



TECHNISCHE
UNIVERSITÄT
WIEN

D I P L O M A R B E I T

**Probabilistic forecasting using Hidden Markov models:
A use-case for household electrical power load forecasting**

ausgeführt am

Institut für
Stochastik und Wirtschaftsmathematik
TU Wien

unter der Anleitung von

Privatdoz. Dipl.-Ing. Dr. Zsolt Saffer

durch

Florian Schimek BSc.

Matrikelnummer: 11717650

Wien, am 14.12.2025

Kurzfassung

Obwohl Hidden Markov Modelle (HMM) in zahlreichen Anwendungsfeldern erfolgreich eingesetzt werden, ist ihre Nutzung für Prognosezwecke bislang vergleichsweise wenig untersucht worden. Aufgrund ihres stochastischen Aufbaus stellen HMM jedoch einen effizienten, mathematisch konsistenten und interpretierbaren Ansatz zur Erstellung probabilistischer Vorhersagen dar – einem Forschungsgebiet mit wachsender Relevanz, das derzeit überwiegend von Machine-Learning Methoden und Black-Box-Modellen dominiert wird.

Diese Arbeit untersucht die Anwendung eines HMM-basierten Prognoseverfahrens zur kurzfristigen Vorhersage der Verteilung von elektrischer Haushaltslasten. Sie erweitert die bestehende Literatur, indem ein allgemeiner theoretischer Rahmen für HMM-basierende Prognosen entwickelt und daraus ein einfacher, zugleich leistungsfähiger Prognosealgorithmus abgeleitet wird. Die Implementierung des Use-Case umfasst eine für HMM neuartige Diskretisierungsmethode, eine umfassende Hyperparameteroptimierung sowie eine detaillierte Modellanalyse, um Stärken und Grenzen des vorgeschlagenen Ansatzes systematisch zu bewerten. Erste empirische Ergebnisse deuten darauf hin, dass die HMM-basierte Prognosemethode etablierte Verfahren übertrifft und die relative Prognosegenauigkeit um bis zu 7% steigern kann. Obwohl der Schwerpunkt dieser Arbeit auf der Theorie, dem Aufbau und der Analyse des einfachen HMM-Prognosemodells liegt, sind die Resultate vielversprechend um HMM-basierende Vorhersagemethode weiter zu erforschen und sie als potenziell wettbewerbsfähige Alternative zu bestehenden kurzfristigen probabilistischen Prognosemethoden zu berücksichtigen.

Abstract

Even though Hidden Markov models (HMMs) have been successfully implemented for various applications, HMM based forecasting has been widely neglected. Due to its stochastic design, HMM offers an efficient, mathematically enclosed, and tractable approach to generate probabilistic forecasts, which is a research field of rising interests dominated by black-box machine learning methods.

This thesis investigates the possibility of applying a HMM forecasting method to predict the distribution of short-term electrical load of households. It contributes to existing literature by introducing a general theoretic framework for HMM based forecasting and derives a simple but efficient HMM forecasting algorithm. The implementation of the use-case includes a for HMM novel discretization type, detailed hyperparameter tuning and model analysis, to investigate the strengths and weaknesses of the proposed model. First tests show that the HMM forecasting method performs widely better than state-of-the-art models, reaching up to 7% improvement in prediction accuracy. While this thesis focuses on the theoretical backing, build-up and exploration of the basic HMM forecasting model, the results are promising to further pursue this topic and consider HMM as a potential, competitive alternative to existing short-term probabilistic forecasting methods.

Danksagung

Ich möchte mich bei meine Betreuer Privatdoz. Dipl.-Ing. Dr. Zsolt Saffer bedanken, mir mit dieser Diplomarbeit die Möglichkeit zu geben meine beiden Forschungsinteressen - die Stochastik und Energie Systeme - zu kombinieren.

Ein herzlicher Dank geht raus nach Vorarlberg an Lukas Moosbrugger, dessen Feedback und Austausch über die letzten zwei Jahre mir immer wieder geholfen haben. Ebenfalls bin ich sehr dankbar für meinen Kollegen und Freunde, die mich während meines Studium unterstützten, und es zu einer wahnsinnig schönen Zeit meines Lebens gemacht hatten. Zu Letzt bedanke ich mich bei meiner Familie, insbesondere meinen Eltern die immer für mich da waren und mich bestens (kulinarisch) versorgten, und meiner Schweseter Babsi für das Korrekturlesen dieser Arbeit.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am Datum

Name des Autors / der Autorin

Contents

| | |
|---|-----------|
| 1. Introduction | 1 |
| 2. Theoretical Foundations | 5 |
| 2.1. Hidden Markov Models | 5 |
| 2.1.1. HMM Foundations | 5 |
| 2.1.2. Algorithms | 7 |
| 2.1.3. HMM Forecasting | 16 |
| 2.2. Probabilistic Forecasting | 19 |
| 2.2.1. Prediction Space | 19 |
| 2.2.2. Calibration and Sharpness | 20 |
| 2.2.3. Scoring Rules and CRPS | 23 |
| 3. Experiment Design | 26 |
| 3.1. Data | 26 |
| 3.2. Model Setup | 27 |
| 3.2.1. Preprocessing | 28 |
| 3.2.2. HMM Training | 29 |
| 3.2.3. HMM Forecasting Algorithm | 30 |
| 3.2.4. Postprocessing | 30 |
| 3.2.5. Hyperparameter Tuning | 30 |
| 3.3. Model Analysis | 32 |
| 3.3.1. Evaluation of Prediction Accuracy | 32 |
| 3.3.2. Calibration via PIT Histograms | 32 |
| 3.3.3. Convergence for Increasing Forecasting Horizon | 33 |
| 3.4. Benchmarks | 33 |
| 3.4.1. Benchmark Setup | 34 |
| 3.4.2. Benchmark Models | 35 |
| 4. Results of Hyperparameter Tuning | 38 |
| 4.1. Optimization of N and M | 38 |
| 4.2. Sensitivity Analysis for T | 41 |
| 5. Results Model Analysis | 43 |
| 5.1. Evaluation of Prediction Accuracy | 43 |
| 5.2. Calibration via PIT Histograms | 44 |
| 5.3. Convergence for Increasing Forecasting Horizon | 45 |
| 6. Results of Benchmarks | 51 |

| | |
|--|-----------|
| 7. Conclusion | 55 |
| Bibliography | 57 |
| A. Results of Hyperparameter Analysis | 61 |
| B. Results of Sensitivity Analysis | 62 |
| C. Results of Benchmark | 64 |

1. Introduction

Hidden Markov Models

The Hidden Markov Model (HMM) is a well-studied statistical model that has evoked great interest and remarkable results in the past decades in both research and practice. In simple words, HMM is a stochastic process which underlies a latent Markov process but can only be observed through probabilistic signals emitted by the states of this hidden process. Originally, it was first studied by Baum and Petrie [3] as *probabilistic functions of Markov chains* and gained significance during the 1980s and 1990s as an application in speech recognition, see Rabiner [35]. During this time, HMMs dominated automated speech recognition research and commercial application, forming the backbone of early voice-controlled interfaces. In general, HMMs are applied for tasks that require decoding, recognition and classification as in DNA analysis and gene recognition [29, 39] or behavior detection [6]. Recently, Makonin, Popowich, Bajić, Gill, and Bartram [32] used HMM to disaggregate electric power load profiles and identify the usage of single electric devices. Over the last years, machine learning techniques like recurrent neural networks or transformers have superseded HMMs especially in large-scale tasks of decoding and classification. Nevertheless, HMM still remains a valuable tool in many applications due to its interpretability, mathematical tractability, and efficiency in low-data settings.

HMM Based Forecasting

In this thesis, our aim is to investigate another possible application of HMM, namely forecasting. In the existing literature, to the best of the author's knowledge, there have been only a few single studies conducted using HMM as a forecasting method. Erlwein, Benth, and Mamon [12] developed a HMM-driven method to model the dynamics of the one-step-a-head prediction for the electrical spot price market for which 2- and 3-states Markov processes are trained to estimate certain dynamic parameters. In Tian and Shen [40], different Markovian models, including 4 HMMs, are compared to predict the probability of the binary event of US recession. Ghasvarian Jahromi, Gharavian, and Mahdiani [13] uses a common approach from speech recognition in which multiple HMM models are trained on wind speed data and then, the currently most fitting model is applied for prediction. Stock market predictions using HMM were studied in Gupta and Dhingra [17] and Hassan and Nath [20], while Khadr [27] developed a HMM based drought forecasting process. However, the literature of HMM in forecasting has several shortcomings.

Foremost, there exists no general theory on HMM as a forecasting method, which is commonly acknowledged and based upon. The theoretical approaches of many papers are

often tightly tailored to their specific model but actually rely on similar ideas, while only a few take completely different methods. This makes it not only difficult to understand in many cases how the forecasts were actually derived, but it also blurs the distinction between already existing methods and novelties in this field.

Secondly, all reviewed papers apply small-scale HMMs (less than 10 hidden states) with predefined hyperparameter settings. While these states are often interpreted by practical groundings (like recession vs economic growth), it is not guaranteed that these assumptions are actually the optimal hyperparameter settings. Ideally, hyperparameter tuning should be conducted like in classical machine learning methodology, and additionally bigger models can be considered which have become feasible with computers nowadays. Thirdly, most literature (with exception of [13]) derive deterministic or point forecasts from their HMM models, which is most often achieved by taking the most probable event as a prediction. However, this approach strips away the HMM method from a significant part of its provided information, since HMM considers the probabilities for all possible outcomes due to its stochastic design. Taking full advantage of this stochastic design results in probabilistic forecasting. Probabilistic forecasts, for example predicting the distribution of a future event, have been applied for many years exclusively in very niche applications like meteorology. However in the last decade, it has generated great interest in many fields due to its capability to measure the uncertainty of forecasting. An example is the area of finance and economics in which probabilistic forecasting offers a great tool for risk assessment. One challenge of applying probabilistic forecasting methods is often the more advanced design and greater computational effort in comparison to their deterministic counterparts. Precisely for this reason, the application of HMM as probabilistic forecasting method offers a great opportunity. Not only does it have a straight forward design that is easy to understand, implement and efficiently train, but also provides directly probabilistic forecasts. Furthermore, in comparison to many black-box machine learning methods, HMMs are described in closed mathematical form and its interpretable parameters offer insight into its decision-making. These benefits can allow the HMM forecasting model to be employed for critical systems in which the model's decision has to be understood or be included in theoretical frameworks for further studies like stochastic optimization systems.

This thesis contributes in filling the three research gaps described above. We derive a general theoretic framework for the application of HMM as a forecasting method, which builds upon and uses similar approaches as the standard HMM work of Rabiner [35]. Based on this general method, we deploy HMM in a use-case of probabilistic forecasting. For the training of the HMM forecasting model, a thorough methodology is developed to investigate the optimal hyperparameter settings, sensitivities and predictive behavior. Here, we study, in addition to the standard hyperparameters of HMM, also different types of discretization in detail and historic window lengths, which is a novelty for HMM.

Use-Case: Probabilistic Load Forecasting for Households

The studied use-case of this work is the short-term, probabilistic forecast of the electrical power load for single households. In detail, we will predict the distribution of a household's electrical power demand 15 minutes ahead using a novel HMM forecasting method. Electrical power load forecasting is an essential task for energy providers and power grid operators to guarantee energy supply and security. In times of energy transition, the rising demand of electricity and increasing level of volatile renewable energy are challenging the current electrical power systems. For this reason, research and companies invest large efforts to find new solutions and innovations to increase the efficiency of the existing systems. The digitalisation of the power systems in the last years, in particular the widespread roll-out of smart meters, contributed to a great increase of available data, literature and new forecasting methods in the field of low voltage energy systems. Haben, Arora, Giasemidis, Voss, and Vukadinović Greetham [18] conducted an encompassing review of low voltage load forecasting, identifying multiple research gaps and challenges, and provides recommendations for future work. One key take away from this review is, that point forecasts are often not sufficient due to the high volatility of demand on the level of low voltage. A similar argument was stated within the *global energy forecasting competitions* (GEFCom), see [22–24]. While the first GEFCom2012 addressed point forecasts, the following competitions GEFCom2014 and GEFCom2017 exclusively targeted probabilistic forecasts.

Probabilistic load forecasting has developed to a popular research field with a rising number of publications in the recent years. Existing statistical methods like *quantile regression* [19] laid the foundation for probabilistic load forecasting. To transfer existing results from developed deterministic forecasting, many ensemble approaches exist. While *quantile regression averaging* applies quantile regression on the output of several point forecasts [30], *dropout* methods generate uncertainty in deterministic machine learning methods [8]. Advanced machine learning methods have been extended to predict probabilistic events directly, even though this significantly increases the computational complexity, summarized in Eren and Küçükdemiral [11]. Wang, Gan, Sun, Zhang, Lu, and Kang [42] developed a *long short-term memory* model to forecast households electrical demand probabilistically, while Xu, Hu, and Fan [43] extended the model with a Bayesian approach. Luo, Zheng, Hong, Luo, and Yang [31] combined *fuzzy logic* with *support vector machines* to predict point and quantile forecasts. In Botman, Lago, Becker, Vanthournout, and Moor [5] a generalized approach is taken to predict the distribution of households electricity demand, in which one general model is trained for a big data set of various households and can be applied for single household electric load prediction also when only little data is available. Furthermore, *general additive models* are capable to explain multiple influences in electrical load forecasting as shown in Zimmermann and Ziel [44]. Our use-case of probabilistic forecasting the load of single households is potentially relevant for two parties. Home owners are confronted with an increase in electric demand in the upcoming future due to the electrification of transport (E-mobility) and heating (heat pumps). Using home energy management systems (HEMS) can optimize the operation of the home system via different components (PV electricity production, battery storage

management, flexible demand, dynamic pricing) for which good demand predictions are valuable. For example, HEMS can utilize probabilistic forecasts to evaluate the risk of crossing a certain power threshold¹ and act accordingly. Essential challenges for home owners to implement forecasting methods are dominantly lack of available data and limited computational resources. The proposed HMM forecasting method counteracts these challenges since it is trained only on historical load data without additional information (like eg. weather data) and efficiently due to its relatively simple mathematical design and algorithms. The second potential target group are distribution grid operators. For a robust operation of any electrical power grid, accurate forecasts are essential to avoid grid congestion and expensive re-dispatching. By aggregating multiple household's forecasts, the implementation of a hierarchical bottom-up forecasting model is made possible by the efficiency of the proposed HMM forecasting method.

