将动态模态分解与集合卡尔曼滤波相结合，用于跟踪与预测

Stephen A Falconer[1],David J.B. Lloyd[1],和 Naratip Santitissadeekorn[1]

[1]萨里大学数学系，英国吉尔福德 GU2 7XH

2023年1月16日

arXiv:2301.05504v1 [math.DS] 13 Jan 2023

摘要

数据同化技术（如集合卡尔曼滤波）已被证明是一种高效且有效的方法，可将含噪数据与数学模型结合起来，对动力系统进行跟踪和预测。然而，在处理高维数据时，许多情况下并没有现成的模型，因此无法应用数据同化技术。本文利用动态模态分解，直接从高维数据中生成动力系统的低维线性模型。该模型由时间模态和空间模态定义，随后可与集合卡尔曼滤波器等数据同化技术结合使用。我们展示了如何将动态模态分解与集合卡尔曼滤波器相结合（称为 DMDEnKF），以便在新数据出现时迭代更新当前状态和时间模态。我们通过合成示例证明，该方法能够跟踪随时间变化的动力系统，并探讨了时间延迟嵌入的应用。随后，我们将 DMDEnKF 应用于美国疾病控制与预防中心的真实季节性流感样疾病数据，并发现，在短期预测方面，DMDEnKF 的表现可与 ILINet竞赛中最优秀的机理模型相媲美。

关键词

动态模态分解；集合卡尔曼滤波器；数据驱动建模；数据同化；动力系统

1 引言

数据同化是指一系列方法，这些方法将海量数据集与复杂的数学模型相结合，以跟踪和预测可能演变或变化的系统 [38]。其应用大多集中在地球科学领域 [51], ，但由于这些技术具有普适性，也已成功应用于医学 [18]到生态学等众多领域 [40]。卡尔曼滤波器 [36] 是工业领域广泛采用的一种数据同化技术 [3]，它能将线性模型的预测与高斯数据进行最优结合。传统上，该方法用于估计模型的状态，但也可以同时对模型参数进行滤波，从而得到

所谓的联合状态-参数估计问题 [33]。如果被滤波的系统是非线性的，则可以采用卡尔曼滤波器的其他形式，例如扩展卡尔曼滤波器[53], 、无迹卡尔曼滤波器 [59] 或集合卡尔曼滤波器（EnKF） [23]。集合卡尔曼滤波器用随机样本集合表示系统状态的分布，随后可利用这些样本估计有用的统计量，例如通过样本协方差估计状态的协方差，或通过样本均值获得状态的点估计 [23] ，因此非常适用于高维问题。所有这些方法都需要系统模型；如果不存在模型，就必须构建一个，而实现这一目标最具普适性的方式是采用数据驱动建模。

动态模态分解（DMD）是一种数据驱动建模方法，可直接从高维数据中识别动力系统内的低维时空模式 [54]。该方法假设状态向量随时间按照线性系统演化，并寻找特征值（时间模态）及其对应特征向量（空间模态）的低维近似。空间模态可理解为将状态变量分解为若干独立分量的模态，这些分量随时间共同线性演化。相应的时间模态则描述空间模态随时间增长、衰减还是保持稳定。DMD 已用于根据测量数据近似多个领域中的动力系统，包括流行病学 [49], 、金融 [43] 和神经科学 [12]。由于其广受欢迎，DMD 已扩展至记录的测量函数呈非线性的系统，形成了扩展/核动态模态分解[61]/[44], ；其中一种扩展方法 Hankel-DMD [2]采用原始可观测量的时间延迟嵌入。在存在测量噪声时，研究表明，标准 DMD 会产生系统偏差：训练期间，它将所有噪声不对称地归因于模型的目标输出测量值，而不归因于其输入 [15]。这种系统偏差促使人们提出了能够直接处理噪声项的 DMD 噪声处理变体 [15], 、前向后向动态模态分解 [15] （通过分别沿时间正向和反向执行 DMD 并结合结果），以及全动态模态分解（TDMD） [31]；后者最小化全最小二乘误差，而非普通最小二乘误差。

本文旨在开发一种算法，随着新数据的不断获取，利用动态模态分解和集合卡尔曼滤波（EnKF）迭代改进其生成的时间模态（特征值）和状态估计。对于随时间从增长行为转变为衰减行为（或反之）的动力系统而言，这将非常有用。虽然可以使用标准卡尔曼滤波器，仅依据 DMD模态估计系统状态，但如果不同时对模型的时间模态进行滤波，那么当系统随时间发生变化时，估计结果很可能会受到影响。现有方法已将 DMD 与卡尔曼滤波器 [45] 或扩展卡尔曼滤波器 [46], 相结合，通过滤波估计整个系统动力学矩阵。相比之下，我们的工作着重于高效跟踪系统的时间模态，并预测系统的未来状态。DMD 会生成线性模型，因此天然适用于卡尔曼滤波器；然而，同时估计系统状态和时间模态时，滤波过程会变成非线性过程。因此，我们需要使用专为非线性模型设计的滤波器，并选择了集合卡尔曼滤波（EnKF），因为它适用性强、可扩展至大型动力系统，且易于实现 [52]。任何能够生成时间模态的 DMD变体都可与 DMDEnKF 框架兼容；为与 EnKF 关于数据中存在噪声的假设保持一致，我们采用 TDMD。同时，我们采用 Hankel-DMD 的全最小二乘版本应用 DMDEnKF，后文称为 Hankel-DMDEnKF，以考察时间延迟嵌入对我们框架的影响。

为展示 DMDEnKF 方法，我们首先在合成数据集上对其进行测试。首先，我们在一个振荡周期不断缩短的简单含噪振荡系统上，使用 DMDEnKF 跟踪系统的时间模态，并将结果与 Hankel-DMDEnKF、其他迭代式DMD变体以及“黄金标准”滤波方法进行比较。接下来，我们模拟一场大流行，并评估 DMDEnKF 跟踪系统时间模态以及生成多步预测的能力。

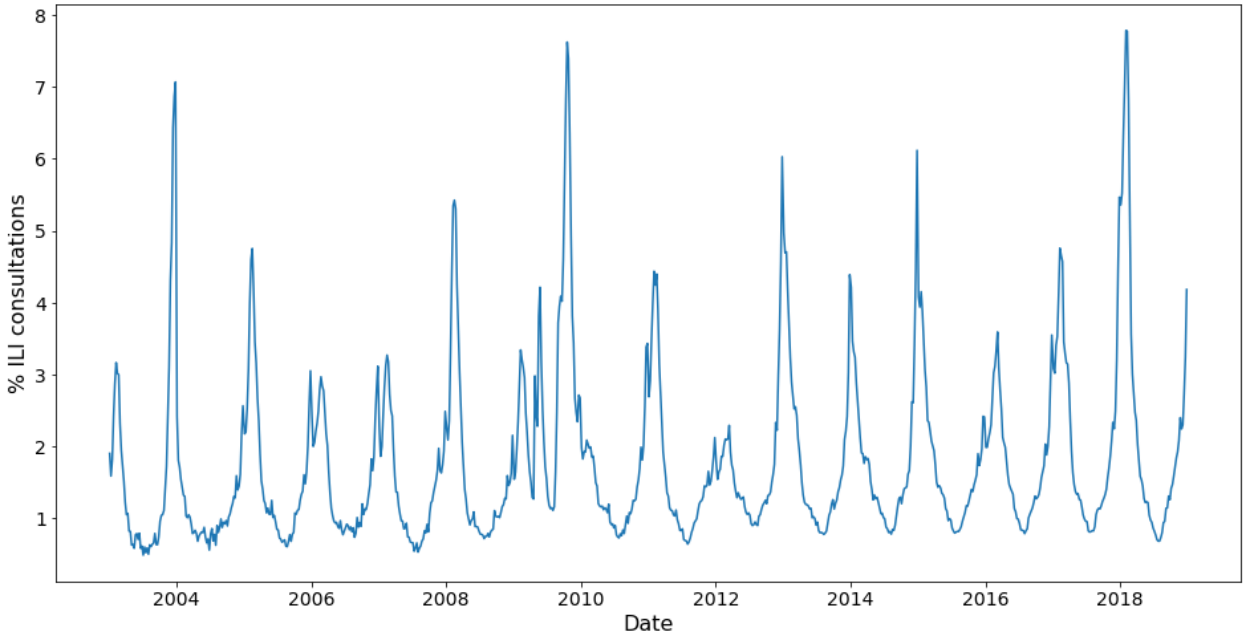


图 1：*2003*年至*2018*年底，美国每周流感样疾病就诊人数占全科医生每周总就诊人数的百分比。数据显示，流感样疾病就诊的年度峰值在大小、时间和形态上各不相同，因此难以用简单的 *SIR* 型模型进行建模。

最后，我们将 DMDEnKF 和 Hankel-DMDEnKF 应用于美国疾病控制与预防中心 (CDC) ILINet 中的美国季节性流感样疾病 (ILI) 实际数据[25] ，如图1所示，以考察它们对 ILI 就诊率的预测能力。流感样疾病定义为发热并伴有咳嗽或咽喉痛，且除流感外没有已知病因 [25] ，每年在美国感染人数为 900 万至 3500 万 [24]。由于流感样疾病十分常见，人们已经开发出许多方法来模拟其传播 [14, 47]，这些模型采用的方法大致可分为机制方法和统计方法[37]。机制方法 [5, 48]会明确假设哪些因素驱动传染病传播，然后再根据数据拟合所提出模型中的参数。这类方法具有高度可解释性的优点，因此有助于理解如何控制疾病传播 [39], ，但也可能建立过于简化的假设[4]。例如，简单的 SIR 型模型难以描述图1所示的特定行为，例如圣诞节期间流感样疾病就诊人数的下降。统计方法 [11, 60]通常更为灵活，因为它们所需的领域特定假设较少；但在 ILINet 数据集上实时应用时，两类方法的预测能力相近 [50]。DMDEnKF 试图在这两类方法之间取得平衡：它完全由数据驱动，因而具有灵活性，同时又能通过相关的 DMD 模态提供一定程度的可解释性。

其余部分　　本文其余部分安排如下。首先，简要概述动态模态分解、Hankel　　-

动态模态分解和集合卡尔曼滤波算法，以确保内容完整。随后，将全面介绍 DMDEnKF 算法。接着，我们将把 DMDEnKF 和 Hankel-DMDEnKF 应用于合成数据，并将其性能与其他现有的迭代式 DMD 变体进行比较。最后，我们将使用 DMDEnKF 和 Hankel-DMDEnKF 处理 ILINet 数据，预测未来最多 4 周的流感样疾病就诊率，并评估其性能。

2 DMDEnKF

2.1 动态模态分解（DMD）

考虑一个 n 维状态 $\mathbf{x}_k \in \mathbb{R}^n$，在等间隔时间 $k = 1, ..., m$ 测量。假设这些时间序列数据由线性动力系统生成，则连续状态 $\mathbf{x}_k$ 和 $\mathbf{x}^{k+1}$ 之间满足

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k \tag{2.1}$$

其中未知矩阵为 $\mathbf{A} \in \mathbb{R}^{n\times n}$。令

$$\mathbf{X} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_{m-1} \\ | & | & & | \end{bmatrix}, \quad \mathbf{X}' = \begin{bmatrix} | & | & & | \\ \mathbf{x}_2 & \mathbf{x}_3 & \dots & \mathbf{x}_m \\ | & | & & | \end{bmatrix}, \tag{2.2}$$

方程（2.1）可简洁地表示为所有连续数据对的形式：

$$\mathbf{X}' = \mathbf{A}\mathbf{X}. \tag{2.3}$$

为了最小化 为最小化均方误差项 $\sum_{k=1}^{m-1} ||\mathbf{x}^{k+1} - \mathbf{A}\mathbf{x}_k||_2^2$，标准动态模态分解定义

$$\mathbf{A} = \mathbf{X}'\mathbf{X}^+, \tag{2.4}$$

其中 $\mathbf{X}^+$ 是$\mathbf{X}$的摩尔-彭若斯伪逆 [6]。高效求解$\mathbf{A}$ 的特征分解是动态模态分解的主要目的，因为这些特征值／特征向量对应于数据中的时空模式。

动态模态分解方法首先对数据矩阵$\mathbf{X}$进行奇异值分解，将其表示为两个实值正交矩阵（若 $\mathbf{X} \in \mathbb{C}^{n\times m}$，则为复数酉矩阵）$\mathbf{U} \in \mathbb{R}^{n\times n}$，$\mathbf{V} \in \mathbb{R}^{m\times m}$ 与一个矩形对角矩阵的乘积，其中非负实数值按降序排列（$\mathbf{\Sigma} \in \mathbb{R}^{n\times m}$），形式如下

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*. \tag{2.5}$$

根据埃克特－杨定理，矩阵的最佳秩 r 近似 [22] 可通过截断其奇异值分解得到，因此，将方程（2.5）截断至适当的秩 r [26]，即可在信息损失最小的情况下压缩数据矩阵，记作

$$\mathbf{X} \approx \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^*. \tag{2.6}$$

通过进行这种压缩，我们隐含地假设高维数据中存在一种低维（$\leq r$）线性结构。

摩尔-彭若斯伪逆可直接由方程（2.5）中计算出的奇异值分解得到，其形式为 $\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{*}$。出于效率考虑，我们使用方程（2.6）中的秩 r 截断矩阵，并令

$$\begin{aligned}\mathbf{A} &= \mathbf{X}'\mathbf{X}^{+}, \\ &\approx \mathbf{X}'\mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^{*}.\end{aligned} \tag{2.7}$$

此时，对$\mathbf{A}$ 的近似仅作用于由 Col($\mathbf{U}_r$) 定义的一个 r 维子空间。因此，我们可以将$\mathbf{A}$ 限制在这个 r 维子空间上（该子空间表示 $\mathbf{X}$ 的最大 r 个 POD 模态），并将受限后的 $\mathbf{A} \in \mathbb{R}^{r\times r}$ 记为

$$\begin{aligned}\tilde{\mathbf{A}} &= \mathbf{U}_r^{*}\mathbf{A}\mathbf{U}_r, \\ &\approx \mathbf{U}_r^{*}\mathbf{X}'\mathbf{V}_r\mathbf{\Sigma}_r^{-1}.\end{aligned} \tag{2.8}$$

˜我们计算$\mathbf{A}$的特征值（λ_i）及其对应的特征向量（$\mathbf{v}_i$），并定义

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_r \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_r \\ | & | & & | \end{bmatrix}. \tag{2.9}$$

重构原算子$\mathbf{A}$ 的特征值/特征向量，有助于深入了解˜系统的结构 [1]，并使我们能够将其向前传播。可以证明，$\mathbf{A}$的特征值（$\mathbf{\Lambda}$）与原算子$\mathbf{A}$ [34], 的特征值相同；然而，恢复原始特征向量更为复杂，可通过投影动态模态分解或精确动态模态分解实现。我们采用 Tu 等人提出的精确动态模态分解方法 [34]，因为该方法能够为所有非零 λ_i 的特征向量求得精确的 DMD模态（$\mathbf{\Phi}$），其中 $\mathbf{\Phi}$ 定义为

$$\mathbf{\Phi} = \mathbf{X}'\mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{W}. \tag{2.10}$$

特征值为零的 DMD模态对系统的动力学没有影响，因此精确动态模态分解的这一限制影响不大。该方法求得$\mathbf{A}$，使得$\mathbf{A}\mathbf{X}$ $=\mathbf{X}'$ 严格成立，前提是$r \geq \text{rank}(\mathbf{X})$以及$\mathbf{X}$ 与$\mathbf{X}'$ 线性一致 [34]。

有了 $\mathbf{\Lambda}$ 和 $\mathbf{\Phi}$，我们可以构造$\mathbf{A}$的 r 维近似，不过仍需找到每个模态的初始相位和振幅。计算该向量（$\mathbf{b}$）的标准方法 [54] 是将初始状态$\mathbf{x}_1$ 表示在 DMD模态构成的基中，即

$$\mathbf{b} = \mathbf{\Phi}^{+}\mathbf{x}_1. \tag{2.11}$$

值得注意的是，也存在其他方法，例如 [13, 35]，这些方法着重于在附加条件下，针对所有数据点优化$\mathbf{b}$。

离散系统的最终解可以写为

$$\mathbf{x}_k = \mathbf{\Phi}\mathbf{\Lambda}^{k}\mathbf{b}. \tag{2.12}$$

In 在本文其余部分，我们将 $\mathbf{\Lambda}$ 称为时间模态，将 $\mathbf{\Phi}$ 称为空间模态 .

