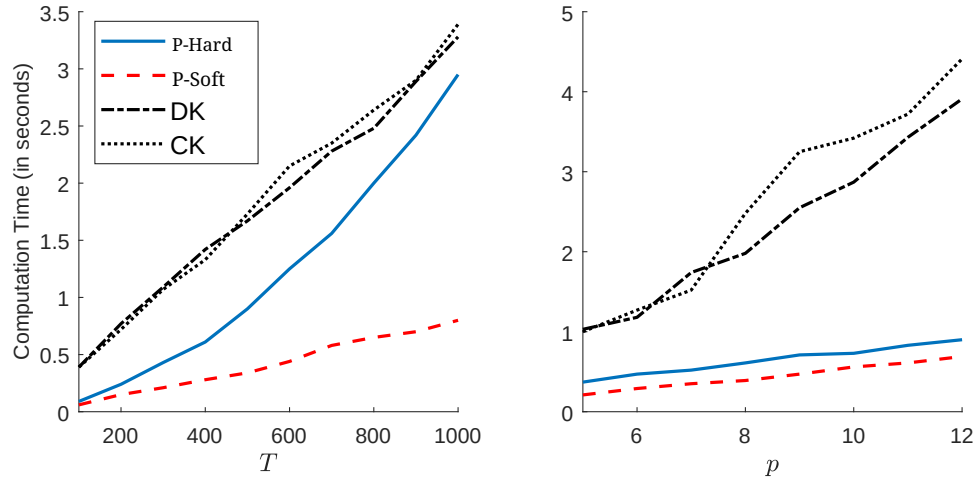


图2：抽取10个样本的计算时间与 T 和 p的关系，其中二者分别表示时期数和滞后项数； $n^m = 5$ 和 $n^o = 10$。四种方法分别为：采用硬性跨期约束的基于精度矩阵的抽样器（P-hard）、采用软约束的基于精度矩阵的抽样器（P-soft）、Schorfheide 和 Song（2015）实现的 Carter 和 Kohn（1994）模拟平滑器（CK），以及 Durbin 和 Koopman（2002）的模拟平滑器。

接下来，图2报告了在不同样本量 T 和滞后阶数 p下，对缺失观测值进行十次抽样的运行时间。虽然两种基于精度矩阵的方法都表现良好，但采用软约束的版本表现明显更好，并且能够很好地扩展到非常大的 T 和 p。还值得一提的是，应用卡尔曼滤波器时，需要重新定义状态，使观测方程仅依赖于当前的（重新定义后的）状态。当 p 较大时，这个新状态向量的维度也很大。这也是基于卡尔曼滤波的方法在 p 较大时计算成本更高的原因之一。相比之下，即使滞后阶数很大，基于精度矩阵的方法的计算成本仍然较低。

3.2 非平衡面板

本节通过涉及非平衡面板的情形，展示所提出的基于精度矩阵的方法的表现；在这类面板中，不同变量在不同的时期缺失。为此，我们使用中等规模的向量自回归模型（$n = 13$， $n^o = 10$， $n^m = 3$），并设定样本量为 $T = 300$，开展一系列模拟。我们使用与上文所述相同的数据生成过程来模拟数据，但此处假设低频变量在选定时期也会缺失（即除其缺失的高频观测值外，还会缺失这些低频观测值）。具体而言，第一个低频变量在最初30期完全缺失；第二个变量从 $t = 150$ 到180期缺失；第三个变量则在最后30期缺失。

表2：使用四种方法进行 $R = 100$ 次重复模拟后，估计缺失观测值的均方误差和平均计算时间：采用硬性跨期约束的拟议基于精度矩阵的方法（P-hard）、采用软跨期约束的基于精度矩阵的方法（P-soft）、Schorfheide 和 Song（2015）实现的 Carter 和 Kohn（1994）模拟平滑器（CK），以及 Durbin 和 Koopman（2002）的模拟平滑器（DK）。

		MSE				计算时间（分钟）			
n^m	n^o	P-hard	P-soft	DK	CK	P-hard	P-soft	DK	CK
3	10	0.004	0.004	0.004	0.009	8	7	40	14

表2报告了估计三个低频变量缺失观测值的均方误差和计算时间。与之前的模拟实验类似，我们在这里也发现，四种方法的均方误差非常接近。与之前一样，

所提出的基于精度矩阵的方法在计算上比两种基于卡尔曼滤波的模拟平滑器都更高效。

接下来，图3将采用四种方法得到的低频变量缺失观测值的后验均值与实际模拟数据进行对比。图中突出显示了每个低频变量完全缺失的时期。正如预期的那样，在这些时期，四种方法对缺失观测值的后验估计非常相似。因此，我们得出结论：基于精度矩阵的方法能够像标准滤波和平滑方法一样处理任意缺失数据模式，而且计算效率更高。

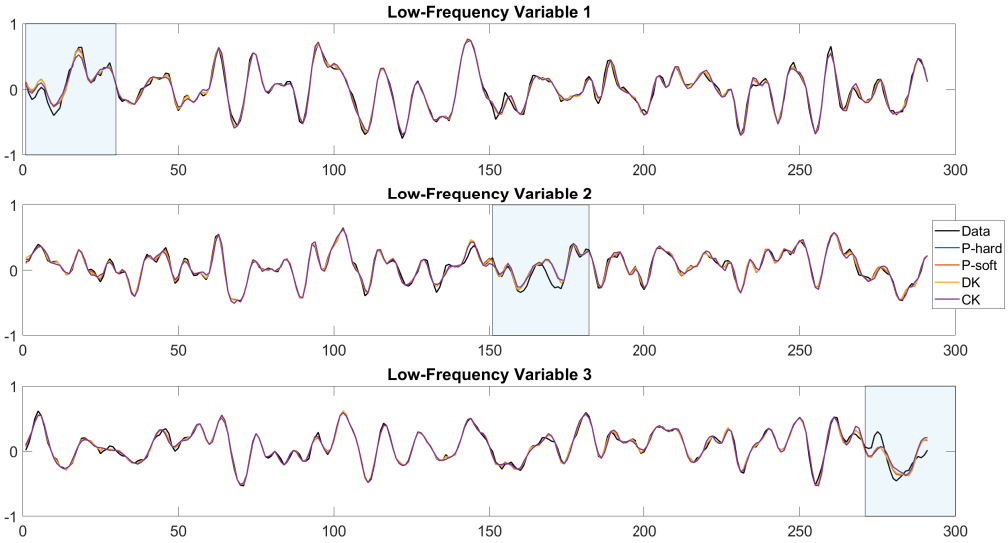


图3：使用四种方法得到的低频变量后验均值：所提出的采用硬性跨期约束的基于精度矩阵的方法（P-hard）、采用软跨期约束的基于精度矩阵的方法（P-soft）、Schorfheide 和 Song（2015）实现的 Carter 和 Kohn（1994）模拟平滑器（CK），以及 Durbin 和 Koopman（2002）的模拟平滑器（DK）。每个面板中的方框标出了低频变量完全缺失的时期。

4 实证应用

我们通过两个实证应用展示所提出的基于精度矩阵的方法，这两个应用涉及两个广受欢迎的大型模型和差异显著的缺失数据模式。在第一个应用中，我们使用一个带随机波动率的大型混频 VAR 来生成实际国内生产总值的每周估计值。在第二个应用中，我们使用带随机波动率的动态因子模型，从非平衡数据集中提取潜在因子。

