TREE-BASED REGRESSION WITHIN A HIDDEN MARKOV MODEL FRAMEWORK

Rouven Michels[1], Timo Adam[2] and Marius Ötting[1]

[1] Department of Business Administration and Economics, Bielefeld University, (e-mail: r.michels@uni-bielefeld.de, marius.oetting@uni-bielefeld.de)

[2] Department of Mathematical Sciences, University of Copenhagen, (e-mail: tiad@math.ku.dk)

ABSTRACT: While tree-based regression methods are popular in practice, they miss a time series component. We thus combine regression trees with hidden Markov models (HMMs) and construct a hybrid model that can effectively capture serial correlation and the complex dependencies between the input and output variables, while also providing interpretable results. In a case study, we demonstrate that such an approach offers a powerful and flexible tool for modeling financial data. However, the presented method can be employed in many more fields, e.g. in ecology or sports.

KEYWORDS: hidden Markov model, regression tree, distributional tree, financial markets.

1 Introduction

Tree-based regression models are a popular machine learning tool as they can capture complex interaction effects and yet can be easily interpreted. Combining these models with hidden Markov models (HMMs), which serve for modelling time-series data with serial correlation, is an approach that uses the strengths of both techniques. The scaffold of this model is the assumption that, for each $t = 1, \ldots, T$, the observed time series data $\{Y_t\}_{t=1,\ldots,T}$ is generated by one of N regression trees built by M input variables. Each of these trees corresponds to one of the N states selected by the hidden state process $\{S_t\}_{t=1,\ldots,T}$. We model the latter by an N-state, first-order Markov chain with initial distribution $\delta_i = \Pr(S_1 = i)$ and state transition probabilities $\gamma_{ij} = \Pr(S_t = j \mid S_{t-1} = i),\ i, j = 1, \ldots, N$. Putting these properties together, this results in a model that probabilistically switches between regression trees.

2 Model fitting with the EM algorithm

To fit the model, we use the EM algorithm (Zucchini *et al.*, 2016). We represent the sequence of states $\{S_t\}_{t=1,\ldots,T}$ by the indicator variables $u_i(t) = I(S_t = i)$ and $v_{i,j}(t) = I(S_{t-1} = i, S_t = j)$, $i, j = 1, \ldots N$, $t = 1, \ldots, T$. Then, we can write the joint log-likelihood of the observation process, $\{Y_t\}_{t=1,\ldots,T}$, and the states (i.e. the complete-data log-likelihood) as

$$\begin{aligned} l(\theta) &= \log\left(\delta_{s_1}\prod_{t=2}^{T}\gamma_{s_{t-1},s_t}\prod_{t=1}^{T}\Pr(Y_t = y_t \mid S_t = s_t)\right) \\ &= \sum_{i=1}^{N} u_i(1)\log(\delta_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T} v_{i,j}(t)\log(\gamma_{i,j}) \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} u_i(t)\log\big(\Pr(Y_t = y_t \mid S_t = i)\big). \end{aligned}$$

The EM algorithm switches between E- and M-Step, i.e. between estimating the $u_i(t)$'s and $v_{i,j}(t)$'s given the current parameter estimates and maximizing the joint log-likelihood $l(\theta)$. We address the problem of local maxima by running the EM algorithm with different starting values.

Still to discuss is the precise form of $p_i(y_t) = \Pr(Y_t = y_t \mid S_t = i)$. For regular HMMs, this expression is given by the density or probability function of the chosen state-dependent distribution. As we do not make any distributional assumption, we have to find an appropriate expression for regression trees. In the following, we will present two possible procedures: The obvious approach is to employ the CART algorithm (Breiman *et al.*, 1984), to use weights according to the actual state probabilities and to fit regression trees by minimizing the corresponding residual sum of squares (Therneau & Atkinson, 2019). Then, we assume $p_i(y_t)$ to be normally distributed where the mean equals the leaf node's means

$$\mu_t = \frac{1}{n_{\tilde{m}_i}} \sum_{j=1,\ldots,T} I(\mathbf{x}_j \in R_{\tilde{m}_i})\, y_j$$

with $\tilde{m}_i \in 1, \ldots, M_i$ being the node for which $\mathbf{x}_t \in R_{\tilde{m}_i}$ and $n_{\tilde{m}_i}$ denoting the number of observations in region $R_{\tilde{m}_i}$ for the tree of state i. Moreover, the standard deviation σ_t is regarded as a hyperparameter to tune. In the second approach, we do not employ classical regression trees but distributional trees which constitute as a specific form of regression trees. The difference is the way of splitting. While for regression trees the splitting rule only optimizes