2.1.1 Hankel-DMD

Hankel-DMD 首先扩充原始测量状态$\mathbf{x}_k \in \mathbb{R}^n$，将前 $d-1$ 个时间步的状态测量值附加到其中

$$h(\mathbf{x}_k) = \begin{bmatrix} \mathbf{x}_k{}^T & \mathbf{x}_{k-1}{}^T & \dots & \mathbf{x}_{k-(d-1)}{}^T \end{bmatrix}^T, \tag{2.13}$$

以形成一个新的状态 $h(\mathbf{x}_k) \in \mathbb{R}^{dn}$。这称为时间延迟嵌入，我们将 d称为延迟嵌入维数。将时间延迟嵌入 $h(\mathbf{x}_k)$作为新的状态，方程（$\mathbf{X}$ 和 $\mathbf{X}'$ 2.2）中的矩阵现在变为

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_d & \dots & \mathbf{x}_{m-1} \\ \vdots & \ddots & \vdots \\ \mathbf{x}_1 & \dots & \mathbf{x}_{m-d} \end{bmatrix}, \quad \mathbf{X}' = \begin{bmatrix} \mathbf{x}_{d+1} & \dots & \mathbf{x}_m \\ \vdots & \ddots & \vdots \\ \mathbf{x}_2 & \dots & \mathbf{x}_{m-(d-1)} \end{bmatrix}. \tag{2.14}$$

利用上述定义的 X 和 X'，Hankel-DMD 的处理过程与标准DMD算法完全相同，可生成特征值 $\mathbf{\Lambda}$、DMD模态 $\mathbf{\Phi}$ 及其初始状态$\mathbf{b}$，如上所述。从 d 开始，应用方程（2.12），并将结果限制在前 n 行，即可重建或预测所有时间步的原始系统。

2.1.2 迭代DMD变体

还有其他专为迭代应用而设计的DMD变体。本文将比较这些方法与 DMDEnKF 在跟踪系统特征值和预测未来状态方面的能力。流式DMD [32] 是对标准DMD算法的改进，可在新数据可用时高效处理；我们首先要比较的是能够感知噪声的变体流式 TDMD[30] 。用于比较的第二种方法是窗口化 DMD [62], ，该方法仅在最近的 w 个数据快照构成的滑动窗口内应用上述标准DMD。最后一种比较方法是在线 DMD [62], ，具体而言，是该算法中对过去测量值的重要性赋予指数衰减权重 ρ 的变体。

2.2 集合卡尔曼滤波（EnKF）

考虑一个受随机扰动影响的离散时间非线性动力系统

$$\mathbf{x}_k = \mathbf{F}(\mathbf{x}_{k-1}) + \mathbf{w}_k, \quad \mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k), \tag{2.15}$$

其中，$\mathbf{F}$ 是非线性函数$\mathbf{F}$ ： $\mathbb{R}^n \to \mathbb{R}^n$，$\mathbf{x}_k \in \mathbb{R}^n$ 是系统的状态，$\mathbf{w}_k \in \mathbb{R}^n$ 是随机扰动，$\mathcal{N}$ 是均值为$\mathbf{0}$ 、协方差矩阵为$\mathbf{Q}_k$的正态分布。描述观测值与系统真实状态之间关系的测量方程如下：

$$\mathbf{y}_k = \mathbf{H}(\mathbf{x}_k) + \mathbf{v}_k, \quad \mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k), \tag{2.16}$$

其中，$\mathbf{H}$: $\mathbb{R}^n \to \mathbb{R}^l$ 是系统的观测算子，$\mathbf{y_k} \in \mathbb{R}^l$ 是对系统的一次观测，$\mathbf{v_k} \in \mathbb{R}^l$ 是观测中的噪声，而 $\mathcal{N}$是均值为$\mathbf{0}$ 、协方差矩阵为$\mathbf{R}_k$的正态分布。我们重点关注与本应用场景相关的情况：$\mathbf{H}$ 为线性算子，因此可表示为矩阵$\mathbf{H} \in \mathbb{R}^{l \times n}$。

一般而言，滤波方法旨在结合状态转移模型（2.15）和观测模型（2.16）的信息，以计算条件密度 $p(\mathbf{x_k}|\mathbf{Y_k})$，其中 $\mathbf{Y_k} = (\mathbf{y_1}, ..., \mathbf{y_k})$。如果$\mathbf{F}$ 和$\mathbf{H}$ 均为线性，且随机扰动服从正态分布 [36]，那么卡尔曼滤波器就是最优滤波器。集合卡尔曼滤波旨在处理线性假设或正态假设（或两者）不成立的滤波问题 [23]。它利用卡尔曼公式将状态集合传播至高概率区域，使集合的离散程度与线性正态模型相一致。

开始集合卡尔曼滤波算法时，需要一个由 N 状态估计值 $\hat{\mathbf{x}}_0^{(1)}$,..., $\hat{\mathbf{x}}_0^{(N)}$ 组成的初始集合。如果没有可用的集合，可以根据初始状态估计值 $\hat{\mathbf{x}}_0$ 和协方差矩阵 $\mathbf{P}_0$，从 $\mathcal{N}(\hat{\mathbf{x}}_0, \mathbf{P}_0)$中进行 N次独立抽样来生成。

算法 随后，集合卡尔曼滤波的执行过程如下 [23]:

步骤 1： 使用方程（2.15）将每个集合成员向前推演至 $i = 1, ..., N$，通过

$$\hat{\mathbf{x}}_{k|k-1}^{(i)} = \mathbf{F}(\hat{\mathbf{x}}_{k-1|k-1}^{(i)}) + \mathbf{w}_k^{(i)}. \tag{2.17}$$

符号 $\hat{\mathbf{x}}_{k|k-1}^{(i)}$ 表示第 i 个集合成员在时间 k 的状态估计值 $\hat{\mathbf{x}}_k^{(i)}$，该估计仅使用截至时间 $k-1$ 的信息；$\hat{\mathbf{x}}_{k-1|k-1}^{(i)}$ 则表示同一集合成员在时间 $k-1$ 的状态估计值，使用截至时间 $k-1$ 的信息。每个 $\mathbf{w}_k^{(i)}$ 均从 $\mathcal{N}(\mathbf{0}, \mathbf{Q}_k)$中独立抽取。现在还可以通过集合的样本协方差来估计当前协方差矩阵，$\hat{}$ $\hat{}$记为 $\mathbf{P}^{k|k-1}$。随后可用它来估计卡尔曼增益矩阵 $\mathbf{K}_k$，即

$$\hat{\mathbf{K}}_k = \hat{\mathbf{P}}_{k|k-1}\mathbf{H}^T(\mathbf{H}\hat{\mathbf{P}}_{k|k-1}\mathbf{H}^T + \mathbf{R}_k)^{-1}. \tag{2.18}$$

步骤 2： 利用公式 (2.16) 计算测量创新。

Fro对于测量值 $\mathbf{y}_k$，我们再次使用 $i = 1, ..., N$，并生成模拟测量值

$$\mathbf{y}_k^{(i)} = \mathbf{y}_k + \mathbf{v}_k^{(i)} \tag{2.19}$$

其中，每个 $\mathbf{v}_k^{(i)}$ 都是从 $\mathcal{N}(\mathbf{0}, \mathbf{R}_k$中独立抽取的样本。这些模拟测量值 $\mathbf{y}_k^{(i)}$ 与公式 ($\mathbf{x}_{k|k-1}^{(i)}$ 2.17) 中的集合成员 $\hat{}$ 组合起来，定义 N 测量创新

$$\mathbf{e}_k^{(i)} = \mathbf{y}_k^{(i)} - \mathbf{H}\hat{\mathbf{x}}_{k|k-1}^{(i)}. \tag{2.20}$$

这些 $\mathbf{e}_k^{(i)}$ 代表从模型预测与测量值之间距离的分布中抽取的样本。

步骤 3： 通过公式 (2.17) 中的模型估计值和公式 (2.20) 中的测量创新，结合由公式 (2.18) 估计的卡尔曼增益，更新每个集合成员的状态估计

$$\hat{\mathbf{x}}_{k|k}^{(i)} = \hat{\mathbf{x}}_{k|k-1}^{(i)} + \hat{\mathbf{K}}_k\mathbf{e}_k^{(i)}. \tag{2.21}$$

我们可以使用 N 个更新后集合成员的均值，生成状态 $\hat{\mathbf{x}}_k$ 的点估计。每当有新的状态测量值可用时，就会重复这一过程；前一个数据点更新后的集合将成为新数据点的初始集合。

我们将前面介绍的两种技术结合起来，构成 DMDEnKF。这种新的混合方法使用动态模态分解为动力系统生成一个低维模型，然后随着新数据的出现，通过集合卡尔曼滤波对其进行迭代改进。

2.3 DMDEnKF

现在，我们介绍如何对系统的时间模态和状态进行滤波，同时固定所选动态模态分解版本在“预热”阶段得到的空间模态。需要指出的是，一旦我们在固定空间模态的同时允许时间模态变化，它们就不再是特征值/特征向量，因此我们将其称为时间模态。考虑一个在规则时间间隔 n测量的维状态 $\mathbf{x}_k \in \mathbb{R}^n$ $k = 1, ..., m$，随后在时刻$k = m + 1$,...进行迭代测量。

算法

步骤 1： 对数据集 $\mathbf{x}_1, ...,\mathbf{x}_m$执行所选版本的动态模态分解，并定义 $\mathbf{X}, \mathbf{X}'$如方程（2.2）中所示，从而得到表达式

$$\mathbf{x}_k = \mathbf{\Phi}\mathbf{\Lambda}^k\mathbf{b}, \tag{2.22}$$

其中

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_r \end{bmatrix}, \quad \mathbf{\Phi} = \begin{bmatrix} | & | & & | \\ \mathbf{d}_1 & \mathbf{d}_2 & \dots & \mathbf{d}_r \\ | & | & & | \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_r \end{bmatrix}, \tag{2.23}$$

并定义 λ_i、$\mathbf{d}_i$、 b_i 分别为保留的 r 个模态中的第 i 个时间模态、动态模态分解模态和初始条件三元组。这一过程充当预热过程，用于生成一个模型，随后我们可以使用集合卡尔曼滤波对其进行滤波。

步骤 2： 定义集合卡尔曼滤波的集合初始化、通过方程（2.15）进行传播以及根据方程（2.16）进行测量所需的矩阵。

首先，将每个 r 时间模态改写为极坐标形式：

$$\lambda_i = \tau_i e^{\theta_i \mathrm{i}}, \tag{2.24}$$

其中 $\tau_i \geq 0$、 $0 \leq \theta_i < 2\pi$ 和 $\mathrm{i}^2 = -1$。由于$\mathbf{x}_k \in \mathbb{R}^n$，DMD 模型谱中的时间模态要么是实数，要么构成共轭复数对。滤波时，我们将时间模态视为随时间变化的参数。不过，必须确保实时间模态保持为实数，且共轭复数对保持完整，这样才能保证模型输出的状态仍为实数。为此，我们将可滤波模型参数 μ_i 定义为以下变量： $i = 1, ..., r$

$$\mu_i = \begin{cases} \tau_i, & \text{if } \theta_i = 0, \text{ or } \nexists\ j \text{ for } j < i \text{ such that } \lambda_j^* = \lambda_i, \\ \theta_i, & \text{otherwise.} \end{cases} \tag{2.25}$$

以这种方式表示时，这些 μ_i 代表模型时间模态在满足状态估计为实数这一附加约束下的所有可能自由度。通过记录初始 DMD 中每个共轭复数对的位置索引，就可以从 μ_i 中恢复 λ_i 表示。虽然这种变换严格来说需要完整的 μ_i 列表，但我们非正式地写作 $\Lambda(\mu_i) = \lambda_i$，以表示从 μ_i 还原为 λ_i。

集合初始化： 现在，我们可以定义一个增广联合参数状态 $\mathbf{z}_0 \in \mathbb{R}^{n+r}$，作为集合卡尔曼滤波的初始状态

$$\mathbf{z_0} = \begin{bmatrix} - & \mathbf{x}_m & - & \mu_1 & \dots & \mu_r \end{bmatrix}^T. \tag{2.26}$$

我们将时间 $m+k$ 的联合参数状态记为 $\mathbf{z}_k \in \mathbb{R}^{n+r}$。为了从该状态生成初始集合，我们首先定义样本协方差 $\mathbf{C} = (1/m)(\mathbf{X}' - \mathbf{\Phi\Lambda\Phi}^{+}\mathbf{X})(\mathbf{X}' - \mathbf{\Phi\Lambda\Phi}^{+}\mathbf{X})^{T}$，它根据自启动动态模态分解中的预测误差表示当前状态不确定性。然后，我们构造初始协方差矩阵

$$\mathbf{P}_0 = \left[\begin{array}{c|c} \mathbf{C} & \mathbf{0} \\ \hline \mathbf{0} & \alpha_2 I_r \end{array}\right], \tag{2.27}$$

其中，$\alpha_2 > 0$, I_r 是 $\mathbf{r}$ 维单位矩阵，而 $\alpha_2 I_r$ 项决定了自启动动态模态分解时间模态的初始不确定性。从 $\mathcal{N}(\mathbf{z}_0, \mathbf{P}_0)$中进行独立抽样，直到集合规模足够大。最佳集合规模因问题而异，增加集合成员可以提高准确度，但会降低计算效率。

传播： 使用记号$\mathbf{z}_k^i$ 表示$\mathbf{z}_k$的第 i个元素，我们将状态$\mathbf{z}_k$ 对应的矩阵 $\mathbf{\Lambda}_{\mathbf{z}_k} \in \mathbb{R}^{r\times r}$ 定义为

$$\mathbf{\Lambda}_{\mathbf{z}_k} = \begin{bmatrix} \Lambda(\mathbf{z}_k^{n+1}) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Lambda(\mathbf{z}_k^{n+r}) \end{bmatrix}. \tag{2.28}$$

集合卡尔曼滤波的传播方程可写为

$$\mathbf{z}_{k+1} = \left[\begin{array}{c|c} \mathbf{\Phi\Lambda}_{\mathbf{z}_k}\mathbf{\Phi}^{+} & \mathbf{0} \\ \hline \mathbf{0} & I_r \end{array}\right] \mathbf{z}_k + \mathbf{w}_k. \tag{2.29}$$

为方便起见，我们引入记号$\mathbf{z}_k^{i:j} \in \mathbb{R}^{j-i+1}$，表示$\mathbf{z}_k$ 中从第 i个到第 j个元素，其中 $i \le j$

$$\mathbf{z}_k^{i:j} = \begin{bmatrix} \mathbf{z}_k^i & \dots & \mathbf{z}_k^j \end{bmatrix}^T. \tag{2.30}$$

方程（2.29）使用动态模态分解框架中的状态 $\mathbf{z}_k^{1:n}$ $\mathbf{x}_{m+k}$，利用标准动态模态分解方程和来自$\mathbf{\Lambda}_{\mathbf{z}_k}$的更新后时间模态向前推进。向量$\mathbf{z}_k^{n+1:n+r}$ 表示当前时间模态估计值的 μ_i 表示形式；除传播方程加入的噪声外，该向量保持不变。虽然我们假定时间模态随时间变化，但无法预先明确参数漂移的方向。向量$\mathbf{w}_k \in \mathbb{R}^{n+r}$是一个正态分布变量$\mathbf{w}_k \sim \mathcal{N}$（$\mathbf{0}$，$\mathbf{Q}_k$），表示系统模型中的不确定性。我们按如下方式构造$\mathbf{Q}_k$，

$$\mathbf{Q}_k = \left[\begin{array}{c|c} \alpha_1 I_n & \mathbf{0} \\ \hline \mathbf{0} & \alpha_2 I_r \end{array}\right], \tag{2.31}$$

其中，α_1 和 α_2 是由用户确定的常数，使得 $\alpha_2 \ll \alpha_1$。这一构造中，$\mathbf{Q}_k$ 是一个对角矩阵，假设$\mathbf{z}_k$ 各元素的模型误差彼此不相关。条件 $\alpha_2 \ll \alpha_1$ 确保动态模态分解系统的状态$\mathbf{z}_k^{1:n}$的变化速度显著快于其时间模态$\mathbf{z}_k^{n+1:n+r}$，因为参数按定义应随时间缓慢变化。此外，对经过滤波的时间模态的模而言，该条件可防止严格为正的模降至 0 以下。

测量： 我们将集合卡尔曼滤波的测量方程写为

$$\mathbf{y}_k = \left[\begin{array}{c|c} I_n & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \mathbf{z}_k + \mathbf{v}_k, \tag{2.32}$$

其中，$\mathbf{y}_k \in \mathbb{R}^n$ 是 DMD 状态 $\mathbf{x}_{m+k}$的观测值，而 $\mathbf{v}_k \in \mathbb{R}^n$ 是一个正态分布变量 $\sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$，表示测量值中的噪声。我们假设可获取完整 DMD 状态$\mathbf{y}_k$ 的新测量值$\mathbf{z}_k^{1:n}$，但无法获取其时间模态$\mathbf{z}_k^{n+1:n+r}$ 的测量值，因为这与用于生成预热 DMD 模型的数据格式一致。我们还假设状态的每个维度上的测量噪声彼此不相关，因此选择对角矩阵$\mathbf{R}_k$。