4.1 每周状态空间混合频率VAR

在第一个应用中，我们展示如何使用所提出的基于精度矩阵的方法，估计一个包含22个变量、频率涵盖每周、每月和季度的贝叶斯 VAR。遵循 Schorfheide 和 Song（2015）的开创性工作，我们在最高观测频率下对所有变量进行建模，并将低频变量在高频下的观测值视为缺失数据。这一方法的一个关键优势是，它能够以更高频率生成低频变量的插值（历史）估计值。例如，Schorfheide 和 Song（2015）使用每月和季度变量来获得月度国内生产总值估计值，而这些估计值是各种应用所需的输入。此外，该方法还可以通过纳入高频变量中的信息，提供更加及时的即时预测。近期采用这一方法的应用包括 Brave、Butters 和 Justiniano（2019），以及 Koop、McIntyre、Mitchell 和 Poon（2020年、2022年）。

我们通过纳入每周宏观经济变量和金融变量，拓展了这一研究方向。通过以周频对所有变量建模，我们能够获得每周国内生产总值估计值。然而，拟合大型混频 VAR 的计算量很大，因为缺失观测值数量众多。此外，由于缺失数据模式不规则（例如，每个季度或每个月包含的周数并不总是相同），传统方法的实现也更加复杂。相比之下，所提出的方法能够轻松处理不规则缺失数据模式，并且可扩展至高维情形。

美国数据集包含16个每周变量（包括粗钢产量、零售额、首次申领失业救济人数以及各种金融变量）、5个月度

变量（如工业生产、消费者价格指数和劳动力市场变量）以及1个季度变量（实际国内生产总值），时间范围为2013年1月至2022年8月。其中，7个每周变量取自 Lewis、Mertens、Stock 和 Trivedi（2022），他们利用这些变量构建了周度经济指数（WEI）；其他变量则取自圣路易斯联邦储备银行的FRED数据库。有关数据的更多详情，请参见附录A。

由于我们的样本涵盖了COVID-19疫情，因此，从实证角度来看，如Hartwig（2021年）、Lenza和Primiceri（2022年）以及Carriero、Clark、Marcellino 和 Mertens（2022年）等近期论文所示，允许存在某种形式的异方差性或非高斯误差十分重要。因此，我们将Carriero、Clark和Marcellino（2016年）提出的共同随机波动率纳入混频 VAR，具体如下：

$$\mathbf{y}_t = \mathbf{b}_0 + \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathrm{e}^{h_t}\boldsymbol{\Sigma}),$$

其中，$\mathrm{y}_t = (\mathrm{y}_t^{o\prime}, \mathrm{y}_t^{m\prime})'$ 是一个由混频数据构成的 $n \times 1$ 向量，y_t^o和y_t^m分别是观测变量和缺失变量的向量。请注意，误差协方差矩阵按共同对数波动率 h_t进行缩放，该波动率被建模为平稳一阶自回归过程：

$$h_t = \phi h_{t-1} + u_t^h, \quad u_t^h \sim \mathcal{N}(0, \sigma_h^2),$$

对于 $t = 2$，……，T，其中 $|\phi| < 1$，初始条件设定为 $h_1 \sim \mathcal{N}(0, \sigma_h^2/(1-\phi^2))$. 在此处，为便于识别，假设该一阶自回归过程的无条件均值为零。[7]

接下来，我们描述模型参数 $\boldsymbol{\beta} = \mathrm{vec}([\mathrm{b}_0, \mathrm{B}_1, \ldots, \mathrm{B}_p]')$，$\boldsymbol{\Sigma}$, ϕ和 σ_h^2的先验分布。具体而言，我们假设以下相互独立的先验分布：

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}}),\ \boldsymbol{\Sigma} \sim \mathcal{IW}(\nu_0, \mathbf{S}_0),\ \phi \sim \mathcal{N}(\phi_0, V_\phi)1(|\phi| < 1),\ \sigma_h^2 \sim \mathcal{IG}(\eta_0, S_0),$$

其中 1(·) 表示 表示指示函数。先验均值向量和协方差矩阵

[7]在初步研究中，我们还实现了一个模型版本，其中采用了 Cogley 和 Sargent（2005）以及 Carriero、Clark 和 Marcellino（2019）提出的更灵活的 Cholesky随机波动率；在该模型中，每个变量都有各自的随机波动率过程。然而，鉴于缺失观测值数量较多，我们发现，在模拟中难以准确确定某些随机波动率过程。相比之下，共同随机波动率表现良好。共同随机波动率的另一个优点是其具有顺序不变性，而 Cholesky随机波动率则不具备这一性质。

VAR系数的先验均值向量和协方差矩阵分别为 $\boldsymbol{\beta}_0$ 和 $\mathbf{V}_\beta$，其设定遵循 Doan、Litterman 和 Sims（1984）以及 Litterman（1986）开创的明尼苏达先验。更具体地说，我们将 $\boldsymbol{\beta}_0 = \mathbf{0}_k$ 设为使系数向零收缩。对于先验协方差矩阵 $\mathbf{V}_\beta$，假设其为对角的，例如

$$\begin{aligned}
\operatorname{Var}(B_{l,ii}) &= \frac{\kappa_1}{l^2},\ l = 1, \ldots, p,\ i = 1, \ldots, n, \\
\operatorname{Var}(B_{l,ij}) &= \frac{\kappa_2 s_i^2}{l^2 s_j^2},\ l = 1, \ldots, p,\ i, j = 1, \ldots, n,\ i \neq j, \\
\operatorname{Var}(b_{0,i}) &= 100 s_i^2,\ i = 1, \ldots, n,
\end{aligned}$$

其中 $B_{l,\ ij}$ 是 $\mathbf{B}_l$ 的（i，j）元素，s_r^2 表示变量 r 在 $r = 1$、.....、n 上的 AR(4)模型残差的样本方差。为实现跨变量收缩，即相比自身滞后项，更强地将其他变量滞后项的系数收缩至零，我们设定 $\kappa_1 = 0.04$ 和 $\kappa_2 = 0.01$。最后，我们设定 $\nu_0 = n + 3$、$\mathbf{S}_0 = \mathbf{I}_n$、$\eta_0 = 10$和 $S_0 = 0.004$，使得 $\boldsymbol{\Sigma}$ 和 σ_h^2 的先验均值分别为 $\mathbf{0.5I}_n$ 和 0.021^2。

我们将所有变量建模为周频变量，而月度或季度变量的缺失观测值通过类似于 Mariano 和 Murasawa（2003，2010）中的跨期约束，与其对应的观测值相联系。与仅涉及季度和月度变量的标准混频设定相比，此处的跨期约束更为复杂，因为不同月份或季度的周数可能不同。具体而言，给定月度/季度变量的发布日期 t，令 $n_{i,\ t}^w$ 表示 t与上一个发布日期之间的周数。则每个观测到的月度/季度变量 $z_{i,\ t}$ 均通过以下跨期约束与缺失观测值 $y_{i,\ t}^m$ 相联系：

$$z_{i,t} = \sum_{s=1}^{2n_{i,t}^w - 1} \left(1(s \leqslant n_{i,t}^w) \frac{s}{n_{i,t}^w} + 1(s > n_{i,t}^w) \frac{2n_{i,t}^w - s}{n_{i,t}^w} \right) y_{i,t-s+1}^m + \varepsilon_{i,t}^z,$$

其中，$\varepsilon_{i,t}^z \sim \mathcal{N}(0, o_i)$ 刻画了对数线性近似误差，并且我们设定 $o_i = 10^{-8}$.

