Statistics and Application 统计学与应用, 2026, 15(5), 192-207
Published Online May 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.155119

Hans汉斯

隐马尔可夫模型参数的变分推断方法研究

张　镭

吉林财经大学统计与数据科学学院，吉林　长春

收稿日期：2026年4月25日；录用日期：2026年5月17日；发布日期：2026年5月27日

摘　要

隐马尔可夫模型因其功能强大且拥有坚实的数理基础而被广泛应用于众多领域。然而，其传统参数估计方法通常基于最大似然估计，难以有效刻画复杂分布结构及模型不确定性，尤其在存在异常值或厚尾特征的金融数据中表现受限。故本文引入变分推断框架，构造了一种基于多元*t*分布的变分隐马尔可夫模型，并给出了详细推导过程。然后，通过数值模拟实验测试了其参数估计与状态预测的正确性与有效性；最后，使用S&P500与CSI300两套真实股指数据进行了实证分析，结果表明VBtHMM模型参数估计更加高效稳健，在应对含异常值或非高斯噪声数据建模中，仍能保持较强的参数估计与状态预测能力。

关键词

多元*t*分布，变分推断，隐马尔可夫模型，股指预测

The Study of Variational Inference Methods for HMMs' Parameters

Lei Zhang

School of Statistics and Data Science, Jilin University of Finance and Economics, Changchun Jilin

Received: April 25, 2026; accepted: May 17, 2026; published: May 27, 2026

Abstract

Hidden Markov Models (HMMs) have been widely used in various fields due to their strong modeling capabilities and sound mathematical foundation. However, traditional parameter estimation of HMMs mainly relies on Maximum Likelihood Estimation (MLE), which fails to effectively capture complex distributional structures and quantify model uncertainty—this limitation is more obvious when modeling financial data with outliers or heavy-tailed distributions. To address this problem, this paper proposes a variational HMM based on multivariate *t*-distribution under the Variational Inference (VI) framework, with detailed mathematical derivations provided. Numerical simulations are

文章引用: 张镭. 隐马尔可夫模型参数的变分推断方法研究[J]. 统计学与应用, 2026, 15(5): 192-207.
DOI: 10.12677/sa.2026.155119

carried out to verify the validity and effectiveness of the proposed model in parameter estimation and latent state prediction. Finally, empirical analyses using real S&P 500 and CSI 300 data show that the VBtHMM achieves more efficient and robust parameter estimation, and maintains strong performance in parameter estimation and state prediction even for data with outliers or non-Gaussian noise.

Keywords

Multivariate *t*-Distribution, VI, HMM, Stock Index Forecasting

1. 引言

波动性和非平稳性是真实股指变化最显著的特性，但早期的股指预测方法主要是基于统计学原理的波动率预测模型，如 ARIMA、GARCH 及其衍生模型等[1] [2]，但这些模型隐含了一个基本假设，即股指时间序列的随机性效应(不可预测性和不确定性)不显著且约束性较强，无法反映波动率的非对称特性，故它们在处理非线性、非平稳数据时存在明显局限[3] [4]。隐马尔可夫模型(Hidden Markov Models, HMM)则具备多方面的优势。例如，它能够通过隐含状态和观测值之间的复杂关系，捕捉时间序列中的非线性和复杂依赖关系[5]。它通过引入多个隐藏状态，能够更细致地描述数据生成过程，适应不同的动态模式[6]。还可根据最新的观测数据动态调整隐藏状态，使模型具有较强的适应性，能够及时反映市场变化[7]。再者，隐含状态可与实际经济或金融市场环境中的不同市场阶段相对应，如繁荣期、萧条期等，从而提供更直观的经济解释[8]。ARIMA 和 GARCH 模型则主要侧重于统计特性解释，缺乏对现实经济或市场状态的直接解释，难以将模型结果与实际市场环境对应起来。因此，本文旨在为处理含异常值、非高斯噪声的现实金融时间序列数据提供一种更具鲁棒性与适应性的建模范式。

本文创新之处为引入新的观测假设并在数学转换处理后将其置于变分推断框架下构造了一种基于多元 *t* 分布的变分隐马尔可夫模型(Variational Bayesian *t*-distribution Hidden Markov Models, VBtHMMs)，且给出了核心步骤的推导过程。同时，在股指预测应用中还构建了一种新的预测方法，以期增强模型在具体应用中的预测能力和适应性。

2. 预备知识

2.1. 隐马尔可夫模型理论

独立随机试验模型最直接的推广就是马尔可夫(Markov)链模型[9]。马尔可夫链(Markov Chain, MC)是一种具有“马尔可夫性”，且存在于离散指数集和状态空间内的随机过程(如图 1)。

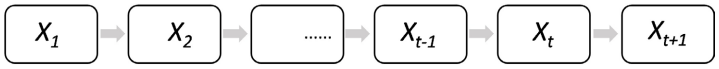

Figure 1. A schematic diagram of a discrete Markov chain
图 1. 一个离散马尔可夫链示意图

HMM 则是在此基础上拓展而来，它通过引入隐藏状态的概念，构建了一个双重随机过程。

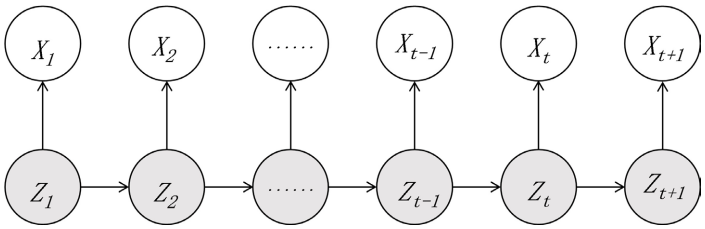

Figure 2. A schematic diagram of HMMs
图 2. 隐马尔可夫过程示意图

如图 2 所示，Z 序列表示无法直接观测到的隐藏状态序列，记为 $Z=\{z_1,z_2,\cdots,z_T\}$，其中隐藏状态集合记为 $S=\{s_1,s_2,\cdots\cdots,s_N\}$。$X$ 序列表示观测序列，记为 $X=\{x_1,x_2,\cdots,x_T\}$，其中可观测状态集合记为 $V=\{v_1,v_2,\cdots,v_M\}$。注意，这里的 N 和 M 未必相等，即可观测状态本身与隐藏状态之间并不存在一一对应关系，且当观测序列服从标准高斯分布时，即为“标准 HMM”模型。A 表示隐藏状态的转移概率矩阵，记为 $A=\{a_{ij}\},a_{ij}=P(z_t=S_j\mid z_{t-1}=S_i)$；$B$ 表示观测生成概率矩阵，记为 $B=\{b_{jm}\},b_{jm}=P(x_t=v_m\mid z_t=S_j)$。隐藏状态初始概率分布记为 $\Pi=\{\pi_i\},\pi_i=P\{z_1=S_i\}$。因此，记 HMM 为 $\Lambda=(N,M,A,B,\Pi)$。

两个基本假设：(1) **马尔可夫性假设**(Markov Assumption)，隐藏状态序列具有马尔可夫性质，即在给定当前状态的条件下，未来状态与过去状态相互独立。形式化地表示为：$P(z_{t+1}\mid z_1,z_2,\cdots,z_t)=P(z_{t+1}\mid z_t)$，其中 z_t 表示 t 时刻的隐藏状态。(2) **观测独立性假设**(Output Independence Assumption)，在给定当前隐藏状态的条件下，当前时刻的观测值与其他所有观测值和隐藏状态相互独立。形式化地表示为：$P(x_t\mid z_1,z_2,\cdots,z_T,x_1,x_2,\cdots,x_T)=P(x_t\mid z_t)$，其中 x_t 表示 t 时刻的观测值。

三个基本问题：(1) **评估问题**(Evaluation Problem)：对某给定 HMM，求产生某个观测序列的概率 $P(X\mid\lambda)$。该问题可通过前向-后向算法求解。(2) **解码问题**(Decoding Problem)：给定 HMMs 和某个已知观测序列，求产生此观测序列可能性最大的状态序列。该问题可通过 Viterbi 算法求解。(3) **学习问题**(Learning Problem)：给定一个观测序列 $X=\{X_1,X_2,\cdots,X_t\}$，求解参数 $\lambda=\{A,B,\pi\}$，使得观测序列产生的概率最大。该问题可通过 Baum-Welch 算法求解。

