Hidden Markov models for multi-scale time series: an application to stock market data

Timo Adam[1,2] and Lennart Oelschlger[2]

[1] University of St Andrews, St Andrews, UK
[2] Bielefeld University, Bielefeld, Germany

E-mail for correspondence: ta59@st-andrews.ac.uk

Abstract: Over the last decades, hidden Markov models have emerged as a versatile class of statistical models for time series where the observed variables are driven by latent states. While conventional hidden Markov models are restricted to modeling single-scale data, economic variables are often observed at different temporal resolutions: an economy's gross domestic product, for instance, is typically observed on a yearly, quarterly, or monthly basis, whereas stock prices are available daily or at even finer temporal resolutions. In this paper, we propose hierarchical hidden Markov models to incorporate such multi-scale data into a joint model, where we illustrate the suggested approach using 16 years of monthly trade volumes and daily log-returns of the Goldman Sachs stock.

Keywords: Hidden Markov models; Multi-scale data; Stock markets; Time series modeling; Temporal resolution.

1 Introduction

Hidden Markov models (HMMs) constitute a versatile class of statistical models for time series where the observed variables are driven by latent states (Zucchini *et al.*, 2016). While the observations can be multivariate, basic HMMs have the limitation that all variables need to be observed at the same temporal resolution. Specifically in economic applications, however, corresponding variables are often observed at different time scales, ranging from yearly data such as economic indices to high-frequency stock market data. Incorporating multiple such variables, with differing sampling rates, into a joint model may help to draw a more comprehensive picture of stock market dynamics, in particular by explicitly distinguishing short-term and long-term variation in volatility. In this paper, we propose hierarchical

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

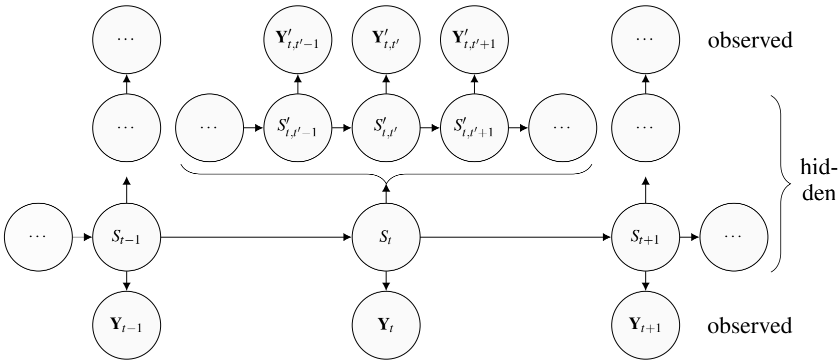

FIGURE 1. Dependence structure of an hierarchical HMM.

HMMs, which originate from the field of machine learning (Fine *et al.*, 1998) and have later been applied in ecology (Leos-Barajas *et al.*, 2017; Adam *et al.*, 2017; Adam *et al.*, 2019), to incorporate such multi-scale data into a joint model. The suggested approach is illustrated by jointly modeling 16 years of monthly trade volumes and daily log-returns of the Goldman Sachs stock.

2 Model formulation and likelihood evaluation

A basic HMM comprises two stochastic processes: a hidden state process $\{S_t\}_{t=1,\dots,T}$ and an observed state-dependent process $\{Y_t\}_{t=1,\dots,T}$. The state process is typically modeled as a discrete-time, N-state Markov chain with initial distribution $\boldsymbol{\delta} = (\delta_i)$, $\delta_i = \Pr(S_1 = i)$, and transition probability matrix (t.p.m.) $\boldsymbol{\Gamma} = (\gamma_{i,j})$, $\gamma_{i,j} = \Pr(S_{t+1} = j | S_t = i)$. The state at time t, $S_t = i$, selects one of N possible distributions, which are denoted by $f(y_t|S_t = i)$, that generates the outcome of the state-dependent process (cf. Zucchini *et al.*, 2016).