混频 VAR 使用马尔可夫链蒙特卡罗方法进行估计。具体而言，给定模型参数和共同随机波动率后，我们使用第2节所述的拟议抽样器，对月度和季度变量的缺失观测值进行抽样。然后，给定这些缺失观测值，即可使用标准算法来

抽样模型参数和共同随机波动率。[8] 为了评估后验抽样器的效率，我们计算了与缺失数据和模型参数相关的低效因子。所有低效因子均小于100（详见附录C），与不含缺失数据的大型向量自回归模型相当。

估计得到的周度GDP增长率波动较大，这并不意外，因为它们是按周频测量的。为便于比较和解读，我们将这些周度估计值转换为更熟悉的季度增长率。具体而言，给定估计得到的周环比GDP增长率 $y_{t,\ j}^{m},t\ =\ 1,\ldots,T$，我们利用跨期约束将其转换为季度增长率：

$$y_{t,j}^{*}=\sum_{s=1}^{n_{j}^{w}}\frac{s}{n_{j}^{w}}y_{t-s,j}^{m}+\sum_{s=n_{j}^{w}+1}^{2n_{j}^{w}-1}\frac{2n_{j}^{w}-s}{n_{j}^{w}}y_{t-s,j}^{m},\tag{13}$$

其中，我们固定 $n_{j}^{w}\ =\ 13$ 周。因此，$y_{t,\ j}^{*}$ 可解释为过去13周的累计GDP增长。图4绘出了这些每周总量GDP增长率的后验均值及相应的68%可信区间。图中还用黑色叉号标出了观测季度GDP增长率。正如预期，所有观测到的季度GDP数值都落在每周总量GDP估计值上——跨期约束确保周度GDP估计值汇总后等于观测到的季度值。

每周总量GDP估计值最显著的特征，是COVID-19疫情暴发之初的大幅下跌，以及随后的反弹。具体而言，疫情导致企业大范围停业，美国实际GDP在2020年第二季度下降了约37%。2020年第三季度，随着经济逐步重启，GDP迅速反弹了约30%。以周频对GDP进行建模的一项关键优势是，在季度GDP发布日期之间，模型能够结合其他周度和月度变量的信息，按周提供GDP估计值。

[8]例如，共同随机波动率可以使用Carriero、Clark和Marcellino（2016）或Chan（2020）的方法进行抽样。VAR系数服从联合高斯分布，可以像Carriero、Chan、Clark和Marcellino（2022）那样联合抽样，也可以逐方程抽样。

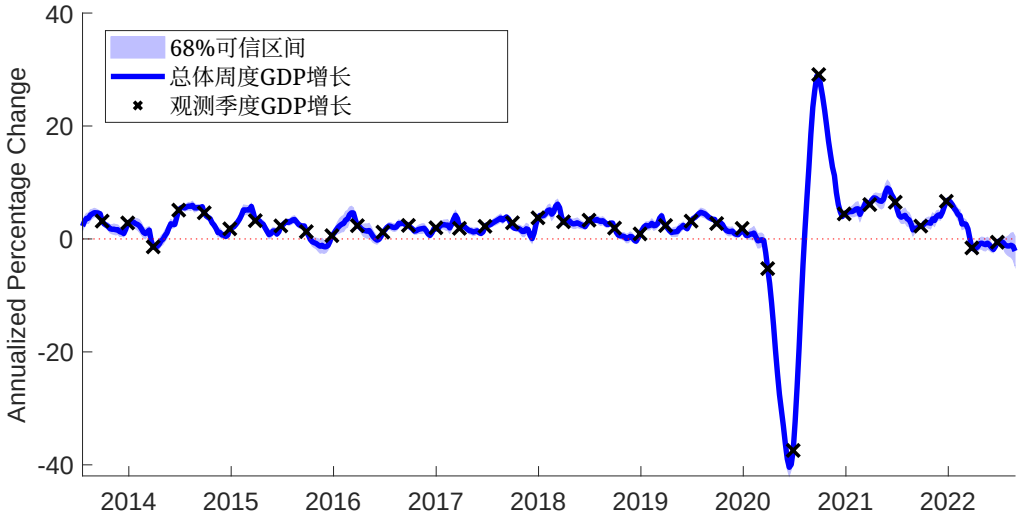


图4：一个季度内GDP增长的总体周度估计值。实线表示总体周度GDP增长的后验均值（年化），阴影区域表示相应的68%可信区间，十字标记表示实际观测季度GDP值。

接下来，我们将每周总量GDP估计值与两个旨在追踪实际经济活动的高频指标进行比较。第一个是 Lewis、Mertens、Stock 和 Trivedi（2022）的每周经济指数，由纽约联邦储备银行每周更新。第二个是 Aruoba、Diebold 和 Scotti（2009年）的商业状况指数，由费城联邦储备银行维护。虽然这两个指标都纳入了大量高观测频率的宏观经济和金融变量，但它们是动态因子模型中的潜在因子。相比之下，我们的混频 VAR 能够直接提供GDP估计值，也更容易解读。

我们分别从纽约联邦储备银行和费城联邦储备银行获取周频的周度经济指数和商业状况指数。随后，使用公式（13）将其转换为季度增长，以便进行比较。图5绘制了每周总量GDP估计值以及这两个指标。由图可见，即使在2020年第二季度和2020年第三季度的极端转折点期间，每周总量GDP也与商业状况指数保持密切一致。[9] 相比之下，周度经济

[9]虽然B usiness Conditions指数的构造使其平均值为零，但平均值

指数在COVID-19疫情爆发及随后立即反弹期间呈现出明显不同的动态。具体而言，周度经济指数表明，美国经济在2020年第二季度从大范围封锁中缓慢复苏。相比之下，每周总量GDP和商业状况指数显示出更为强劲的反弹。这种差异的一个潜在原因是，每周总量GDP和商业状况指数都纳入了季度GDP数据，而周度经济指数没有。因此，后者可能只能反映美国经济中特定行业的经济活动，而前两者则通过GDP数据中的信息跟踪整个美国经济。

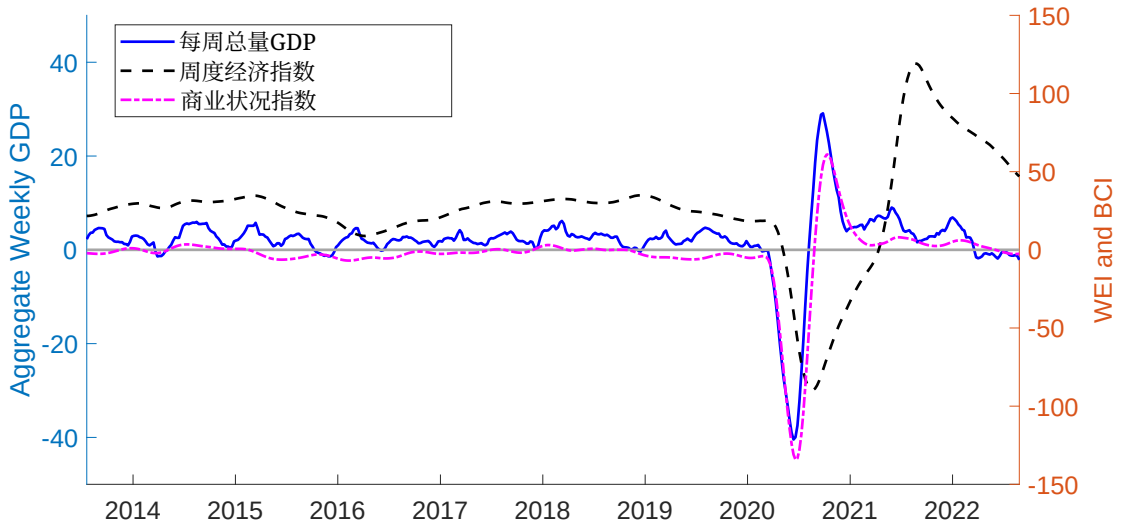


图5：每周总量GDP估计值（实线）、Lewis、Mertens、Stock 和 Trivedi（2022）的周度经济指数（虚线），以及 Aruoba、Diebold 和 Scotti（2009年）的商业状况指数（点划线）的图示。