2.2. 变分推断方法的原理

变分推断(Variational Inference, VI)是一种用于近似复杂概率分布的数学方法。它通过将推断问题转化为优化问题，寻找一个简单分布族中的最佳分布来近似目标后验分布[10] [11]。在贝叶斯统计中，给定观测数据 $\boldsymbol{x}$，隐变量 $\boldsymbol{z}$ 的后验分布 $p(\boldsymbol{z}\mid\boldsymbol{x})$ 由贝叶斯定理给出：$p(\boldsymbol{z}\mid\boldsymbol{x})=\dfrac{p(\boldsymbol{x}\mid\boldsymbol{z})\,p(\boldsymbol{z})}{p(\boldsymbol{x})}$。

变分推断的核心思想是：首先，选择一个参数化的概率分布族 $q(\boldsymbol{z})$，称为变分分布(variational distribution)，通常比真实后验分布更简单。然后，通过最小化变分分布 $q(\boldsymbol{z})$ 与真实后验分布 $p(\boldsymbol{z}\mid\boldsymbol{x})$ 之间的 *Kullback-Leibler* (*KL*)散度来找到最佳近似。其中，*KL* 散度主要用于量化两个分布之间的差异程度：

$$KL\left(q(\boldsymbol{z})\,\|\,p(\boldsymbol{z}\mid\boldsymbol{x})\right)=\int q(\boldsymbol{z})\log\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}\mid\boldsymbol{x})}\mathrm{d}\boldsymbol{z}$$

目标是最小化 $KL\left(q(\boldsymbol{z})\,\|\,p(\boldsymbol{z}\mid\boldsymbol{x})\right)$，但直接最小化 *KL* 散度实际上并不可行，因为它依赖于未知的后验分布 $p(\boldsymbol{z}\mid\boldsymbol{x})$。故在变分推断中又引入证据下界(Evidence Lower Bound, ELBO)来间接解决这个问题。证据下界(ELBO)的推导过程为：首先从 *KL* 散度的定义出发：

$$KL\left(q(\boldsymbol{z})\,\|\,p(\boldsymbol{z}\mid\boldsymbol{x})\right)=\mathbb{E}_{q(\boldsymbol{z})}\left[\log\frac{q(\boldsymbol{z})}{p(\boldsymbol{z}\mid\boldsymbol{x})}\right]=\mathbb{E}_{q(\boldsymbol{z})}\left[\log q(\boldsymbol{z})\right]-\mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{z}\mid\boldsymbol{x})\right]$$

利用贝叶斯定理 $p(\boldsymbol{z}\mid\boldsymbol{x})=\dfrac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x})}$，将其代入上式后可得：

$$\mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{z}\mid\boldsymbol{x})\right]=\mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x},\boldsymbol{z})\right]-\log p(\boldsymbol{x})$$

因此，$KL\left(q(\boldsymbol{z})\|p(\boldsymbol{z}\mid\boldsymbol{x})\right)=\mathbb{E}_{q(\boldsymbol{z})}\left[\log q(\boldsymbol{z})\right]-\mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x},\boldsymbol{z})\right]+\log p(\boldsymbol{x})$。

重新排列各项位置后可得：

$$\log p(\boldsymbol{x})=\mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x},\boldsymbol{z})\right]-\mathbb{E}_{q(\boldsymbol{z})}\left[\log q(\boldsymbol{z})\right]+KL\left(q(\boldsymbol{z})\|p(\boldsymbol{z}\mid\boldsymbol{x})\right)$$

此时，定义证据下界(ELBO)为：

$$\mathrm{ELBO}(q)=\mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x},\boldsymbol{z})\right]-\mathbb{E}_{q(\boldsymbol{z})}\left[\log q(\boldsymbol{z})\right]$$

于是 $\log p(\boldsymbol{x})=\mathrm{ELBO}(q)+KL\left(q(\boldsymbol{z})\|p(\boldsymbol{z}\mid\boldsymbol{x})\right)$。由于 $\log p(\boldsymbol{x})$ 是一个常数(证据)，且 KL 散度是非负的，故 ELBO 是 $\log p(\boldsymbol{x})$ 的下界。这时，最大化 ELBO 就等价于最小化 KL 散度，记住这一点十分重要。

接着是变分分布的选择问题。为了简化优化难度，常选择平均场变分族(Mean-field variational family)。其中，假设隐变量 $\boldsymbol{z}$ 的各分量相互独立(这一点非常重要)，即 $q(\boldsymbol{z})=\prod_{i=1}^{n}q_i(z_i)$。其中 $q_i(z_i)$ 是每个隐变量分量的分布。最后是优化(最大化) ELBO 问题，通常使用坐标上升法。以平均场假设为例，优化过程涉及迭代更新每个 $q_i(z_i)$。从证据下界的表达式可看出，第一项是联合分布的期望，第二项是变分分布的熵。在平均场假设下，ELBO 可以分解为：

$$\mathrm{ELBO}(q)=\mathbb{E}_{q(\boldsymbol{z})}\left[\log p(\boldsymbol{x},\boldsymbol{z})\right]-\sum_{i}\mathbb{E}_{q_i(z_i)}\left[\log q_i(z_i)\right]$$

通过固定其他 $q_j\,(j\neq i)$，优化 $q_i(z_i)$ 时，最优解为：

$$q_i^{*}(z_i)\propto\exp\left(\mathbb{E}_{q_{-i}}\left[\log p(\boldsymbol{x},\boldsymbol{z})\right]\right)$$

其中 $\mathbb{E}_{q_{-i}}$ 表示对除 z_i 外所有隐变量的期望。这个过程反复迭代，同时监控 ELBO，直至其收敛。

3. VBtHMMs 理论

3.1. VBtHMMs 框架设计

设计一个变分隐马尔可夫模型框架主要从模型设定、辅助变量引入、共轭先验分布设定、变分后验数学推导和算法实现流程五个方面展开[12]。

首先，需明确设定模型的各个要素：观测变量(观测序列)：$X=\{x_1,\cdots,x_T\}$，其中 $x_t\in\mathbb{R}^3$。隐变量(隐状态序列)：$Z=\{z_1,\cdots,z_T\}$，其中 $z_t\in\{1,\cdots,K\}$，K 为隐状态数。状态转移概率矩阵 $A=\left[a_{ij}\right]$，其中 $a_{ij}=p\left(z_t=S_j\mid z_{t-1}=S_i\right)$。初始状态分布 $\pi=(\pi_1,\cdots,\pi_K)$，$\pi_k=p(z_1=k)$。此外，对于观测模型中状态 k，为控制尾部厚度设定观测 x_t 的分布为自由度 u 的多元 t 分布。然而，由于多元 t 分布的概率密度函数结构复杂，不便于代数处理，故将其转换为高斯–伽马的混合形式[13]：

$$T(x\mid\mu,\Sigma,u)=\int N\left(x\mid\mu,\lambda^{-1}\Sigma\right)\cdot Gam(\lambda\mid u/2,u/2)\mathrm{d}\lambda$$

因此，我们就相当于为每个观测 x_t 引入了辅助标量变量 $\lambda_t>0$，使得

$$p(x_t,\lambda_t\mid z_t=k)=N\left(x_t\mid\mu_k,\lambda_t^{-1}\Sigma_k\right)\cdot Gam(\lambda_t\mid u/2,u/2)$$

此时，完整数据就可以表示为 $\{X,Z,\lambda\}$，其中 $\lambda=\{\lambda_1,\cdots,\lambda_T\}$。相应的完整联合分布为

$$p(X,Z,\lambda,\theta)=p(\pi)\,p(A)\prod_{k=1}^{K}p(\mu_k,\Sigma_k)\cdot p(Z\mid\pi,A)\cdot\prod_{t=1}^{T}p(x_t,\lambda_t\mid z_t)$$

其中 $\theta=\left\{\pi,A,\{\mu_k,\Sigma_k\}_{k=1}^{K}\right\}$ 表示模型中的所有参数。然后，需要设定模型各要素的共轭先验分布：