By exploiting this relatively simple dependence structure, the likelihood can be written as a matrix product,

$$\mathcal{L}^{\text{HMM}}(\boldsymbol{\theta}|y_1,\dots,y_T) = \boldsymbol{\delta}\mathbf{P}(y_1)\prod_{t=2}^{T}\boldsymbol{\Gamma}\mathbf{P}(y_t)\mathbf{1}, \tag{1}$$

where $\mathbf{P}(y_t) = \text{diag}\big(f(y_t|S_t = 1),\dots,f(y_t|S_t = N)\big)$ and $\mathbf{1}$ denotes a column vector of ones (cf. Zucchini *et al.*, 2016).

Hierarchical HMMs extend the model structure outlined above in that they distinguish between processes operating at different time scales (cf. Figure 1 for an illustration of the model structure): the coarse-scale state at time t, $S_t = i$, selects among N possible distributions for the coarse-scale observations, which are denoted by y_t (e.g. the trade volume observed for month

t), and N possible HMMs (each of which has its own t.p.m. $\mathbf{\Gamma}'_i$) for the fine-scale observations, which are denoted by $\mathbf{y}'_t$ (e.g. all daily log-returns observed during month t). The likelihood then follows as

$$\mathcal{L}^{\text{HHMM}}(\boldsymbol{\theta}|y_1,\ldots,y_T,\mathbf{y}'_1,\ldots,\mathbf{y}'_T) = \boldsymbol{\delta}\mathbf{P}(y_1,\mathbf{y}'_1)\prod_{t=2}^{T}\mathbf{\Gamma}\mathbf{P}(y_t,\mathbf{y}'_t)\mathbf{1}, \quad (2)$$

where $\mathbf{P}(y_t,\mathbf{y}'_t) = \text{diag}\big(f(y_t|S_t=1)\mathcal{L}^{\text{HMM}}(\boldsymbol{\theta}|\mathbf{y}'_t,S_t=1),\ldots,f(y_t|S_t=N)\mathcal{L}^{\text{HMM}}(\boldsymbol{\theta}|\mathbf{y}'_t,S_t=N)\big)$. Estimation of the model parameters is typically carried out by numerical likelihood maximization (cf. Adam *et al.*, 2019).

3 Application to stock market data

To investigate stock market dynamics at different time scales, we jointly model 16 years of monthly trade volumes and daily log-returns of the Goldman Sachs stock. The data cover 4,026 working days (192 months) between January 1, 2004, and December 31, 2019. For the trade volumes, we assumed gamma distributions, while for the log-returns, scaled t-distributions (as preferred over Normal distributions by Akaike's information criterion) were considered.

The estimated state-dependent distributions of monthly trade volumes, displayed in the top left panel of Figure 2, reveal three different market regimes: while coarse-scale states 1 and 2 capture low and moderate trade volumes (inactive and moderately active market), respectively, state 3 relates to high trade volumes (active market).

The t.p.m. associated with the coarse-scale state process was estimated as

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.984 & 0.016 & 0.000 \\ 0.043 & 0.900 & 0.057 \\ 0.000 & 0.282 & 0.718 \end{pmatrix},$$

which implies the stationary distribution $(0.687, 0.261, 0.053)$, indicating that about 69 %, 26 %, and 5 % of the monthly trade volumes were generated in coarse-scale states 1, 2, and 3, respectively. Notably, in 2007, when a sudden increase in interest rates for inter-bank credits marked the beginning of the global financial crisis, the decoded time series displayed in the top right panel of Figure 2 reveals a switch from coarse-scale state 1 (inactive market) to 2 (moderately active market). In September 2008 (when the Lehman Brothers collapse marked the peak of the global financial crisis), we observe a switch from coarse-scale state 2 to 3 (active market).

The estimated state-dependent distributions of daily log-returns are displayed in the middle panel of Figure 2: depending on the coarse-scale state that is active in month t, the log-returns' volatility is determined by the fine-scale HMM associated with the two distributions displayed in either the left, the middle, or the right panel, respectively.