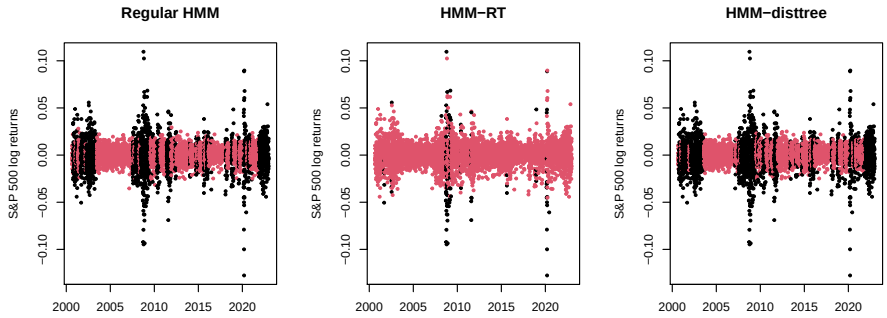

Figure 1. *The time series of log-returns of the S&P 500 from 30th August, 2000 until 30th December, 2022 are displayed. The most likely states under the corresponding model (left panel: Regular HMM; middle panel: HMM-RT; right panel: HMM-disttree) are colorized.*

according to the means between the leaf nodes, for distributional trees the data are split into homogeneous groups with respect to a full parametric distribution (Schlosser *et al.*, 2019). Like in the first approach, we replace $p_i(y_t)$ with the density of a normal distribution, however, the standard deviation is no longer a hyperparameter. We fit such distributional trees using the R package *disttree* (Schlosser *et al.*, 2021).

3 Application to financial data

To illustrate the usefulness of the proposed approach, we consider a case study on financial data. In financial markets, the terms "bullish" and "bearish" describe the overall sentiment of the market participants towards a particular asset or the market as a whole. In an HMM context, we can use these two terms as proxies for latent states. A bullish market is characterized by a calm period of moderately rising prices, while a bearish market is marked by nervousness and oscillating, but mostly falling prices. We apply the presented methods to the daily S&P 500 log-returns from 30.08.2000 – 30.12.2022 as the observed time series and use two input variables, the daily oil and gold log-returns.

After fitting both models to the data, we use the Viterbi algorithm (see Zucchini *et al.*, 2016) for state decoding. We can see in Figure 1 (middle panel) that the classical regression tree approach is not able to capture the bullish and bearish markets as the model switches between states within these market

phases. In contrast, the distributional tree recognizes calm and nervous markets (right panel of Figure 1) which builds the basis for further analysis, e.g. the prediction of future log-returns. When comparing the distributional tree method to a regular HMM with a normal distribution as the state-dependent distribution (left panel of Figure 1), significant similarities can be observed. However, in the presence of more covariates, the distributional tree regression method automatically chooses variables and interactions (see Schlosser *et al.*, 2019) and, thus, circumvents the usual selection problems.

4 Discussion

Using tree-based regression in the framework of HMMs presents a promising approach for modeling complex data sets with a wide range of input variables. Specifically, our findings indicate that employing distributional trees in the EM algorithm outperforms classical regression trees in this context. Differences in other distribution parameters than the mean (such as the standard deviation) can only be captured by distributional trees, which provide much more flexibility without being computationally more costly. In particular, the HMM-RT approach is twice as fast, but also requires cross-validation via the standard deviation, which is why in the end the HMM-disttree method is more efficient.

The approach presented herein should be considered as merely a starting point for establishing connections between HMMs and machine learning algorithms within the regression domain. For instance, the combination of HMMs and random forests could potentially mitigate concerns related to overfitting.

References

BREIMAN, L., FRIEDMAN, J., OLSHEN, R., & STONE, C. 1984. *Classification and Regression Trees*. New York: Wadsworth.

SCHLOSSER, L., HOTHORN, T., STAUFFER, R., & ZEILEIS, A. 2019. Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain. *The Annals of Applied Statistics*, **13**(3), 1564 – 1589.

SCHLOSSER, L., LANG, M.N., HOTHORN, T., & ZEILEIS, A. 2021. *disttree: Trees and Forests for Distributional Regression*. R package version 0.2-0.

THERNEAU, T., & ATKINSON, B. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1–15.

ZUCCHINI, W., MACDONALD, I. L., & LANGROCK, R. 2016. *Hidden Markov Models for Time Series: An Introduction Using R*. New York: CRC press.