初始状态 $\pi\sim Dir(\alpha_0)$、状态转移概率矩阵中每一行 $a_i\sim Dir(\beta_{i0})$。对每个状态 k，观测参数 $\Sigma_k\sim W^{-1}(W_0,u_0)$，逆 Wishart 分布，$\mu_k\mid\Sigma_k\sim N(\mu_0,\kappa_0^{-1}\Sigma_k)$。

为了变分推断现实可行，我们作出平均场假设：

$$q(Z,\lambda,\theta)=q(Z)q(\lambda)\prod_{k=1}^{K}q(\mu_k,\Sigma_k)q(\pi)\prod_{i=1}^{K}q(a_i)$$

然后，通过坐标上升变分推断(CAVI)最大化证据下界(ELBO)。作出这一假设后，需进行必要的误差分析，主要包括变分逼近的紧凑性和参数估计的偏差问题[14]。

接上，这里涉及几个关键推导步骤，具体包括以下四个方面：

(1) 隐状态后验 $q(z_t=k)$，同时定义责任变量 $\gamma_{tk}=q(z_t=k)$，在变分 E 步中，将 λ_t 用其期望 $\hat{\lambda}_t=E_q[\lambda_t]$ 替代，观测似然近似为 $x_t\mid z_t=k\approx N(\mu_k,\hat{\lambda}_t^{-1}\Sigma_k)$，取发射概率为

$$b_{tk}=\left|\hat{\Sigma}_k\right|^{-1/2}\hat{\lambda}_t^{d/2}\exp\left(-\frac{\hat{\lambda}_t}{2}(x_t-\hat{\mu}_k)^{\mathrm{T}}\hat{\Sigma}_k^{-1}(x_t-\hat{\mu}_k)\right)$$

然后再通过变分前向–后向算法计算：

$$\gamma_{tk}\propto\alpha_t(k)\beta_t(k)$$

$$\xi_{tij}\propto\alpha_{t-1}(i)a_{ij}b_{tj}\beta_t(j)$$

(2) 辅助变量后验 $q(\lambda_t)=Gam(a_t,b_t)$，其中

$$a_t=\frac{u+d}{2}$$

$$b_t=\frac{1}{2}\left(u+\sum_{k=1}^{K}\gamma_{tk}\cdot E_q\left[(x_t-\mu_k)^{\mathrm{T}}\Sigma_k^{-1}(x_t-\mu_k)\right]\right)$$

$$\hat{\lambda}_t=E_q[\lambda_t]=\frac{a_t}{b_t}$$

(3) 观测参数后验 $q(\mu_k,\Sigma_k)$，同时定义 $N_k=\sum_{t=1}^{T}\gamma_{tk}$，$\bar{x}_k=\frac{1}{N_k}\sum_{t=1}^{T}\gamma_{tk}x_t$。此时，后验分布仍是高斯–逆威沙特分布：$q(\mu_k,\Sigma_k)=N(\mu_k\mid\mu_{k,post},\kappa_{k,post}^{-1}\Sigma_k)\cdot W^{-1}(\Sigma_k\mid W_{k,post},u_{k,post})$。

(4) 初始状态与转移概率后验

$$q(\pi)=Dir(\alpha_k=\alpha_{0k}+\gamma_{1k})$$

$$q(a_i)=Dir\left(\beta_{ij}=\beta_{0ij}+\sum_{t=2}^{T}\xi_{tij}\right)$$

而期望为 $\hat{\pi}_k=\frac{\alpha_k}{\sum_j\alpha_j}$，$\hat{a}_{ij}=\frac{\beta_{ij}}{\sum_l\beta_{il}}$。

经过上述理论准备工作后，算法实现流程为：

(1) 初始化参数：$\mu_k,\Sigma_k,\lambda_t,\pi,A$ (其中需注意：$u_0\geq 2$，$\kappa_0>0$，保证 W_0 始终正定)；

(2) E 步：用当前参数计算 γ_{tk},ξ_{tij} (变分前向－后向算法)，更新 $q(\lambda_t)$，计算 $\hat{\lambda}_t=\frac{a_t}{b_t}$；

(3) M 步：更新 $q(\mu_k,\Sigma_k)$ 得到 $\mu_{k,post},W_{k,post},u_{k,post}$，更新 $q(\pi),q(A)$；

(4) 重复前面的 E 步和 M 步，直至算法收敛(即监控 ELBO 下界，当其变化小于人为设定的阈值 10^{-6} 时停止迭代)。

3.2. VBtHMMs 数学推导

在隐马尔可夫模型中，设观测序列为 $X=\{x_1,x_2,\cdots,x_T\}$，其中每个观测点 $x_t\in\mathbb{R}^3$ 是一个三维向量，T 是序列的总长度。又设隐状态序列为 $Z=\{z_1,z_2,\cdots,z_T\}$，其中每个隐状态 $z_t\in\{1,2,\cdots,K\}$，K 表示隐状态的总数。$z_t=k$ 表示在时刻 t，系统处于第 k 个隐状态。状态转移概率矩阵 $A=\left[a_{ij}\right]\in\mathbb{R}^{K\times K}$，其元素 $a_{ij}=p\left(z_t=s_j\mid z_{t-1}=s_i\right)$ 表示从状态 i 转移到状态 j 的概率，同时满足概率约束：$\sum_{j=1}^{K}a_{ij}=1$ 且 $a_{ij}\geq 0$。初始概率分布 $\pi=(\pi_1,\pi_2,\cdots,\pi_K)$，其中 $\pi_k=p(z_1=k)$ 且满足概率约束：$\sum_{k=1}^{K}\pi_k=1$ 且 $\pi_k\geq 0$。

在标准 HMM 中通常使用正态分布作为观测序列模型假设，但这个假设对异常值非常敏感且过于理想化。因此，对具有厚尾或非高斯噪声特性的金融观测数据，本文假设其服从多元 t 分布，其概率密度函数为：

$$T(x\mid\mu,\Sigma,u)=\frac{\Gamma\left(\frac{u+d}{2}\right)}{\Gamma\left(\frac{u}{2}\right)u^{d/2}\pi^{d/2}|\Sigma|^{1/2}}\left[1+\frac{1}{u}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right]^{-\frac{u+d}{2}}$$

这里 $d=3$ 表示数据维度。当 $u\to\infty$ 时，t 分布收敛于高斯分布；当 u 较小时，t 分布具有更厚的尾部，对异常值更鲁棒。对于状态 k，观测 x_t 的条件分布为：

$$p(x_t\mid z_t=k)=T(x_t\mid\mu_k,\Sigma_k,u)$$

其中：$\mu_k\in\mathbb{R}^3$：状态 k 的均值向量，$\Sigma_k\in\mathbb{R}^{3\times 3}$：状态 k 的协方差矩阵，$u=4$ 设定为观测序列的自由度以控制尾部厚度[15]。根据上节框架设计，多元 t 分布应转换为高斯-Gamma 分布的混合形式：

$$T(x\mid\mu,\Sigma,u)=\int_0^{\infty}N\left(x\mid\mu,\lambda^{-1}\Sigma\right)\cdot Gam\left(\lambda\mid\frac{u}{2},\frac{u}{2}\right)\mathrm{d}\lambda$$

证明：

$$\begin{aligned}
\text{等式右边}&=\int_0^{\infty}N\left(x\mid\mu,\lambda^{-1}\Sigma\right)\cdot Gam\left(\lambda\mid\frac{u}{2},\frac{u}{2}\right)\mathrm{d}\lambda\\
&=\int_0^{\infty}\frac{1}{(2\pi)^{d/2}\left|\lambda^{-1}\Sigma\right|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^{\top}\left(\lambda\Sigma^{-1}\right)(x-\mu)\right)\cdot\frac{\left(\frac{u}{2}\right)^{u/2}}{\Gamma\left(\frac{u}{2}\right)}\lambda^{\frac{u}{2}-1}\exp\left(-\frac{u\lambda}{2}\right)\mathrm{d}\lambda\\
&=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\cdot\frac{\left(\frac{u}{2}\right)^{u/2}}{\Gamma\left(\frac{u}{2}\right)}\int_0^{\infty}\lambda^{d/2}\exp\left(-\frac{\lambda}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right)\cdot\lambda^{\frac{u}{2}-1}\exp\left(-\frac{u\lambda}{2}\right)\mathrm{d}\lambda\\
&=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\cdot\frac{\left(\frac{u}{2}\right)^{u/2}}{\Gamma\left(\frac{u}{2}\right)}\int_0^{\infty}\lambda^{\frac{u+d}{2}-1}\exp\left(-\frac{\lambda}{2}\left[(x-\mu)^{\top}\Sigma^{-1}(x-\mu)+u\right]\right)\mathrm{d}\lambda
\end{aligned}$$