The choice of applying the proposed HMM forecasting method for probabilistic forecasting the electrical power load of households is well justified by following summarizing arguments. The novel approach of HMM forecasting method offers great opportunities in the field of energy forecasting, especially probabilistic load forecasting, because its stochastic design provides an efficient alternative to existing probabilistic forecasting methods. Further, probabilistic forecasts show great benefits over point estimations for high volatile data such as household's electric demand. Lastly, the proposed HMM forecasting method satisfies the requirements for potential applications of household's electrical power forecasting.

Contribution and Structure of the Thesis

Following bullet points highlight the contribution of this thesis to current research:

- Development of a general theoretic framework for HMM based forecasting.
- Deployment of a novel probabilistic HMM forecasting method on households electrical power load.
- Detailed analysis of the prediction model's hyperparameter and behavior, including a novel discretization method for HMM.

The presented thesis is structured as follows. Chapter 2 covers the theoretic foundations required, including the necessary theory of HMM and probabilistic forecasting. Notably in section 2.1.3, the theoretic framework for HMM based forecasting and a forecasting algorithm are introduced. Chapter 3 discusses the experiment design and methodology of the use-case of the HMM forecasting method for probabilistic load prediction. In chapter 4 and 5, the results of the hyperparameter tuning and model analysis are presented respectively. Benchmarks are conducted in chapter 3.4 in order to compare the proposed HMM forecasting method to other state-of-the-art probabilistic load forecasting methods.

¹In Austrian's upcoming law (EIWG-Gesetz), home owners with PV production will be penalized by a fee, if their current supply into the grid exceeds 20kW.

2. Theoretical Foundations

2.1. Hidden Markov Models

HMM has proven to be a valuable extension of the general Markov processes because of its capability to model signals and other volatile data. This has been shown in its wide application for speech recognition. The most popular paper and standard reference in this field is Rabiner [35] summarizing the main challenges of HMM, presenting the most important results and algorithms and gives an introduction to speech recognition applications. Other HMM related work builds upon the results and theory discussed in that paper. In this thesis, we exploit a relatively unexplored application of HMM as a forecasting method, for which, however, also already existing algorithms are required. For this reason, we give an encompassing introduction to the theory and present existing algorithms to solve important problems regarding HMM. In the last section, we derive the general theory for HMM forecasting and formulate an efficient HMM forecasting algorithm.

2.1.1. HMM Foundations

Starting with the notion of HMM, there are N possible hidden states forming the *state space* S and M different observation values forming the *observation space* O ,

$$\begin{aligned}
 S &= \{s_1, s_2, \dots, s_N\} \\
 O &= \{o_1, o_2, \dots, o_M\}.
 \end{aligned}$$

We restrict the discussion to discrete HMM with finite state and observation space, thus $N, M \in \mathbb{N}$. Denote the sets of indexes by $I_S = \{1, \dots, N\}$ and $I_O = \{1, \dots, M\}$. Let $T \in \mathbb{N}$ be the time horizon and its corresponding index set $I_T = \{1, \dots, T\}$. The HMM consists of two dependent stochastic processes, a hidden and an observable one. For each time instance, the hidden process $\{Z_t, t \in I_T\}$ takes on values in the state space, while the observable process $\{X_t, t \in I_T\}$ is located in the observation space,

$$\begin{aligned}
 (Z_1, Z_2, \dots, Z_T) &\in S^T \\
 (X_1, X_2, \dots, X_T) &\in O^T.
 \end{aligned}$$

To distinguish between random variables and the event of random variables happening, we introduce the following compact notation. The event of Z_t taking the value s_{i_t} is denoted by \dot{Z}_t ,

$$\dot{Z}_t = \{Z_t = s_{i_t}\},$$

for a fixed $i_t \in I_S$ that describes the state index at time t . All events until time instance t are summed up by the event vector $\dot{\mathbf{Z}}_t$,

$$\dot{\mathbf{Z}}_t = (\dot{Z}_1, \dot{Z}_2, \dots, \dot{Z}_t).$$

Analogously for the observable process, a fixed $i_k \in I_O$ describes the observation index at time t ,

$$\dot{X}_t = \{X_t = o_{k_t}\},$$

In addition to $\dot{\mathbf{X}}_t$, the vector $\dot{\mathbf{X}}'_t$ describes all occurred events after time instance t (including time instance t),

$$\begin{aligned}\dot{\mathbf{X}}_t &= (\dot{X}_1, \dot{X}_2, \dots, \dot{X}_t) \\ \dot{\mathbf{X}}'_t &= (\dot{X}_t, \dot{X}_{t+1}, \dots, \dot{X}_{T-1}, \dot{X}_T).\end{aligned}$$

The HMM is then entirely characterized through the following three properties.

(A1) Markov property:

$$\mathcal{P}(\dot{Z}_{t+1} | \dot{\mathbf{Z}}_t) = \mathcal{P}(\dot{Z}_{t+1} | \dot{Z}_t)$$

(A2) Homogeneity:

$$\mathcal{P}(Z_{t+1} = s_j | Z_t = s_i) = \mathcal{P}(Z_t = s_j | Z_{t-1} = s_i) = \dots = \mathcal{P}(Z_2 = s_j | Z_1 = s_i)$$

(A3) Independent observation assumption:

$$\mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-1}, \dot{\mathbf{X}}'_{t+1}, \dot{\mathbf{Z}}_T) = \mathcal{P}(\dot{X}_t | \dot{Z}_t)$$

The properties **A1** and **A2** determine the process Z to be a homogeneous Markov chain. Thus, a future event depends on the past events only through the present event. The third property **A3** implies that the observation is influenced only by the current hidden state. Due to these properties, a HMM can be parameterized by stochastic matrices and vectors similar to the transition matrix of standard Markov chains. The hidden Markov process $\{Z_t, t \in I_T\}$ is determined by the $N \times N$ -dimensional *transition matrix* \mathbb{A} with elements a_{ij} ,

$$a_{ij} = \mathcal{P}(Z_t = s_j | Z_{t-1} = s_i) \quad \text{for } i, j = 1, \dots, N,$$

and by the N -dimensional *initial distribution vector* π with

$$\pi_i = \mathcal{P}(Z_1 = s_i) \quad \text{for } i = 1, \dots, N.$$

The $N \times M$ -dimensional *observation matrix* \mathbb{B} describes the emission probabilities of the observation process $\{X_t, t \in I_T\}$ with elements b_{ik} ,

$$b_{ik} = \mathcal{P}(X_t = o_k | Z_t = s_i) \quad \text{for } i = 1, \dots, N \text{ and } k = 1, \dots, M.$$

These three parameters fully describe a finite HMM, summarized by $\lambda = (\mathbb{A}, \mathbb{B}, \pi)$.

2.1.2. Algorithms

The three strong properties **A1**, **A2**, **A3** allow us to compute HMM related problems very efficiently. Before we present common algorithms, we first identify the three main questions regarding HMM formulated by Rabiner [35]:

1. **Evaluation:**

Given an HMM and observation sequence $\dot{\mathbf{X}}_{\mathbf{T}}$, what is the probability that this sequence occurred,

$$\mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}})?$$

2. **Decoding:**

Given a HMM and observation sequence $\dot{\mathbf{X}}_{\mathbf{T}}$, what is the most probable state sequence $\dot{\mathbf{Z}}_{\mathbf{T}}$ which emits the given observations,

$$\arg \max_{\dot{\mathbf{Z}}_{\mathbf{T}}} \mathcal{P}(\dot{\mathbf{Z}}_{\mathbf{T}} | \dot{\mathbf{X}}_{\mathbf{T}})?$$

3. **Training:**

Given observation sequence $\dot{\mathbf{X}}_{\mathbf{T}}$, what are the optimal parameters λ of the HMM to describe the given observations,

$$\arg \max_{\lambda} \mathcal{P}_{\lambda}(\dot{\mathbf{X}}_{\mathbf{T}})?$$

Due to the three properties of HMMs, all of these quantities can be efficiently computed. In this section, we discuss solutions and algorithms for the problem of evaluation and training, which is relevant for the application of HMM as a forecasting tool. The question of decoding can also be relevant for the analysis of existing forecasting models to develop a deeper understanding of the underlying mechanisms, which however we do not discuss further in this thesis.

To show the computational complexity of the above problems, we first make an exemplary naive approach of calculating the evaluation problem. Given $\dot{\mathbf{X}}_{\mathbf{T}}$, we reformulate the desired probability using the conditional probability over all possible hidden state sequences $\mathcal{Z}_{\mathcal{T}}$,

$$\begin{aligned} \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}}) &= \sum_{\dot{\mathbf{Z}}_{\mathbf{T}} \in \mathcal{Z}_{\mathcal{T}}} \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}}, \dot{\mathbf{Z}}_{\mathbf{T}}) \\ &= \sum_{\dot{\mathbf{Z}}_{\mathbf{T}} \in \mathcal{Z}_{\mathcal{T}}} \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}} | \dot{\mathbf{Z}}_{\mathbf{T}}) \mathcal{P}(\dot{\mathbf{Z}}_{\mathbf{T}}). \\ &= \sum_{\dot{\mathbf{Z}}_{\mathbf{T}} \in \mathcal{Z}_{\mathcal{T}}} \left(\prod_{t=1}^T \mathcal{P}(\dot{X}_T | \dot{Z}_T) \right) \left(\mathcal{P}(\dot{Z}_1) \prod_{t=2}^T P(\dot{Z}_t | \dot{Z}_{t-1}) \right) \\ &= \sum_{\dot{\mathbf{Z}}_{\mathbf{T}} \in \mathcal{Z}_{\mathcal{T}}} \left(\prod_{t=1}^T b_{i_t k_t} \right) \left(\pi_{i_1} \prod_{t=2}^T a_{i_{t-1} i_t} \right). \end{aligned}$$

The set of all possible hidden state sequences is defined formally by

$$\mathcal{Z}_T = \{(Z_t = s_{i_t})_{t \in \{1, \dots, T\}} \text{ for } (i_t)_{t \in \{1, \dots, T\}} \in I_S^T\}.$$

Since I_S has N elements, there are N^T different possible hidden state sequences. Thus, evaluating $\mathcal{P}(\dot{\mathbf{X}}_T)$ has a non-polynomial complexity of order $\mathcal{O}(N^T \cdot 2T)$ and for larger models or longer time horizons reaches quickly the limits of computational feasibility. However, due to its Markovian behavior, the evaluation of $\mathcal{P}(\dot{\mathbf{X}}_T)$ can be done more efficiently by building up an iterative procedure based on its three properties. This is called the *forward algorithm* presented in the following.

We define the quantity $\alpha_t(i)$ as

$$\alpha_t(i) = \mathcal{P}(\dot{\mathbf{X}}_t, Z_t = s_i), \quad \text{for } i \in I_S, t \in \{1, \dots, T\}.$$

For $t \geq 2$, the joint probability $\alpha_t(j)$ can be expressed by its predecessors $\alpha_{t-1}(\cdot)$ and the parameters of the given HMM,

$$\begin{aligned} \alpha_t(j) &= \mathcal{P}(\dot{\mathbf{X}}_t, Z_t = s_j) \\ &= \sum_{i \in I_S} \mathcal{P}(\dot{\mathbf{X}}_t, Z_t = s_j, Z_{t-1} = s_i) \\ &= \sum_{i \in I_S} \mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-1}, Z_t = s_j, Z_{t-1} = s_i) \mathcal{P}(Z_t = s_j | \dot{\mathbf{X}}_{t-1}, Z_{t-1} = s_i) \\ &\quad \cdot \mathcal{P}(\dot{\mathbf{X}}_{t-1}, Z_{t-1} = s_i) \\ &= \sum_{i \in I_S} \mathcal{P}(\dot{X}_t | Z_t = s_j) \mathcal{P}(Z_t = s_j | Z_{t-1} = s_i) \mathcal{P}(\dot{\mathbf{X}}_{t-1}, Z_{t-1} = s_i) \\ &= \sum_{j \in I_S} b_{jk_t} \cdot a_{ij} \cdot \alpha_{t-1}(i). \end{aligned}$$

The computation of the first joint probabilities $\alpha_1(\cdot)$ requires the initial distribution π ,

$$\alpha_1(i) = \mathcal{P}(\dot{X}_1, Z_1 = s_i) = \mathcal{P}(\dot{X}_1 | Z_1 = s_i) \mathcal{P}(Z_1 = s_i) = b_{ik_1} \cdot \pi_i.$$

Using the quantities $\alpha_T(\cdot)$, the probability targeted in the evaluation problem is calculated by

$$\mathcal{P}(\dot{\mathbf{X}}_T) = \sum_{i \in I_S} \mathcal{P}(\dot{\mathbf{X}}_T, Z_T = s_i) = \sum_{i \in I_S} \alpha_T(i).$$

This is the standard forward algorithm, discussed in [35], and has a computational complexity of order $\mathcal{O}(N^2 \cdot 3T)$, which is significantly lower than that of the naive approach. However, this algorithm also has its drawbacks in practice. If we look closely at the definition of the joint probability $\mathcal{P}(\dot{\mathbf{X}}_T)$, it can be observed that with increasing time horizon T , this probability is becoming smaller and in regular cases converges to 0. It implies that with a significantly large time horizon, this quantity will be smaller than any

for computer representable number and thus not feasible to compute. For this reason, we present a version of the forward algorithm adapted to conditional probabilities which do not vanish with increasing T and allows for the computation of the log-likelihood of the evaluation problem.