4.2 非平衡数据集的动态因子模型

为展示所提出方法的通用性，在第二个应用中，我们在另一种适用于处理大型数据集的热门模型——动态因子模型——的背景下，考察一种不同类型的缺失数据模式。更具体地说，我们使用 McCracken 和 Ng（2016）的 FRED-MD 数据集，通过动态

样本期间的GDP增长约为2%。因此，这两个序列的水平存在差异。

因子模型（含随机波动率）提取共同因子。我们聚焦于 COVID-19 疫情时期，并使用2020年3月至2022年9月的数据版本实时估计共同因子。每个数据版本都包含 128 个月度变量，但其中许多变量存在缺失值。这些缺失值主要来自两个方面：一些较新构建的变量在样本初期缺少观测值，以及由于发布滞后而导致样本末期出现缺失值，这种情况通常称为参差边缘。

令 $\mathrm{y}_t = (\mathrm{y}_t^{o\prime}, \mathrm{y}_t^{m\prime})'$ 表示月度变量的 $n \times 1$ 向量，其中 y_t^o 是观测变量的 n_t^o 向量，y_t^m 是时点 t 存在缺失值的变量的 n_t^m 向量。此外，令 f_t 表示潜在动态因子的 $k \times 1$ 向量。由于 COVID-19 异常值的影响，我们在因子和特异误差中均纳入随机波动率，并考虑以下动态因子模型：

$$
\begin{aligned}
\mathbf{y}_t &= \mathbf{A}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \\
\boldsymbol{\varepsilon}_t &= \boldsymbol{\Psi}_1 \boldsymbol{\varepsilon}_{t-1} + \cdots + \boldsymbol{\Psi}_p \boldsymbol{\varepsilon}_{t-p} + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{\Sigma}_t), \\
\mathbf{f}_t &= \boldsymbol{\Phi}_1 \mathbf{f}_{t-1} + \cdots + \boldsymbol{\Phi}_q \mathbf{f}_{t-q} + \boldsymbol{\varepsilon}_t^{\mathbf{f}}, \quad \boldsymbol{\varepsilon}_t^{\mathbf{f}} \sim \mathcal{N}(\mathbf{0}_k, \boldsymbol{\Omega}_t),
\end{aligned}
$$

其中 $\boldsymbol{\Psi}_1$、……、$\boldsymbol{\Psi}_p$、$\boldsymbol{\Phi}_1$、……、$\boldsymbol{\Phi}_q$ 为对角矩阵，$\boldsymbol{\Sigma}_t = \mathrm{diag}(\mathrm{e}^{h_{1,t}}$、……、$\mathrm{e}^{h_{n,t}})$，$\boldsymbol{\Omega}_t = \mathrm{diag}(\mathrm{e}^{h_{n+1,t}}$、……、$\mathrm{e}^{h_{n+k,t}})$。假设 $n+k$ 对数波动率过程遵循相互独立的随机游走：

$$
h_{i,t} = h_{i,t-1} + \varepsilon_{i,t}^h, \quad \varepsilon_{i,t}^h \sim \mathcal{N}(0, \sigma_{h,i}^2),\ i = 1, \ldots, n+k,
$$

其中初始条件 $h_{1,0}$、……、$h_{n+k,0}$ 被视为未知参数。按照 Antolin-Diaz、Drechsel 和 Petrella（2021）的做法，我们设定 $p = q = 2$。

根据 Bai 和 Ng（2002）提出的 PC_p 准则，McCracken 和 Ng（2016年）发现，他们的数据集的最优因子数为 8。因此，我们设定 $k = 8$。为进行识别，假定因子载荷矩阵 A 为下三角矩阵，且对角线元素均设为 1。随后，我们谨慎选取前 8 个变量，以便于解释潜在因子。具体而言，我们借鉴 McCracken 和 Ng（2016年）的研究结果，使用在各个因子上的载荷最大的变量。[10] 他们的结果表明，第一因子解释了工业生产中相当大一部分

[10]这 8 个变量 分别是 ‘usgood’、‘t10yffm’、‘cusr0000sac’、‘aaa’、‘gs5’、‘ipcongd’、‘S&P: indust’ 和

波动以及许多劳动力市场变量的波动，说明它反映了整体经济状况。第二因子对利率差波动的解释力尤为突出，而第三和第四因子则分别对价格和利率的波动具有较强的解释力。

接下来，我们指定模型参数的先验分布。令 a 表示因子载荷矩阵 A 中的自由元素，并令 $\boldsymbol{\psi}$ 和 $\boldsymbol{\phi}$ 分别表示由 $\boldsymbol{\Psi}_i, i = 1, \ldots, p$ 和 $\boldsymbol{\Phi}_j, j = 1, \ldots, q$ 的对角元素组成的向量。然后，考虑 a、$\boldsymbol{\psi}$ 和 $\boldsymbol{\phi}$ 上的以下独立先验分布：

$$\mathbf{a} \sim \mathcal{N}(\mathbf{a}_0, \mathbf{V}_\mathbf{a}), \quad \boldsymbol{\psi} \sim \mathcal{N}(\boldsymbol{\psi}_0, \mathbf{V}_{\boldsymbol{\psi}}) 1(|\boldsymbol{\psi}| < 1), \quad \boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\phi}_0, \mathbf{V}_{\boldsymbol{\phi}}) 1(|\boldsymbol{\phi}| < 1),$$

其中，示性函数确保元素 ψ_i和 $\phi_j, i = 1, \ldots, np, j = 1, \ldots, kq$ 的绝对值小于 1。我们将先验均值 $\mathbf{a}_0$、$\boldsymbol{\psi}_0$ 和 $\boldsymbol{\phi}_0$ 设为零，并将先验协方差矩阵设为 $\mathbf{V}_\mathbf{a} = \mathbf{I}_r$，其中 $r = rn - r(r+1)/2$、$\mathbf{V}_{\boldsymbol{\psi}} = 0.01\mathbf{I}_{np}$ 和 $\mathbf{V}_{\boldsymbol{\phi}} = 0.01\mathbf{I}_{nq}$。最后，对于随机波动率方程中的参数，我们假设

$$\sigma^2_{h,i} \sim \mathcal{IG}(\nu_{h,i}, S_{h,i}), \quad h_{i,0} \sim \mathcal{N}(m_{h,i}, V_{h_{i,0}}), \; i = 1, \ldots, n + k,$$

其中我们设定 $m_{h,i} = 0, V_{h_{i,0}} = 0.01$, $\nu_{h,i} = 3$ 以及 $S_{h,i} = 1$，使得 $h_{i,0}$和 $\sigma^2_{h,i}$ 的先验均值分别为 0 和 0.5。

这个具有非平衡面板的动态因子模型可以使用马尔可夫链蒙特卡罗方法进行估计。更具体地说，给定模型参数和潜在因子后，我们使用所提出的抽样器模拟缺失值，如第2节所述。然后，给定抽样得到的缺失值，可使用标准算法从模型参数和潜在因子的全条件分布中抽样；例如，参见 Antolin-Diaz、Drechsel 和 Petr