然后，令 $a=\frac{u+d}{2}$， $b=\frac{(x-\mu)^{\top}\Sigma^{-1}(x-\mu)+u}{2}$，则

$$\text{等式右边}=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\cdot\frac{\left(\frac{u}{2}\right)^{u/2}}{\Gamma\left(\frac{u}{2}\right)}\cdot\frac{\Gamma\left(\frac{u+d}{2}\right)}{b^{a}}$$

$$=\frac{\Gamma\left(\frac{u+d}{2}\right)}{\Gamma\left(\frac{u}{2}\right)(2\pi)^{d/2}|\Sigma|^{1/2}}\cdot\frac{\left(\frac{u}{2}\right)^{u/2}}{\left[\frac{(x-\mu)^{\top}\Sigma^{-1}(x-\mu)+u}{2}\right]^{\frac{u+d}{2}}}$$

$$=\frac{\Gamma\left(\frac{u+d}{2}\right)}{\Gamma\left(\frac{u}{2}\right)u^{d/2}\pi^{d/2}|\Sigma|^{1/2}}\left[1+\frac{1}{u}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right]^{-\frac{u+d}{2}}$$

等式左边 $=T(x\mid\mu,\Sigma,u)$，等式得证。这里为每个观测 x_t 引入了一个辅助标量变量 $\lambda_t>0$，使得：

$$p(x_t,\lambda_t\mid z_t=k)=N\left(x_t\mid\mu_k,\lambda_t^{-1}\Sigma_k\right)\cdot Gam\left(\lambda_t\mid\frac{u}{2},\frac{u}{2}\right)$$

该式也就等价于：

$$p(\lambda_t\mid z_t=k)=Gam\left(\lambda_t\mid\frac{u}{2},\frac{u}{2}\right)$$

$$p(x_t\mid\lambda_t,z_t=k)=N\left(x_t\mid\mu_k,\lambda_t^{-1}\Sigma_k\right)$$

易见，这个形式不仅使整个推断过程保持解析可解，而且在后续工作中可充分利用高斯分布的共轭先验这一性质。然后，可以定义完整数据为 $\{X,Z,\lambda\}$，其中，$\lambda=\{\lambda_1,\lambda_2,\cdots,\lambda_T\}$ 为辅助变量序列。在完整数据空间中，所有变量的联合分布此时都拥有了更简洁的表达形式，在贝叶斯框架下完整数据的联合分布可被分解为：

$$p(X,Z,\lambda,\theta)=p(\theta)\cdot p(Z\mid\theta)\cdot p(X,\lambda\mid Z,\theta)$$

其中，$\theta=\left\{\pi,A,\{\mu_k,\Sigma_k\}_{k=1}^{K}\right\}$ 表示模型中的所有参数。更具体地，可以写成：

$$p(X,Z,\lambda,\theta)=p(\pi)\,p(A)\prod_{k=1}^{K}p(\mu_k,\Sigma_k)\cdot p(Z\mid\pi,A)\cdot\prod_{t=1}^{T}p(x_t,\lambda_t\mid z_t,\mu_k,\Sigma_k)$$

这种分解方法严格遵循了概率论的链式法则，并利用了模型中的条件独立性假设。

根据贝叶斯统计理论，若后验分布与先验分布为同一类型分布，那么先验分布与后验分布就被称为共轭分布，且先验分布被称为似然函数的共轭先验分布。由于指数型分布都具有共轭先验分布，故本文将所需用到的先验分布均设定为指数型分布。

首先，设定初始状态分布 π 的先验分布为 *Dirichlet* 分布[16]：$\pi\sim Dir(\alpha_0)$，其概率密度函数为：

$$p(\pi)=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_{0k}\right)}{\prod_{k=1}^{K}\Gamma(\alpha_{0k})}\prod_{k=1}^{K}\pi_k^{\alpha_{0k}-1}$$

其中 $\alpha_0=(\alpha_{01},\alpha_{02},\cdots,\alpha_{0K})$ 为超参数，*Dirichlet* 分布为多项分布的共轭先验分布。转移概率矩阵 A 的先验

分布则设定为每一行独立的 *Dirichlet* 分布：$a_i \sim Dir\left(\beta_{i0}\right)$，其概率密度函数为：

$$p(A)=\prod_{i=1}^{K}\frac{\Gamma\left(\sum_{j=1}^{K}\beta_{0ij}\right)}{\prod_{j=1}^{K}\Gamma\left(\beta_{0ij}\right)}\prod_{j=1}^{K}a_{ij}^{\beta_{0ij}-1}$$

其中 $\beta_{i0}=\left(\beta_{0i1},\beta_{0i2},\cdots,\beta_{0iK}\right)$ 为超参数，这就保证了每一行仍是有效的分布函数。

接着将观测参数 $\{\mu_k,\Sigma_k\}$ 的先验分布设定为高斯－逆 *Wishart* 分布，即 $\Sigma_k \sim W^{-1}\left(W_0,u_0\right)$，$\mu_k \mid \Sigma_k \sim N\left(\mu_0,\kappa_0^{-1}\Sigma_k\right)$，其概率密度函数分别为：

$$p\left(\Sigma_k\right)=\frac{\left|W_0\right|^{u_0/2}}{2^{u_0 d/2}\Gamma_d\left(\frac{u_0}{2}\right)}\left|\Sigma_k\right|^{-\frac{u_0+d+1}{2}}\exp\left(-\frac{1}{2}tr\left(W_0\Sigma_k^{-1}\right)\right)$$

$$p\left(\mu_k \mid \Sigma_k\right)=\frac{1}{\left(2\pi\right)^{d/2}\left|\kappa_0^{-1}\Sigma_k\right|^{1/2}}\exp\left(-\frac{\kappa_0}{2}\left(\mu_k-\mu_0\right)^{\top}\Sigma_k^{-1}\left(\mu_k-\mu_0\right)\right)$$

这里 Γ_d 表示多元 *Gamma* 函数。在多元高斯分布的均值向量 μ_k 和协方差矩阵 Σ_k 均未知时，高斯－逆 *Wishart* 分布正是其共轭先验分布[17]。根据马尔可夫性质，隐状态序列的条件分布就可以写成：

$$p\left(Z \mid \pi,A\right)=p\left(z_1 \mid \pi\right)\prod_{t=2}^{T}p\left(z_t \mid z_{t-1},A\right)=\pi_{z_1}\prod_{t=2}^{T}a_{z_{t-1},z_t}$$

它表示初始状态是由 π 决定的，后续每个状态仅依赖于其前一个状态。最后，对观测变量(包括辅助变量)的条件分布，这里假设对于每个时间点 t，给定隐状态 $z_t=k$，观测变量和辅助变量的联合分布为：

$$p\left(x_t,\lambda_t \mid z_t=k\right)=N\left(x_t \mid \mu_k,\lambda_t^{-1}\Sigma_k\right)\cdot Gam\left(\lambda_t \mid \frac{u}{2},\frac{u}{2}\right)$$

将其展开为显示表达形式：

$$p\left(x_t,\lambda_t \mid z_t=k\right)=\frac{\lambda_t^{d/2}}{\left(2\pi\right)^{d/2}\left|\Sigma_k\right|^{1/2}}\exp\left(-\frac{\lambda_t}{2}\left(x_t-\mu_k\right)^{\top}\Sigma_k^{-1}\left(x_t-\mu_k\right)\right)$$
$$\cdot\frac{\left(\frac{u}{2}\right)^{u/2}}{\Gamma\left(\frac{u}{2}\right)}\lambda_t^{\frac{u}{2}-1}\exp\left(-\frac{u\lambda_t}{2}\right)$$