We define two quantities $\bar{\alpha}$ and $\tilde{\alpha}$ similar to the original α ,

$$\bar{\alpha}_t(i) = \mathcal{P}(\dot{X}_t, Z_t = s_i | \dot{\mathbf{X}}_{t-1})$$

and

$$\tilde{\alpha}_t(i) = \mathcal{P}(Z_t = s_i | \dot{\mathbf{X}}_t),$$

for $i \in I_S$ and $t \in \{1, \dots, T\}$. On the one hand, $\bar{\alpha}_t(i)$ is computed by the predecessors $\tilde{\alpha}_{t-1}(\cdot)$,

$$\begin{aligned} \bar{\alpha}_t(i) &= \mathcal{P}(\dot{X}_t, Z_t = s_i | \dot{\mathbf{X}}_{t-1}) \\ &= \sum_{j \in I_S} \mathcal{P}(\dot{X}_t, Z_t = s_i, Z_{t-1} = s_j | \dot{\mathbf{X}}_{t-1}) \\ &= \sum_{j \in I_S} \mathcal{P}(\dot{X}_t, Z_t = s_i | \dot{\mathbf{X}}_{t-1}, Z_{t-1} = s_j) \mathcal{P}(Z_{t-1} = s_j | \dot{\mathbf{X}}_{t-1}) \\ &= \sum_{j \in I_S} \mathcal{P}(\dot{X}_t | Z_t = s_i) \mathcal{P}(Z_t = s_i | Z_{t-1} = s_j) \mathcal{P}(Z_{t-1} = s_j | \dot{\mathbf{X}}_{t-1}) \\ &= \sum_{j \in I_S} b_{ikt} \cdot a_{ji} \cdot \tilde{\alpha}_{t-1}(j). \end{aligned}$$

Considering the fact of

$$\begin{aligned} \mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-t}) &= \sum_{i \in I_S} \mathcal{P}(Z_t = s_i, \dot{X}_t | \dot{\mathbf{X}}_{t-t}) \\ &= \sum_{i \in I_S} \bar{\alpha}_i(t), \end{aligned}$$

we can express on the other hand $\tilde{\alpha}_t(i)$ by $\bar{\alpha}_t(\cdot)$,

$$\begin{aligned} \tilde{\alpha}_t(i) &= \mathcal{P}(Z_t = s_i | \dot{\mathbf{X}}_t) \\ &= \frac{\mathcal{P}(\dot{X}_t, Z_t = s_i | \dot{\mathbf{X}}_{t-t})}{\mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-t})} \\ &= \frac{\bar{\alpha}_t(i)}{\sum_{j \in I_S} \bar{\alpha}_j(t)}. \end{aligned}$$

As mentioned, the probability $\mathcal{P}(\dot{\mathbf{X}}_T)$ induces underflow errors. However, the log-likelihood of the evaluation problem, $\log \mathcal{P}(\dot{\mathbf{X}}_T)$ lies generally within the computationally feasible range. The log-likelihood is calculated by splitting up the original problem into

subproblems,

$$\log \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}}) = \log \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}}) = \log \prod_{t=1}^T \mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-1}) = \sum_{t=1}^T \log \mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-1}),$$

and using the fact that $\mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-t}) = \sum_{j \in I_S} \bar{\alpha}_j(t)$ as shown above. Additionally, consider the following relation between $\tilde{\alpha}$ and α ,

$$\begin{aligned} \tilde{\alpha}_t(i) &= \mathcal{P}(Z_t = s_i | \dot{\mathbf{X}}_t) \\ &= \frac{\mathcal{P}(Z_t = s_i, \dot{\mathbf{X}}_t)}{\mathcal{P}(\dot{\mathbf{X}}_t)} \\ &= \frac{\mathcal{P}(\dot{\mathbf{X}}_t, Z_t = s_i)}{\sum_{j \in I_S} \mathcal{P}(\dot{\mathbf{X}}_t, Z_t = s_j)} \\ &= \frac{\alpha_t(i)}{\sum_{j \in I_S} \alpha_j(t)}. \end{aligned}$$

This shows that $\tilde{\alpha}_t(i) = c_t \cdot \alpha_t(i)$ with the constant factor $c_t = \frac{1}{\sum_{j \in I_S} \alpha_j(t)}$. Thus, $\tilde{\alpha}_t$ is a scaled version of α_t , which becomes relevant in the training problem.

Adapted forward algorithm

1. *Initialization:* ($t = 1$)

$$\begin{aligned} \bar{\alpha}_1(i) &= b_{ik_1} \cdot \pi_i \quad \text{for } i \in I_S \\ \mathcal{P}(\dot{X}_1) &= \sum_{j \in I_S} \bar{\alpha}_1(j) \\ \tilde{\alpha}_1(i) &= \bar{\alpha}_1(i) / \mathcal{P}(\dot{X}_1) \quad \text{for } i \in I_S \end{aligned}$$
2. *Iteration:* ($t = 2, \dots, T$)

$$\begin{aligned} \bar{\alpha}_t(i) &= \sum_{j \in I_S} b_{ikt} \cdot a_{ji} \cdot \bar{\alpha}_{t-1}(j) \quad \text{for } i \in I_S \\ \mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-1}) &= \sum_{j \in I_S} \bar{\alpha}_j(t) \\ \tilde{\alpha}_t(i) &= \bar{\alpha}_t(i) / \mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-1}) \quad \text{for } i \in I_S \end{aligned}$$
3. *Termination:*

$$\log \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}}) = \sum_{t=1}^T \log \mathcal{P}(\dot{X}_t | \dot{\mathbf{X}}_{t-1})$$

return $\log \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}})$

In addition, the evaluation can also be computed with the so-called *backward algorithm*. It utilizes the HMM properties similar as the forward algorithm but iterates from the other temporal direction, thus backwards. Both of these algorithms are needed later on to solve the training problem.

Let $\beta_t(i)$ be the following quantity for $t \in \{1, \dots, T-1\}$ and $i \in I_S$,

$$\beta_t(i) = \mathcal{P}(\dot{\mathbf{X}}'_{t+1} | Z_t = s_i).$$

The probabilities $\beta_t(\cdot)$ are described recursively and initiated by setting $\beta_T(\cdot)$ to one,

$$\beta_T(i) = 1.$$

Then,

$$\begin{aligned} \beta_t(i) &= \mathcal{P}(\dot{X}_{t+1}, \dot{\mathbf{X}}'_{t+2} | Z_t = s_i) \\ &= \sum_{j \in I_S} \mathcal{P}(\dot{X}_{t+1}, \dot{\mathbf{X}}'_{t+2} | Z_t = s_i, Z_{t+1} = s_j) \mathcal{P}(Z_{t+1} = s_j | Z_t = s_i) \\ &= \sum_{j \in I_S} \mathcal{P}(\dot{X}_{t+1} | Z_{t+1} = s_j) \mathcal{P}(\dot{\mathbf{X}}'_{t+2} | Z_{t+1} = s_j) \mathcal{P}(Z_{t+1} = s_j | Z_t = s_i) \\ &= \sum_{j \in I_S} b_{jk_{t+1}} \cdot \beta_{t+1}(j) \cdot a_{ij}. \end{aligned}$$

Since $\dot{\mathbf{X}}_T = \dot{\mathbf{X}}'_1$, the desired probability of the evaluation problem is then computed by,

$$\begin{aligned} \mathcal{P}(\dot{\mathbf{X}}_T) &= \mathcal{P}(\dot{\mathbf{X}}'_1) \\ &= \sum_{i \in I_S} \mathcal{P}(\dot{\mathbf{X}}'_1 | Z_1 = s_i) \mathcal{P}(Z_1 = s_i) \\ &= \sum_{i \in I_S} \mathcal{P}(\dot{X}_1 | Z_1 = s_i) \mathcal{P}(\dot{\mathbf{X}}'_2 | Z_1 = s_i) \mathcal{P}(Z_1 = s_i) \\ &= \sum_{i \in I_S} b_{ik_1} \cdot \beta_1(j) \cdot \pi_i. \end{aligned}$$

Similar to the forward algorithm, the standard version of the backward algorithm as described above induces numerical difficulties in practice and therefore, demands an adaptation. To avoid infeasibility, $\beta_t(\cdot)$ is scaled with the constant $c_t = \sum_{j \in I_S} \beta_t(j)$ for $t \in \{1, \dots, T-1\}$ and $c_T = 1$. summarized by $\tilde{\beta}_t(\cdot)$,

$$\tilde{\beta}_t(i) = \frac{\beta_t(i)}{c_t} = \frac{\mathcal{P}(\dot{\mathbf{X}}'_{t+1} | Z_t = s_i)}{\sum_{j \in I_S} \mathcal{P}(\dot{\mathbf{X}}'_{t+1} | Z_t = s_j)}$$

This scaling implies that

$$\sum_{i \in I_S} \tilde{\beta}_t(i) = 1,$$

and thus assures feasibility for all $\tilde{\beta}_t(i)$. The factor c_t has no meaningful interpretation as the scaling factor of the adapted forward algorithm and also cannot be directly computed.

However, due to the above result, the ratio c_t/c_{t+1} is calculated as follows,

$$\begin{aligned}
 1 &= \sum_{i \in I_S} \tilde{\beta}_t(i) \\
 &= \sum_{i \in I_S} \frac{\beta_t(i)}{c_t} \\
 &= \sum_{i \in I_S} \frac{\sum_{j \in I_S} b_{jk_{t+1}} \cdot a_{ij} \cdot \beta_{t+1}(j)}{c_t} \\
 &= \sum_{i \in I_S} \frac{\sum_{j \in I_S} b_{jk_{t+1}} \cdot a_{ij} \cdot \tilde{\beta}_{t+1}(j) \cdot c_{t+1}}{c_t} \\
 &= \frac{c_{t+1}}{c_t} \sum_{i \in I_S} \sum_{j \in I_S} b_{jk_{t+1}} \cdot a_{ij} \cdot \tilde{\beta}_{t+1}(j).
 \end{aligned}$$

Hence,

$$\frac{c_t}{c_{t+1}} = \sum_{i \in I_S} \sum_{j \in I_S} b_{jk_{t+1}} \cdot a_{ij} \cdot \tilde{\beta}_{t+1}(j).$$

The log-likelihood is then derived by

$$\begin{aligned}
 \log \mathcal{P}(\dot{\mathbf{X}}_T) &= \log \left(\sum_{i \in I_S} b_{ik_1} \cdot \pi_i \cdot \beta_1(j) \right) \\
 &= \log \left(\sum_{i \in I_S} b_{ik_1} \cdot \pi_i \cdot \tilde{\beta}_1(j) \cdot c_1 \right) \\
 &= \log \left(\sum_{i \in I_S} b_{ik_1} \cdot \pi_i \cdot \tilde{\beta}_1(j) \cdot \prod_{t=1}^{T-1} \frac{c_t}{c_{t+1}} \right) \\
 &= \log \left(\sum_{i \in I_S} b_{ik_1} \cdot \pi_i \cdot \tilde{\beta}_1(j) \right) + \sum_{t=1}^{T-1} \log \frac{c_t}{c_{t+1}}.
 \end{aligned}$$

Adapted backward algorithm

1. *Initialization:* ($t = T$)
 $\tilde{\beta}_T(i) = 1$
 $c_T = 1$
 2. *Iteration:* ($t = T - 1, \dots, 1$)
 $\frac{c_t}{c_{t+1}} = \sum_{i \in I_S} \sum_{j \in I_S} b_{jk_{t+1}} \cdot a_{ij} \cdot \tilde{\beta}_{t+1}(j)$
 $\beta_t(i) = \sum_{j \in I_S} b_{jk_{t+1}} \cdot a_{ij} \cdot \tilde{\beta}_{t+1}(j)$
 3. *Termination:*
 $\log \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}}) = \log \left(\sum_{i \in I_S} b_{ik_1} \cdot \pi_i \cdot \tilde{\beta}_1(j) \right) + \sum_{t=1}^{T-1} \log \frac{c_t}{c_{t+1}}$
- return* $\log \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}})$

Turning now to the more advanced problem of training HMMs, we introduce the *Baum-Welch algorithm*. The objective is to find the optimal parameters of an HMM to fit a given observation sequence. The Baum-Welch algorithm is a special case of the expectation maximization (EM) algorithm, which is an iterative approach to find the local maxima of the likelihood for the parameter settings of a statistical model. Each iteration is divided into two steps, the expectation and the maximization step. While in the expectation step, the current expectational values under fixed parameters are computed, these expectational values are then used in the maximization step to re-estimate the next parameters. The reader is referred to Dempster, Laird, and Rubin [10] for a detailed discussion of EM algorithms and to Baum [2] for a formal derivation of the Baum-Welch algorithm. In this work, we illustrate the main ideas of Baum-Welch and give an intuitional approach. Let $\xi_t(i, j)$ be the conditional probability of being in the hidden state s_i at the time instance t and transitioning to state s_j under the given observation sequence $\dot{\mathbf{X}}_{\mathbf{T}}$,

$$\xi_t(i, j) = \mathcal{P}(Z_t = s_i, Z_{t+1} = s_j \mid \dot{\mathbf{X}}_{\mathbf{T}}).$$

Due to the fact of

$$\begin{aligned} \mathcal{P}(Z_t = s_i, Z_{t+1} = s_j, \dot{\mathbf{X}}_{\mathbf{T}}) &= \\ &= \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}} \mid Z_t = s_i, Z_{t+1} = s_j) \mathcal{P}(Z_t = s_i, Z_{t+1} = s_j) = \\ &= \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{t}} \mid Z_t = s_i) \mathcal{P}(\dot{\mathbf{X}}'_{\mathbf{t}+1} \mid Z_{t+1} = s_j) \mathcal{P}(Z_{t+1} = s_j \mid Z_t = s_i) \mathcal{P}(Z_t = s_i) \\ &= \mathcal{P}(\dot{\mathbf{X}}_{\mathbf{t}}, Z_t = s_i) \mathcal{P}(\dot{\mathbf{X}}'_{\mathbf{t}+2} \mid Z_{t+1} = s_j) \mathcal{P}(\dot{X}_{t+1} \mid Z_{t+1} = s_j) \mathcal{P}(Z_{t+1} = s_j \mid Z_t = s_i) \\ &= \alpha_t(i) \cdot \beta_{t+1}(j) \cdot b_{jk_{t+1}} \cdot a_{ij}, \end{aligned}$$

we can express $\xi_t(i, j)$ as the following,

$$\begin{aligned}\xi_t(i, j) &= \frac{\mathcal{P}(Z_t = s_i, Z_{t+1} = s_j, \dot{\mathbf{X}}_{\mathbf{T}})}{\mathcal{P}(\dot{\mathbf{X}}_{\mathbf{T}})} \\ &= \frac{\mathcal{P}(Z_t = s_i, Z_{t+1} = s_j, \dot{\mathbf{X}}_{\mathbf{T}})}{\sum_{i' \in I_S} \sum_{j' \in I_S} \mathcal{P}(Z_t = s_{i'}, Z_{t+1} = s_{j'}, \dot{\mathbf{X}}_{\mathbf{T}})} \\ &= \frac{\alpha_t(i) \cdot a_{ij} \cdot b_{jk_{t+1}} \cdot \beta_{t+1}(j)}{\sum_{i' \in I_S} \sum_{j' \in I_S} \alpha_t(i') \cdot a_{i'j'} \cdot b_{j'k_{t+1}} \cdot \beta_{t+1}(j')}.\end{aligned}$$

Where the quantities $\alpha_t(i)$ and $\beta_t(j)$ are used from the forward and backward algorithm. Since $\alpha_t(\cdot)$ and $\beta_t(\cdot)$ are both used in the nominator as well as the denominator, they can be exchanged by its scaled versions $\tilde{\alpha}_t(\cdot)$ and $\tilde{\beta}_t(\cdot)$.