步骤 3 状态测量 $\mathbf{x}_{m+k}$ 在时刻 $k = 1,\ldots$ 正被迭代生成。通过将$\mathbf{y}_k = \mathbf{x}_{m+k}$ 设为每个新测量值，我们可以迭代应用集合卡尔曼滤波，生成$\mathbf{z}_k$ 的混合估计，该估计结合了来自 $\mathbf{z}_{k-1}$ 的模型预测和带噪测量值 $\mathbf{y}_k$。第 2.2 节简要概述了集合卡尔曼滤波如何实现这一点，更详细的说明见 [42]。

步骤 4： 只需取$\mathbf{z}_k$ 的前 n 个元素$\mathbf{z}_k^{1:n}$，即可重建原始系统的状态$\mathbf{x}_{m+k}$ 。此外，还可以根据$\mathbf{z}_k$ 预测时刻 $m+k$ 之后 p 步的状态：

$$\mathbf{x}_{m+k+p} = \mathbf{\Phi}\mathbf{\Lambda}_{\mathbf{z}_k}^p \mathbf{\Phi}^+ \mathbf{z}_k^{1:n}. \tag{2.33}$$

Hankel-DMDEnKF的算法定义方式完全相同，唯一的区别在于，Hankel-DMD应用于“预热”周期，而非标准DMD。

3 合成应用

3.1 与其他迭代式DMD变体的比较

为了测试DMDEnKF，我们首先将其应用于由具有时变特征值的合成系统生成的数据，以跟踪这些特征值。该系统的动力学由二维旋转矩阵控制，其旋转角度 θ_k 在 500 个时间步内从 $\pi/64$ 线性增加到 $\pi/8$。因此，该系统的状态$\mathbf{x}_k$ 演化可写为

$$\mathbf{x}_{k+1} = \begin{bmatrix} \cos(\theta_k) & -\sin(\theta_k) \\ \sin(\theta_k) & \cos(\theta_k) \end{bmatrix} \mathbf{x}_k, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{3.1}$$

其中 $\theta_k = \pi/64 + (k-1)(7\pi/64)/499$ 且 $k = (1,\ldots,500)$。

我们假设每个时间步都能获得状态的含噪测量值 $\mathbf{y}_k$ ，因此

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{v}_k, \quad \mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_2), \tag{3.2}$$

其中，每次实验 $\sigma = 0.05$或 0.5，分别用于模拟低水平或高水平的测量噪声。500 个 $\mathbf{y}_k$ 值（见图 2）用于训练 DMDEnKF 和 Hankel-DMDEnKF，其中前 100 个时间步用于执行 DMDEnKF 算法步骤 1 中所述的预热过程，以生成方程（2.23）所述的输出。

我们还将在该数据集上训练第 2.1 节末尾所述的 DMD 迭代变体（流式 TDMD[1], 窗口化 DMD 和在线 DMD），以比较它们跟踪

[1]由于合成数据集较小，对数据应用批处理方法在计算上是可行的。因此，我们没有应用真正的流式 TDMD 算法，而是使用截至当前时间步的所有数据执行批处理 TDMD，作为流式 TDMD 的替代方案，并采用 PyDMD 库 [19]中的代码。由于流式 TDMD 对 TDMD 的结果进行近似，二者唯一的差异是流式 TDMD 算法中额外的数据压缩步骤，因此我们认为这种替代是可以接受的。

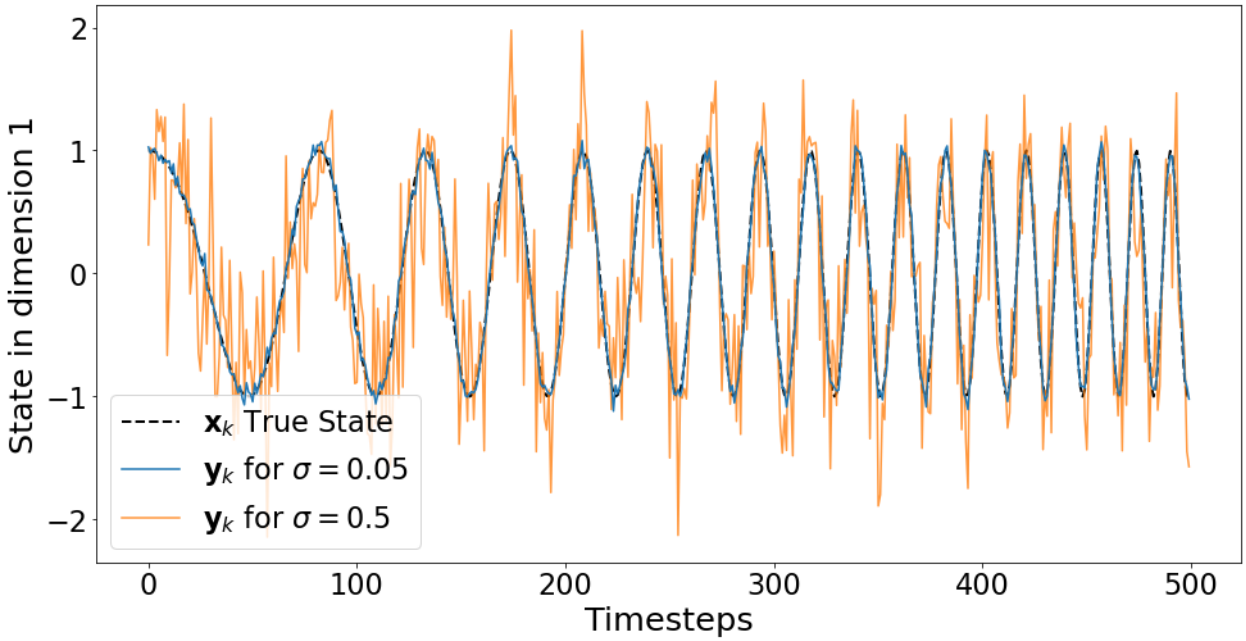


图 2：线性递增的特征值辐角的合成系统时间序列，显示状态的第一维在无测量噪声、低测量噪声（$\sigma = 0.05$）和高测量噪声（$\sigma = 0.5$）下的情况。

系统时变特征值的能力与 DMDEnKF 的能力。对于窗口化 DMD 算法，我们将 DMD 替换为 TDMD，使该方法能够有效处理数据中的噪声，以下将这两种方法的结合称为窗口化 TDMD。为实现在线 DMD，我们使用其创建者在此处提供的代码 [29]。计算参数设置如下：窗口化 TD MD 的窗口大小 $w = 10$，在线 DMD 的指数衰减率 $\rho = 0.9$，Hankel-DMDEnKF 的延迟嵌入维数 $d = 50$，以及 DMDEnKF 的预热时间步数 $m = 100$，如前所述。

在每个时间步 k，系统的真实特征值都可以写成模-辐角形式：

$$\lambda_k = 1e^{\pm\theta_k i}, \tag{3.3}$$

对于模型有定义的每个时间步，其对系统特征值的估计值也可以写成：

$$\hat{\lambda}_k = \hat{\tau}_k e^{\pm\hat{\theta}_k i}. \tag{3.4}$$

我们首先比较各方法对恒定特征值模（$\hat{}\tau_k - 1$）的估计误差。针对每个 σ 值，我们生成了 1000 组该合成系统的数据，并收集了预热期之后每个时间步上各方法所估计的特征值模和辐角与其真实值之间的差异。在某次运行的某个时间步中，如果任何一种方法未能将特征值识别为共轭复数对，则使用主导特征值的模进行模误差计算。特征值模估计的平均误差见表1。

在所有测量噪声水平下，流式 TDMD 对特征值模的估计最准确。这是因为该方法假设系统平稳，因此对每个数据点的重要性赋予相同权重；在估计恒定参数时，这种做法效果良好。在表 1的第一列所示的低测量噪声水平下，窗口化 TDMD、在线 DMD、DMDEnKF 和 Hankel-DMDEnKF 的表现相近

迭代DMD 变体	平均特征值模误差 $\sigma = 0.05$	$\sigma = 0.5$
窗口化 TDMD	9.82×10^{-3}	1.39
在线 DMD	6.04×10^{-3}	3.06×10^{-1}
流式 TDMD	2.31×10^{-4}	2.50×10^{-3}
DMDEnKF	8.07×10^{-3}	1.89×10^{-2}
Hankel-DMDEnKF	9.49×10^{-3}	1.38×10^{-2}

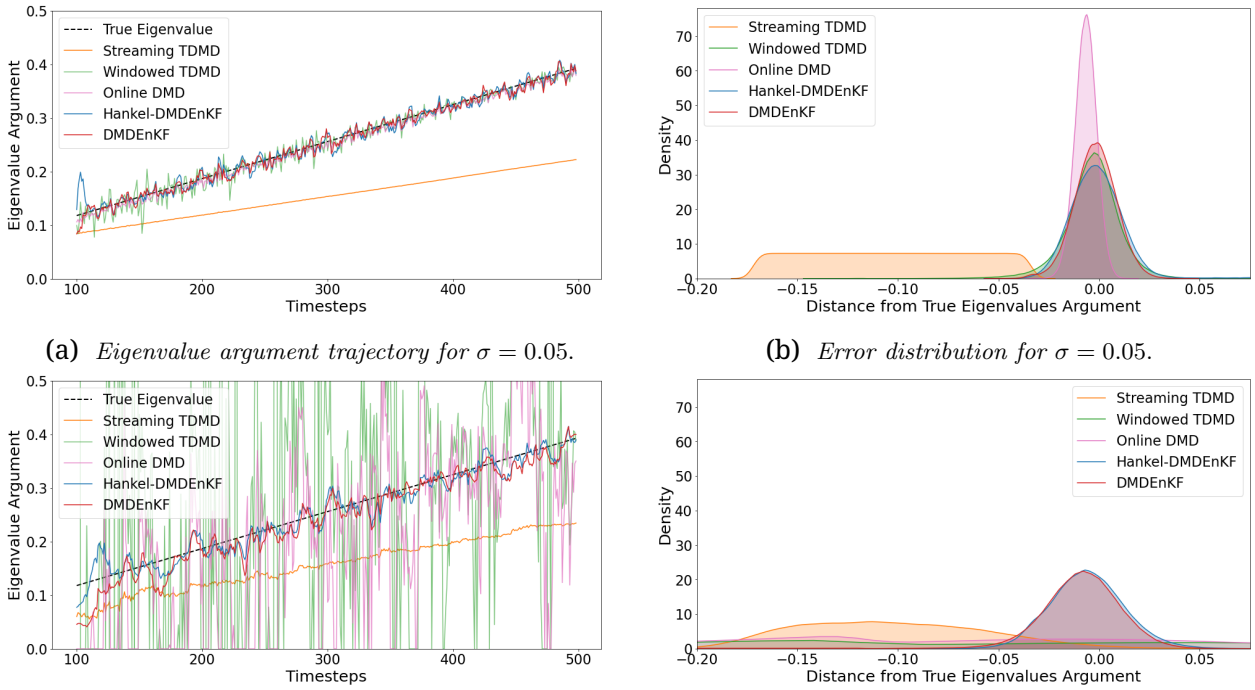
表1： 在全部 *1000* 次运行的所有时间步中，各迭代 *DMD* 变体对合成系统特征值模估计所得的平均绝对误差。测量噪声设置为低水平 $\sigma = 0.05$ （左）或高水平 $\sigma = 0.5$ （右）。流式 *TDMD* 的误差显著低于所有其他方法；随着噪声水平升高，窗口化 *TDMD* 和在线 *DMD* 的误差显著增大，远高于 *DMDEnKF* 和 *Hankel-DMDEnKF* 的误差。

，平均特征值模误差均低于 0.01。由于预测未来状态时，特征值模误差会呈指数增长，这 4 种方法能够给出可接受的短期预测，但随着预测时域延长，预测结果很快就会偏离真实状态。对于表1第二列所示的高噪声水平，窗口化 TDMD 和在线 DMD 的特征值模估计显著退化，因此不适用于此情形下的预测。不过，DMDEnKF 和 Hankel-DMDEnKF 的误差仍然相当小，但依然比流式 TDMD 产生的误差高一个数量级。

ˆ 从预热期结束起，在一次运行过程中，各方法特征值辐角估计（θ_k）的典型轨迹见图 3a 和 3c。图中绘出了全部 1000 次运行中各方法特征值辐角估计的 ˆ 误差分布（$\theta_k - \theta_k$），见图3b和 3d。

如图所示，在较低噪声水平下，图 3a 和 3b中展示的结果表明，5 种方法平均而言都低估了系统的特征值辐角。对此结果并不意外，因为特征值辐角随时间增大，这意味着除每种方法可用的最后一个数据对之外，其他数据对都是在特征值辐角小于其当前值时生成的。流式 TDMD 的表现最差，这同样是因为它对每个数据点都赋予相等权重；不过在本例中，这反而成了缺点，因为它妨碍模型适应反映参数变化的新数据。窗口化 TDMD、在线 DMD、DMDEnKF 和 Hankel-D MDEnKF 的表现相近。在线 DMD 的误差分布更集中，但其偏差略大于窗口化 TDMD。这表明，与窗口化 TDMD 采用的硬截断相比，在线 DMD 的软阈值处理能够减少测量噪声引起的模型波动性。然而，出于同样的原因，在线 DMD 对新测量值的适应速度慢于窗口化 TDMD，因此其估计值相对于系统真实的特征值辐角存在更大的低估偏差。在这一噪声水平下，DMDEnKF 和 Hankel-DMDEnKF 的表现与窗口化 TDMD 非常相似；不过，如有需要，可以调整 DMDEnKF 的系统不确定性矩阵的幅度，以平衡模型创新速度与模型波动性，从而生成更接近在线 DMD 的误差分布。

在图3c 和3d所示的较高噪声水平下，窗口化 TDMD 和在线 DMD 的性能显著下降。对近期样本赋予更大的权重，使这些方法



(a) *Eigenvalue argument trajectory for* $\sigma = 0.05$.

(b) *Error distribution for* $\sigma = 0.05$.

(c) *Eigenvalue argument trajectory for* $\sigma = 0.5$.

(d) *Error distribution for* $\sigma = 0.5$.

图 3： 各迭代 *DMD* 变体对合成系统特征值辐角的估计结果。左图展示了在 *1* 次实验运行过程中，每个时间步的特征值辐角的典型轨迹；右图展示了在全部 *10* 次运行的整个过程中所有时间步上，真实系统特征值辐角与估计的特征值辐角之差的误差分布。测量噪声分别设置为低水平（$\sigma = 0.05$，上图）或高水平（$\sigma = 0.5$，下图）。在测量噪声较低时，*DMDEnKF* 和 *Hankel-DMDEnKF* 的误差与在线 *DMD* 和窗口化 *TDMD* 相近；但在测量噪声较高时，它们对特征值辐角的跟踪精度远高于后两者。

能够快速适应系统参数的变化，但随着噪声增加，这会导致各自模型出现极端波动。与低噪声情况相比，流式 TDMD 的性能没有太大变化；尽管仍落后于真实系统的取值，但由于对所有数据点进行对称处理，它在一定程度上不受噪声影响。在 DMDEnKF 框架中明确纳入测量噪声的优势由此显现：在这一噪声水平下，DMDEnKF 和 Hankel-DMDEnKF 是所测试的唯一能够准确估计特征值辐角的技术。

此外，我们在这里看到了 DMDEnKF 和 Hankel-DMDEnKF 性能上的第一个显著差异：DMDEnKF 的误差分布具有一条细长的尾部，一直延伸到 $-\pi/8$，而 Hankel-DMDEnKF 的误差分布中没有这一现象。这些额外误差是由 DMDEnKF 方法中的 DMD 预热过程偶尔未能将系统的特征值识别为共轭复数对所导致的（根据经验，这种情况发生的概率为 $\sim 3\%$），原因是数据中的噪声增大。一旦发生这种情况，DMDEnKF 就会彻底失效，因为无论滤波多少个未来时间步，集合卡尔曼滤波都无法根据其公式（2.25）从实特征值生成复特征值。可以通过以下方式缓解 DMDEnKF 的这种失效。如果在滤波阶段，DMDEnKF 产生的误差在较长时间内被认为过大（例如超过给定阈值），那么可以在扩展后的数据集上重新运行 DMD 预热过程，该数据集由原始预热数据以及

迄今为止在滤波步骤中使用的新增可用数据组成。通过在预热过程中纳入更多数据，预热 DMD 模型更有可能成功地将数据中的信号分量与测量噪声区分开来，从而生成结构与真实系统特征值相同的特征值。时间延迟嵌入使 DMD 算法中的奇异值分解步骤对测量噪声更加稳健 [16]。因此，尽管 Hankel-DMDEnKF 在滤波阶段同样受限于其能够生成的特征值，但在全部 1000 次合成数据运行中，自启动 Hankel-DMD 都能够将系统的特征值识别为共轭复数对，所以这对 Hankel-DMDEnKF 并不是问题。