（2021年）和 Chan（2022年）。为评估后验抽样器的效率，我们计算与缺失数据和模型参数相关的低效因子（详见附录C）。特别是，所有低效因子都小于 100，且与具有平衡面板的动态因子模型的低效因子相当。

我们使用涵盖 COVID-19 疫情时期的 FRED-MD 数据版本（从2020年3月至2022年9月）估计动态因子模型。对于每个数据版本，我们首先按照 McCracken 和 Ng（2016年）的建议对序列进行转换。

'exszusx'。在我们考虑的数据版本中，这些变量没有缺失值。

按照惯例，随后我们对每个序列进行标准化，使其均值为 0、方差为 1。如前所述，每个数据版本都包含128个每月变量，但其中许多变量在样本开头或结尾处存在缺失值。为便于比较，我们还使用不含缺失值的变量估计动态因子模型，即在每个数据版本中剔除任何存在缺失值的变量。在所考察的数据版本中，平均约有24个变量存在缺失值，因此未纳入因子估计。

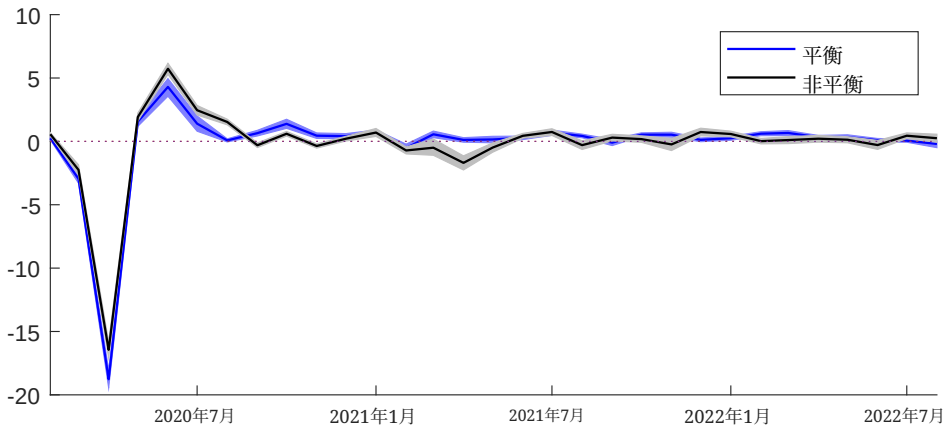


图6：使用平衡和非平衡数据集的动态因子模型得到的第一因子滤波估计。黑线表示使用全部128个变量（含缺失数据）得到的第一因子后验均值，灰色阴影区域表示相应的68%可信区间。蓝线表示仅使用无缺失值的变量得到的后验均值，蓝色阴影区域表示相应的68%区间。

图6绘出了使用平衡与非平衡数据集的动态因子模型中第一因子的滤波估计。这两种估计总体上相似，且均很好地反映了 COVID-19疫情爆发之初的显著下滑及随后反弹，证实第一因子捕捉到了整体经济状况。不过，两者也存在明显差异，尤其是在封锁后全经济范围重新开放所对应的峰值和谷值处。具体而言，使用全部128个变量得到的因子估计显示，下滑幅度较小（-16.4与 -18.8相比），而随后反弹更加明显（5.7 对 4.3）。这些差异可能源于若干订单和库存变量存在缺失值，例如消费品新订单（“acogno”）和资本品新订单（“andenox”）在

样本开始时的总商业库存和样本结束时的总商业库存（“businvx”）。

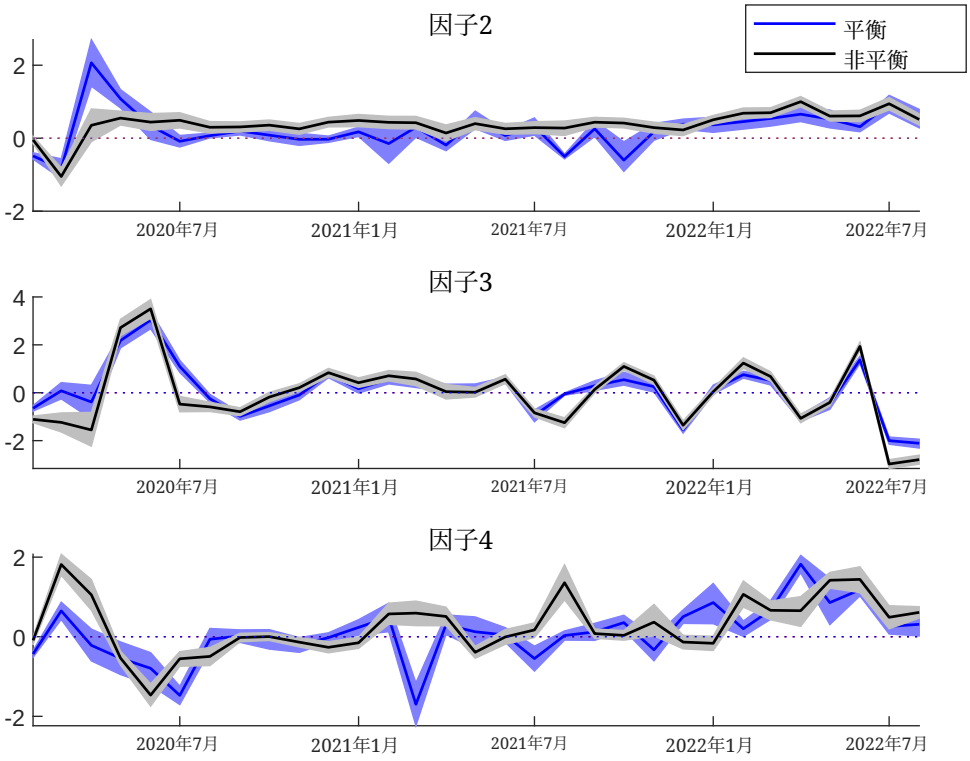


图7：使用平衡与非平衡数据集的动态因子模型得出的第二、第三和第四个因子的滤波估计。黑线表示使用全部128个变量（含缺失数据）得到的因子后验均值，灰色阴影区域表示相应的68%可信区间。蓝线表示仅使用无缺失值的变量得到的后验均值，蓝色阴影区域表示相应的68%区间。

接下来，我们在图7中报告第二、第三和第四因子的滤波估计。使用平衡数据集与非平衡数据集得到的因子估计之间存在更显著的差异。具体而言，最突出的差异出现在第四因子上，该因子尤其能够很好地解释利率变化。这些差异可能反映出许多存在缺失值的变量在第四因子上的载荷较大。例如，若干私人新建住房许可证变量存在缺失值，而这些变量可能包含有关利率的有用信息

利率。忽略这些变量可能无法完整呈现利率的发展情况。[11]

总而言之，这些结果表明，在实证分析中省略存在缺失值的变量，可能会导致对整体经济状况的动态以及利率或价格的共同变动产生失真的认识。这进一步凸显了所提出的缺失值插补方法的实用性：该方法灵活且适用于大型模型。

5 结论

我们提出了一种新颖且高效的方法，可适用于任意条件高斯状态空间模型和具有任意缺失数据模式的数据集，并能一步完成所有缺失观测值的抽样。通过一系列蒙特卡洛模拟，我们证明，在各种设置下，所提出的方法在计算效率上都优于标准的基于卡尔曼滤波的方法。我们还展示了如何将该方法应用于两个宏观经济实证案例：一个涉及大型混频 VAR，另一个涉及采用非平衡数据集的动态因子模型。这两个实证案例都说明，将更多信息（来自高频指标或含有缺失值的变量）纳入宏观经济分析是有益的。