因此，对于整个序列可以得到：

$$\prod_{t=1}^{T}p\left(x_t,\lambda_t \mid z_t\right)=\prod_{t=1}^{T}\sum_{k=1}^{K}\mathbb{I}\left(z_t=k\right)\cdot p\left(x_t,\lambda_t \mid z_t=k\right)$$

其中，$\mathbb{I}(\cdot)$ 为示性函数，即设 $Z_t \subseteq \mathbb{K}$，记其为：

$$\mathbb{I}_{Z_t}\left(k\right)=\begin{cases}1, & \text{若 } z_t \in k,\\ 0, & \text{若 } z_t \notin k.\end{cases}$$

且满足约束 $\sum_{t=1}^{T}z_{tk}=1$。

由此得到完整数据的联合分布为：

$$p\left(X,Z,\lambda,\theta\right)=p\left(\pi\right)p\left(A\right)\prod_{k=1}^{K}p\left(\mu_k,\Sigma_k\right)\cdot\left[\pi_{z_1}\prod_{t=2}^{T}a_{z_{t-1},z_t}\right]\cdot\prod_{t=1}^{T}\left[\mathcal{N}\left(x_t \mid \mu_{z_t},\lambda_t^{-1}\Sigma_{z_t}\right)\cdot Gam\left(\lambda_t \mid \frac{u}{2},\frac{u}{2}\right)\right]$$

将其展开为显式形式：

$$p(X,Z,\lambda,\theta)=\underbrace{\left[\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_{0k}\right)}{\prod_{k=1}^{K}\Gamma\left(\alpha_{0k}\right)}\prod_{k=1}^{K}\pi_k^{\alpha_{0k}-1}\right]}_{\text{初始分布 }\pi\text{ 的 Dirichlet 先验}}\cdot\underbrace{\left[\prod_{i=1}^{K}\frac{\Gamma\left(\sum_{j=1}^{K}\beta_{0ij}\right)}{\prod_{j=1}^{K}\Gamma\left(\beta_{0ij}\right)}\prod_{j=1}^{K}a_{ij}^{\beta_{0ij}-1}\right]}_{\text{转移矩阵 }A\text{ 的 Dirichlet 先验}}$$

$$\cdot\prod_{k=1}^{K}\left[\frac{\left|W_0\right|^{u_0/2}}{2^{u_0 d/2}\Gamma_d\left(u_0/2\right)}\left|\Sigma_k\right|^{-\frac{u_0+d+1}{2}}\exp\left(-\frac{1}{2}tr\left(W_0\Sigma_k^{-1}\right)\right)\right.$$

$$\left.\cdot\frac{1}{\left(2\pi\right)^{d/2}\left|\kappa_0^{-1}\Sigma_k\right|^{1/2}}\exp\left(-\frac{\kappa_0}{2}\left(\mu_k-\mu_0\right)^{\top}\Sigma_k^{-1}\left(\mu_k-\mu_0\right)\right)\right]\cdot\underbrace{\left[\pi_{z_1}\prod_{t=2}^{T}a_{z_{t-1},z_t}\right]}_{\text{隐状态序列的 Markov 链}}$$

$$\cdot\underbrace{\prod_{t=1}^{T}\left[\frac{\lambda_t^{d/2}}{\left(2\pi\right)^{d/2}\left|\Sigma_{z_t}\right|^{1/2}}\exp\left(-\frac{\lambda_t}{2}\left(x_t-\mu_{z_t}\right)^{\top}\Sigma_{z_t}^{-1}\left(x_t-\mu_{z_t}\right)\right)\right.}_{\text{基于尺度混合的观测似然}}\cdot\underbrace{\left.\frac{\left(u/2\right)^{u/2}}{\Gamma\left(u/2\right)}\lambda_t^{\frac{u}{2}-1}\exp\left(-\frac{u\lambda_t}{2}\right)\right]}_{\text{尺度参数 }\lambda_t\text{ 的 Gamma 先验}}$$

易见，该式囊括了所有变量和参数的所有信息。至此，下面可直接给出模型参数的变分后验分布：

$$q\left(\pi\right)=Dir\left(\alpha_k=\alpha_{0k}+\gamma_{1k}\right)$$

$$q\left(a_i\right)=Dir\left(\beta_{ij}=\beta_{0ij}+\sum_{t=2}^{T}\xi_{tij}\right)$$

$$q\left(\mu_k,\Sigma_k\right)=\mathcal{N}\left(\mu_k\mid\mu_k^{post},\left(\kappa_k^{post}\right)^{-1}\Sigma_k\right)\cdot\mathcal{W}^{-1}\left(\Sigma_k\mid W_k^{post},u_k^{post}\right)$$

$$q\left(\lambda_t\right)=Gam\left(a_t=\frac{u+d}{2},b_t=\frac{1}{2}\left[u+\sum_{k=1}^{K}\gamma_{tk}\left(x_t-\mu_k^{post}\right)^{\top}\hat{\Sigma}_k^{-1}\left(x_t-\mu_k^{post}\right)\right]\right)$$

对于每个状态 k，超参数计算为：

$$N_k=\sum_{t=1}^{T}\gamma_{tk}$$

$$\overline{x}_k=\frac{1}{N_k}\sum_{t=1}^{T}\gamma_{tk}x_t$$

$$\kappa_k^{post}=\kappa_0+N_k$$

$$\mu_k^{post}=\frac{\kappa_0\mu_0+N_k\overline{x}_k}{\kappa_0+N_k}$$

$$u_k^{post}=u_0+N_k$$

$$W_k^{post}=W_0+\sum_{t=1}^{T}\gamma_{tk}\hat{\lambda}_t\left(x_t-\mu_k^{post}\right)\left(x_t-\mu_k^{post}\right)^{\top}+\frac{\kappa_0 N_k}{\kappa_0+N_k}\left(\overline{x}_k-\mu_0\right)\left(\overline{x}_k-\mu_0\right)^{\top}$$

注意：为确保后验期望 $\mathbb{E}\left[\Sigma_k\right]$ 存在，先验自由度 u_0 的设定值应该至少大于 2，从而保证后验自由度 $u_k^{post}=u_0+N_k>d+1=4$ 成立。

由此，隐藏状态序列的变分后验分布为：

$$q_z(Z) \propto \hat{\pi}_{z_1} \prod_{t=2}^{T} \hat{a}_{z_{t-1},z_t} \prod_{t=1}^{T} \left(\left\| \hat{\Sigma}_{z_t} \right\|^{-1/2} \hat{\lambda}_t^{d/2} \exp\left(-\frac{\hat{\lambda}_t}{2} \left(x_t - \hat{\mu}_{z_t} \right)^{\top} \hat{\Sigma}_{z_t}^{-1} \left(x_t - \hat{\mu}_{z_t} \right) \right) \right)$$

其中，模型参数估计表达式分别为：

$$\hat{\pi}_k = \frac{\alpha_{0k} + \gamma_{1k}}{\sum_{j=1}^{K} \left(\alpha_{0j} + \gamma_{1j} \right)}$$

$$\hat{a}_{ij} = \frac{\beta_{0ij} + \sum_{t=2}^{T} \xi_{tij}}{\sum_{l=1}^{K} \left(\beta_{0il} + \sum_{t=2}^{T} \xi_{til} \right)}$$

$$b_{tk} = \left\| \hat{\Sigma}_k \right\|^{-1/2} \hat{\lambda}_t^{d/2} \exp\left(-\frac{\hat{\lambda}_t}{2} \left(x_t - \hat{\mu}_k \right)^{\top} \hat{\Sigma}_k^{-1} \left(x_t - \hat{\mu}_k \right) \right)$$

$$\hat{\mu}_k = \mu_k^{post} = \frac{\kappa_0 \mu_0 + N_k \bar{x}_k}{\kappa_0 + N_k}$$

$\hat{\Sigma}_k = \dfrac{W_k^{post}}{u_k^{post} - d - 1}$，其中 $u_k^{post} = u_0 + N_k > d + 1$

$\hat{\lambda}_t = \dfrac{a_t}{b_t}$，其中 $a_t = \dfrac{u+d}{2}$，$b_t = \dfrac{1}{2}\left[u + \sum_{k=1}^{K} \gamma_{tk} \left(x_t - \hat{\mu}_k \right)^{\top} \hat{\Sigma}_k^{-1} \left(x_t - \hat{\mu}_k \right) \right]$