Further, consider the probability of being in state i during the time instance t given $\dot{\mathbf{X}}_{\mathbf{T}}$, thus¹

$$\gamma_t(i) = \mathcal{P}(Z_t = s_i | \dot{\mathbf{X}}_{\mathbf{T}}) = \sum_{j \in I_S} \mathcal{P}(Z_t = s_i, Z_{t+1} = s_j | \dot{\mathbf{X}}_{\mathbf{T}}) = \sum_{j \in I_S} \xi_t(i, j).$$

Given the quantities $\xi_t(\cdot, \cdot)$ and $\gamma_t(\cdot)$, useful information on the hidden process is available. The expected number of transitions from state s_i to state s_j is equivalent to the sum of all $\xi_t(i, j)$ over $t \in \{1, \dots, T-1\}$,

$$\mathbb{E}(\#transitions\ from\ s_i\ to\ s_j) = \sum_{t=1}^{T-1} \xi_t(i, j).$$

The expected number of visits in state s_i is described by the sum of all $\gamma_t(i)$ over $t \in \{1, \dots, T\}$,

$$\mathbb{E}(\#visits\ in\ s_i) = \sum_{t=1}^T \gamma_t(i).$$

Similarly, the expected number of transitions from state s_i is the sum of all $\gamma_t(i)$ over $t \in \{1, \dots, T-1\}$,

$$\mathbb{E}(\#transitions\ from\ s_i) = \sum_{t=1}^{T-1} \gamma_t(i).$$

The above equations represent the expectation step of the Baum-Welch algorithm.

¹In case $t = T$, $\gamma_T(j)$ is computed by $\sum_{i \in I_S} \xi_{T-1}(i, j)$.

In the second step, the maximization step, the parameters of the HMM are re-estimated. Hereby, the results from the expectation step and the probabilistic concept of counting occurrences are used the following way,

$$\begin{aligned}\pi'_i &= \mathcal{P}(Z_1 = s_i | \dot{\mathbf{X}}_{\mathbf{T}}) = \gamma_1(i) \\ a'_{ij} &= \frac{\mathbb{E}(\# \text{transitions form } s_i \text{ to } s_j)}{\mathbb{E}(\# \text{transitions from } s_i)} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\ b'_{ik} &= \frac{\mathbb{E}(\# \text{visits in } s_i \text{ and observe } o_k)}{\mathbb{E}(\# \text{visits in } s_i)} = \frac{\sum_{t=1}^T \mathbb{1}_{\{o_{k_t}=o_k\}} \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)}.\end{aligned}$$

Baum [2] shows that the likelihood of the re-estimated parameters $\lambda' = (\pi', \mathbb{A}', \mathbb{B}')$ is always greater or equal to the likelihood of the previous parameter setting λ . By iterating this procedure, the parameters are improved and thus, provides an applicable training method. However, it only converges against a local minimum, not guaranteeing an optimal solution to the training problem. Nonetheless, it has been shown in practice that in general this local minimum is sufficient and additionally not very sensitive to different initial settings. One possible reason is that the parameter space is highly symmetric due to interchangeability of the hidden states. Therefore, we choose the initial parameters at random. The algorithm terminates if the log-likelihood has converged or a maximum number of iterations is reached.

Baum-Welch algorithm

1. *Initialization*

$\lambda = (\pi, \mathbb{A}, c)$ is chosen randomly

For i in $1, \dots, \text{maxIteration}$:

2. *Expectation step*

$$\begin{aligned}\xi_t(i, j) &= \frac{\tilde{\alpha}_t(i) \cdot a_{ij} \cdot b_{j k_{t+1}} \cdot \tilde{\beta}_{t+1}(j)}{\sum_{i' \in I_S} \sum_{j' \in I_S} \tilde{\alpha}_t(i') \cdot a_{i' j'} \cdot b_{j' k_{t+1}} \cdot \tilde{\beta}_{t+1}(j')} \\ \gamma_t(i) &= \sum_{j \in I_S} \xi_t(i, j)\end{aligned}$$

3. *Maximization step*

$$\begin{aligned}\pi'_i &= \gamma_1(i) \\ a'_{ij} &= \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\ b'_{ik} &= \frac{\sum_{t=1}^T \mathbb{1}_{\{o_{k_t}=o_k\}} \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)}\end{aligned}$$

4. *Termination condition*

if $\mathcal{P}_\lambda(\dot{\mathbf{X}}_{\mathbf{T}}) = \mathcal{P}_{\lambda'}(\dot{\mathbf{X}}_{\mathbf{T}})$: terminate
else: set $\lambda = \lambda'$

return λ

2.1.3. HMM Forecasting

This section introduces the theoretic reasoning behind HMM forecasting methods. Given the efficient algorithms of the previous section, this task is derived straight forward. Additional, while we aim to use the resulting probabilities for probabilistic forecasts, we show possible methods to apply HMM in point forecasting.

The general goal of the probabilistic forecasting problem is to predict with which probabilities future events are occurring. Specifying a fixed future time horizon $H \in \mathbb{N}$ and given the historic observation sequence $\dot{\mathbf{X}}_{\mathbf{T}}$, we aim to calculate the probability

$$\mathcal{P}(X_{T+H} = o_k | \dot{\mathbf{X}}_{\mathbf{T}})$$

for all possible observations $k \in I_O$. Due to the HMM properties, these probabilities are calculated by tracing back all possible hidden state sequences probabilities leading up to the forecasting time instance.

First, we break down the initial probability into the following,

$$\begin{aligned} \mathcal{P}(X_{T+H} = o_k | \dot{\mathbf{X}}_{\mathbf{T}}) &= \sum_{j \in I_S} \mathcal{P}(X_{T+H} = o_k | \dot{\mathbf{X}}_{\mathbf{T}}, Z_{T+H} = s_j) \mathcal{P}(Z_{T+H} = s_j | \dot{\mathbf{X}}_{\mathbf{T}}) \\ &= \sum_{j \in I_S} b_{jo_k} \mathcal{P}(Z_{T+H} = s_j | \dot{\mathbf{X}}_{\mathbf{T}}) \\ &= \sum_{j \in I_S} b_{jo_k} \sum_{i \in I_S} \mathcal{P}(Z_{T+H} = s_j | Z_T = s_i) \mathcal{P}(Z_T = s_i | \dot{\mathbf{X}}_{\mathbf{T}}) \end{aligned}$$

From Markov theory, we can identify the probability $\mathcal{P}(Z_{T+H} = s_j | Z_T = s_i)$ as the *H-step-transition* probability $a_{ij}^{(H)}$, see Saffer [36]. This quantity is computed using the *H-step-transition matrix* $\mathbb{A}^{(H)}$ for which the Chapman-Kolmogorov equation shows that

$$\mathbb{A}^{(H)} = A^H.$$

Further, the second probability term is the already known quantity $\tilde{\alpha}_T(i)$ from the forward algorithm, $\tilde{\alpha}_T(i) = \mathcal{P}(Z_T = s_i | \dot{\mathbf{X}}_{\mathbf{T}})$. Let $\tilde{\alpha}_T$ denote a row vector with elements $\tilde{\alpha}_T(i)$. Then we can formulate equation 2.1 in matrix notation.

$$\begin{aligned} \mathcal{P}(X_{T+H} = o_k | \dot{\mathbf{X}}_{\mathbf{T}}) &= \sum_{j \in I_S} b_{jo_k} \sum_{i \in I_S} a_{ij}^{(H)} \tilde{\alpha}_T(i) \\ &= \sum_{j \in I_S} b_{jo_k} [\tilde{\alpha}_T \mathbb{A}^H]_j \\ &= \tilde{\alpha}_T \mathbb{A}^H \mathbb{B}_{.o_k} \end{aligned} \tag{2.1}$$

Hereby, $[\tilde{\alpha}_T \mathbb{A}^H]_j$ denotes the j -element of the row vector $\tilde{\alpha}_T \mathbb{A}^H$ and $\mathbb{B}_{.o_k}$ is the o_k -column of the observation matrix \mathbb{B} . We introduce the M -dimensional row vector θ representing the desired distribution vector over the observation space

$$\theta = [\mathcal{P}(X_{T+H} = o_k | \dot{\mathbf{X}}_{\mathbf{T}})]_{k \in S_O}.$$

Considering equation 2.1.3, the solution for the forecasting problem can then be written in the following compact form,

$$\theta = \tilde{\alpha}_T \cdot \mathbb{A}^H \cdot \mathbb{B}.$$

In this equation, we can observe well the mechanism of the HMM forecasting procedure. The vector $\tilde{\alpha}_T$ is the distribution of the hidden Markov process during the time step T (representing all available information of historical data). If we multiply this distribution by the transition matrix \mathbb{A} , we derive the probabilities for the position of the hidden Markov process of the next time step. Repeating this matrix multiplication h -times, we result in the distribution of the hidden Markov process at the time instance $T + H$. In the final step, we multiply the observation matrix \mathbb{B} that results in the desired distribution of the observable process.

Summarizing the forecasting process within an algorithm consists therefore of two steps: First, computing $\tilde{\alpha}_T$ via the adapted forward algorithm and second, calculating θ through matrix multiplication.

HMM Forecasting Algorithm

1. Step:
Calculate $\tilde{\alpha}_T$ via the adapted forward algorithm.

2. Step:
 $\theta = \tilde{\alpha}_T \cdot \mathbb{A}^H \cdot \mathbb{B}$

return θ

The computational complexity of a single isolated forecast is of the order $O(N^2 \cdot (T + H))$. Even though this computational effort is already viable, the forecasting efficiency can be improved by the following thoughts. Given that in the forecasting step the parameters and thus N are fixed, the two influencing factors are the historic window length T and the forecasting horizon H .

First, discussing the historical factor, recent historic events greatly influence the future outcome, while events longer ago do not bring much additional information. Although using all available data maximizes the available information, the question is how much historic data is truly needed to include all essential information and achieve equally good predictions. Depending strongly on the parameters of the specific HMM, the historic window length T can be reduced to a reasonable length accordingly. In the upcoming chapters of this work, we present an exemplary analysis of the correct choice of T . Furthermore in practice, when new forecasts are generated for each new time instances (for example at time instance T and then at $T + 1$) the previous calculated quantities $\tilde{\alpha}_T$ can be reused for the computation of the new $\tilde{\alpha}_{T+1}$, which cuts down the computational time significantly.²

Secondly, turning to the future horizon H , we can assess the quality of the forecast.

²For scientific reproducibility, this idea is not applied in the experiments of this thesis.

Generally, the farther ahead in the future the predicting event is, the less precise the forecast it will be. In the HMM setting, this is determined by the characteristics of the transition matrix \mathbb{A} . Given that the hidden Markov process fulfills certain conditions³, the process converges to the stationary distribution $\delta = \delta \cdot \mathbb{A}$. Thus, also the forecast $\tilde{\alpha}_T \cdot \mathbb{A}^H \cdot \mathbb{B}$ converges with increasing H towards a fixed distribution (which is the empirical distribution of the observation process). Implying that with increasing H , the informative value of the forecast decreases and thus, putting an upper bound to the reasonable usage of HMM as a forecasting method.

The predicted probabilities $\mathcal{P}(X_{T+H} = o_k | \dot{\mathbf{X}}_T)$ contain all information available through the HMM model, and therefore, the resulting distribution vector θ represents the prediction with the maximum amount of information. While in the present thesis, we apply θ directly as a probabilistic forecast, these probabilities can also be utilized as a basis for other types of forecasts. This can be described by the abstract mapping function f using θ as input,

$$f(\theta).$$

Depending on the application, the observation space and objective, f can take on various forms. Here, we present some suitable use cases in the following table 2.1.3, while Gneiting and Katzfuss [14] thoroughly discuss the theoretical reasoning and properties of statistical functionals of probabilistic forecasts.

| Function f | Application |
|---|--|
| arg max | point forecast for categorical data |
| \mathbb{E} , median | point forecast for numerical data |
| \mathbb{V} , quantiles, confidence interval | measure of uncertainty of the forecast |

Table 1.: Overview of functions and their applications.

³For irreducible, aperiodic and recurrent Markov chains.

2.2. Probabilistic Forecasting

Forecasts aim to predict a future event as accurately as possible. In classic deterministic forecasting, one possible outcome is predicted and the discrepancy between forecast and realization - the prediction error - is measured and in machine learning context also minimized. This approach suggests that the exact result can be predicted with perfect prediction. However, information about future events is by nature incomplete or not fully available, and thus the events themselves are necessarily subject to uncertainty. So, it can be beneficial to acknowledge the stochastic nature of forecasts and look at the problem from a probabilistic perspective. The goal of probabilistic forecasting is to model the uncertain behavior of the future event. This can be done by approximating its probability distribution or, in a more simple form, predicting confidence intervals or quantiles of the desired distribution. A popular example are weather forecast with rain probabilities (e.g. 80% chance of rain tomorrow). In general, meteorology has been and still is the driving field of applied probabilistic forecasting, see Murphy and Winkler [33], since weather forecasts always considered meteorological events part of probabilistic processes. Nowadays, probabilistic forecasting also finds applications like e.g. in financial and economic risk management [16], prediction of low-probability events such as earthquakes [26] or floods [9], and optimization of energy systems [34, 41].

In this section, we summarize the elementary theory of probabilistic forecasting based on the work of Gneiting and Katzfuss [14], who provide an overarching introduction to probabilistic forecasting. The section is structured into the theory of *prediction spaces*, the theoretic foundation, followed by the two most important concepts, *calibration and sharpness*, and finally concluded with *scoring rules*, the evaluation metrics for probabilistic forecasts.

2.2.1. Prediction Space

The theoretic framework for probabilistic forecasting is build on the idea of the *prediction space*, which was first introduced by Gneiting and Katzfuss [14] and developed further in Strähl and Ziegel [38]. The general problem is to predict the distribution for a random event. For our purpose sufficient, we discuss a simplified version of the prediction space based on two components.

First, we consider a one dimensional, real-valued random variable Y on the probability space (Ω, A, P) with a cumulative distribution function (CDF). Y is the future event, we aim to predict, and the probability space denotes the underlying probabilistic process. The second component is the prediction described by the random variable F on the same probability space. It takes a value in the space \mathcal{F} of all continuous CDFs and measurable to the sub σ -algebra $A' \subseteq A$, hence $\mathcal{F} = \{\hat{F} \text{ is CDF} : \hat{F} \text{ continuous} \wedge \hat{F} \text{ } A'\text{-measurable}\}$. Thus, F is a random CDF depending on the information set A' , which can be interpreted as the information available to the predictor like historic data or expertise. We will use the naming 'predictive distribution' for the CDF-valued random quantity F further on. The event Y and prediction F are connected in the sense that the predictor constructs F in such a way with the available information A' that it represents the unknown distribution

of Y as closely as possible.⁴ Formalizing this idea, we introduce the phrase *relative ideal* in the following definition.