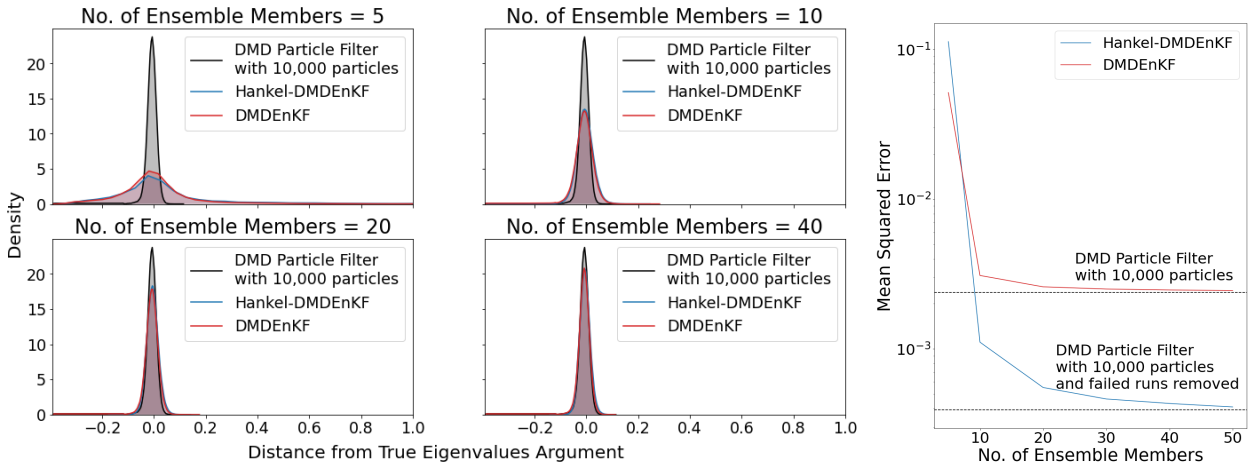
3.2 与采用粒子滤波器的 DMD 进行比较

在将 DMDEnKF 的性能与其他迭代式DMD变体进行比较后，我们现在重点评估算法的滤波组件。由于线性 DMD 模型作用于模型状态和特征值时，在滤波器中表现为非线性，因此我们将 EnKF 滤波器与粒子滤波器进行比较。对于具有非高斯噪声的一般非线性模型，已有研究表明，随着粒子数量趋于无穷大，粒子滤波器 [27] 会收敛到最优滤波器 [17]。不过，粒子滤波器仅适用于低维系统，因为所需粒子数量会随着状态维度的增加而近似呈指数增长 [57]。因此，我们将 DMDEnKF 和 Hankel-DMDEnKF 与 DMD 加粒子滤波器进行比较，并将后者视为估计的“黄金标准”，以评估 EnKF 在非线性滤波问题中的表现。

我们使用与上一小节相同的合成系统（3.1），该系统具有线性递增的特征值辐角，用于生成测量噪声水平较高的数据（$\sigma = 0.5$）；其轨迹见图 2。同样，对于 Hankel-DMDEnKF，时滞嵌入维度 $d = 50$，并使用前 100 个时间步训练自启动 DMD 模型，再用接下来的 400 个时间步对状态和自启动模型的特征值进行滤波。

因此，DMDEnKF 的滤波状态维度为 4（2 个状态维度和 2 个时间模态），而 Hankel-DMDEnKF 的滤波状态维度为 102（100 个状态维度和 2 个时间模态）。为生成“黄金标准”解，在滤波步骤中，我们使用包含 10,000 个粒子的粒子滤波器，并在有效样本量低于粒子数的一半时进行多项式重要性重采样 [20]，以避免样本退化 [21]。在 DMDEnKF 和 Hankel-DMDEnKF 的滤波步骤中，我们使用不同数量的集合成员（N）运行集合卡尔曼滤波，观察随着 N 增加，它们的估计均值和协方差是否会趋近于粒子滤波器集合的均值和协方差。对于每个 N 取值，我们生成了 1000 组该合成数据，分别应用粒子滤波器和集合卡尔曼滤波，并收集各方法在每个时间步的特征值辐角估计误差。

如图 4a所示，采用 10,000 个粒子的 DMD粒子滤波器生成的误差分布极为集中，但略有偏差，估计值低于真实特征值辐角。这正如上一小节所述，是系统的特征值辐角不断增大的预期结果。误差分布中还有一条细长的尾部，一直延伸至 $-\pi/8$。这同样是因为启动DMD有时无法识别共轭复特征值对，导致粒子滤波器陷入特征值为实数这一错误假设。



(a) *Error distributions.* (b) *Mean squared errors.*

图4: 合成系统特征值辐角估计值的误差分布（左）和均方误差（右）。这些估计值由使用不同数量集合成员（N）的 *DMDEnKF* 和 *Hankel-DMDEnKF* 得出，并与使用 *10,000* 个粒子的粒子滤波器所得结果进行比较。增加 N 可迅速使 *DMDEnKF* 和 *Hankel-DMDEnKF* 的误差水平接近各自“黄金标准”所产生的误差水平。

当集合成员数量较少（$N = 5$）时，DMDEnKF 和 Hankel-DMDEnKF 的中心值与“黄金标准”相近。然而，两者产生的分布范围要大得多，且两端都有长尾，这表明集合成员数量如此之少时，方法缺乏稳健性。集合成员数量仅略微增加至 $N = 10$ 后，两种方法都变得更加稳定；尽管其方差仍大于粒子滤波器，但来自 $N = 5$ 的长正向尾部已消失。增加到 $N = 20$ 时也呈现类似趋势：集合成员越多，误差分布越集中。此时，Hankel-DMDEnKF 的分布与 DMDEnKF 和 DMD粒子滤波器的分布有所不同，表现为前文提到的、没有持续存在的细长负向尾部。到了 $N = 40$，从图形尺度上看，DMDEnKF、Hankel-DMDEnKF 和“黄金标准”的主要峰值几乎无法区分；DMDEnKF 和 DMD粒子滤波器则都有一条细长的负向尾部。

图 4b展示了 DMDEnKF 和 Hankel-DMDEnKF 预测的特征值辐角均方误差如何受到集合成员数量变化的影响。对于 DMDEnKF，随着 N 增加，误差最初会急剧下降；然而，在这个规模较小的合成系统中，超过 $N = 20$后，收益很快便会递减。当达到 $N = 50$时，DMDEnKF 的均方误差仅比“黄金标准”高 $\sim 3\%$，而使用的粒子数量却少了 200 倍。将 Hankel-DMDEnKF 与“黄金标准”比较时，DMD粒子滤波器的特征值估计误差受到某些运行结果的偏斜：在这些运行中，启动DMD未能识别出共轭复特征值对，而 Hankel-DMD 在这些合成示例中没有遇到这一问题。为尽可能公平地比较滤波方法，我们先剔除所有启动DMD以这种方式失败的运行，然后再次计算 DMD 粒子滤波器的均方误差，并将其记录在图 4b中。随着集合规模增大，Hankel-DMDEnKF 的误差也呈现出类似的下降趋势，但收益逐渐递减；达到$N = 50$ 时，其均方误差与重新计算的 DMD 粒子滤波器得分相差不到 5%。我们的结果表明，至少在这个简单的合成案例中，集合卡尔曼滤波是解决 DMDEnKF 框架中非线性滤波问题的一种高效且有效的方法。

3.3 跟踪合成生成的大流行病

最后，我们在合成数据上测试 DMDEnKF 的性能；这些数据旨在模拟一场简单的大流行病，其中状态$\mathbf{x}_k$ 表示 3 种不同人口类别中的感染程度。系统的动力学由矩阵$\mathbf{A}$ $\in \mathbb{R}^{3\times3}$ 控制，该矩阵由我们随机生成，其非负元素均从均匀分布中抽取 $\mathcal{U}[0,1)$。$\mathbf{A}$ 的第（i,j）个元素表示在时间 k，类别 j 中的感染程度将如何影响时间 $k+1$类别 i中的感染程度。为了控制合成大流行病是蔓延还是消退，我们接着定义一个新矩阵$\hat{}\mathbf{A} = \mathbf{A}\ \lambda_1$，其中 λ_1 是$\mathbf{A}$的最大特征值，从而确保谱半径$\rho(\hat{}\mathbf{A}) = 1$。引入常数 γ后，我们可以用 $\gamma\ \hat{}\mathbf{A}$ 替换$\mathbf{A}$，使状态在$\gamma > 1$ 时增长，在 $\gamma < 1$时衰减。在实验的 1000 个时间步内，我们将 γ 从 1.01 线性递减至 0.99，以模拟一场大流行病。所用初始状态为全 1 向量。生成合成数据的系统可写为

$$\mathbf{x}_{k+1} = \gamma_k \hat{\mathbf{A}} \mathbf{x}_k, \quad \mathbf{x}_1 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T, \tag{3.5}$$

其中状态 $\mathbf{x}_k \in \mathbb{R}^3$, $\gamma_k = 1.01 - {}^{0.02}(k-1)\ 999$ 以及 $k = (1,...,1000)$。我们假设无法获取系统的真实状态 $\mathbf{x}_k$，而只能获得噪声测量值 $\mathbf{y}_k$，其定义为

$$\mathbf{y}_k = \mathbf{x}_k + \mathbf{v}_k, \tag{3.6}$$

控制测量噪声水平的常数 σ 设为 $\sigma = 0.05$ 表示低噪声，设为 $\sigma = 0.5$ 表示高噪声，如（3.2）所示。图 5 展示了一次运行过程中系统三个状态维度的取值以及相应的可用测量值。

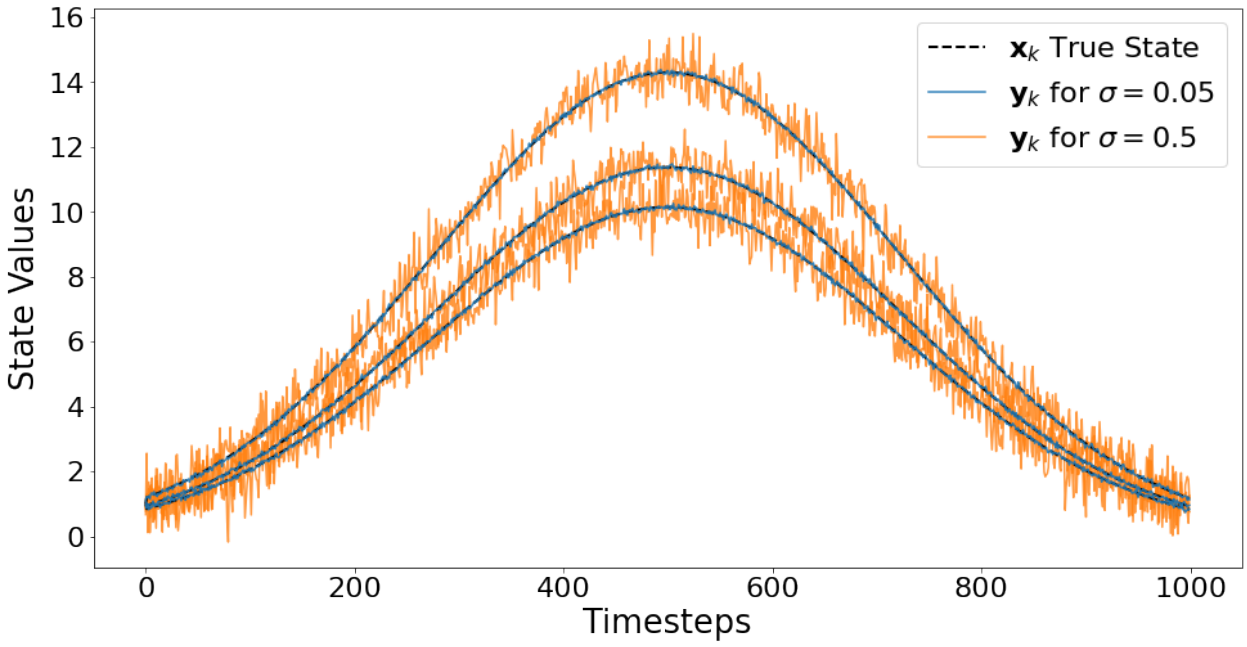


图5： 模拟大流行病的合成系统的时间序列，展示状态的全部3个维度，其中测量噪声为无、低（$\sigma = 0.05$）和高（$\sigma = 0.5$）。

测试的全部五种 DMD变体均采用与第3.1节合成实验中相同的计算参数。唯一的细微差别是，流式 TDMD、窗口化 TDMD、DMDEnKF 和 Hankel-DMDEnKF 通过移除最小奇异值来截断数据，以减少因该方向上的信噪比通常非常低而导致的模型不稳定性

。在线 DMD 未对数据进行任何截断，因为该方法并非为此而设计；不过，似乎并未因此出现任何稳定性问题。

前 100 个测量值（$\mathbf{y}_1$，…，$\mathbf{y}_{100}$）用于初始化模型；随着每个新的数据点（$\mathbf{y}_{100}$，…，$\mathbf{y}_{1000}$）依次输入模型，模型会生成提前 50 步的预测（$\hat{}\mathbf{x}_{150}$，…，$\hat{}\mathbf{x}_{1050}$）。我们生成了 1000 个数据点，但标准流感季通常持续约 20 周。因此，我们选择提前预测 50 步，以便在更现实的时间尺度上模拟提前 1 周进行预测。随后可以计算相对预测误差 $\hat{}e_k = \frac{\|\mathbf{x}_k - \ _k\|}{\|\ _k\|}\hat{}\mathbf{xx}$，计算范围为 $k =$（150，…，1000），这些误差的平均值是我们用来评估每种方法在一次运行过程中预测能力的主要指标。在低噪声和高噪声水平下，我们分别进行了 1000 次运行；低噪声情况下提前 50 步的预测平均运行相对误差的经验累积分布（$\sigma = 0.05$）如图 6所示。

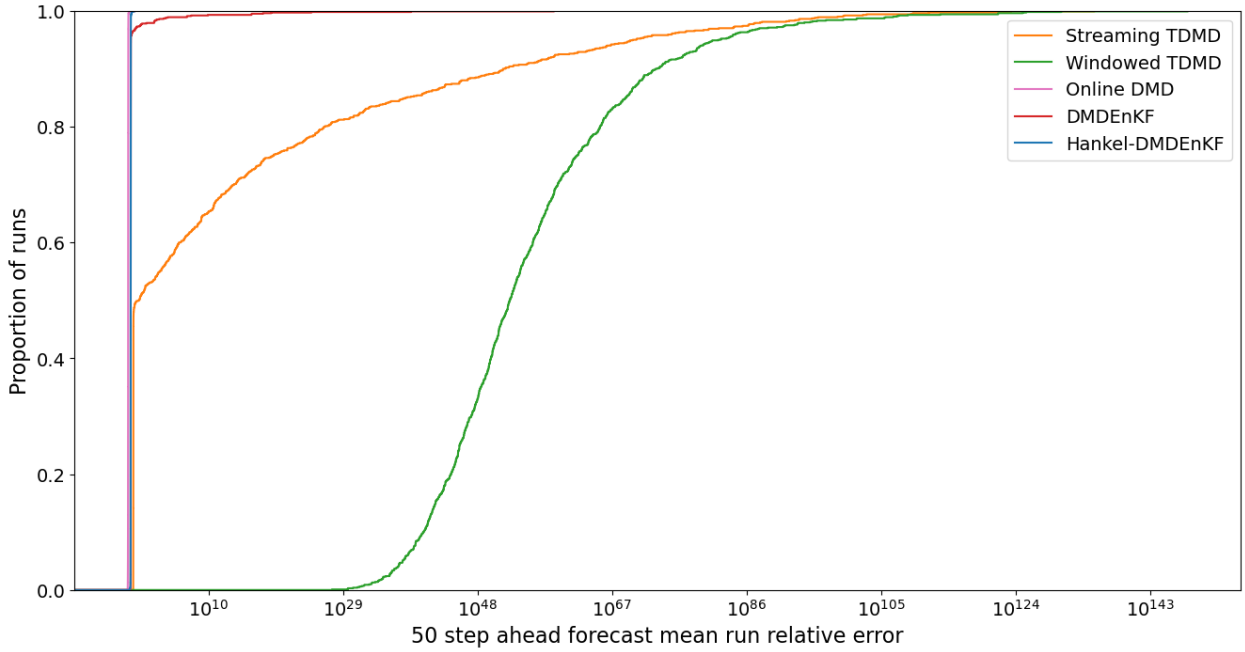
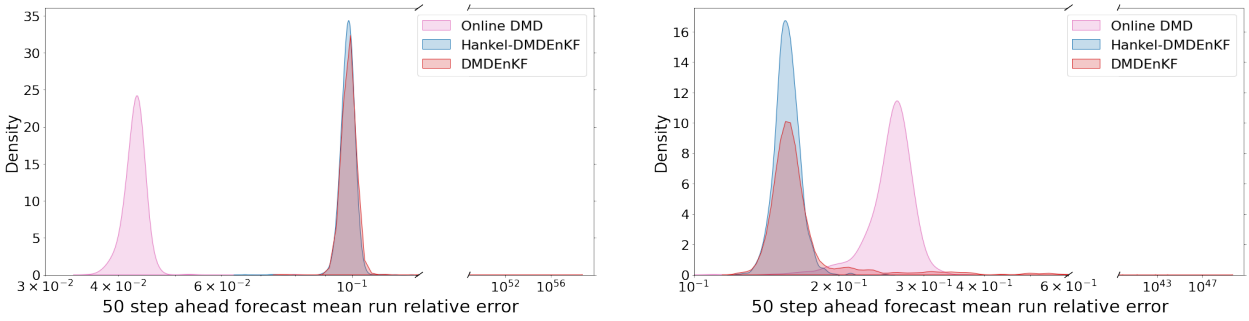