4. 数值模拟与实证分析

4.1. 数值模拟

数值模拟的首要目标就是要验证改进模型的正确性，即实现过程是否严格遵循数学推导。其次就是验证模型对厚尾数据的建模能力。我们设计三种场景(见表 1)进行模拟测试，并以此形成纵向对照。数据生成模拟器先验参数设置见表 2，模拟数据由一个离散时间 HMM 生成，其发射分布为多元 t 分布，并在干净数据生成后叠加了均匀分布异常值污染机制。具体模型及参数为 $\pi \sim Dirichlet(\alpha_1, \cdots, \alpha_K)$，其中 $\alpha_k = 2$，等价于对 K 个独立的 $Gamma(2,1)$ 变量进行归一化。$A_{i,\cdot} \propto (g_{i1}, \cdots, g_{ii}+5, \cdots, g_{iK})$，其中 $g_{ij} \sim Gamma(2,1)$。发射分布参数 $\{\mu_k, \Sigma_k\}_{k=1}^{K}$ 各维度独立。$\mu_{k,j} \sim \mathcal{N}(8k, 1.5^2), j = 1, \cdots, d$，状态间基底均值按 8，16，24 线性拉开。对 Σ_k 构造随机矩阵 $C_k \in \mathbb{R}^{d \times d}$，元素 $C_{k,ij} \sim \mathcal{N}(0, 0.5^2)$，尺度矩阵为 $\Sigma_k = C_k C_k^{\top} + I_d$，确保严格正定。给定 $Z_t = k$，观测向量 $X_t \mid Z_t = k \sim t_d(\nu = u = 4, \delta = \mu_k, \Sigma = \Sigma_k)$。注：实际协方差矩阵为 $\dfrac{\nu}{\nu - 2}\Sigma_k$ (当 $\nu > 2$ 时存在)。污染数量 $N_{out} = \lfloor T \cdot \rho \rfloor$；污染索引 $\mathcal{I}_{out} \sim U(\{1, \cdots, T\}, N_{out})$ (无放回抽样)；污染分布，令 $R = 5 \cdot \max_{t,j} \left| X_{t,j}^{clean} \right|$，对 $\forall t \in \mathcal{I}_{out}, j \in \{1, \cdots, d\}$，$X_{t,j}^{contaminated} \sim \mathcal{U}(-R, R)$。

同时，为了测试的完整性，还需加入标准高斯 HMM 与之进行同样的全过程测试，并以此形成横向对照，这里以温和场景为例进行说明。

Table 1. Simulated scenarios
表 1. 模拟场景

场景类型	观测分布	目的	特征说明
基准场景	标准 t 分布	验证模型的正确性	不含异常值
温和场景	95% t 分布	测试温和异常值下稳健性	含 5%异常值
极端场景	90% t 分布	测试极端异常值下拟合表现	含 10%异常值

Table 2. Prior parameter setting
表 2. 先验参数设置

参数	设定值	说明
α_0	1.0 (可调)	初始状态 *Dirichlet* 先验
β_0	1.0 (可调)	转移矩阵 *Dirichlet* 先验
μ_0	数据均值	观测均值先验
κ_0	0.1 (可调)	均值先验强度
W_0	样本协方差	协方差尺度矩阵
u_0	3.0 (可调)	逆 *Wishart* 先验自由度(确保 $u_0 > 2$)

状态对齐(识别)性能看混淆矩阵中每个主对角元素的值，其大小反映了其相对应状态的对齐程度。也就是说，主对角元素值越大，则它所代表的状态对齐率越高。混淆矩阵从上往下每一行分别表示真实状态 1、2、3，从左往右每一列分别表示预测状态 1、2、3。

$$\begin{bmatrix} 230 & 1 & 4 \\ 1 & 168 & 4 \\ 2 & 89 & 1 \end{bmatrix}$$

图 3 是 VBtHMM 模型数值模拟测试结果降维后的二维平面图。从图中可以清晰看到，红色代表隐状态 1、绿色代表隐状态 2、蓝色代表隐状态 3。左侧图表示真实隐状态的数量和位置，右侧图表示测试后模型状态预测的数量与位置。若右侧图与左侧图的相同位置的圆点颜色相同，则表示该位置所示隐状态预测正确。

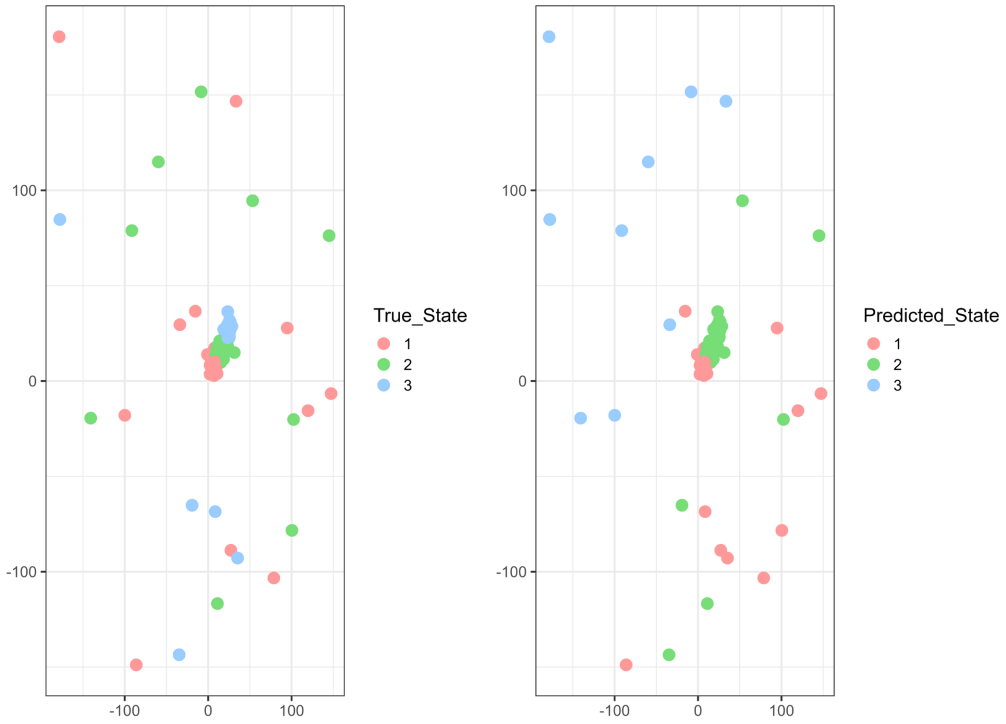

Figure 3. The distribution of state prediction in mild scenarios
图 3. 温和场景下状态预测分布情况

从该场景混淆矩阵的元素来看，除了状态 3 预测较差，主对角线其他元素均显著大于非主对角线位置的元素。原因在于样本不平衡，采集的样本中隐藏状态 3 的样本量相对太少，仅占 18.4%，这就导致模型很难学习到其特征。在模拟测试中平衡样本状态当然并不难做到(如可调大“小状态”的先验分布参数 κ_0)，但这里没有必要，因为现实金融数据天然就是不平衡的。同理，三种测试场景下的状态预测性能表现情况如表 3 所示。

从表 3 可见，ELBO 收敛稳定，VBtHMM 模型的正确性得到验证。在异常值占比升高过程中，其性能下降较为平缓，而与之对比的标准高斯 HMM 模型则出现性能骤降表现，造成这种差异的原因正是高斯分布对异常值敏感，且其协方差估计会显著变大，而 VBtHMM 模型中引入的辅助变量 λ 会自发降低异常值的权重，加之变分推断框架提供了正则化效果以防止模型过拟合，这均符合理论预期。

Table 3. Performance of state prediction
表 3. 状态预测性能表现

类别	ELBO	收敛迭代次数	预期区间	VBtHMM	HMM
0%异常值(基准)	−3722.2154	31	(0.85,0.95)	0.8732	0.8320
5%异常值(温和)	−4088.2257	76	(0.75,0.85)	0.7980	0.4259
10%异常值(极端)	−5511.5437	85	(0.60,0.70)	0.6350	0.3446