Definition 2.1. The predictive distribution F is ideal relative to the information set A' if $F = \mathcal{L}(Y|A')$, where $\mathcal{L}(\cdot)$ denotes the conditional distribution of the random variable.

2.2.2. Calibration and Sharpness

In this section, we discuss beneficial properties of our predictive distribution F . Since in practice, it is not possible to observe the conditional distribution $\mathcal{L}(Y|A')$, we require different forms of assessment for the predictive distribution.

The first main idea is that the future event should occur with the same probabilities as predicted. If we could repeat the same experiment multiple times with a fixed predicted distribution $\hat{F} \in \mathcal{F}$, we would use the empiric distribution of Y and compare it to \hat{F} . However in general, a repetition is not possible since for each time instance the available information and thus also F changes. To overcome this problem, the *probability integral transform* (PIT) is introduced.⁵

Definition 2.2. The random variable $F(Y)$ is called the PIT of the predictive distribution F .

The PIT is a transform of the original random variable in the interval $[0, 1]$ and is the value that the predictive CDF attains at the observed event. Assuming that F is the true distribution of Y , we see that the PIT $F(Y)$ follows a uniform distribution. Furthermore, if we repeat the experiment for a different time instance, the resulting PIT should still be uniformly distributed. This property is generalized by the concept of *calibration* and elaborated farther in the following definition and theorem.

Definition 2.3. Given the random variable Y and the predictive distribution F .

- (i) The predictive distribution F is called *marginally calibrated* if $\mathbb{E}[F(y)] = \mathcal{P}(Y \leq y)$.
- (ii) The predictive distribution F is called *probabilistically calibrated* if its PIT is uniformly distributed, thus $F(Y) \sim U(0, 1)$.

Theorem 2.4. If the predictive distribution F is ideal relative to the information set A' , then it is also marginally and probabilistically calibrated.

Proof. Since F is relative ideal to A' , we can describe it as $F(y) = \mathcal{P}(Y \leq y | A')$. It follows the proof for marginal calibration,

$$\mathbb{E}[F(y)] = \mathbb{E}[\mathcal{P}(Y \leq y | A')] = \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{Y \leq y\}} | A']] = \mathbb{E}[\mathbf{1}_{\{Y \leq y\}}] = \mathcal{P}(Y \leq y).$$

⁴In this abstraction, the forecast underlies the same probabilistic mechanism as the predicted event because it is a random variable on the same probability space. One can argue against this modeling idea, since the forecast will be already known before the random event happening, however stoic philosophers would confidently confirm this approach.

⁵For the case of discrete random variables, the PIT is formulated similar to definition 2.2 with technical additions for handling discontinuities.

Hereby, the characterization of the probability function via the expectancy of the indicator function $\mathbb{1}$, and the tower rule was applied.

For the probabilistic calibration, we need the fact that F takes a value in \mathcal{F} , which implies a continuous, monotonically increasing, surjective function. Thus, there also exists a monotonic increasing, right inverse function F^{-1} , hence

$$F \circ F^{-1} = id, \tag{2.2}$$

whereby id denotes the identity function. Considering the push-forward measure of P by Y onto \mathbb{R} , F is also almost surely strictly monotonic and injective. Hence, the right inverse F^{-1} is also almost surely left inverse,

$$F^{-1} \circ F = id \quad a.s. \tag{2.3}$$

It follows for $s \in [0, 1]$,

$$\begin{aligned} \mathcal{P}(F(Y) \leq s) &= \mathcal{P}(F^{-1}(F(Y)) \leq F^{-1}(s)) \\ &= \mathcal{P}(Y \leq F^{-1}(s)) \\ &= \mathbb{E}[\mathcal{P}(Y \leq F^{-1}(s) | A')] \\ &= \mathbb{E}[F(F^{-1}(s))] \\ &= \mathbb{E}[s] = s. \end{aligned}$$

Respectively, the equations are argued by the monotony of F^{-1} , equation (2.3), the tower rule, the marginal calibration of F and equation (2.2). Thus, we have shown that $F(Y)$ follows a uniform distribution. \square

The theorem 2.4 also holds for non-continuous random variables, which is shown in Brockwell [7]. This result provides us with a necessary, but not sufficient, condition to assess whether a prediction is possibly relatively ideal. That is, by checking the PIT for uniform distribution. There are several statistical tests to check uniformity, such as the Kolmogorov-Smirnov test or the Chi-squared test. More practical, however, is to assess the PIT graphically with the standard method of plotting the histogram of the PIT for sufficient many PIT sample.⁶ In example, each 10% interval (0-10%, 10%-20%,...) should occur in 10% of all cases, which implies that each bar of the PIT histogram should have roughly the same height at 10%. The advantage over statistical numeric test is that this approach provides additional information of the behavior of the prediction. Characteristic errors such as skewness and over-/underdispersion can be easily spotted in the PIT histogram, see exemplary figures 1.

A uniformly distributed PIT, however, does not imply that the prediction F is ideal to the information set A' . Uniformity also applies for predictions which are still calibrated but do not use all of the available information. In mathematical terms, it is the case for a prediction relative ideal to another sub-sigma A^* algebras that is coarser than A' ,

⁶Also, the auto-covariance function of PIT is regularly used to study historic dependencies of forecasts when using time series data.

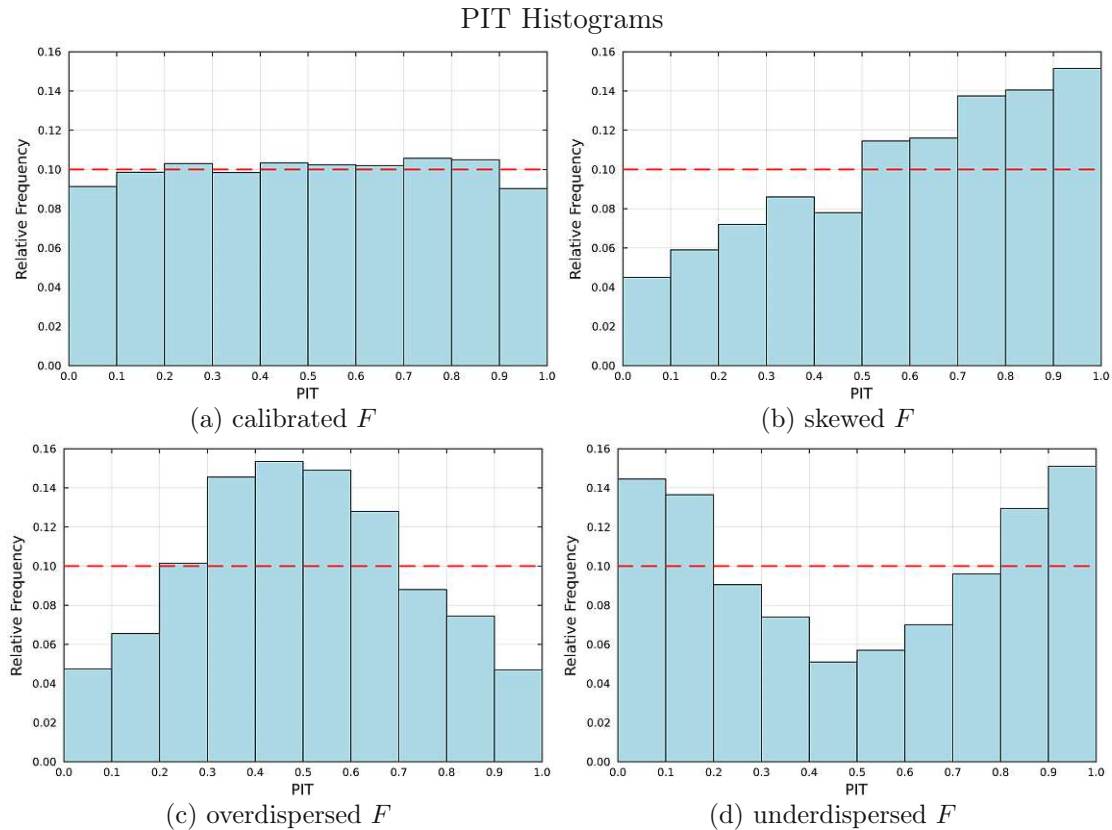


Figure 1.: PIT histogram 1a shows a uniform distributed PIT. In figure 1b the higher intervals occur more often than predicted, while lower intervals occur less often, indicating that the predictive distribution is shifted to the right compared to the underlying distribution. The PIT of figure 1c does not obtain sufficiently often the extreme ends of the distribution implying the predictive distribution to be overdispersed. Whereas in figure 1d, the opposite is the case.

thus $A^* \subset A'$. In example, the unconditional distribution of Y fulfills the condition of a uniformly distributed PIT and is also ideal to the trivial sub-sigma Algebra $A^* = \{\emptyset, \Omega\}$. However, it does not utilize any additional information given by A' . The question is, how can we assess that the maximum of available information was used?

The essential effect of utilizing more information is that uncertainty decreases, thus the predictive distribution becomes generally more concentrated. In probabilistic forecasting, the so-called attribute *sharpness* is used to describe the level of uncertainty. The "sharper" a distribution is, the less uncertain the predictive event is. Given now a set of calibrated predictive distribution, the prediction, which uses the most information, is also the one which maximizes the sharpness. Sharpness can be measured by known metrics like the variance, confidence intervals, or the entropy of a distribution, for example the normal

distribution $N(0, 1)$ is sharper than $N(0, 2)$. Notably, sharpness only takes the predictive distribution into account and does not consider the true realization of the future event. Thus, it evaluates exclusively the predictive distribution.

In summarizing this section, we present the dominating principle of probabilistic forecasting formulated by Gneiting and Katzfuss [14]:

“Probabilistic forecasting has the general goal of maximizing the sharpness of the predictive distributions, subject to calibration.”

2.2.3. Scoring Rules and CRPS

This section discusses how probabilistic forecasts are evaluated numerically. It introduces the concept of *proper scoring rules* and gives a prominent example, the *continuous ranked probability score* (CRPS).

The evaluation metrics of probabilistic forecasts are not as straightforward as those applied in point forecasting. While in point forecasting, the distance between the point prediction and observed value is measured, the aim of the distribution forecast is to match the true underlying distribution, which, however, only "materializes" in one observation per instance. Hence, the task is to measure the distance between distribution forecast and observation, and with sufficiently many evaluations, derive comparable results. These measures are called scoring rules, see Nowotarski and Weron [34] for a detailed approach.

Definition 2.5. The function $S : \mathcal{F} \times \mathbb{R} \rightarrow \mathbb{R}$ is called a scoring rule.

The scoring rule $S(\hat{F}, y)$ assigns the distribution \hat{F} and the realization y of Y a numerical value. Note that in contrast to the previous section, we consider now a fixed distribution \hat{F} , which results from the predictive distribution F .

For the following property definition, another notation of the scoring rule is needed. Given a random variable Y , which follows the distribution $G \in \mathcal{F}$, the expression $S(\hat{F}, G)$ denotes

$$S(\hat{F}, G) = \mathbb{E}_G[S(\hat{F}, Y)].$$

In order to be applicable for the evaluation of probabilistic forecasts, a scoring rule must be *proper*.

Definition 2.6. A scoring rule is *proper* if for all $\hat{F}, G \in \mathcal{F}$

$$S(G, G) \leq S(\hat{F}, G)$$

holds. It is *strictly proper* if $S(G, G) = S(\hat{F}, G)$ implies $\hat{F} = G$.

Thus, a proper scoring rule is designed in such a way that for the true underlying distribution the score is minimized in expectation.

In probabilistic load forecasting, the Continuous Ranked Probability Score (CRPS) has asserted itself in many applications, notably here is the GEFCom2014 summarized in Hong, Pinson, Fan, Zareipour, Troccoli, and Hyndman [23]. Its original form is defined by

$$\text{CRPS}(\hat{F}, y) = \int_{-\infty}^{\infty} (\hat{F}(x) - \mathbb{1}_{\{y < x\}})^2 dx.$$

The elementary idea is to calculate on the left side of the observation the area underneath the predicted distribution and on the right side the area above the predicted distribution, see figure 2. The value is minimized, if the observation lies "central" and the distribution is as sharp as possible. In this way, the CRPS takes the two major concepts of probabilistic forecasting, calibration and sharpness, into account.

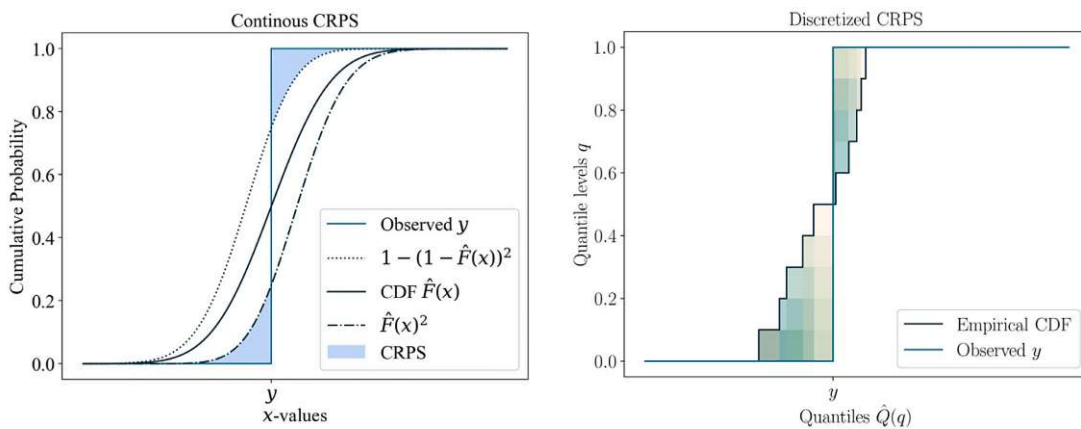


Figure 2.: Graphical example of the continuous and discretized CRPS, reproduced from Botman, Lago, Becker, Vanthournout, and Moor [5]

Interestingly, the CRPS can be equivalently formulated in the following form,

$$\text{CRPS}(\hat{F}, y) = \mathbb{E}(|X - y|) - \frac{1}{2} \mathbb{E}(|X - X'|),$$

for which X and X' are two independent random variables subject to the distribution \hat{F} . The first term of this representation represents the absolute difference between the random variable and the observed realization and can be associated with the calibration. The second term quantifies the volatility of the predicted distribution, and represents the sharpness. With this representation, it is also visible that the CRPS is a generalized form of the mean absolute error (MAE). To wit, if \hat{F} describes a point forecast, thus a trivial distribution with only one event assigned probability 1, the CRPS is equal the MAE.