[11]滤波估计 的第五至第八个因子报告于附录C中。我们还发现了显著的差异 平衡与非平衡数据集的估计值。

附录A：混频 VAR 的数据集

本附录详细介绍混频 VAR 实证应用所用的 22变量数据集。具体而言，表3列出了这22个变量及其转换方式。前七个每周变量取自 Lewis、Mertens、Stock 和 Trivedi（2022），其余变量则来自 FRED数据库。样本期为2013年1月至2022年8月。

表3：混频 VAR 应用中变量及其相应变换方式列表。

变量	频率	FRED 代码	变换
燃料	每周	-	$100\Delta\ln\left(\frac{x_t}{x_{t-52}}\right)$
粗钢产量	每周	-	$100\Delta\ln\left(\frac{x_t}{x_{t-52}}\right)$
零售销售均值	每周	-	水平
电力公用事业产出	每周	-	$100\Delta\ln\left(\frac{x_t}{x_{t-52}}\right)$
美国铁路运输量	每周	-	$100\Delta\ln\left(\frac{x_t}{x_{t-52}}\right)$
持续申领失业救济人数	每周	CCSA	$\ln x_t$
首次申领失业救济人数	每周	ICSA	$\ln x_t$
美国普通汽油所有配方价格	每周	美国普通汽油所有配方价格	$100\Delta\ln\left(\frac{x_t}{x_{t-52}}\right)$
原油价格	每周	WCOILWTICO	$\ln x_t$
全国金融状况指数	每周	NFCI	水平
标准普尔500指数	每周	标准普尔500指数	$100\Delta\ln\left(\frac{x_t}{x_{t-1}}\right)$
美国1年期国债收益率	每周	DGS1	水平
美国10年期国债收益率	每周	DGS10	水平
穆迪精选Baa级公司债券收益率	每周	WBAA	水平
穆迪精选Aaa级公司债券收益率	每周	WAAA	水平
VIX	每周	VIXCLS	水平
工业生产	每月	INDPRO	$100\Delta\ln\left(\frac{x_t}{x_{t-1}}\right)$
城镇居民消费价格指数	每月	CPIAUCSL	$100\Delta\ln\left(\frac{x_t}{x_{t-1}}\right)$
失业率	每月	UNRATE	水平
非农就业人数总计	每月	PAYEMS	$100\Delta\ln\left(\frac{x_t}{x_{t-1}}\right)$
制造业平均每周工时	每月	AWHMAN	$\frac{x_t}{10}$
实际国内生产总值	季度	GDPC1	$400\Delta\ln\left(\frac{x_t}{x_{t-1}}\right)$

附录B：Durbin-Koopman模拟平滑器

要实现Durbin-Koopman模拟平滑器，我们首先需要将VAR(5)改写为其伴随形式：

$$\mathbf{s}_t = \mathbf{F}_0 + \mathbf{F}_1 \mathbf{s}_{t-1} + \mathbf{v}_t, \quad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}_{5n}, \mathbf{\Omega}), \tag{14}$$

其中，$\mathrm{s}_t = (\mathrm{y}_t', \mathrm{y}_{t-1}', \mathrm{y}_{t-2}', \mathrm{y}_{t-3}', \mathrm{y}_{t-4}')$为 $5n \times 1$ 状态向量，且

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Sigma} & \mathbf{0}_{n \times 4n} \\ \mathbf{0}_{4n \times n} & \mathbf{0}_{4n \times 4n} \end{bmatrix}, \quad \mathbf{F}_0 = \begin{bmatrix} \mathbf{b}_0 \\ \mathbf{0}_{4n \times 1} \end{bmatrix}, \quad \mathbf{F}_1 = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \mathbf{B}_4 & \mathbf{B}_5 \\ \mathbf{I}_n & \mathbf{0}_n & \mathbf{0}_n & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{I}_n & \mathbf{0}_n & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n & \mathbf{I}_n & \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{0}_n & \mathbf{0}_n & \mathbf{I}_n & \mathbf{0}_n \end{bmatrix}.$$

当高频变量和低频变量均被观测到时，测量方程为

$$\mathbf{y}_t = \mathbf{\Lambda}_1 \mathbf{s}_t. \tag{15}$$

当仅观测到高频变量时，测量方程变为

$$\mathbf{y}_t^o = \mathbf{\Lambda}_2 \mathbf{s}_t. \tag{16}$$

矩阵 $\mathbf{\Lambda}_1$ 和 $\mathbf{\Lambda}_2$ 包含跨期约束：

$$\mathbf{\Lambda}_1 = \begin{bmatrix} \mathbf{I}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} \\ \mathbf{0}_{n^m} & \frac{1}{3}\mathbf{I}_{n^m} & \mathbf{0}_{n^o} & \frac{2}{3}\mathbf{I}_{n^m} & \mathbf{0}_{n^o} & \mathbf{I}_{n^m} & \mathbf{0}_{n^o} & \frac{2}{3}\mathbf{I}_{n^m} & \mathbf{0}_{n^o} & \frac{1}{3}\mathbf{I}_{n^m} \end{bmatrix},$$
$$\mathbf{\Lambda}_2 = \begin{bmatrix} \mathbf{I}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} & \mathbf{0}_{n^o} & \mathbf{0}_{n^m} \end{bmatrix}.$$