4.2. 实证分析

4.2.1. 数据说明与预测方法

本文选取标普 500 指数和沪深 300 指数两套大盘股指(从 2010 年到 2023 年所有交易日的)有效数据作为样本，再分别将其分为两部分，其中 2010 年至 2021 年的(3300*4 个)股指开盘价、最高价、最低价和收盘价数据作为训练集，2022 年 1 月至 2023 年 8 月的 400 个(交易日)收盘价数据作为测试集。

为了更直观地展现 VBtHMM 模型的预测性能，这里将构建新的股指预测方法，区别于标准 HMM 的经典股指预测方法[18]，这里不再赘述。由于真实的金融市场数据中包含的隐藏状态数未知且同一时段不同市场的隐藏状态数也未必相同，所以在实证分析中需构建某种预测方法来间接判断模型预测性能。在实务中，人们常关注股指每日的开盘价 $Open_t$、最高价 $High_t$、最低价 Low_t 和收盘价 $Close_t$，这里对它们做一个数学转换：

$$R_t^{close}=\frac{Close_t-Open_t}{Open_t}\times 100\%$$

$$R_t^{high}=\frac{High_t-Open_t}{Open_t}\times 100\%$$

$$R_t^{low}=\frac{Low_t-Open_t}{Open_t}\times 100\%$$

本文对状态转移过程进行加权处理[19]，即取 $a_{s_t s_{t+1}}$ 为权重，重新计算每个观测值的概率分布。令

$$f\left(V_{t+1}\right)=\sum_{s_{t+1}=j}^{S} a_{s_t s_{t+1}} * b_{s_{t+1}}\left(V_{t+1}\right)。$$

其中，$f\left(V_{t+1}\right)$表示下一时刻隐藏状态下观测值的概率密度函数；$b_{s_{t+1}}\left(V_{t+1}\right)$表示$t+1$时刻隐藏状态对应生成的观测值的概率密度函数。然后，再计算观测值所处区间的概率 $P\left(a<V_{t+1}\le b\right)$：

$$P\left(a<V_{t+1}\le b\right)=\int_a^b f\left(V_{t+1}\right)\mathrm{d}x$$

在计算 $P\left(a<V_{t+1}\le b\right)$之前，需对收盘价相对变化率区间进行等宽离散化处理。由于每日股指涨跌幅

限制在±10%，所以将该区间分为 100 个等宽小区间，再通过上式重新计算每个值所在小区间的概率值，再取概率值最大那个小区间的中位数为预测值 $\hat{V}_{t+1}$，同时假设 $t+1$ 时刻的开盘价变化率等于 t 时刻的收盘价变化率($R_{t+1}^{Open}=R_t^{Close}$)，最后再通过 $Close_{t+1}=Close_t*\left(1+\hat{V}_{t+1}\right)$ 转换得出 $t+1$ 时刻的收盘价。最后，本文采用均方根误差 RMSE (Root Mean Square Error)和 MAPE (Mean Absolute Percentage Error，即平均绝对百分比误差)来表示 VBtHMM 模型在真实股指数据上的预测性能，计算方法为

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\%$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}$$

其中， $\hat{y}_i$ 表示预测值， y_i 表示真实值，n 则表示测试数据集的样本容量。

4.2.2. 股指预测结果

隐藏状态数 K 的确定在实证中十分关键，需利用变分推断方法的损失函数－自由能与 ELBO 存在相反数的关系[20] [21]。故若能找到最小自由能就能相应确定最大的 ELBO，从而使近似后验概率与真实值之间的 KL 距离达到最小。K 与 ELBO 在 CSI 300 (左)和 S&P 500 (右)上的自由能随 K 值变化情况见表 4。

Table 4. Evolution of free energy with K
表 4. 自由能随 K 变化过程

隐状态数 K	迭代次数	Free Energy (−ELBO)	隐状态数 K	迭代次数	Free Energy (−ELBO)
3	327	1094.27	3	343	856.3245
4	84	408.74	4	79	386.5634
5	61	73.67	5	50	−81.9863
6	148	−490.53	6	151	−387.4325
7	71	−587.22	7	217	−699.9827
8	57	−660.70	8	75	−802.4524
9	133	−710.95	9	139	−840.3401
10	**290**	**−722.57**	10	393	−859.9741
11	42	−688.08	**11**	**377**	**−899.8502**
12	298	−659.49	12	248	−810.8762
13	154	−631.20	13	193	−788.0436
14	226	−593.14	14	263	−760.9873
15	47	−577.32	15	57	−703.9847

利用 VBtHMM 模型分别对前述两个数据集进行实证的结果如图 4 和图 5 所示。图中纵坐标表示股价指数，横坐标表示 2022 年 1 月~2023 年 8 月的 400 个交易日期。易见，红绿两条线走势存在少许不一致，但绝大部分几乎一致吻合，说明模型预测性能较好。同理，还应加入标准 HMM 进行实证分析以形成横向对照，其结果如图 6 和图 7 所示。易见，标准 HMM 在两套股指测试集上的误差明显较大，其预测性能不及 VBtHMM 模型，实证结果也证明缺乏实际理论支撑的经典方法对金融股指数据的预测能力有限。这也恰好说明本文构建的 VBtHMM 模型在建模含异常值、非高斯噪声的金融股指数据时更具鲁棒性与适应性。最后，为了定量判断二者预测性能，下面分别计算出各自的两个误差指标，如表 5 所示。

Figure 4. CSI 300 Index trend forecast comparison
图 4. 沪深 300 指数走势预测对比图

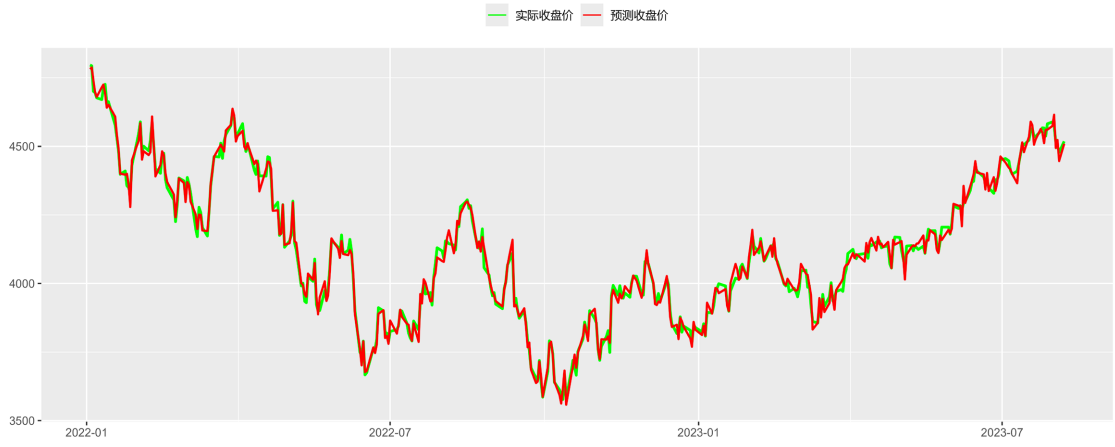

Figure 5. S&P 500 Index trend forecast comparison
图 5. 标普 500 指数走势预测对比图

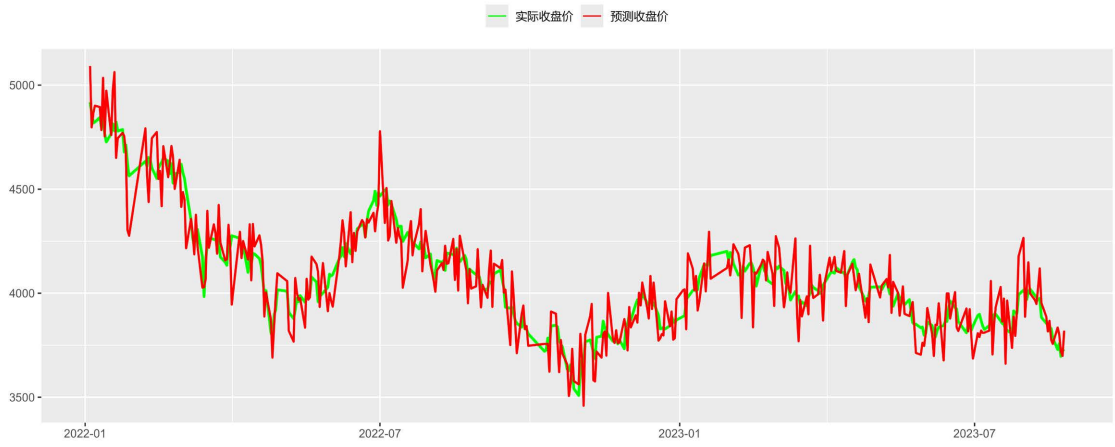

Figure 6. CSI 300 Index trend forecast comparison by HMM
图 6. HMM 对沪深 300 指数走势预测对比图