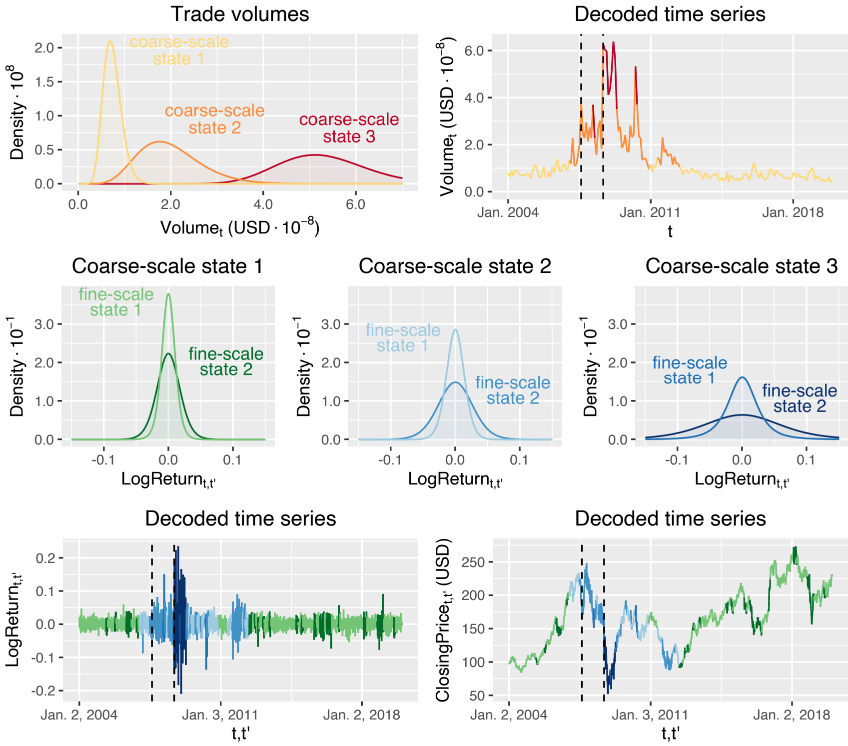

FIGURE 2. Estimated state-dependent distributions and decoded time series of monthly trade volumes, daily log-returns, and closing prices of the Goldman Sachs stock. Dashed lines in the top-right and the bottom panel indicate important events associated with the global financial crisis.

The t.p.m.s associated with the fine-scale state processes were estimated as

$$\hat{\boldsymbol{\Gamma}}'_1 = \begin{pmatrix} 0.993 & 0.007 \\ 0.034 & 0.966 \end{pmatrix}, \; \hat{\boldsymbol{\Gamma}}'_2 = \begin{pmatrix} 0.993 & 0.007 \\ 0.024 & 0.976 \end{pmatrix}, \; \hat{\boldsymbol{\Gamma}}'_3 = \begin{pmatrix} 0.915 & 0.085 \\ 0.029 & 0.971 \end{pmatrix},$$

which imply the stationary distributions $(0.823, 0.177)$, $(0.779, 0.221)$, and $(0.255, 0.745)$. According to the fitted model, when coarse-scale state 1 (inactive market) is active (about 67 % of the time) then the marginal distribution of the log-returns under the fitted model has standard deviation 0.013. When coarse-scale state 3 (active market) is active (about 5 % of the time), then the log-returns' volatility is about five times higher: the corresponding marginal distribution has standard deviation 0.065.
Quantile-quantile-plots and sample autocorrelation functions (ACFs) of ordinary normal pseudo-residuals for monthly trade volumes and daily log-returns are displayed in Figure 3. While, in principle, more flexible state-