However, computing the original version of the CRPS can be computationally complicated; therefore, an approximation of the CRPS is often used in practice. For this reason, we consider a third equivalent description of the CRPS using another prominent scoring rule, the *Pinball loss*,

$$\text{discrete CRPS}(\hat{F}, y) = \int_0^1 \text{Pinball}(\hat{Q}(q), y, q) \, dq$$

with

$$\text{Pinball}(\hat{Q}(q), y, q) = \begin{cases} q \cdot (y - \hat{Q}(q)), & \text{if } y \geq \hat{Q}(q) \\ (q - 1) \cdot (y - \hat{Q}(q)), & \text{if } y < \hat{Q}(q), \end{cases}$$

where q denotes a quantile and $\hat{Q}(q)$ the predicted value for q of the predicted distribution \hat{F} . The applied approximation of the CRPS is then defined by the following, which we name the *discrete CRPS*,

$$\text{discrete CRPS}(\hat{F}, y) = \sum_{q \in \mathcal{Q}} \text{Pinball}(\hat{Q}(q), y, q).$$

The set of all quantiles \mathcal{Q} defines how accurate the discrete CRPS approximates the continuous CRPS. A usual choice are all one-percent quantiles, $\mathcal{Q} = \{0.01, 0.02, \dots, 0.99\}$. To achieve comparable results, multiple evaluations are needed and then the mean-CRPS is computed as a statistical valid inference.

3. Experiment Design

This chapter provides details on how HMM can be used in the specific use case of forecasting probabilistically the electrical power load of households. The concrete task is to predict the distribution of the household's electricity demand 15 minutes into the future, only given historical load data. For this reason, we develop an implementation of the HMM forecasting method to predict the electrical load distribution based on the theory presented in the previous chapter. The data set of real, historic, electrical power load of households is used and the applied data processing discussed. The model setup includes a clear structured methodology for building up the HMM forecasting model and a detailed hyperparameter tuning. Furthermore, we conduct a model analysis to study the predictive performance and behavior of the presented HMM model.

3.1. Data

The data source for the experiments presented in this work are measurements from the research project Wind-Solar-Heat Pump District (WPuQ) described by Schlemminger, Ohrdes, Schneider, and Knoop [37]. This dataset contains the electrical power load data of 38 single-family houses in northern Germany including additional information on PV-production, separated heat pump load, weather data, as well as household-specifics and distribution grid measurements. The electrical power consumption was measured over two and a half years during the period of 2018 to 2020 in 10 second intervals. Missing measurements are either labeled as missing value or in cases of missing data gaps smaller than one day linearly interpolated. The authors also provide aggregated data over longer time intervals (1, 15 and 60 minutes) in which the mean power values of the original measurements are taken.

For this thesis, the experiments are conducted on 5 out of the 38 households of the WPuQ data set. These 5 households were randomly selected with the condition that they do not have PV production and only a minimal number of values are interpolated (less than 1% of the total data). The selection of households are described in table 3.1 and denoted with HH1-HH5 in this paper.

For each of these households, one single time series is taken, namely the total¹ active power load, in 15 minutes resolution. The 15 minutes resolution is chosen because this is the current standard resolution of electricity data in most European and especially Austrian energy systems (like eg. EPEX spot market, smart meter data transmission, etc.). In addition, each of the time series is normalized by max-normalization. Thus, each

¹Total is in this context of the sum of all three measured phases of the households electric circuit.

| Selected Household | WPuQ household ID |
|--------------------|-------------------|
| HH1 | SFH3 |
| HH2 | SFH4 |
| HH3 | SFH9 |
| HH4 | SFH12 |
| HH5 | SFH14 |

Table 2.: Mapping between selected households and WPuQ household IDs

value lies in the range of 0 and 1.² We split the data into a training, validation and test data sets. On the training data set, the forecasting models will be trained, the validation data set is used to find the optimal hyperparameter setting, and the test set is applied for evaluation and benchmarking. To avoid seasonal biases, the training data set includes all values of year 2019, which is the first full year with available data. The second year, 2020, is used for validation and testing. Again, to avoid seasonal biases in the evaluation, every odd-numbered month of 2020 (January, March, May,...) is assigned for validation while every even-numbered month of 2020 (February, April, June,...) is assigned to the test data set.

To emphasize the importance of unbiased data sets, Figure 3 demonstrates the time series pattern and differences concerning day times and seasonality. We refer to the paper of Schlemminger, Ohrdes, Schneider, and Knoop [37] for an in-depth analysis of the data characteristics of household’s electrical power demand.

3.2. Model Setup

This section provides insights into the essential steps of the HMM forecasting model applied in the use-case. The target is to predict the distribution of the future electric demand of a household. As a proposed method, the discrete HMM is chosen as presented in section 2.1, which allows a non-parametric approach. Another possibility for future research is the continuous HMM which is potentially more efficient in generating a distribution forecast but further assumptions and prerequisites have to be studied, like the emission distribution family. In order to apply discrete HMMs for this continuous regression problem, preprocessing and postprocessing of the data is required. Also, we assume training and forecasting as isolated problems and no learning during the forecasting process is allowed. Thus, we only train the model once and use this model for multiple forecasts, although more data may be available in future time steps. The general HMM forecasting pipeline is therefore broken down into preprocessing, training, forecasting, and post-processing.³ In addition, and as standard machine learning methodology recommends, we conduct a hyperparameter tuning to obtain the optimal hyperparameter settings whereas most existing literature in HMM forecasting neglected this essential step.

²No negative values are feasible since households with PV production are excluded.

³This methodology is not restricted to load forecasting and can be applied for any other one-dimensional signal.

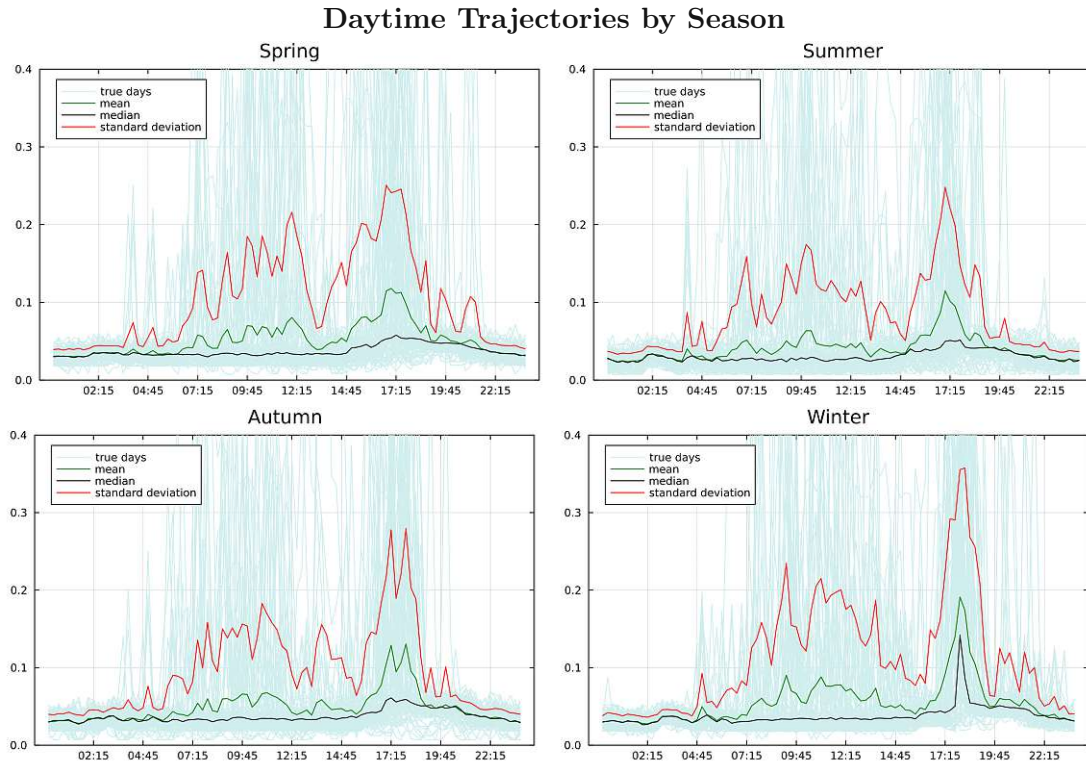


Figure 3.: In these figures, the daytime trajectories of all days for each season are displayed in gray together with statistics like the mean, median, and mean plus standard deviation for HH1. All four seasons show similar load profiles with obvious differences during day and night time. Even though the profiles of different seasons have a similar shape, the variability of the data is significantly greater during the winter months compared to the summer months.

3.2.1. Preprocessing

The historical data is represented by a time series with real valued elements in the power unit *watts*. During preprocessing, this data is discretized from the space of real numbers \mathbb{R} to the discrete observation space O . There are three variabilities of the discretization process: The number of observations M , the selection of observations in O , and the mapping from \mathbb{R} to O . We assume M as a model hyperparameter, which has to be optimized. The remaining two factors are determined by the classic discretization principle. The original observation space \mathbb{R} is divided into disjunct intervals called bins, with each bin having one representative value. In the discretization process, each real-valued observation is then replaced by the representative value of the corresponding bin in which it is lying. Since the choice of the intervals and the representative values can have a significant impact on the models performance, we propose two different discretization

methods and study the effect of both discretization types in section 5

Discretization Type A: Equal-mass bins

The bins are selected so that each bin includes the same number of observations. More precisely, the interval limits of the bins are selected based on the empiric quantiles⁴. As representative value, the median of each bin is chosen. Therefore, high frequented areas of the observation space have a higher resolution than low frequented areas. This discretization method benefits the stochastic design of the HMM model, because the resulting transition and observation probabilities will be balanced due to the balanced occurrences of each discrete observation.

Discretization Type B: Equidistant Bins

The original observation space \mathbb{R} is discretized into bins with equal ranges (like eg. 0-100 Watts, 100-200 Watts, 200-300 Watts, etc.). The representative value of each bin is then the middle point of the bin's intervals. This is the standard discretization method and often desirable from an applied perspective.

3.2.2. HMM Training

The application of HMM to model a one-dimensional signal like electricity demand is designed so that the signal is represented by the observation process $(X_t)_{t \in \{1, \dots, T\}}$ while the hidden process $(Z_t)_{t \in \{1, \dots, T\}}$ is unknown. In our case, we can measure the total amount of electricity demand of a household but do not know the underlying process. This process is influenced by the inhabitants, the electric devices and other factors like weather or season. The advantage of HMM is that this hidden process does not need to be specified. During the training phase, the optimal parameters of the model are generated without deeper understanding of the process. Via decoding algorithms, the most probable hidden state sequences can be identified and interpreted in a post-training analysis. For example, low energy emitting states can be connected to night time hours or times when nobody is at home. The only condition for the hidden state process is to satisfy the Markov conditions which are quite strong.

During the training phase, the parameters λ of the HMM are trained with the Baum-Welch algorithm given a specific hyperparameter setting⁵ and the historic observation sequence $\dot{\mathbf{X}}_T$. The Baum-Welch algorithm is run using the fixed settings of random initial parameters and a maximum number of iterations set to 100 which has been shown to be sufficient. Since the training algorithm relies strongly on the empirical occurrences of the observations, in the selection of the training data, data-specific bias should be avoided. In the case of household's electric power load, it is known that there are strong daytime

⁴For example, when the number of discrete observations is fixed to N , the n 'th bin has as left interval limit the $(n-1)/N$ 'th empiric quantile and as the right limit the n/N 'th empiric quantile.

⁵The hyperparameters of the proposed HMM forecasting method are the type of discretization, the size of the observation space M , the number of hidden states N and the historic window length T . We will explain and conduct a detailed hyperparameter analysis later on.

and seasonal differences. Therefore, only data including full days and years should be included.

3.2.3. HMM Forecasting Algorithm

Given a trained HMM with parameters λ , we apply the probabilistic forecasting algorithm and calculated the desired distribution θ as explained in 2.1.3. The two relevant parameters for forecasting are the historic window length T and the future horizon H . For the first parameter T we will use a fixed length while tuning the other hyperparameters and conduct a sensitivity analysis later to verify the choice. The future horizon H is set to 1, thus, it will be a one-step-ahead prediction 15 minutes into the future. Even though it is the most simple forecast, we can derive a lot of information about the quality of the forecasting model because predictions further ahead build upon the same information as the one-step-ahead prediction.

3.2.4. Postprocessing

The resulting prediction from the HMM forecast is the discrete distribution vector θ over the discretized observation space O , however, the original problem requested a distribution forecast in the continuous space \mathbb{R} . For this reason, the discrete distribution is re-converted to a continuous CDF with the following method. Recalling that the distribution vector θ represents the probabilities of the corresponding bins, we presume that values inside each bin have equal densities. Assuming the bins are sorted ascending, the probability of X_{T+H} being smaller than the right interval limit of bin k is the sum of the probabilities of all bins l lower and equal to bin k ,

$$\mathcal{P}(X_{T+H} \leq \text{right interval limit of bin } k) = \sum_{l=1}^k \theta_l.$$

This leads to a piecewise-linear CDF, where the above defined supporting points are connected by straight lines as shown exemplary in graph 4.

3.2.5. Hyperparameter Tuning

The general goal of the hyperparameter tuning is to find the best HMM forecasting models for one-step-ahead predictions. For this reason, not only the optimal parameters are trained, but also the right choice of hyperparameter has to be studied. There are four different hyperparameters that result from the methodology: the discretization method, the number of hidden states N , the number of observations M , and the historic window length for forecasting T . For each discretization method, the optimal settings of N and M are elaborated with a hyperparameter analysis, while the correct choice of T is verified afterwards with a sensitivity analysis. All experiments are conducted for each household and discretization type independently, resulting in two final HMM forecasting models per household.