我们遵循 Jaroci´nski（2015），并对式（14）、（15）和（16）应用Durbin-Koopman模拟平滑器。

附录C：补充结果

本附录展示了两项应用的更多实证结果。首先，我们在图 8 中报告了使用平衡和非平衡 FRED-MD数据集得到的动态因子模型中因子 5–8 的滤波估计。

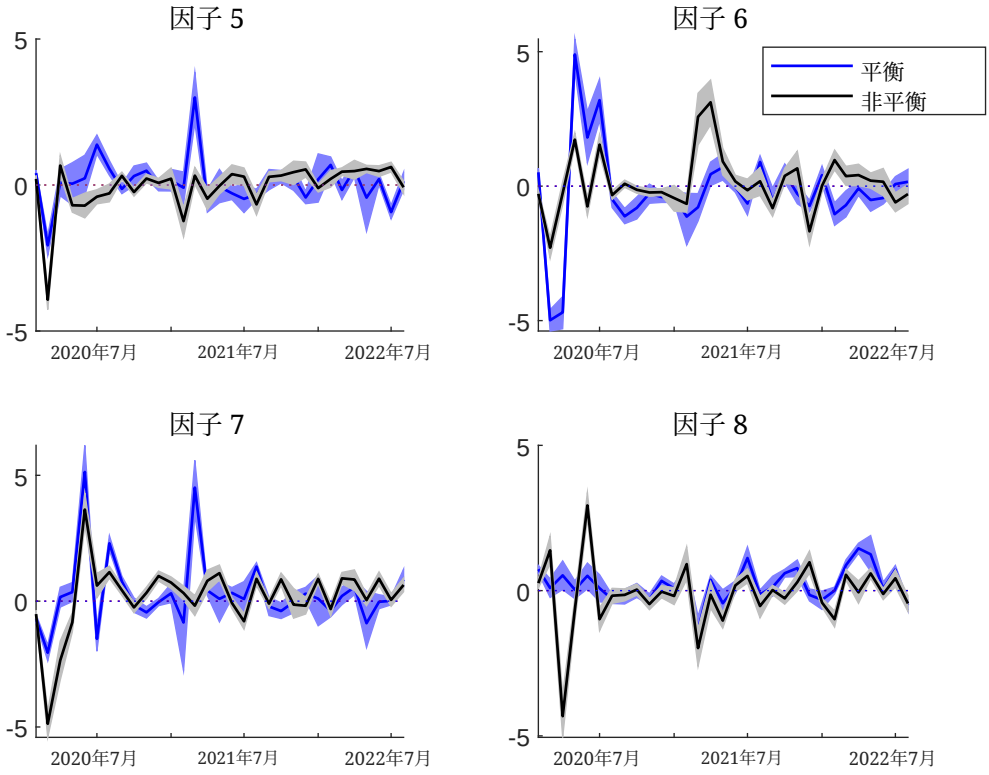


图 8：使用平衡与非平衡数据集的动态因子模型中，第五至第八个因子的滤波估计。黑线表示使用全部 128 个变量（含缺失数据）得到的因子后验均值，灰色阴影区域表示相应的 68%可信区间。蓝线表示仅使用无缺失值的变量得到的后验均值，蓝色阴影区域表示相应的 68%区间。

接下来，图 9–11 报告了两个实证应用中 MCMC 样本的低效因子。具体而言，图 9 绘制了周度状态空间混频向量自回归模型中缺失数据及其他模型参数的低效因子。为更简洁地呈现信息，这里使用低效因子的箱线图。

每个箱体中间的线表示中位数，下线和上线分别表示第 25 和第75百分位数。须线延伸至最大值和最小值。

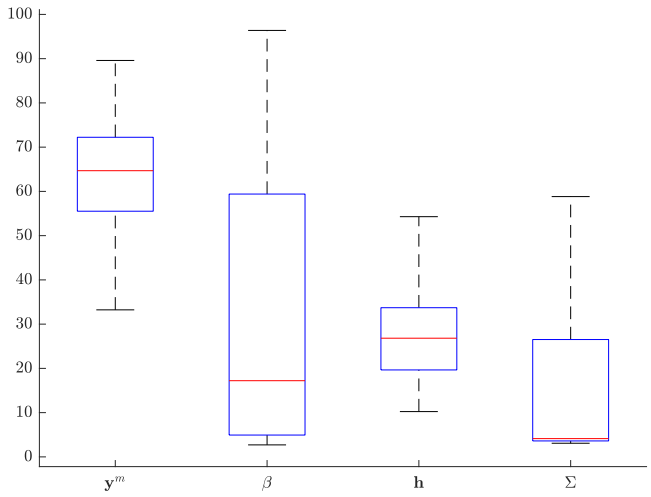


图 9：周度状态空间混频向量自回归模型中，缺失数据 $\mathbf{y}^m$ 以及其他模型参数 $\boldsymbol{\beta}$、h 和 $\boldsymbol{\Sigma}$对应后验抽样的低效因子箱线图。

图10和图11分别展示了非平衡面板和平衡面板动态因子模型的低效因子。

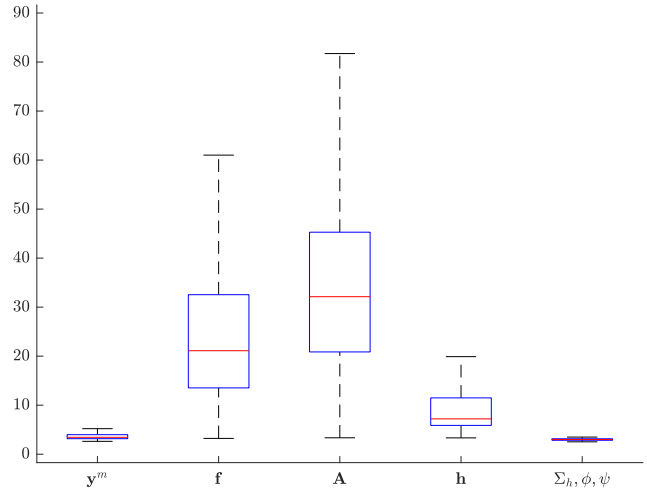


图10：非平衡面板动态因子模型中，与缺失数据 $\mathbf{y}^m$ 及其他模型参数 f、A、h、 $\boldsymbol{\sigma}_h^2$、$\boldsymbol{\psi}$ 和 $\boldsymbol{\phi}$ 的后验抽样相对应的低效因子箱线图。

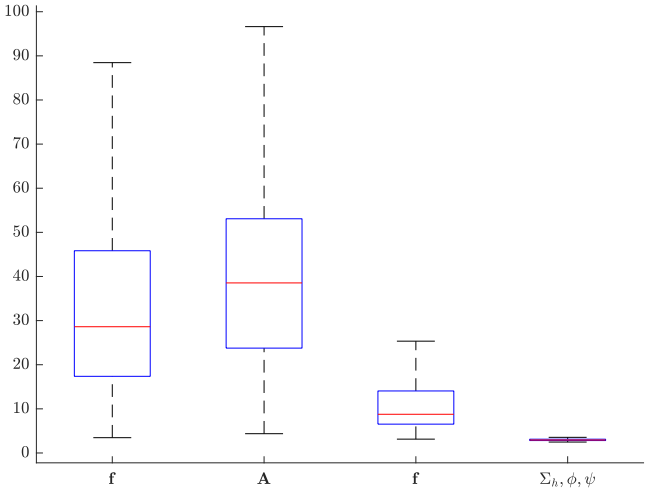


图11：平衡面板动态因子模型中，与 f、A、h、 $\boldsymbol{\sigma}_h^2$、 $\boldsymbol{\psi}$ 和 $\boldsymbol{\phi}$ 的后验抽样相对应的低效因子箱线图。

参考文献

Antolin-Diaz, J., T. Drechsel, and I. Petrella(2021)：《经济活动临近预测的进展：长期趋势、重大冲击与新数据》，.Aruoba, S. B., F. X. Diebold, and C. Scotti (2009)：《商业状况的实时测量》，《商业与经济统计杂志》，27(4)，417–427。Bai, J., and S. Ng (2002)：《确定近似因子模型中的因子数量》，《计量经济学》，70(1)，191–221。Bańbura, M., D. Giannone, and L. Reichlin (2010)：《大型贝叶斯向量自回归》，《应用计量经济学杂志》，25(1)，71–92。Beyeler, S., and S. Kaufmann (2021)：《简化形式的因子增强型向量自回归——利用稀疏性纳入有意义的因子》，《应用计量经济学杂志》，即将出版。Brave, S. A., R. A. Butters, and A. Justiniano(2019)：《利用混频贝叶斯向量自回归预测经济活动》，《国际预测学杂志》，35(4)，1692–1707。Carriero, A., J. C. C. Chan, T. E. Clark, and M. G. Marcellino(2022)：《〈含随机波动率和非共轭先验的大型贝叶斯向量自回归〉勘误》，《计量经济学杂志》，227(2)，506–512。Carriero, A., T. E. Clark, M. Marcellino, and E. Mertens(2022)：《利用随机波动率处理贝叶斯向量自回归中的 COVID-19 异常值》，《经济学与统计学评论》，即将出版。Carriero, A., T. E. Clark, and M. G. Marcellino(2016)：《大型贝叶斯向量自回归中的共同漂移波动率》，《商业与经济统计杂志》，34(3)，375–390。(2019)：《含随机波动率和非共轭先验的大型贝叶斯向量自回归》，《计量经济学杂志》，212(1)，137–154。Carter, C. K., and R. Kohn (1994)：《状态空间模型的吉布斯抽样》，《生物统计学》，81，541–553。Chan, J. C. C. (2013)：《移动平均随机波动率模型及其在通胀预测中的应用》，《计量经济学杂志》，176(2)，162–172。(2017)：《时变参数均值中的随机波动率模型：在通胀建模中的应用》，《商业与经济统计杂志》，35(1)，17–28。