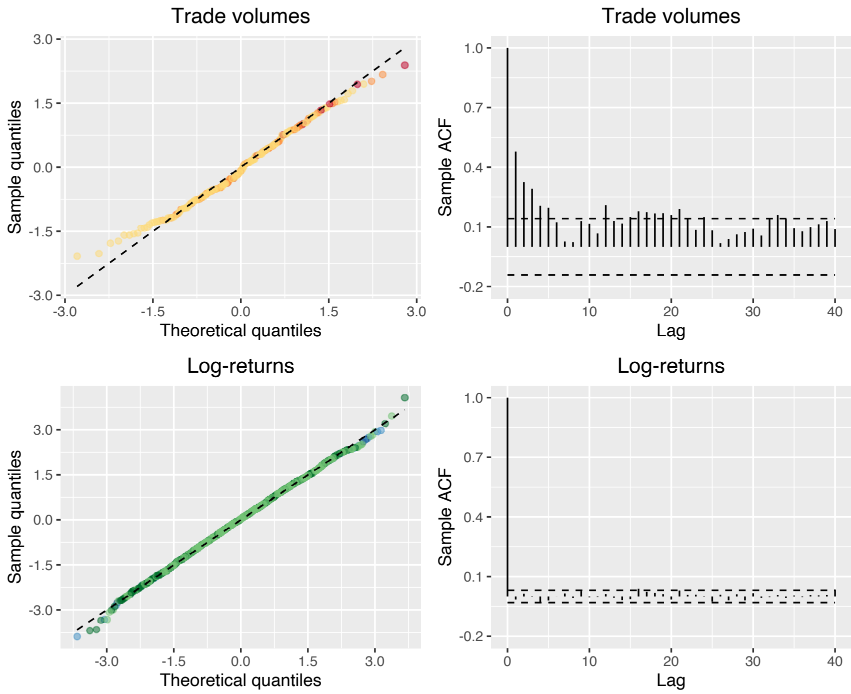

FIGURE 3. Quantile-quantile-plots (left panel) and sample ACFs (right panel) of normal ordinary pseudo-residuals for monthly trade volumes (top panel) and daily log-returns (bottom panel). Overall, the plots indicate some minor lack of fit with regard to the marginal distribution of the trade volumes and the serial correlation in the trade volumes' series.

dependent distributions (especially for the trade volumes) could be used to improve the model fit (cf. Langrock *et al.*, 2018), we consider the goodness of fit to be satisfactory and trade some minor lack of fit against a more complex model formulation, which facilitates the interpretation of the fitted model.

4 Conclusions

The results presented in this paper indicate that coarse-scale market dynamics strongly affect the stochastic properties of other processes operating at finer time scales. By explicitly modeling such multi-scale processes, hierarchical HMMs may help to draw a more comprehensive picture of stock market dynamics, to more accurately quantify risks conditional on the coarse-scale market regime, and ultimately to improve our understanding of the market agents' behavior.

References

Adam, T., Leos-Barajas, V., van Beest, F.M., and Langrock, R. (2017). Using hierarchical hidden Markov models for joint inference at multiple temporal scales. *Proceedings of 32nd International Workshop on Statistical Modelling*, **2**, 181 – 184.

Adam, T., Griffiths, C.A., Leos-Barajas, V., Meese, E.N., Lowe, C.G., Blackwell, P.G., Righton, D., and Langrock, R. (2019). Joint modeling of multi-scale animal movement data using hierarchical hidden Markov models. *Methods in Ecology and Evolution*, **10**(9), 1536 – 1550.

Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden Markov model: analysis and applications. *Machine Learning*, **32** (1), 41 – 62.

Langrock, R., Adam, T., Leos-Barajas, V., Mews, S., Miller, D.L., and Papastamatiou, Y.P. (2018). Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, **72**(3), 179 – 200.

Leos-Barajas, V., Gangloff, E.J., Adam, T., Langrock, R., van Beest, F.M., Nabe-Nielsen, J., and Morales, J. (2017). Hierarchical generalized linear models. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 232 – 248.

Zucchini, W., MacDonald, I.L., and Langrock, R. (2016). *Hidden Markov models for time series: an introduction using R.* Boca Raton: Chapman and Hall/CRC.