

基于高斯隐马尔可夫模型的金融市场波动机制检测

学号: 253108110079 姓名: 刘孟繁

2025 年 12 月 9 日

摘要

本研究旨在利用高斯隐马尔可夫模型 (Gaussian HMM) 对标准普尔 500 指数 (S&P 500) 的历史数据进行建模，以识别金融市场中不可观测的隐状态。通过 Baum-Welch 算法进行参数估计，并利用 Viterbi 算法进行最大后验概率 (MAP) 推断。实验结果表明，该模型能有效捕捉 2000 年互联网泡沫、2008 年金融危机及 2020 年新冠疫情等重大市场波动事件。

1 数据集描述

本研究选取美国股市最具代表性的 **标准普尔 500 指数 (S&P 500)** 日交易数据作为观测对象。

- **时间跨度:** 2000 年 1 月 1 日至 2023 年 12 月 31 日，覆盖了完整的几个经济周期。
- **数据预处理:** 由于股票收盘价是非平稳序列，直接建模效果较差。我们计算 **对数收益率 (Log Returns)** 作为模型的观测序列 X 。设 P_t 为第 t 日的收盘价，则观测值 x_t 定义为：

$$x_t = \ln(P_t) - \ln(P_{t-1}) \quad (1)$$

- **数据链接:** 数据通过 Yahoo Finance API 获取，公开页面地址：Yahoo Finance - S&P 500。

2 模型建立 (Gaussian HMM)

我们构建一个具有 K 个隐状态的连续型隐马尔可夫模型。其中，**隐状态** (Z_t) 满足 $Z_t \in \{1, \dots, K\}$ ，表示 t 时刻的市场机制 (Regime)。在本实验中设 $K = 3$ ，分别对应：低波动、中波动、高波动。隐状态序列遵循一阶马尔可夫性。

观测值 (X_t) 满足 $X_t \in \mathbb{R}$ ，为当日的对数收益率。最后，**发射概率**则是假设在给定隐状态 $Z_t = k$ 的条件下，观测值服从单变量高斯分布：

$$P(x_t | Z_t = k) = \mathcal{N}(x_t; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_t - \mu_k)^2}{2\sigma_k^2}\right) \quad (2)$$

训练模型参数 λ : 包括初始状态概率 π 、状态转移矩阵 A 以及高斯分布参数 $\theta = \{\mu_k, \sigma_k^2\}_{k=1}^K$ 。

3 算法实现 (Baum-Welch Algorithm)

由于隐状态不可见，无法直接使用 MLE。我们采用 EM 算法的 HMM 特例——**Baum-Welch 算法**进行参数训练。伪代码如 Algorithm 1 所示。

Algorithm 1 Baum-Welch 算法 (Gaussian HMM 参数估计)

Require: 观测序列 $X = \{x_1, \dots, x_T\}$, 状态数 K

Ensure: 模型参数 $\lambda = (\pi, A, \mu, \Sigma)$

- 1: 初始化 λ (随机或 K-Means)
 - 2: **repeat**
 - 3: // E-Step (期望步):
 - 4: 利用 Forward-Backward 算法计算 $\alpha_t(i)$ 和 $\beta_t(i)$
 - 5: 计算后验概率 $\gamma_t(i) = P(Z_t = i | X, \lambda)$
 - 6: 计算双状态后验 $\xi_t(i, j) = P(Z_t = i, Z_{t+1} = j | X, \lambda)$
 - 7: // M-Step (最大化步):
 - 8: 更新初始概率: $\pi_i = \gamma_1(i)$
 - 9: 更新转移矩阵: $A_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$
 - 10: 更新高斯参数 (加权 MLE):
$$\mu_k = \frac{\sum_{t=1}^T \gamma_t(k) x_t}{\sum_{t=1}^T \gamma_t(k)}, \quad \sigma_k^2 = \frac{\sum_{t=1}^T \gamma_t(k) (x_t - \mu_k)^2}{\sum_{t=1}^T \gamma_t(k)}$$
 - 11: **until** 对数似然函数收敛
 - 12: **until** 对数似然函数收敛
 - 13: **return** λ
-

4 推断表现与讨论 (Inference & Discussion)