图6：各迭代 *DMD* 变体提前 *50* 步预测的平均单次运行相对误差累积分布。对实验的每次运行，均在所有时间步上计算平均相对误差；此处展示了测量噪声较低（$\sigma = 0.05$）时 *1000* 次运行的结果。由于提前 *50* 步进行预测会导致误差呈指数级累积，某些方法的预测误差范围较广。*DMDEnKF*、*Hankel-DMDEnKF* 和在线 *DMD* 所产生的误差比流式 *TDMD* 和窗口化 *TDMD* 小数个数量级。

累积误差分布中第一个值得注意的特点，是某些方法的预测误差范围很广。这是因为，50 步超前预测是通过训练每个模型进行 1 步超前预测，再将训练好的模型应用于数据 50 次而得到的。因此，预测误差会呈指数级累积，1 步预测时域内的微小误差经过 50 次迭代后可能变得非常大。考察各个方法可以发现，窗口化 TDMD 的表现最差。这是由于前文所述的测量噪声下的不稳定性，而这种不稳定性是由每次仅考虑一小部分数据造成的。增大窗口大小（w）这一计算参数可以降低这种不稳定性；然而，随着 w 增大，模型跟踪随时间变化的系统的能力会减弱。流式 TDMD 的误差位居第二，这是因为该方法假设系统平稳，这一假设妨碍了它在新数据不断出现时正确适应系统变化的特征值，

。在大多数情况下，在线 DMD、DMDEnKF 和 Hankel-DMDEnKF 的表现都相近且出色。这三种方法的累积误差分布都紧密集中在较低的误差值附近；然而，在少数几次运行中，DMDEnKF 变得不稳定并产生了较大的误差。显然，即使在噪声水平较低时，对于这类系统，在线 DMD、DMDEnKF 和 Hankel-DMDEnKF 的预测表现也远优于窗口化 TDMD 和流式 TDMD。因此，我们现在仅关注表现最佳的这三种方法，以便在合适的误差尺度上对它们进行全面比较。



(a) $\sigma = 0.05$ (b) $\sigma = 0.5$

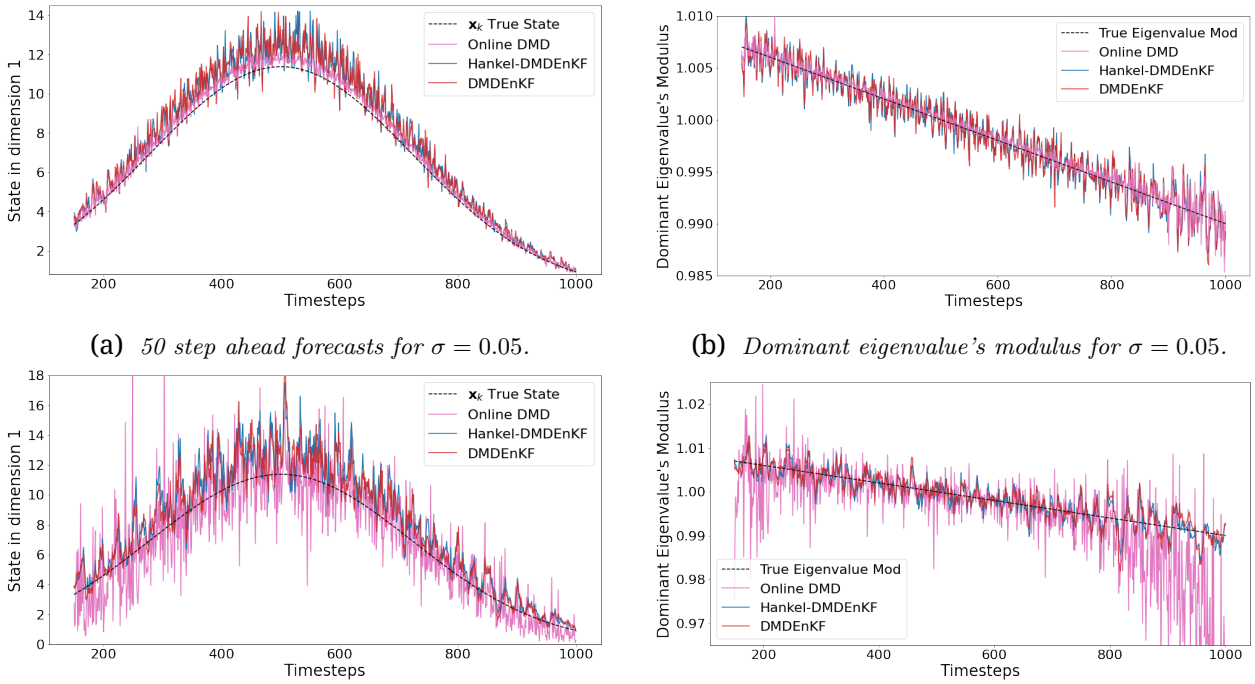
图 7： 在低 $\sigma = 0.05$ （左）和高$\sigma = 0.5$ （右）测量噪声水平下，经过 *1000* 次运行得到的在线 *DMD*、*DMDEnKF* 和 *Hankel-DMDEnKF* 提前 *50* 步预测的平均运行相对误差分布。两种噪声水平下的误差相近；测量噪声较低时，在线 *DMD* 表现更佳，而测量噪声较高时，*DMDEnKF/Hankel-DMDEnKF* 表现更佳。

在图 7 中，我们绘制了在线 DMD、DMDEnKF 和 Hankel-DMDEnKF 在测量噪声水平较低和较高时，提前 50 步预测的平均运行相对误差分布。在噪声水平较低时，在线 DMD 的误差峰值低于 DMDEnKF 和 Hankel-DMDEnKF；然而，随着噪声水平升高，峰值的相对位置发生反转，DMDEnKF/Hankel-DMDEnKF 成为表现更好的方法。在两种噪声水平下，DMDEnKF 的误差分布峰值都与 Hankel-DMDEnKF 的峰值位于相同位置，但其峰值密度较低，这是因为 DMDEnKF 的误差分布长尾中包含额外的概率质量，而 Hankel-DMDEnKF 的误差分布中则没有这部分概率质量。

DMDEnKF 分布尾部出现的这些异常大的误差，是因为自启动动态模态分解过程未能生成与系统真实动力学足够相似的模型。正如第一个合成示例中简要提到的，如果预热 DMD 模型不够准确，就可能导致集合卡尔曼滤波无法有效同化新数据，进而使 DMDEnKF 灾难性失效。在这个示例中，由于第二大奇异值方向上的信噪比往往较低，预热期内系统动力学（$\mathbf{A}$）和测量噪声（$\mathbf{v}_k$）的一次不利随机抽样，可能导致 DMD 的第二个特征值出现较大误差。根据经验，使用四分位距法检测异常预测误差后发现，对于 $\sigma = 0.05$，这种 DMDEnKF 失效发生的概率为 5.5%。随着新数据被处理，这些误差会在滤波步骤中持续存在，而其他方法则能够更有效地从较差的初始模型估计中恢复。初始 DMD 所生成模型的质量取决于预热数据的质量和数量，因此在较高噪声水平下，这个问题会加剧，发生频率也会高得多（对于 $\sigma = 0.5$，发生概率为 21.9%）。可以通过

增加预热阶段的时间步数来在一定程度上缓解这一问题，具体如第3.1节末尾所述。不过，与窗口化 TDMD 类似，由于假设系统随时间变化，当预热期过长时，批量 DMD 的平稳性假设会逐渐遭到更严重的违背，因此可能存在一个效果开始适得其反的临界点。

与 DMDEnKF 不同，Hankel-DMDEnKF 和在线 DMD 的误差分布没有长尾现象，并且在全部 1000 次运行中表现始终良好。在两种噪声水平下，它们的误差分布方差相近，其中 Hankel-DMDEnKF 的误差比在线 DMD 的误差略微集中。因此，在本例中，区分各方法性能的主要因素是平均误差：低噪声下，在线 DMD 是首选方法；高噪声下，Hankel-DMDEnKF（或者在自启动动态模态分解未发生灾难性失败时采用 DMDEnKF）更准确。由于这两种方法各有实用之处但特点不同，我们分别针对低测量噪声和高测量噪声生成一条典型数据轨迹（其中 DMDEnKF 不会失败）。然后，我们考察每次运行过程中各模型的提前 50 步预测和主导特征值估计如何变化，如图 8 所示。



(a) *50 step ahead forecasts for* $\sigma = 0.05$.

(b) *Dominant eigenvalue's modulus for* $\sigma = 0.05$.

(c) *50 step ahead forecasts for* $\sigma = 0.5$.

(d) *Dominant eigenvalue's modulus for* $\sigma = 0.5$.

图8： 在一次运行过程中，测量噪声水平较低（$\sigma = 0.05$）和较高（$\sigma = 0.5$）时，在线 *DMD*、*DMDEnKF* 和 *Hankel-DMDEnKF* 对状态第一维数值进行提前 *50* 步预测（左），以及对主导特征值当前模的估计值（右）的典型轨迹。低噪声下，在线 *DMD* 对未来 *50* 步的预测更准确；高噪声下，*DMDEnKF/Hankel-DMDEnKF* 的预测更准确。然而，当信噪比较低时（实验开始和结束时），在线 *DMD* 的特征值估计会变得不稳定。

首先，观察图 8a中的低噪声预测，可以清楚地看出，在线 DMD 生成的预测比 DMDEnKF 和 Hankel- DMDEnKF 的预测更加稳健，也更接近真实状态的数值。鉴于在这一噪声水平下，在线 DMD 在图 7a所示的误差分布中具有较低的平均误差，这一结果并不意外。随着噪声增大，图 8c 中的预测表明，DMDEnKF 和 Hankel-DMDEnKF 成为更准确的方法；然而，在线

DMD 的预测仍相当稳定，看起来仍是可行的预测方案。

分析图中的特征值估计 8b和 8d，我们可以看到，在数据的中间部分 $k = (250, ..., 750)$，在线 DMD 能够有效跟踪主导特征值。然而，在数据集的开头和结尾，状态较小，因此每个新数据点的信号分量相对于测量噪声也较小，此时在线 DMD 的特征值估计会逐渐变得更加不稳定。在低噪声情况下，这并不是问题，因为在线 DMD 的估计值明显比 DMDEnKF/Hankel-DMDEnKF 的估计值更准确；因此，即使在数据表现较差的部分，其估计值仍然优于或至少不逊于 DMDEnKF 的估计值。然而，在噪声较大的情况下，在线 DMD 在数据集的开头和结尾对主导特征值的估计远不如 DMDEnKF 和 Hankel-DMDEnKF 稳健。在传染病暴发的流行病学背景下（本合成示例试图模拟这一情形），科学家通常会尝试利用少量初始感染病例产生的含噪数据来计算基本再生数（R_0）[58] 。如果 $R_0 > 1$，那么若不加以控制，感染病例数将呈指数增长；如果 $R_0 < 1$，感染病例数则会自然衰减至 0。在此示例中，借助 DMDEnKF/Hankel-DMDEnKF，可以利用其早期特征值估计的稳定性，迅速判断初始时 $R_0 > 1$，并采取任何必要的措施；而在线 DMD 则需要更长时间，且感染水平更高，才能持续判断 R_0 是否高于或低于增长/衰减阈值。

4 季节性流感样疾病应用

4.1 问题设置

基于DMD的方法此前已应用于传染病数据 [49]。在这种情况下，DMD模态可视为静态空间模态，用于构建降阶模型，其中只有振幅和频率随时间变化 [9]。因此，流感样疾病（ILI）数据建模是DMDEnKF/Hankel-DMDEnKF极具潜力的应用场景。

美国疾病控制与预防中心的ILINet数据 [25] 记录了美国每周流感样疾病全科医生（GP）就诊次数，以及全科医生就诊总次数，后者可用于对流感样疾病数据进行归一化。我们使用从2003年初（全年数据首次可用的一年）至2018年末的数据子集，如图1所示。随后，我们将每周数据按人口统计群体划分，包括4个年龄组（0-4、5-24、25-24、 65+）以及10个卫生与公众服务部（HHS）地区。每个地区包括以下地点：

- 第1区 - 康涅狄格州、缅因州、马萨诸塞州、新罕布什尔州、罗得岛州和佛蒙特州。
- 第2区 - 新泽西州、纽约、波多黎各和美属维尔京群岛。
- 第3区 - 特拉华州、哥伦比亚特区、马里兰州、宾夕法尼亚州、弗吉尼亚州和西弗吉尼亚州。
- 第4区 - 阿拉巴马州、佛罗里达州、佐治亚州、肯塔基州、密西西比州、北卡罗来纳州、南卡罗来纳州和田纳西州。
- 第5区 - 伊利诺伊州、印第安纳州、密歇根州、明尼苏达州、俄亥俄州和威斯康星州。

- 第6区 - 阿肯色州、路易斯安那州、新墨西哥州、俄克拉何马州和得克萨斯州。
- 第7区 - 艾奥瓦州、堪萨斯州、密苏里州和内布拉斯加州。
- 第8区 - 科罗拉多州、蒙大拿州、北达科他州、南达科他州、犹他州和怀俄明州。
- 第9区 - 亚利桑那州、加利福尼亚州、夏威夷州和内华达州。
- 第10区 - 阿拉斯加州、爱达荷州、俄勒冈州和华盛顿州。

这40个分层均有流感样疾病（ILI）就诊数据，但全科医生（GP）就诊总数据仅按区域提供。为了按年龄细分某一区域的总就诊量，我们利用人口普查数据进行线性插值，以估算特定周美国人口的年龄人口结构。随后，根据各年龄组占总人口的比例，将该区域的总就诊量分配给各年龄组。此方法假设所有年龄组就诊全科医生的可能性相近；这一假设或许并不完全合理，但我们认为，对于展示 DMDEnKF 在真实世界数据上的应用而言，这种方法已经足够。

4.2 构建预热 DMD 模型

因此，DMDEnKF中使用的流感样疾病数据格式为每周一个40维向量，记录各人口统计群体中流感样疾病就诊占全科医生就诊总数的百分比。这些数据存在于 $\mathbb{R}_{+}^{40}$ 然而，动态模态分解是在 $\mathbb{R}^{40}$中计算模态的。因此，为确保模型的估计值为非负值，我们首先通过加上一个小常数（$c = 1$）对数据进行变换，然后对每个元素取自然对数。对于Hankel-DMDEnKF，随后使用之前的99个时间步（$d = 100$）对这些变换后的数据进行延迟嵌入，从而在 $\mathbb{R}^{4000}$中形成一个状态。我们将截至2012年底的数据用作DMDEnKF/Hankel-DMDEnKF预热 DMD 过程的训练数据；这些过程在DMDEnKF算法的步骤 1中进行了详细说明。然后，对2013年至2018年剩余年份的数据进行滤波。变换并居中的数据，以及标记预热过程结束和滤波开始位置的分割线，如图9所示。

我们最初选择将 DMDEnKF 截断为 8 个 DMD模态，以展示该方法。下文将讨论改变截断对 DMDEnKF 方法的影响；但在保留 8 个 DMD模态时，通过保留更多模态而在数据中额外保留的方差量已显著减少。图 14a 的累积方差图中出现的“肘部”对此进行了说明，该点表示图线从方差快速增加过渡到 r 更为平缓的上升。我们还将 Hankel-DMDEnKF 截断为 8 个 DMD模态，以便更直接地比较这两种变体。

各个自启动动态模态分解过程识别出的频率所对应的谱以及主要 DMD/Hankel-DMD 模态见图 10。图 10a和 10c 所示的所有特征值的模均为 ~ 1，这意味着在这两种情况下，各模态预计都会在数据中持续存在，而不会呈指数增长。两种方法的谱之间的主要差异在于，Hankel-DMD 识别出的最主要模态周期为一年，而 DMD 未能检测到任何周期为一年的模态。流感样疾病就诊的年度峰值每年大致出现在相近的时间，这表明数据包含一个周期为一年的强模态；傅里叶分析也支持这一结论 [10] ，并同样将一年识别为主导周期。因此，DMD 漏掉了数据中存在的年度模态，而 Hankel-DMD 能够检测到该模态；