——(2020)：《大型贝叶斯向量自回归：灵活的克罗内克误差协方差结构》，《商业与经济统计杂志》，38(1)，68–79。(2022)：《大型贝叶斯向量自回归的随机波动率设定比较》，《计量经济学杂志》，即将出版。Chan, J. C. C., E. Eisenstat, and G. Koop (2016)：《大型贝叶斯向量自回归移动平均模型》，《计量经济学杂志》，192(2)，374–390。Chan, J. C. C., and I. Jeliazkov (2009)：《状态空间模型中的高效模拟与积分似然估计》，《国际数学建模与数值优化杂志》，1(1)，101–120。Chan, J. C. C., G. Koop, and S. M. Potter (2013)：《趋势通胀的新模型》，《商业与经济统计杂志》，31(1)，94–106。Chib, S., E. Greenberg, and I. Jeliazkov (2009)：《内生性与样本选择情形下半参数模型的估计》，《计算与图形统计学杂志》，18，321–348。Chib, S., and I. Jeliazkov (2006)：《二元纵向数据半参数动态模型的推断》，《美国统计学会杂志》，101(474)，685–700。Clark, T. E. (2011)：《基于含随机波动率的贝叶斯向量自回归的实时密度预测》，《商业与经济统计杂志》，29(3)，327–341。Cogley, T., and T. J. Sargent (2005)：《漂移与波动率：二战后美国的货币政策与结果》，《经济动态评论》，8(2)，262–302。Cong, Y., B. Chen, and M. Zhou (2017)：《超平面截断多元正态分布的快速模拟》，《贝叶斯分析》，12(4)，1017–1037。Cross, J., C. Hou, and A. Poon (2019)：《利用大型贝叶斯向量自回归进行宏观经济预测：全局—局部先验与稀疏性的假象》，《国际预测学杂志》，36(3)，899–915。Cross, J., and A. Poon (2016)：《预测澳大利亚宏观经济变量的结构变化与厚尾事件》，《经济建模》，58，34–51。D'Agostino, A., L. Gambetti, and D. Giannone (2013)：《宏观经济预测与结构变化》，《应用计量经济学杂志》，28，82–101。Dimitrakopoulos, S., and M. Kolossiatis (2020)：《移动平均随机波动率模型的贝叶斯分析：金融时间序列的均值效应建模与杠杆效应》，《计量经济学评论》，39(4)，319–343。

Doan, T., R. Litterman, and C. Sims (1984)：“使用现实先验分布进行预测和条件投影”，《计量经济学评论》，3(1)，1–100。Durbin, J., and S. J. Koopman (2002)：“一种简单高效的状态空间时间序列分析模拟平滑器”，《生物统计学》，89，603–615。Eckert, F., P. Kronenberg, H. Mikosch, and S. Neuwirth(2020)：“利用替代性高频数据追踪经济活动”，KOF 工作论文，488，1–35。Fu, B. (2020)：“菲利普斯曲线的斜率是否随时间变化？来自未观测成分模型的证据”，《经济建模》，88，320–340。Grant, A. L., and J. C. C. Chan (2017a)：“产出趋势—周期分解的贝叶斯模型比较”，《货币、信贷与银行学期刊》，49(2-3)，525–552。(2017b)：“调和产出缺口：未观测成分模型与霍德里克-普雷斯科特滤波器”，《经济动态与控制期刊》，75，114–121。Hartwig, B. (2021)：“COVID-19时期的贝叶斯向量自回归模型与先验校准”，见 SSRN 3792070。Hauber, P., and C. Schumacher(2021)：“存在缺失观测值时基于精度的抽样：因子模型应用”，讨论论文 11/2021，德国联邦银行。Hou, C. (2020)：“通货膨胀与通货膨胀不确定性之间的时变关系”，《牛津经济学与统计学通报》，82(1)，83–124。Jarociński, M.(2015)：“关于实现 Durbin 和 Koopman 模拟平滑器的说明”，《计算统计与数据分析》，91，1–3。Kastner, G. (2019)：“高维情形下的稀疏贝叶斯时变协方差估计”，《计量经济学杂志》，210(1)，98–115。Kastner, G., and F. Huber (2020)：“超高维稀疏贝叶斯向量自回归模型”，《预测学期刊》，39(7)，1142–1165。Kaufmann, S., and C. Schumacher (2019)：“采用与顺序无关的事后众数识别进行稀疏动态因子模型的贝叶斯估计”，《计量经济学杂志》，210(1)，116–134。Koop, G. (2013)：“利用中型和大型贝叶斯向量自回归模型进行预测”，《应用计量经济学杂志》，28(2)，177–203。

Koop, G., S. McIntyre, J. Mitchell, and A. Poon (2020)：“英国区域产出增长：自1970年以来更及时、频率更高的估计”，《应用计量经济学杂志》，35(2)，176–197。(2022)：“美国月度国内生产总值的协调估计”，《商业与经济统计杂志》，即将出版。Lenza, M., and G. E. Primiceri(2022)：“如何在2020年3月之后估计向量自回归模型”，《应用计量经济学杂志》，即将出版。Lewis, D. J., K. Mertens, J. H. Stock, and M. Trivedi(2022)：“COVID-19之后的宏观经济波动建模”，《应用计量经济学杂志》，37(4)，667–687。Litterman, R. (1986)：“贝叶斯向量自回归预测——五年的经验”，《商业与经济统计杂志》，4，25–38。Mariano, R. S., and Y. Murasawa (2003)：“基于月度和季度序列的商业周期新同步指数”，《应用计量经济学杂志》，18(4)，427–443。(2010)：“同步指数、共同因子与月度实际国内生产总值”，《牛津经济学与统计学通报》，72(1)，27–46。McCausland, W. J., S. Miller, and D. Pelletier(2011)：“状态空间模型的模拟平滑：计算效率分析”，《计算统计与数据分析》，55(1)，199–212。McCracken, M. W., and S. Ng (2016)：“FRED-MD：宏观经济研究月度数据库”，《商业与经济统计杂志》，34(4)，574–589。Pettenuzzo, D., R. Sabbatucci, and A. Timmermann (2023)：“COVID-19疫情期间的股息暂停与现金流：动态计量经济模型”，《计量经济学杂志》，即将出版。Primiceri, G. E.(2005)：“时变结构向量自回归与货币政策”，《经济研究评论》，72(3)，821–852。Rue, H. (2001)：“高斯马尔可夫随机场的快速抽样及其应用”，《皇家统计学会杂志：B辑（方法论）》，63(2)，325–338。Rue, H., and L. Held(2005)：《高斯马尔可夫随机场：理论与应用》。CRC出版社。Schorfheide, F., and D. Song (2015)：“利用混频 VAR 进行实时预测”，《商业与经济统计杂志》，33(3)，366–380。

Stock, J. H., and M. W. Watson (2016年)：《核心通胀与趋势通胀》，《经济学与统计学评论》，98(4)，770–784。

Zhang, B., J. C. C. Chan, and J. L. Cross(2020年)：《带有 ARMA 创新项的随机波动率模型：在七国集团通胀预测中的应用》，《国际预测学杂志》，36(4)，1318–1328。