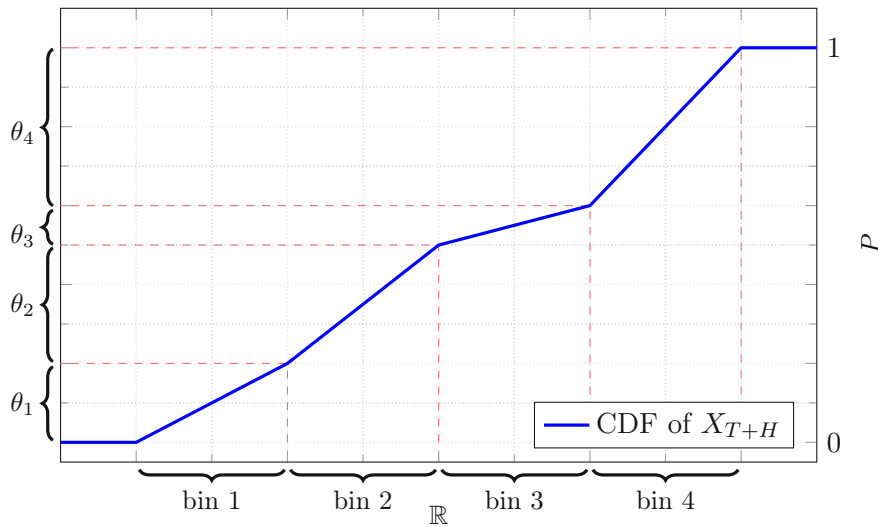


Figure 4.: Continuization of the distribution vector θ with four exemplary bins.

Optimization of N and M

For the hyperparameter tuning, we perform a grid search to find the optimal settings of the number of hidden states N and observations M for each household and discretization type A and B . For each hyperparameter setting, the optimal parameters of the HMM model are trained via the Baum-Welch algorithm on the test data set of the household. The models are then evaluated using the validation data set and computing the mean-CRPS of the one-step-ahead prediction for each validation time instance. Hereby, we fix the historic window length T to be 100, which is a length of slightly more than one day considering the data time resolution of 15 minutes. The tested values for the grid search of N and M additional to the other fixed hyperparameters are summarized in table 3. All combinations of these hyperparameter settings are evaluated resulting in the training and testing of 36 models per household and discretization type.

| Hyperparameter | Values |
|----------------|-------------------------------------|
| N | 10, 20, 30, 40, 50, 60, 70, 80, 100 |
| M | 25, 50, 100, 200 |
| T | 100 |

Table 3.: Overview of hyperparameter settings used in the hyperparameter analysis.

Sensitivity Analysis for T

In the sensitivity analysis, we investigate the effect of change in the historic window length T on the models forecasting performance. Hereby, we aim to verify the choice of T of

the hyperparameter analysis, but also to get insights into the HMM forecasting behavior. Since the historic window length T is only relevant during the prediction step and not during training, the models do not have to be retrained. Building onto the previous results, we use the best performing models for each household and discretization type, which we will refer to as *optimal models* for the rest of the paper. With a variable size of T for the sliding window, the optimal models are evaluated once again by computing the one-step-ahead prediction for each time instance of the validation data set and calculating the mean-CRPS. The tested values are described in table 4 and the results are analyzed visually.

| Hyperparameter | Values |
|----------------|----------------------------------|
| N | fixed to optimal |
| M | fixed to optimal |
| T | 1, 2, 3, 5, 10, 25, 50, 100, 200 |

Table 4.: Overview of hyperparameter settings used in the sensitivity analysis.

3.3. Model Analysis

In order to gain better understanding of the model’s characteristics, we investigate the predictive performance and power. First, the general performance of each optimal model is evaluated, before we study the predicted distributions in detail. For this purpose, the calibration of the models are examined, and then the behavior of the model for multi-step-ahead predictions is studied. All experiments of the model analysis are conducted separately on the test data set of the households and for each optimal model.

3.3.1. Evaluation of Prediction Accuracy

Similarly to hyperparameter tuning, the models are evaluated using the discrete mean-CRPS. Thus, for each optimal model, the one-step-ahead distribution is predicted for each time instance of the corresponding household’s test data set and evaluated. The result will be two evaluated models for each household, on which we can deduct a good comparison of the discretization types.

3.3.2. Calibration via PIT Histograms

While the mean-CRPS already considers calibration, it is also influenced by the sharpness of the predicted distribution. To gain more insight into the predictive power of the models, we generate the PIT histograms for the optimal models, as presented in section 2.2.2. Thus, for each time instance of the test data set, the PIT value is calculated between the predicted distribution and the true realization, and then the histogram of these PIT values is generated. We structure the PIT histogram into 10 sections, displaying the frequencies for each 10% interval (0-10%, 10-20%, etc.). This provides a visual judgment

of the model’s calibration, which is sufficient for our purpose. In literature, there exist some additional numeric and theoretic in-depth analytic methods like the reliability index [15] or calibration vs sharpness decompositions of the CRPS [1].

3.3.3. Convergence for Increasing Forecasting Horizon

The last study of our model investigates the model behavior for a forecasting horizon larger than one, as applied until this point in the thesis. For a first understanding of the models behavior, we compute the mean-CRPS with increasing forecasting horizon for the optimal models and display the results visually. We are not focusing on the numerical performance of the models, however, we want to know how far into the future our HMM models can be reasonably used. For this reason, we investigate the convergence of the predicted distribution θ with increasing forecasting horizon H . Convergence means that the predicted distribution does not change after a certain point in the future. To illustrate the idea of convergence, figure 5 shows a typical situation of a multi-step-ahead distributional forecast. Generally speaking, the convergence of a distributional forecast can be meaningful in certain applications. However, knowing that there are daytime patterns within the power load data, the convergence of the forecast only shows the limitation of the HMM forecasting method which cannot model long-term patterns.

Formulating the idea of convergence in mathematical terms, the stochastic vector θ lies on a hyperplane in the N -dimensional real space. We denote θ_h as the predicted distribution h steps into the future, for $h \in \{1, \dots, H\}$. The time step h' is defined as the *convergence point* if for all h greater than equal to h' , θ_h lies inside the epsilon-neighborhood around the empiric distribution of the observation process,

$$\|\theta_h - (\delta \cdot \mathbb{B})\|_1 < \epsilon \quad \text{for all } h \geq h'.$$

Here, δ denotes the stationary distribution of \mathbb{A} , and thus, $(\delta \cdot \mathbb{B})$ is the empiric distribution. The L^1 metric is chosen, so that the choice of epsilon defines the absolute difference between the stochastic vectors. For example with $\epsilon = 0.01$, convergence is achieved when the predicted distribution vectors differ only up to 1% in total. We experiment with two different epsilon settings, with $\epsilon = 0.1$ for a rough convergence of the distributional shape and with $\epsilon = 0.01$ for a detailed convergence.

3.4. Benchmarks

In the last experiments, we compare the proposed HMM forecasting model to existing state-of-the-art probabilistic load forecasting methods. The field of load forecasting is diverse, and the consensus of the literature is that there does not exist one ultimate best model, but rather depends on the specific application and target. A good overview of the different load forecasting methods is given by Haben, Voß, and Holderbaum [19]. Key differences in load forecasting models are namely the forecasting horizon, the voltage level and the application. Our target here is to benchmark with models as similar

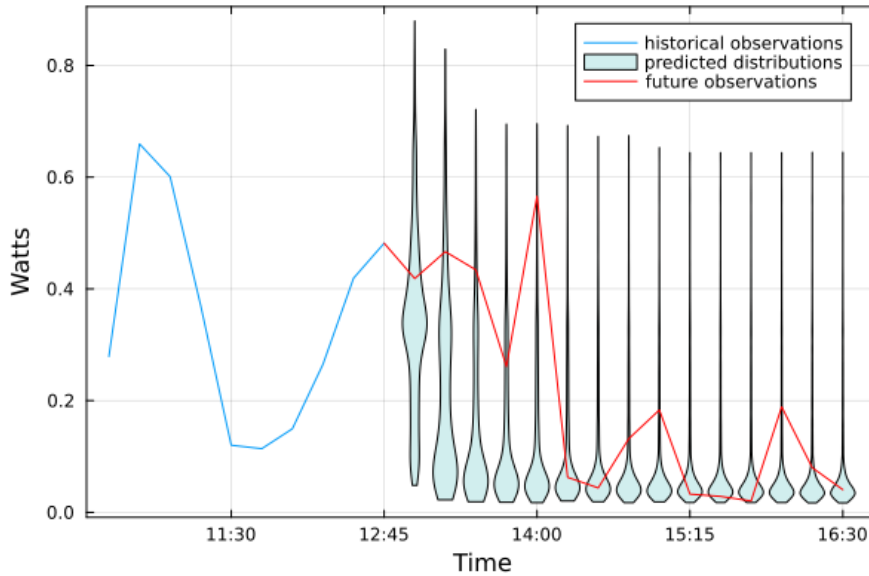


Figure 5.: Example of a typical situation of a multi-step-ahead distribution forecast. Given the historic observation, a distribution forecast is computed for each time instance in the future illustrated by the multiple violin plots. Additionally, the true realization of the future time instances were added to the plot afterwards. Examining the shape of the single distributions, the convergence is visible. From time instance 14:00 onward, the violins do not change significantly and from 15:15 onward, they are visually not differentiable.

applicable as the presented HMM forecasting method. Namely, the benchmark models have to be probabilistic, short-term, low-voltage load forecasting model and relatively easy implementable, meaning that both the theoretic and computational complexity of the model is limited.

In this section, we clearly state our benchmark framework and then present the benchmark models including two heuristic baselines, one statistic method and one advanced machine learning model.

3.4.1. Benchmark Setup

The setup for the benchmarks is similar to the experiments conducted in the previous experiments. The general goal is to predict the distribution⁶ for the electric power load one-step-ahead. The available information is restricted to historical load data and temporal information, which can be utilized for each benchmark model differently. The historical load data is generally represented by

$$(O_t)_{t=1,\dots,T}$$

⁶Or a distribution approximation with quantiles.

for a model specific historic window length $T \in \mathbb{N}$. The available temporal information are daytime or month, which is commonly used to model daytime patterns and seasonality. For this reason, we introduce the two functions $time(\cdot)$ and $month(\cdot)$, which define a mapping from the observation's daytime and month of the year onto the interval $[0, 2\pi]$ respectively. This is a common machine learning technique to enable cyclical encoding via $\sin(\cdot)$ and $\cos(\cdot)$, see Bishop [4].

The benchmarks are conducted on the households electrical load data discussed in section 3.1. Thus, the models are trained on the trainings data set and evaluated over the test data set for each household separately. The validation data set can be utilized for model specific validation tasks. The evaluation is obtained in the same way as in section 3.3.1. For each instance of the test data set, the 99 quantiles for 0.01, 0.02, ..., 0.99 are calculated for the one-step-ahead distribution forecast and evaluated with the discrete CRPS. For each model and household, one mean-CRPS value is computed separately. Furthermore, we then compare the results of the benchmark models with the optimal HMM models by their relative improvement to the baseline.

3.4.2. Benchmark Models

In this thesis four benchmark models for probabilistic short-term load forecasting are investigated. Similar models were studied in Botman, Lago, Becker, Vanthournout, and Moor [5], while the reader is referred to Hong and Fan [21] for a detailed discussion of different probabilistic load forecasting methods.

Persistence Model

The most forward baseline method in point forecasting is the so-called *naive* or *persistence* model. The idea is that the predicted value of the future is equal to the value of the presence. To apply this concept as a probabilistic forecast, we use the deterministic naive forecast as a basis and add to this forecast a distribution. This distribution is constructed by applying the naive point forecast on the training set, taking all the error terms for each quarter hour of the day separately and computing the empiric error distributions. Thus, one of the 96 empiric error distribution $error_{qh}$ for the specific quarter hour qh is computed based on the set $\{o_t - o_{t-1} : time(o_t) = qh\}$. The resulting predicted distribution of the persistence model for observation o_t consists of the previous observation plus the error distribution of the corresponding quarter hour of the day, to wit $o_{t-1} + error_{time(o_t)}$.

Historical Sampling

The second baseline method is also based on the empiric distribution of each quarter hour of the day. In historical sampling, all values with the same quarter hour of each day are grouped together and the empirical distribution for the quarter hour is calculated. Thus, the empirical distribution for the specific quarter hour qh is calculated based on the set $\{o_t : time(o_t) = qh\}$. In the forecasting step, the empiric distribution of the corresponding quarter hour of the forecasting time instance is applied as the predicted distribution.

In Botman, Lago, Becker, Vanthournout, and Moor [5], this simple method has shown effective results in forecasting load data up to 24 hours ahead. Notably, this method only utilizes temporal information for forecasting and does not take recent historical load data into account.

Linear Quantile Regression

Linear quantile regression (LinQR) is a similar estimation method to classic linear regression. The target is to find a linear relation between the predictor variable and the response variable. Consider the general linear regression problem,

$$\hat{\beta} = \arg \min_{\beta} \sum_{i \in I} c(y_i, x_i \cdot \beta).$$

Here, the aim is to find the optimal parameter column vector $\hat{\beta}$, so that the sum of a general cost function $c(\cdot)$ between the observation y_i and the linear transformation of the explanatory row vector x_i is minimized over a general index set I . In standard linear regression, the cost function $c(\cdot)$ is the mean squared error, which leads to a closed algebraic solution. However in LinQR, the pinball loss for a fixed quantile is taken as a cost function as defined in section 2.2.3. This results in a linear programming problem, which has to be solved algorithmically. For each quantile, a regression model has to be trained individually. However, due to the independent training of different quantiles, *quantile crossing* can happen, meaning that it possible that quantile predictions for a smaller quantile are greater than bigger quantiles. For this reason of multiple quantile regressions, the predicted quantiles are sorted in ascending order in post-processing.

For the benchmark model, we trained a LinQR model for each quantile in for 0.01, 0.02, ..., 0.99 using the Python package scikit-learn and the HiGHS solver [25]. For the predictor variables, we apply a similar feature selection as in Botman, Lago, Becker, Vanthournout, and Moor [5]. The explanatory vector consists of historical data up to 10 time steps into the past, 2 shifted values of the exact daytime of one and two weeks prior to the forecasting time instance, and temporal data of the daytime and month of the year encoded with sine and cosine to retain the cyclic characteristic. Thus, if the load observation o_t is to be predicted, the explanatory vector is defined by

$$x_t = [o_{t-1}, o_{t-2}, \dots, o_{t-10}, o_{t-96.7}, o_{t-96.14}, \sin(\text{time}(o_t)), \cos(\text{time}(o_t)), \sin(\text{month}(o_t)), \cos(\text{month}(o_t))].$$

Long short-term memory network

Long short-term memory networks (LSTM) are considered one of the most competitive neural network designs and achieve great results in various fields, also for energy forecasting. Here, we rely on an adapted model of Wang, Gan, Sun, Zhang, Lu, and Kang [42] which was further developed by Xu, Hu, and Fan [43].