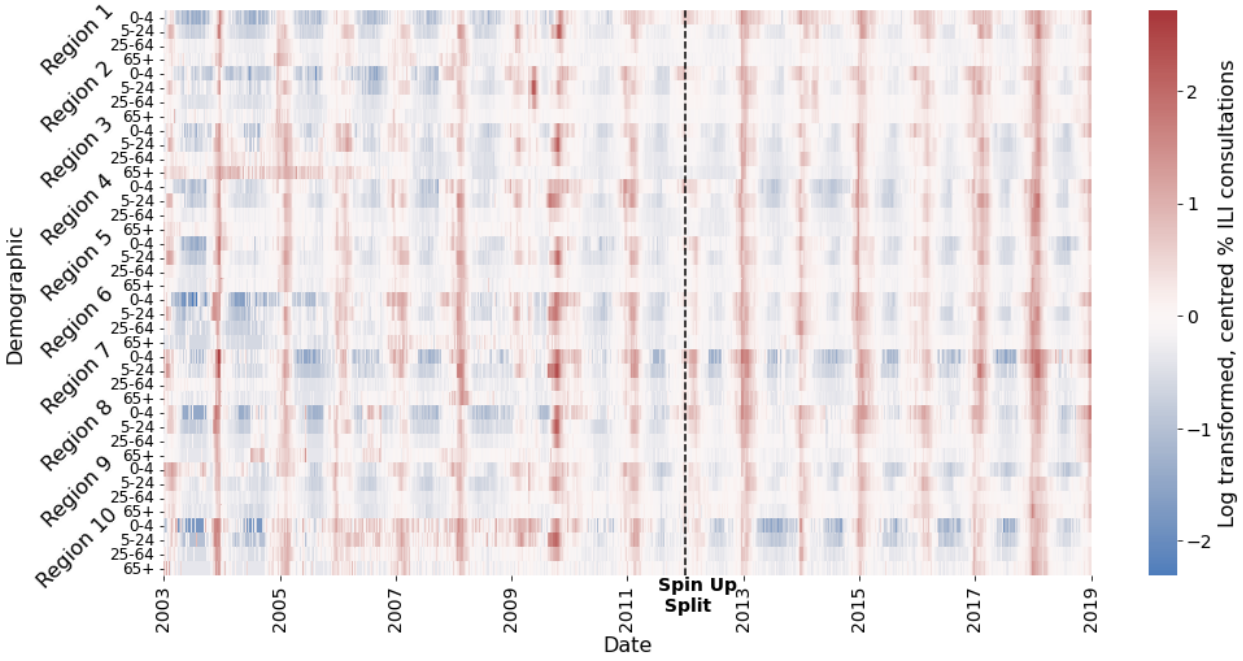
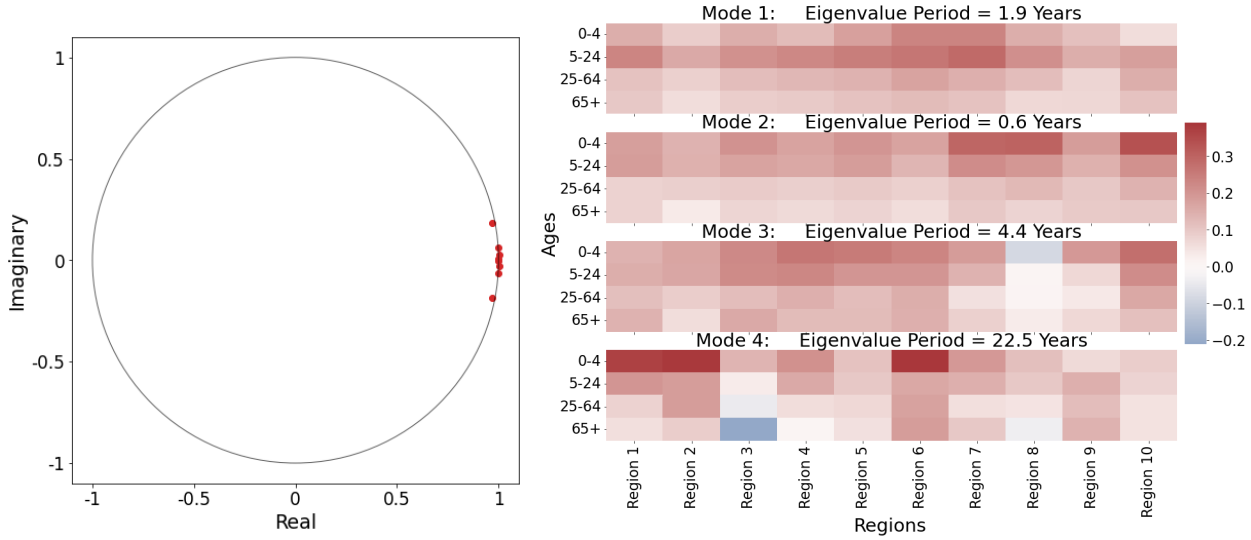


图9： 美国*4*个年龄组和*10*个*HHS*区域中，流感样疾病就诊占每周全科医生就诊总数的百分比，经对数变换并中心化处理。图*1* 中大小、时间和形状各异的峰值，在此处显示为宽度和强度各异的垂直红色区域，覆盖了大多数人口统计群体。

这很可能是因为 Hankel-DMD 如前所述，对测量噪声具有更强的鲁棒性。图10b 和10d所示的主要 DMD 和 Hankel-DMD 模态结构有两个明显规律。首先，各分层通常同步变化。这一点表现为：每个 DMD 模态中的绝大多数元素，以及 Hankel-DMD 模态中属于同一延迟嵌入周（以穿过模态的垂直切片表示）的元素，符号相同。这意味着所有年龄和地区的流感样疾病就诊百分比大致在同一时间上升或下降。其次，较年轻年龄组的方差更高。这一点可从各地区前两行元素的绝对值通常大于后两行看出。从图10b中可以直观地看出，DMD 模态具有这一特征。在图10d中，各地区的年龄组按年龄升序排列，因此 Hankel-DMD 模态中每个地区都呈现出更多颜色较深、宽度为 2 的水平线，随后重复出现颜色较浅、宽度也为 2 的水平线；这一现象在 10 个地区中依次出现，体现了该效应。这表明，年轻人群的流感样疾病就诊百分比峰值更尖、谷值更深，而 25 岁及以上人群的就诊率则更为稳定。

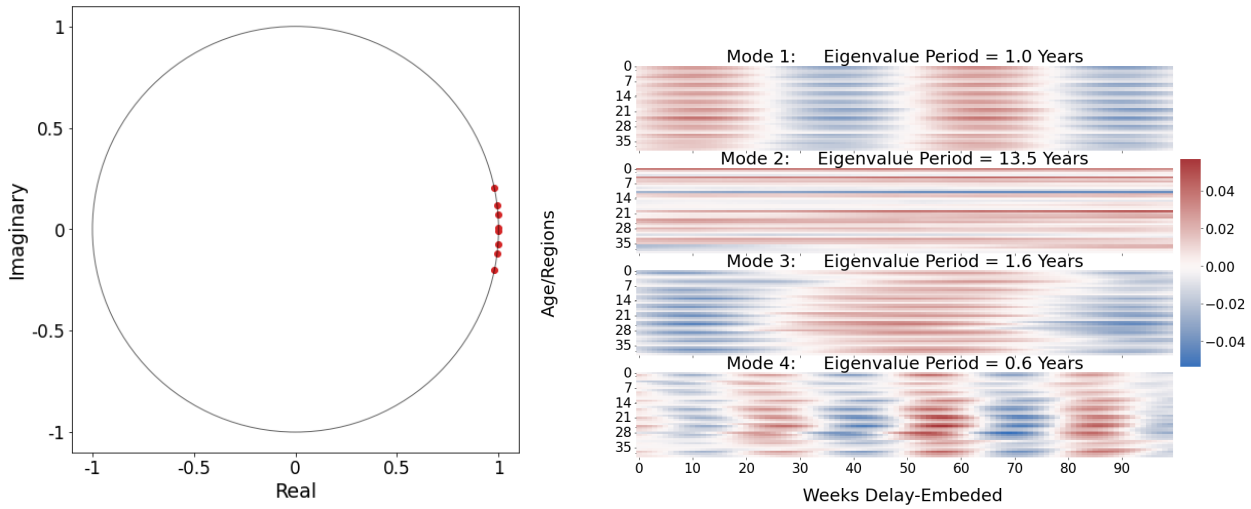
4.3 应用滤波器

随后，利用自启动动态模态分解和预热Hankel-DMD分别得到的空间模态和时间模态，对剩余数据应用DMDEnKF/Hankel-DMDEnKF的滤波步骤。即使对于最先进模型而言，要对ILINet数据生成始终优于简单历史基线预测的未来4周流感样疾病预测也很困难 [50]。因此，为了测试DMDEnKF/Hankel-DMDEnKF，我们在进行预测时采用4周的预测时域。在图11中，DMDEnKF和Hankel-DMDEnKF对流感样疾病就诊总数的预测



(a) *DMD Eigenvalue Spectrum.*

(b) *Dominant DMD Modes.*



(c) *Hankel-DMD Eigenvalue Spectrum.*

(d) *Dominant Hankel-DMD Modes.*

图10： 对图中数据应用动态模态分解（上）/*Hankel-DMD*（下）得到的特征值谱（左）及按主导性降序排列的 *DMD*模态（右）*9* （截至预热日期）。在这两种情况下，所有特征值都近似位于单位圆上，且对于大多数人口统计群体，主导模态的符号相同，其幅值随年龄而变化。*DMD*模态更易于解释，但 *Hankel-DMD* 能识别出周期为 *1* 年的模态，而 *DMD* 无法识别。

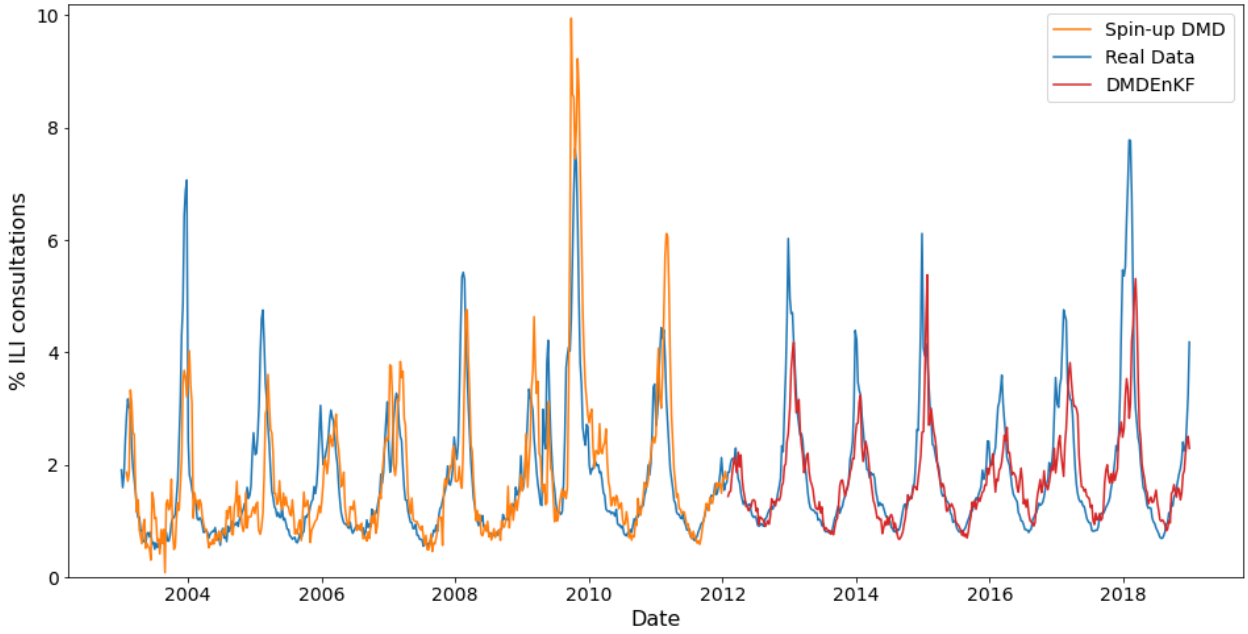
每周全科医生就诊总数的百分比可以完整显示。直到 2012 年，我们仅使用预热型动态模态分解模型生成提前 4 周的重构：通过取 4 周前的状态测量来估计每个状态，然后对其迭代应用模型 4 次。图 11a 所示的提前 4 周 DMD重构比 Hankel-DMD重构捕捉到了更多数据波动，但这些高频波动也可能表明测量值中的噪声产生了影响。图11b所示的 Hankel-DMD重构对噪声的敏感度低得多，但未能识别数据中的尖锐峰值，这表明它可能存在过度平滑。从 2012 年起开始进行滤波，并按照 DMDEnKF 算法中的说明，使用公式（ 2.33 ）生成预测。DMDEnKF 预测变得显著更加稳定，而 Hankel-DMDEnKF 预测在捕捉数据真实形态方面有所改善，但两者的预测都存在一定程度的滞后。

在数据的第二部分中，模型开始生成实际预测，因为 DM-DEnKF 只能获取截至预测目标日期前 4 周的数据。因此，我们在这一部分将模型的表现与历史基线进行比较。历史基线预测的生成方式与以下方法类似： [8],从数据中提取此前每一年同一周的流感样疾病就诊率（不包括 2009 年大流行年份），然后通过高斯核密度估计（KDE）得出当前周就诊率的概率分布 [55]。KDE 带宽根据 Silverman经验法则确定 [56], ；需要点估计时，则取该分布的中位数。比较结果见图 12 和表2。值得注意的是，尽管我们使用的是预测日期前 4 周的数据，但实际上这些数据往往会经过修订，因此当前版本的 ILINet数据未必能在实时情况下获取 [50]。

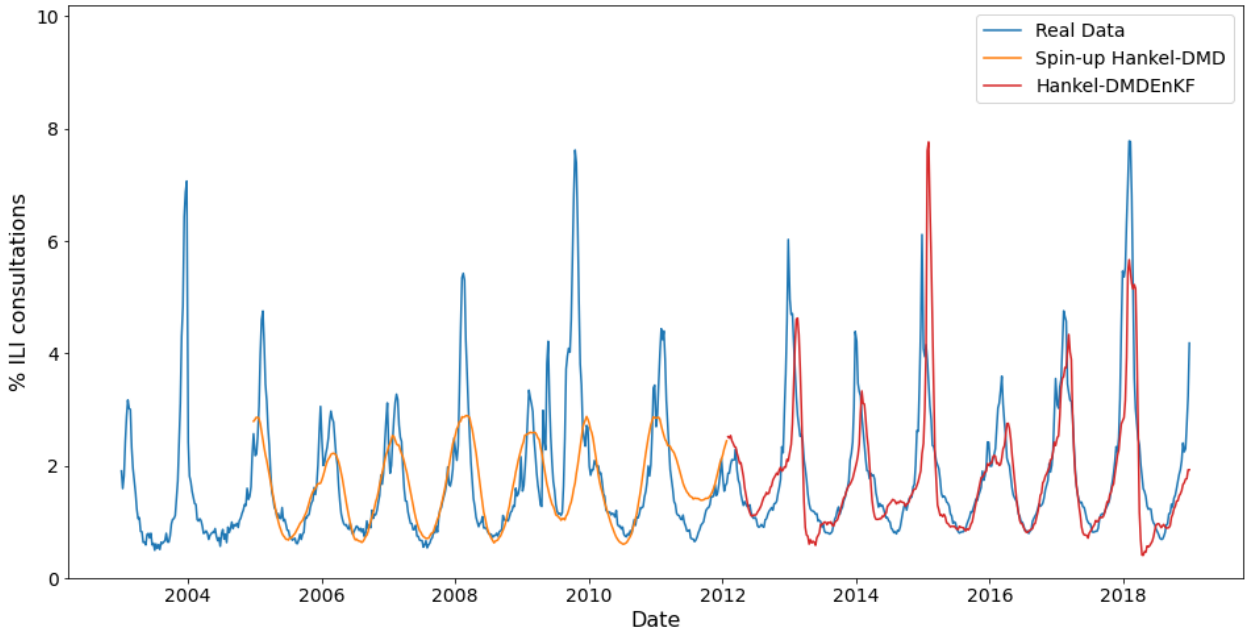
4.4 评估 DMDEnKF 的性能

图 12 以图形方式展示了 DMDEnKF、Hankel-DMDEnKF 和历史基线的提前 4 步预测。DMDEnKF 更成功地捕捉了每个流感季高峰的形态和高度，但往往会预测得偏晚；历史基线则始终低估高峰水平，不过对高峰出现时间的预测相当准确。Hankel-DMDEnKF 的预测比 DMDEnKF 更平滑，但无法捕捉高峰形态中的细节。我们还在图12中绘制了 DMDEnKF 和 Hankel-DMDEnKF 预测的 95%置信区间，这些区间是利用集合卡尔曼滤波框架中维护并传播的集合生成的。真实数据始终位于 DMDEnKF 的置信区间内，而 Hankel-DMDEnKF 的置信区间则并非如此。DMDEnKF 的置信区间显著宽于 Hankel-DMDEnKF 的置信区间，这是因为 Hankel-DMD 对噪声具有鲁棒性，也就是说，当集合通过模型传播时，大量概率质量会集中在状态空间的一小块区域内。这会导致 Hankel-DMDEnKF 低估系统中的不确定性，因此一些真实数据值会落在其 95%置信区间的边界之外。