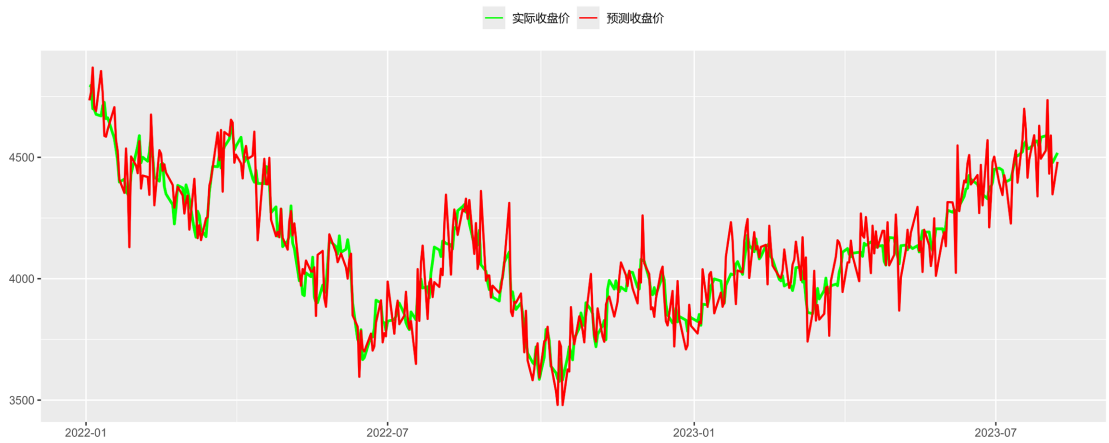

Figure 7. S&P 500 Index trend forecast comparison by HMM
图 7. HMM 对标普 500 指数走势预测对比图

Table 5. Comparison of empirical forecasting performance
表 5. 实证预测性能对比

市场类别	误差指标	VBtHMM	HMM
CSI 300 指数	MAPE	0.00787	0.01960
CSI 300 指数	RMSE	41.0391	102.2067
S&P 500 指数	MAPE	0.00432	0.01780
S&P 500 指数	RMSE	32.3546	92.2158

从表 5 中易见，纵向上二者各自分别在 CSI 300 和 S&P 500 两套股指数据上的预测性能存在相似性，即在 CSI 300 指数上的预测精度明显均低于在 S&P 500 指数上的预测精度，这种差异恰恰反映了中美两国股市确实属于不同的市场环境。横向上，VBtHMM 预测性能明显强于标准 HMM，说明前者在含异常值、非高斯噪声的金融时间序列数据上拥有显著的建模优势。

5. 结论

标准 HMM 的参数估计方法难以有效刻画复杂分布结构及模型不确定性，它在常呈现出“尖峰厚尾”特征的真实金融时间序列数据中表现受限，其观测假设(服从正态分布)过于理想化。本文提出的 VBtHMM 模型使参数估计更加稳健，在应对异常值或非高斯噪声数据建模中，仍能保持较好的参数估计与状态预测能力。

当然，本文也存在不足之处。如若将自由度 u 也视为随机变量，那么就会破坏高斯－伽马混合形式的共轭结构，使得推断不再具有可解析性。而本文将其固定为常数的做法又会使其变为了不可学习参数。理论上，模型必定存在无法自适应数据尾部厚度的情况，导致出现欠拟合或过拟合。在更新 $q(\lambda_t)$ 时，为简化计算与实现而采用了点估计近似处理。这一粗暴简化方式会导致模型在小样本下参数不确定性被完全忽略，最终导致推断偏差大等，这些问题必定是未来研究中需要克服的难点。

参考文献

[1] 贺本岚. 股票价格预测的最优选择模型[J]. 统计与决策, 2008(6): 135-137.
[2] 魏宇. 沪深 300 股指期货的波动率预测模型研究[J]. 管理科学学报, 2010, 13(2): 66-76.

[3] Pagan, A. (1996) The Econometrics of Financial Markets. *Journal of Empirical Finance*, **3**, 15-102. https://doi.org/10.1016/0927-5398(95)00020-8

[4] Bollerslev, T., Chou, R.Y. and Kroner, K.F. (1992) ARCH Modeling in Finance: A Review of the Theory and Empirical Evidence. *Journal of Econometrics*, **52**, 5-59. https://doi.org/10.1016/0304-4076(92)90064-x

[5] Hamilton, J.D. (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, **57**, 357-384. https://doi.org/10.2307/1912559

[6] Rabiner, L.R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77**, 257-286. https://doi.org/10.1109/5.18626

[7] Nystrup, P., Madsen, H. and Lindström, E. (2017) Long Memory of Financial Time Series and Hidden Markov Models with Time-Varying Parameters. *Journal of Forecasting*, **36**, 989-1002. https://doi.org/10.1002/for.2447

[8] Oelschläger, L. and Adam, T. (2023) Detecting Bearish and Bullish Markets in Financial Time Series Using Hierarchical Hidden Markov Models. *Statistical Modelling*, **23**, 107-126. https://doi.org/10.1177/1471082x211034048

[9] 方兆本, 缪柏其. 随机过程[M]. 第 3 版. 北京: 科学出版社, 2011: 26.

[10] Bishop, C.M. and Nasrabadi, N.M. (2006) Pattern Recognition and Machine Learning. Springer, 462-474.

[11] Murphy, K.P. (2012) Machine Learning: A Probabilistic Perspective. MIT Press, 731-746.

[12] McGrory, C.A. and Titterington, D.M. (2009) Variational Bayesian Analysis for Hidden Markov Models. *Australian & New Zealand Journal of Statistics*, **51**, 227-244. https://doi.org/10.1111/j.1467-842x.2009.00543.x

[13] Andrews, D.F. and Mallows, C.L. (1974) Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **36**, 99-102. https://doi.org/10.1111/j.2517-6161.1974.tb00989.x

[14] Turner, R.E. and Sahani, M. (2011) Two Problems with Variational Expectation Maximisation for Time Series Models. In: Barber, D., Cemgil, A.T. and Chiappa, S., Eds., *Bayesian Time Series Models*, Cambridge University Press, 104-124. https://doi.org/10.1017/cbo9780511984679.006

[15] Villa, C. and Rubio, F.J. (2018) Objective Priors for the Number of Degrees of Freedom of a Multivariatetdistribution and Thet-Copula. *Computational Statistics & Data Analysis*, **124**, 197-219. https://doi.org/10.1016/j.csda.2018.03.010

[16] Blei, D.M. and Jordan, M.I. (2006) Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, **1**, 121-143. https://doi.org/10.1214/06-ba104

[17] Murphy, K.P. (2007) Conjugate Bayesian Analysis of the Gaussian Distribution (Technical Report). University of British Columbia. https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf

[18] Hassan, M.R. and Nath, B. (2005) Stock Market Forecasting Using Hidden Markov Model: A New Approach. 5*th International Conference on Intelligent Systems Design and Applications* (ISDA'05), Warsaw, 8-10 September 2005, 192-196. https://doi.org/10.1109/isda.2005.85

[19] 张旭东, 黄宇方, 等. 基于离散型隐马尔可夫模型的股票价格预测[J]. 浙江工业大学学报, 2020, 48(2): 148-153.

[20] Dayan, P., Hinton, G. E., Neal, R.M. and Zemel, R.S. (1995) The Helmholtz machine. *Neural Computation*, **7**, 889-904.

[21] Neal, R.M. and Hinton, G.E. (1998) A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants. In: Jordan, M.I., Ed., *Learning in Graphical Models*, Kluwer Academic Publishers, 355-368.