模型训练完成后，我们主要关注 **Decoding** (解码) 问题，即寻找最可能的隐状态序列 $Z_{1:T}^* = \arg \max P(Z_{1:T} | X_{1:T})$ 。利用 **Viterbi** 算法对测试集进行推断，结果如图 1 所示。

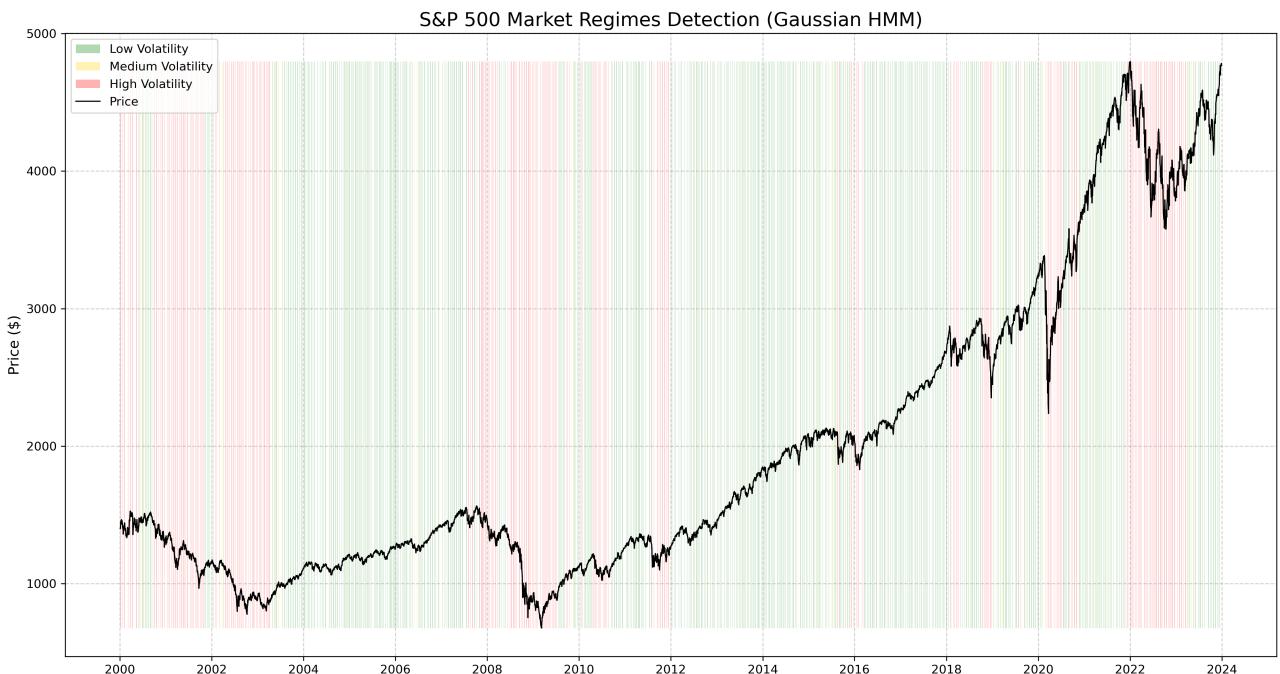


图 1: S&P 500 市场机制检测结果。背景颜色表示推断出的隐状态: 绿色 = 低波动 (牛市), 黄色 = 中波动, 红色 = 高波动 (熊市/危机)。

4.1 结果分析

1. **市场机制的有效划分：**模型成功将市场划分为不同的“体制”。绿色区域（低波动）对应长期平稳上涨区间；红色区域（高波动）对应剧烈下跌或震荡区间。
2. **重大事件捕捉：**
 - **2008 年金融危机：**图中 2008 年区间出现了大面积连续的红色色块，准确反映了市场恐慌。
 - **2020 年新冠疫情：**2020 年初出现了一条极窄但极显著的红色竖条，反映了当时市场的瞬时熔断。
3. **波动率聚类：**推断出的状态切换呈现“块状”分布，说明状态具有较高的持续性（转移矩阵对角线元素值较大），符合金融时间序列的波动率聚类特征。

4.2 结论

Gaussian HMM 有效地从收益率数据中提取了隐含的市场风险状态，MAP 推断结果与历史金融危机高度吻合。