为了对性能进行数值比较，我们采用了为流感季节预测协作挑战赛（FluSight）设计的指标。在该挑战赛中，多支团队提交了对即将到来的流感季全国及美国卫生与公众服务部区域级别每周 ILINet 咨询率的预测 [7], [8]。FluSight 挑战赛评估模型生成提前 1 至 4 周预测的能力



(a) *DMDEnKF 4-week ahead forecast.*



(b) *Hankel-DMDEnKF 4-week ahead forecast.*

图 11：使用 *DMDEnKF*（上）和 *Hankel-DMDEnKF*（下）对 *4* 周后的流感样疾病就诊量占每周全科医生就诊总量的百分比进行预测。*DMD*重构能够很好地捕捉数据的形状，但不够稳定；*Hankel-DMD*重构对噪声的敏感性较低，但过于平滑。*DMDEnKF* 和 *Hankel-DMDEnKF* 的预测有助于减轻各自重构中的这些问题，但两者的预测都存在一定程度的滞后。

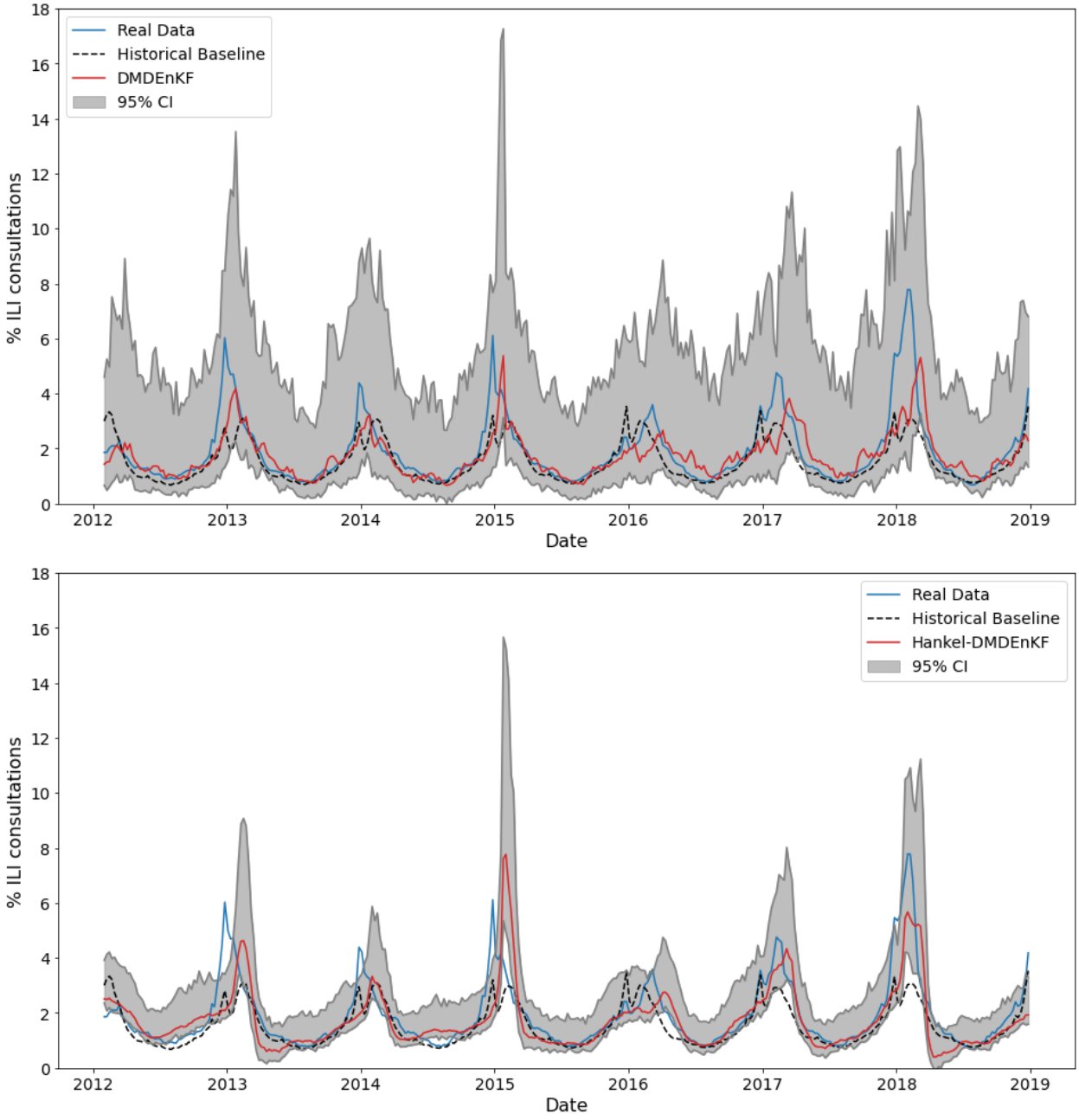


图12：流感样疾病就诊人数占每周全科医生就诊总数的百分比，使用 *DMDEnKF*（上图）和 *Hankel-DMDEnKF*（下图）对未来 *4* 周进行预测。图中还显示了每项预测的 *95%* 置信区间以及历史基线预测。*Hankel-DMDEnKF* 的预测比 *DMDEnKF* 的预测更加平滑，但两者都存在一定滞后。真实数据始终位于 *DMDEnKF* 的置信区间内，但并非始终位于 *Hankel-DMDEnKF* 的置信区间内，不过这可能是因为 *DMDEnKF* 的置信区间明显宽于 *Hankel-DMDEnKF* 的置信区间。

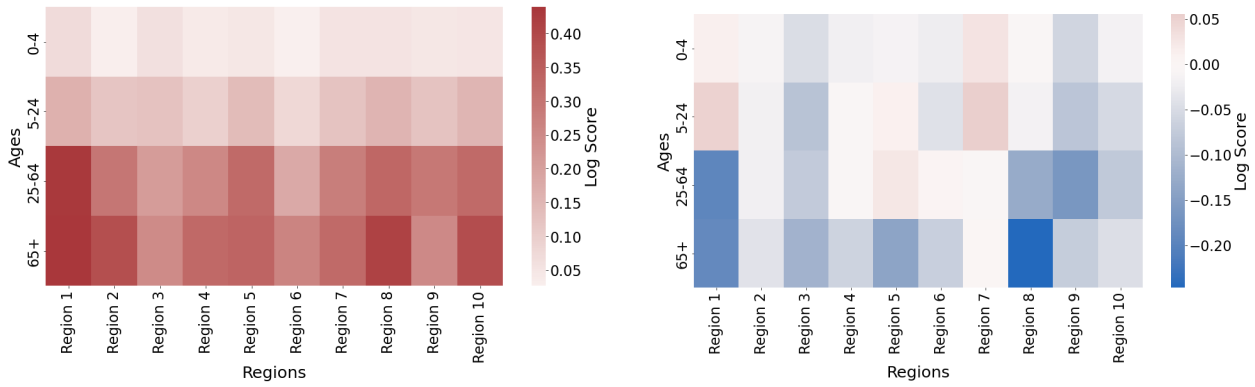
预测	对数评分	均方误差
历史基线	0.28	1.24
提前 1 周的 DMDEnKF	0.49	0.33
提前 2 周的 DMDEnKF	0.38	0.61
提前3周的DMDEnKF	0.32	0.87
提前4周的DMDEnKF	0.27	1.16
提前1周的Hankel-DMDEnKF	0.41	0.49
提前2周的Hankel-DMDEnKF	0.33	0.70
提前3周的Hankel-DMDEnKF	0.29	0.97
提前4周的Hankel-DMDEnKF	0.23	1.26

表2： *DMDEnKF* 和 *Hankel-DMDEnKF* 在不同预测时域下的对数评分和均方误差，以及历史基线预测。对于最多提前*4*周的预测时距，*DMDEnKF* 的对数评分和均方误差均高于历史基线，其预测技能与历史基线相近。在这些较短的预测时距内，*Hankel-DMDEnKF* 的两项指标始终不及 *DMDEnKF*。评分根据 *2012/13*至*2017/18*这*6*个流感季计算得出。

在一个流感季期间被称为短期目标，以及流感季开始前的其他较长期目标，即季节性目标。DMDEnKF 主要用于跟踪和短期预测，因此我们仅关注这些短期目标的预测。为此，我们采用了两种不同的指标：一种是对 FluSight 挑战赛略作调整的对数概率度量（对数评分），如 [50] 所用；另一种是均方误差，因其广泛应用于回归问题而被采用。对数评分表示模型预测准确的概率几何平均值，其中准确是指预测值与真实流感样疾病就诊率的差距在 $+/-0.5$ 以内。对数评分越高，预测效果越好。指标根据从第40周到次年第20周的数据计算，以优先评估流感季期间的预测；我们采用了2012/13至2017/18这6个完整流感季的数据。

全国范围内的历史基线预测以及 DMDEnKF/Hankel-DMDEnKF 的预测结果见表 2。正如预期，随着预测时间延长，两个 DMDEnKF 的准确性都会下降。在提前4周以内的预测时距内，DMDEnKF 的对数评分和均方误差均优于历史基线，并达到与之相近的预测技能。对于提前5周或更久的预测，DMDEnKF 在这两项指标上都无法胜过历史基线。在这些较短的预测时距内，Hankel-DMDEnKF 在两项指标上的表现始终不及 DMDEnKF。FluSight 挑战赛中排名前 3 的统计模型和机理模型，其提前4周预测的对数评分分别为 0.32 和 0.3，因此，DMDEnKF 的预测能力低于（但与之相当于）当前最先进的流感样疾病模型。随着预测时域延长至提前12周，DMDEnKF 的预测评分持续单调下降；而 Hankel-DMDEnKF 提前9至12周的对数评分不低于提前5至8周的评分。因此，短期预测更适合采用 DMDEnKF，而在更长时间尺度上进行预测时，Hankel-DMDEnKF 更具优势。

图13展示　　ws 4 周期前 DMDEnKF 预测的对数评分，以及这些评分与



(a) *DMDEnKF 4-week ahead forecast.*

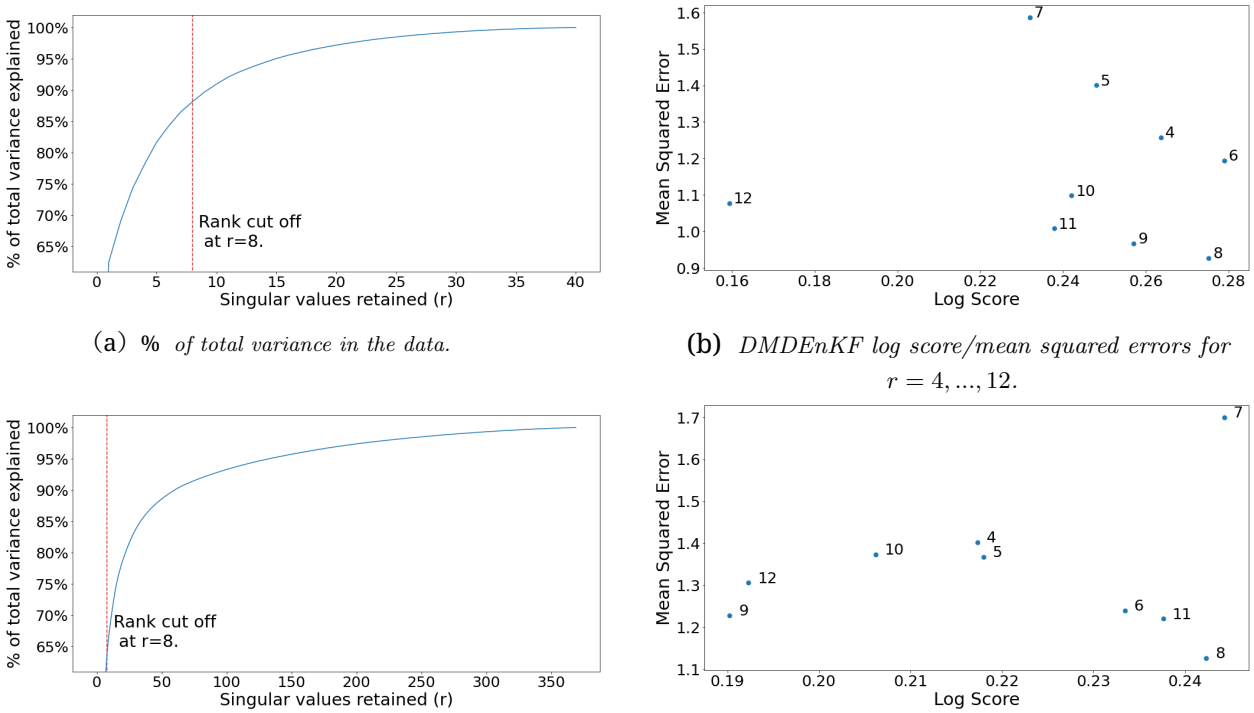
(b) *DMDEnKF 4-week ahead forecast - historical baseline.*

图 13： *DMDEnKF*提前*4*周预测在所有年龄和地区上的对数评分（左），以及减去历史基线预测的对数评分后的相同评分（右）。*Hankel-DMDEnKF*在所有年龄和地区上的评分与*DMDEnKF*相近，因此为避免重复，我们未列出其细分结果。上图中，随着年龄组逐渐向下，红色整体上颜色越深，表明*DMDEnKF*对年龄较大的年龄组预测得更准确。下图中红色/蓝色区域分布各异，表明*DMDEnKF*与历史基线的预测能力孰优孰劣，取决于所预测的年龄和地区；在大多数地区，历史基线的评分更高。

历史基线预测在年龄和地区层面取得的评分相比如何。Hankel-DMDEnKF 在所有年龄和地区的评分都与 DMDEnKF 相似，因此其细分情况也与 DMDEnKF 相当类似；为避免冗余，我们不再列出。观察 DMDEnKF 的对数评分，可以发现一个显著规律：年龄较大的年龄组评分更高，因此比年轻的人口统计群体更容易预测。去除历史基线的评分后，这一规律不再持续存在，表明这更可能是数据的一般特征，而非 DMDEnKF 建模方法的特定优势。不同地区的可预测性也存在显著差异。例如，第1区是 DMDEnKF 和历史基线最容易预测的地区，这与 [50]中的发现一致。不过，在该地区的四个年龄组中，DMDEnKF 仅有两个年龄组的预测优于历史基线。根据 [50] ，与历史基线预测相比，使用模型进行地区预测所带来的总体改进幅度最大，也出现在第1区，因此人们原本会预期 DMDEnKF 在全部四个年龄组中的表现都优于历史基线。由于对数评分很大程度上受预测不确定性大小的影响，DMDEnKF 使用的协方差矩阵对于该地区而言可能过大。因此，按地区设定 DMDEnKF 方差的大小，或许能带来更好的结果和更准确的预测。第6区是 DMDEnKF 预测表现最差的地区，历史基线预测略为准确。同样，这与 [50] 中的结果一致，其中第6区是各模型评分最低的地区。不过，在该研究中，与历史基线预测相比，使用模型在第6区带来的改进幅度位居第二。因此，对于该地区，DMDEnKF 采用更大的方差或许更为合适，以反映其更高的不确定性；这进一步支持了今后按地区调整方差的想法。

4.5 改变截断秩

在分析了采用 8 个 DMD 模态的 DMDEnKF 和 Hankel-DMDEnKF 的流感样疾病预测后，我们现在通过图r考察不同截断秩（r）对其性能的影响。14。从图 14a可以看出，在数据中主观判断“肘点”的过程可能会使观察者将合适的截断秩确定为低至 4，或高至 12。应用 Gavish和Donoho 的算法来确定最优截断阈值 [26]，得到的截断秩也为 12，因此我们将重点考察 r 在 4 到 12 区间内的取值。



（a）% *of total variance in the data.*

(b) *DMDEnKF log score/mean squared errors for* $r = 4, ..., 12.$

（c）% *of total variance in the delay-embedded data.*

(d) *Hankel-DMDEnKF log score/mean squared errors for* $r = 4.....12$。

图14：左侧显示了数据（上）和延迟嵌入数据（下）中总方差的%，其取决于所保留的奇异值数量（r）。数据中在$r = 8$ 附近出现了一个“肘部”，我们据此选择截断位置；然而，确定“肘部”的确切位置具有主观性，也可以认为其位于 $r = 4$ 至 $r = 12$之间。右侧显示了使用不同 r值，通过*DMDEnKF*（上）和*Hankel-DMDEnKF*（下）生成的提前*4*步预测的对数评分和均方误差。在这两种情况下，当 $r = 8$时，对数评分最大且均方误差最小。

图 14b 和 14d 展示了用于衡量 DMDEnKF 和 Hankel-DMDEnKF 预测能力的指标如何随 r变化。理想的预测应具有较高的对数评分，这反映出概率分布相对集中且准确；同时均方误差较低，表明点估计接近流感样疾病就诊的实际百分比。对于这两种方法，$r = 8$使对数评分达到最大、均方误差达到最小，这表明它是我们模型截断时的最优秩。对于 $r = 4$，我们测试的是最简单的模型，因此不确定性较低，对数评分相对较高；然而，该模型过于简单，无法恰当地刻画系统，因此均方误差较高。增大 r后，DMDEnKF 能够更自由地捕捉系统的复杂性，从而更准确地表示真实动力学，因此通常会降低均方误差。当特征值的数量——

增加过多时，模型便会开始拟合系统中的噪声，从而对未来预测产生负面影响，如 $r > 8$的均方误差上升所示。增大 r 也赋予 DMDEnKF 更多自由度，这意味着模型包含更多参数，每个参数都具有一定程度的不确定性。这会使整体预测概率分布变得更加分散；当模型精度的提升不再足以抵消这种影响（最高至 $r = 8$）时，预测的对数评分就会降低。

5 结论

总而言之，我们定义了两种新算法：DMDEnKF和Hankel-DMDEnKF，分别将动态模态分解和Hankel动态模态分解与集合卡尔曼滤波相结合，以便在新数据可用时更新动力系统的状态和时间模态估计值。将这两种算法应用于具有时变参数且测量噪声较低的简单合成系统时，它们在跟踪系统的时变参数和预测未来状态方面，与所测试的其他迭代式DMD变体表现相近。随着测量噪声增大，DMDEnKF在这两项指标上的表现均优于其他受测方法，且Hankel-DMDEnKF的预测比DMDEnKF的预测更稳定。两种DMDEnKF的表现均与其对应的DMD粒子滤波器相近（DMD粒子滤波器是对DMDEnKF算法的改动，即用粒子滤波器替代集合卡尔曼滤波器），但所需的集合成员数量显著更少。使用美国疾病控制与预防中心（CDC）的数据，对美国各年龄组和HHS区域的流感样疾病进行预测时，DMDEnKF在未来3周内的预测比历史基线预测更准确，在未来4周内的预测准确度则大致相当。对于这些短期目标，Hankel-DMDEnKF的预测准确度低于DMDEnKF，但总体而言，其预测更加稳定。此外，Hankel-DMDEnKF能够识别出一个周期为1年的模态，该模态在数据中十分明显，但DMDEnKF未能识别。两种DMDEnKF的预测能力均低于当前最先进的流感样疾病模型。

将扩展/核动态模态分解应用于DMD预热阶段，是对DMDEnKF的一种自然扩展。这使该算法能够更有效地应用于测量状态呈非线性变化的动力系统。与仅将观测值作为系统状态 $\mathbf{x}_k$不同，这些变体将一组定义在可观测量上的函数 $\mathbf{g}(\mathbf{x}_k$作为状态 n。集合卡尔曼滤波器非常适合这种组合，因为与其他卡尔曼滤波方法相比，其计算效率受状态维度增长的影响较小 [23]。如何选择函数集合 $\mathbf{g}(\mathbf{x}_k$作为非线性系统的嵌入，以便有效利用动态模态分解，是一个值得进一步研究的方向。已有许多方法提出了生成$\mathbf{g}(\mathbf{x}_k$的方法，例如采用深度学习 [41] 或储备池计算[28], ；这仍是一个很有前景的未来研究方向。

代码可用性

用于生成本文结果的代码可在 https://github.com/falconical/DMDEnKF 获取。

数据可用性声明

使用的所有数据 用于生成本文结果的代码将在收到合理请求后提供。 .

致谢

本研究得到 UKRI 的支持，其博士培训合作伙伴奖学金为 Stephen Falconer 的博士学位研究提供了资助。他还要感谢 Nadia Smith 和来自国家物理实验室的 Spencer Thomas 所进行的宝贵讨论。

参考文献

[1] H. Abdi, 《特征分解：特征值与特征向量》，《测量与统计百科全书》，（*2007*年）。

[2] H. Arbabi and I. Mezić, 《遍历理论、动态模态分解及 *Koopman* 算子谱性质的计算》，《*SIAM* 应用动力系统期刊》，*16*（*2017*），第*2096–2126*页。[3] F. Auger M. Hilairet J. Guerrero E. Monmasson T. Orlowska-Kowalskaand S. Katsura, 《卡尔曼滤波器的工业应用：综述》，《*IEEE* 工业电子学汇刊》，*60*（*2013*），第*5458*页。[4] R. Baker J.-M. Peña J. Jayamohan and A. Jérusalem, 《机理模型与机器学习：一场值得生物学界参与的较量？》，《生物学通讯》，*14*（*2018*），第*20170660*页。[5] D. Balcan V. Colizza B. Gonçalves H. Hu J. J. Ramasco and A. Vespignani, 《多尺度流动网络与传染病的空间传播》，《美国国家科学院院刊》，*106*（*2009*），第*21484–21489*页。

[6] J. C. A. Barata and M. S. Hussein, 《摩尔–彭罗斯伪逆：理论教程式综述》，《巴西物理学杂志》，*42*（*2011*），第*146–165*页。[7] M. Biggerstaff D. Alper M. Dredze S. Fox I. C.-H. Fung K. S. HickmannB. Lewis R. Rosenfeld J. Shaman M.-H. Tsou et al., 《美国疾病控制与预防中心“预测*2013–2014*年流感季挑战赛”的结果》，《*BMC*传染病》，*16*（*2016*），第*1–10*页。[8] M. Biggerstaff M. Johansson D. Alper L. C. Brooks P. Chakraborty D. C.Farrow S. Hyun S. Kandula C. McGowan N. Ramakrishnan R. RosenfeldJ. Shaman R. Tibshirani R. J. Tibshirani A. Vespignani W. Yang Q. Zhangand C. Reed, 《美国流感季预测合作项目第二年的结果》，《流行病》，*24*（*2018*），第*26–33*页。[9] D. A. Bistrian G. Dimitriu and I. M. Navon, 《使用动态模态分解方法处理流行病学数据》，《*AIP*会议论文集》，*2164*（*2019*），第*080002*页。

[10] R. N. Bracewell and R. N. Bracewell，《傅里叶变换及其应用》，第*31999*卷，麦格劳-希尔，纽约，*1986*年。[11] L. Brooks D. Farrow S. Hyun R. Tibshirani and R. Rosenfeld，《使用经验贝叶斯框架灵活建模流行病》，*PLOS*计算生物学，*11*（*2014*）。[12] B. W. Brunton L. A. Johnson J. G. Ojemann and J. N. Kutz，《使用动态模态分解从大规模神经记录中提取时空相干模式》，《神经科学方法杂志》，*258*（*2016*），第*1–15*页。[13] K. Chen J. Tu and C. Rowley，《动态模态分解的变体：边界条件、*Koopman* 算子与傅里叶分析》，《非线性科学杂志》，*22*（*2012*）。[14] J.-P. Chretien D. George J. Shaman R. Chitale and F. McKenzie，《人群流感预测：范围综述》，《公共科学图书馆·综合》，*9*（*2014*），第*e94130*页。[15] S. T. M. Dawson M. S. Hemati M. O. Williams and C. W. Rowley，《表征并校正传感器噪声对动态模态分解的影响》，《流体实验》，*57*（*2016*）。[16] A. de Cheveigné and J. Z. Simon，《基于时间平移主成分分析的去噪》，《神经科学方法杂志》，*165*（*2007*），第*297–305*页。[17] P. Del Moral，《非线性滤波：相互作用粒子求解》，《法国科学院学报·数学系列*I*》，*325*（*1997*），第*653–658*页。[18] M. D'Elia L. Mirabella T. Passerini M. Perego M. Piccinelli C. Vergara andA. Veneziani，《变分资料同化在计算血流动力学中的应用》，《建模、仿真与应用》，*5*（*2012*）。[19] N. Demo M. Tezzele and G. Rozza，《*PyDMD*：*Python* 动态模态分解》，《开放源代码软件杂志》，*3*（*2018*），第*530*页。[20] R. Douc and O. Cappé，《粒子滤波重采样方案的比较》，载于 *ISPA 2005*：《第四届国际图像与信号处理及分析研讨会论文集》，*2005*年，*IEEE*，*2005*年，第*64–69*页。[21] A. Doucet and A. Johansen，《粒子滤波与平滑教程：十五年后》，《非线性滤波手册》，*12*（*2009*）。[22] C. Eckart and G. Young，《用另一个低秩矩阵逼近一个矩阵》，《心理计量学》，*1*（*1936*），第*211–218*页。[23] G. Evensen，《集合卡尔曼滤波器：理论表述与实际实现》，《海洋动力学》，*53*（*2003*），第*343–367*页。[24] S. I. Flu，《美国通过疫苗接种避免的流感病例、就诊、住院及死亡估计》，*Prevent*，（*2007*），第*2006–2007*页。[25] C. for Disease Control and Prevention，《美国疾病控制与预防中心流感监测系统：目的与方法》。*https://www.cdc.gov/flu/weekly/overview.htm*，*2021*。访问日期：*2021-04-27*。

[26] M. Gavish and D. L. Donoho，奇异值的最优硬阈值为 *4/√3*，*2014*年。[27] N. Gordon D. Salmond and A. F. M. Smith，非线性/非高斯贝叶斯状态估计的新方法，*IEE Proceedings F*（雷达与信号处理），*140*（*1993*），第 *107–113(6)* 页。
[28] M. Gulina and A. Mauroy，使用储备池计算机逼近 *Koopman* 算子的两种方法，*Chaos: An Interdisciplinary Journal of Nonlinear Science*，*31*（*2021*），第 *023116* 页。
[29] C. W. R. Hao Zhang，在线 *DMD GitHub* 仓库。*https://github.com/haozhg/odmd*，*2020*年。访问日期：*2021*年*1*月*19*日。[30] M. Hemati E. Deem M. Williams C. W. Rowley and L. N. Cattafesta，利用抗噪动态模态分解变体改进分离控制，载于第*54*届 *AIAA* 航空航天科学会议，*2016*年，第 *1103* 页。[31] M. S. Hemati C. W. Rowley E. A. Deem and L. N. Cattafesta，对含噪数据集的应用型 *Koopman* 谱分析中的动态模态分解进行去偏，理论与计算流体力学，*31*（*2017*），第 *349–368* 页。[32] M. S. Hemati M. O. Williams and C. W. Rowley，适用于大型流式数据集的动态模态分解，*Physics of Fluids*，*26*（*2014*），第 *111701* 页。[33] R. Isermann and M. Münchhof，卡尔曼滤波的状态与参数估计，*Springer Berlin Heidelberg*，柏林，海德堡，*2011*年，第 *539–551* 页。[34] T. Jonathan H. R. Clarence W. L. Dirk M. B. Steven L. and K. J. Nathan，论动态模态分解：理论与应用，*Journal of Computational Dynamics*，*1*（*2014*），第 *391* 页。[35] M. R. Jovanović P. J. Schmid and J. W. Nichols，稀疏促进动态模态分解，*Physics of Fluids*，*26*（*2014*），第 *024103* 页。[36] R. E. Kalman，线性滤波与预测问题的新方法，*Journal of Basic Engineering*，*82*（*1960*年），第 *35–45* 页。[37] S. Kandula T. Yamana S. Pei W. Yang H. Morita and J. Shaman，流感样疾病预测中机理方法与统计方法的评估，*Journal of The Royal Society Interface*，*15*（*2018*），第 *20180174* 页。[38] K. J. H. Law A. M. Stuart and K. C. Zygalakis，数据同化：数学导论，*2015*年。[39] J. Lessler and D. Cummings，传染病机理模型及其对公共卫生的影响，*American Journal of Epidemiology*，*183*（*2016*），第 *kww021* 页。[40] Y. Luo K. Ogle C. Tucker S. Fei C. Gao S. LaDeau J. S. Clark and D. SSchimel，数据丰富时代的生态预测与数据同化，*Ecological Applications*，*21*（*2011*），第 *1429–1442* 页。

[41] B. Lusch J. N. Kutz and S. L. Brunton，用于非线性动力学通用线性嵌入的深度学习，《自然·通讯》，*9*（*2018*）。[42] J. Mandel，集合卡尔曼滤波器简明教程，*2009*年。
[43] J. Mann and J. N. Kutz，用于金融交易策略的动态模态分解，*2015*年。
[44] I. G. K. "Matthew O. Williams" "Clarence W. Rowley"，一种用于数据驱动的 *Koopman* 谱分析的基于核的方法，*Journal of Computational Dynamics*，*2*（*2015*），第 *247* 页。[45] T. Nonomura H. Shibata and R. Takaki，利用卡尔曼滤波器进行参数估计的动态模态分解，《*AIP*进展》，*8*（*2018*），第 *105106* 页。[46] T. Nonomura H. Shibata and R. Takaki，基于扩展卡尔曼滤波器的动态模态分解，用于同步系统辨识与去噪，《公共科学图书馆·综合》，*14*（*2019*），第 *e0209836* 页。[47] E. O. Nsoesie J. S. Brownstein N. Ramakrishnan and M. V. Marathe，关于流感暴发动力学预测研究的系统综述，《流感与其他呼吸道病毒》，*8*（*2014*），第 *309–316* 页。[48] D. Osthus K. S. Hickmann P. C. Caragea D. Higdon and S. Y. D. Valle，使用状态空间 *SIR* 模型预测季节性流感，《应用统计年鉴》，*11*（*2017*），第 *202–224* 页。[49] J. Proctor and P. Welkhoff，使用动态模态分解从传染病数据中发现动态模式，《国际健康》，*7*（*2015*），第 *139–45* 页。
[50] N. G. Reich L. C. Brooks S. J. Fox S. Kandula C. J. McGowan E. Moore D. Osthus E. L. Ray A. Tushar T. K. Yamana M. Biggerstaff M. A. Johansson R. Rosenfeld and J. Shaman，美国季节性流感预测的多年度、多模型协作评估，《美国国家科学院院刊》，*116*（*2019*），第 *3146–3154* 页。[51] R. H. Reichle，地球科学中的数据同化方法，《水资源进展》，*31*（*2008*），第 *1411–1418* 页。[52] R. H. Reichle J. P. Walker R. D. Koster and P. R. Houser，陆面资料同化中的扩展卡尔曼滤波与集合卡尔曼滤波，《水文气象学杂志》，*3*（*2002*年*12*月*1*日），第 *728–740* 页。[53] M. Ribeiro and I. Ribeiro，卡尔曼滤波器与扩展卡尔曼滤波器：概念、推导及性质，*2004*年*4*月。
[54] P. Schmid and J. Sesterhenn，数值与实验数据的动态模态分解，载于美国物理学会流体动力学分会会议摘要，第*61*卷 *APS* 会议摘要，*2008*年*11*月，第 *MR.007* 页。
[55] S. J. Sheather，密度估计，《统计科学》，*19*（*2004*），第 *588–597* 页。

[56] B. W. Silverman，统计学与数据分析中的密度估计，Routledge，2018年。

[57] C. Snyder, T. Bengtsson, P. Bickel, and J. L. Anderson，高维粒子滤波的障碍，《月度天气评论》，136（2008年），第4629–4640页。[58] P. Van den Driessche，传染病模型的再生数，《传染病建模》，2（2017年），第288–303页。
[59] E. A. Wan and R. Van Der Merwe，用于非线性估计的无迹卡尔曼滤波器，载于《IEEE 2000年信号处理、通信与控制自适应系统研讨会论文集》（目录编号：00EX373），2000年，第153–158页。[60] Z. Wang, P. Chakraborty, S. R. Mekaru, J. S. Brownstein, J. Ye, and N. Ra-makrishnan，用于流感样疾病病例数预测的动态泊松自回归，载于第21届ACM SIGKDD知识发现与数据挖掘国际会议论文集，KDD '15，纽约，纽约州，美国，2015年，计算机协会，第1285–1294页。[61] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley，数据驱动的*Koopman*算子近似：扩展动态模态分解，《非线性科学杂志》，25（2015年），第1307–1346页。[62] H. Zhang, C. W. Rowley, E. A. Deem, and L. N. Cattafesta，面向时变系统的在线动态模态分解，2017年。