The simplified model presented here consists of two phases, visualized in figure 6. In the

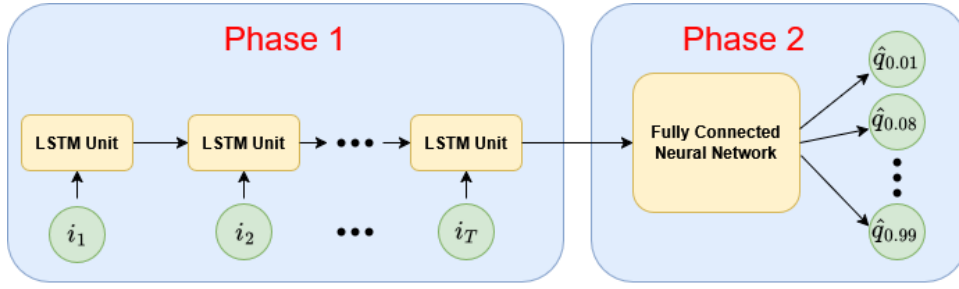


Figure 6.: Overall structure of the pinball loss guided LSTM.

LSTM phase, which iterates over the input time series $\{i_t\}_{t=1,\dots,T}$ to recognize relevant short- and long-term information. This information is then forwarded to a multilayered fully-connected neural network, which utilizes the output of the LSTM phase and translates it to the desired quantile forecast vector.

For backpropagation, the pinball loss as defined in section 2.2.3 is used as the error metric. To limit the computational effort, we design the model to predict 15 quantiles reaching from 0.01% to 0.99% in 0.07% gaps and interpolate the missing quantiles linearly. For each prediction, the LSTM model receives a time series of input vectors reaching back 2 days implying the historic window length $T = 192$. The single input vectors consist of the historic load data and the temporal data equal to the one used in the LinQR, thus for $n \in \{1, \dots, T\}$

$$i_n = [o_{t-T+n-1}, \sin(\text{time}(o_{t-T+n-1})), \cos(\text{time}(o_{t-T+n-1})), \sin(\text{month}(o_{t-T+n-1})), \sin(\text{month}(o_{t-T+n-1}))]$$

if the distribution of observation o_t is to be predicted. We implemented the model in Julia using the Flux package with the ADAMS optimizer set to the learning rate of 0.001 [28] and applied the hyperparameter setting described in table 5. During the training phase, the maximum number of epochs is set to 20 with a batch size of 1024.

| Hyperparameter | Value |
|--------------------|-------|
| LSTM unit size | 64 |
| Neural unit 1 size | 128 |
| Neural unit 2 size | 64 |

Table 5.: Hyperparameter setting for the LSTM model.

4. Results of Hyperparameter Tuning

4.1. Optimization of N and M

For each household and both discretization types, 36 HMM models with different hyperparameter settings were trained and evaluated. Figures 7-11 visualize the results of the hyperparameter tuning for each household. In each figure, a plot for each discretization type shows the mean-CRPS value in dependence of different settings of hidden state numbers, while models with equal number of observation are visualized by equal color. For each plot, the hyperparameter setting inducing the lowest mean-CRPS value is chosen for the optimal model. In appendix A, the reader finds a complete table with the exact numerical results of the hyperparameter tuning.

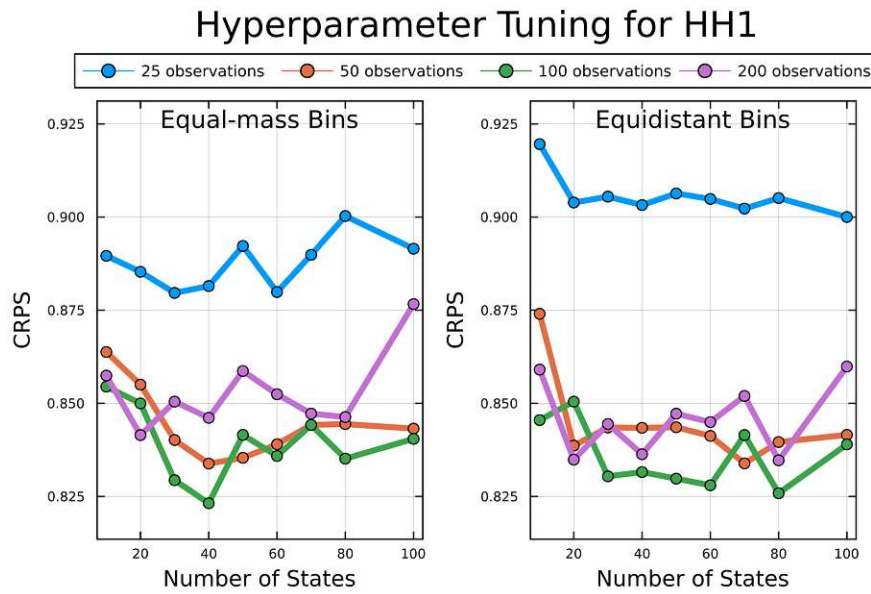


Figure 7.: The optimal models are chosen by the minimal mean-CRPS value, which is for discretization type A the model with 100 observations and 40 hidden states and for discretization type B the model with 100 observations and 80 hidden states.

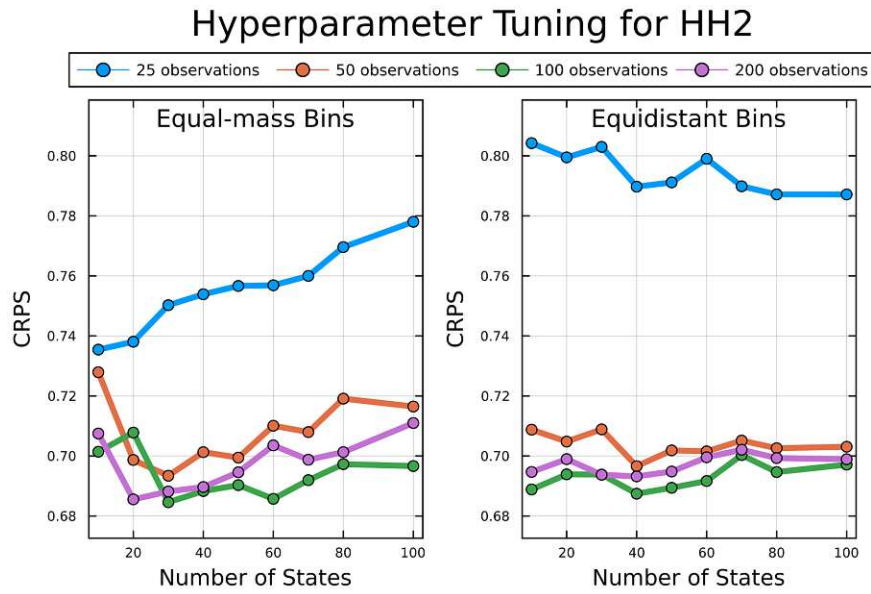


Figure 8.: For discretization type A, two hidden states settings ($N = 30$ and $N = 60$) achieve similar optimal results for models with 100 observations. The development of the mean-CRPS for increasing number of states have a similar trajectory between models with equal number of observation, while the setting with 100 observation achieves the optimal result at 40 hidden states.

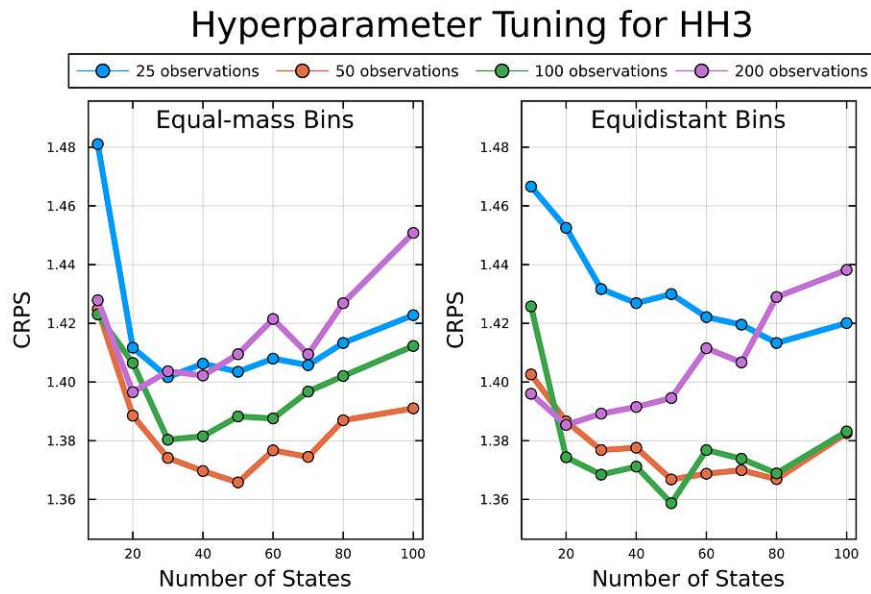


Figure 9.: HH3 has the highest mean-CRPS overall. Also it is the only household, where a model with 50 observation outperforms the rest, in the case of equal-mass bins with $N = 50$. For the other discretization type, the optimal model is set by $N = 50$ and $M = 100$

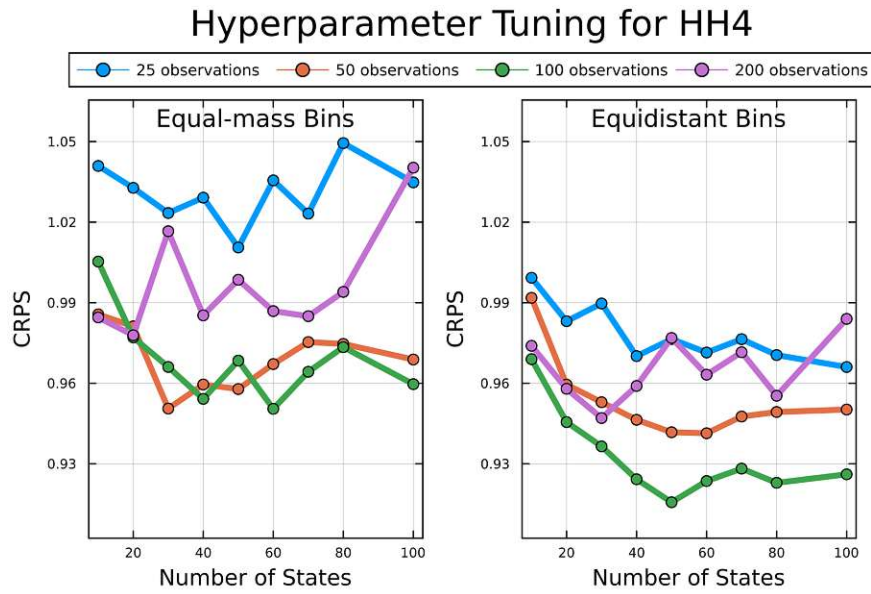


Figure 10.: For discretization type A, two different models achieve a similar optimal mean-CRPS value, with the model for $N = 60$ and $M = 100$ is surpasses by a small margin. For discretization type B, models with 100 observation are consistently better and optimal at $N = 50$.

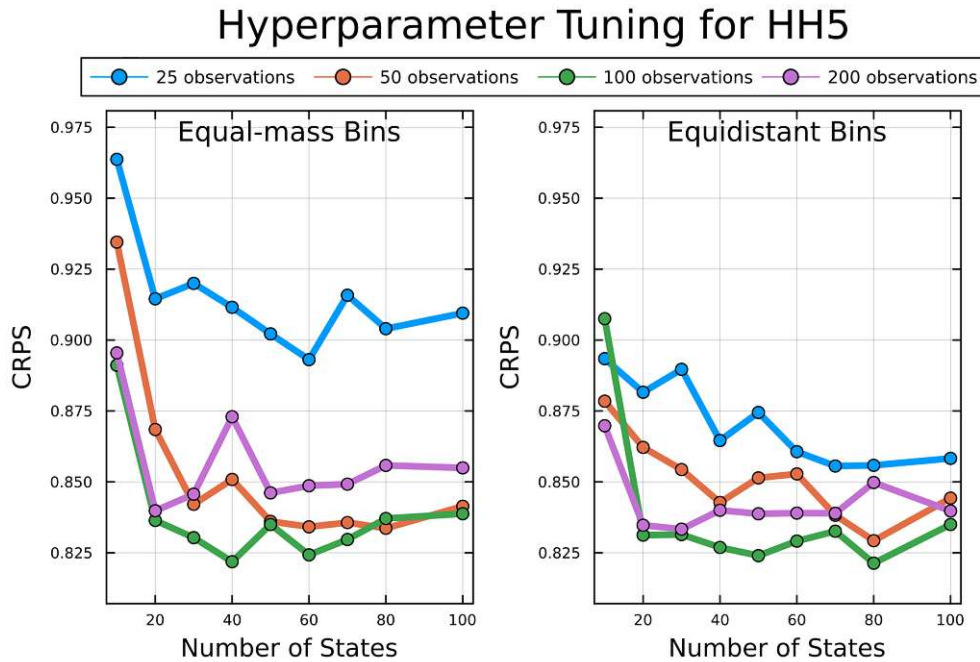


Figure 11.: HH5 shows similarities to HH1 in the aspect of optimal hyperparameter setting and mean-CRPS value for both discretization types ($N=40$, $M=100$ for type A; $N=80$, $M=100$ for type B).

A summary of the numeric results is given in table 6, which highlights the hyperparameter settings of the optimal models with their corresponding mean-CRPS values. From the hyperparameter analysis figures, the following conclusions are drawn.

| Household | N | M | CRPS | Household | N | M | CRPS |
|-----------|----|-----|--------|-----------|----|-----|--------|
| HH1 | 40 | 100 | 0.8232 | HH1 | 80 | 100 | 0.8259 |
| HH2 | 30 | 100 | 0.6846 | HH2 | 40 | 100 | 0.6874 |
| HH3 | 50 | 50 | 1.3658 | HH3 | 50 | 100 | 1.3587 |
| HH4 | 60 | 100 | 0.9505 | HH4 | 50 | 100 | 0.9157 |
| HH5 | 40 | 100 | 0.8219 | HH5 | 80 | 100 | 0.8213 |

(a) Discretization Type A (b) Discretization Type B

Table 6.: The hyperparameter setting of the optimal models for each household.

First, the number of observations have a significant impact on the models performance. The choice of discretizing the observation space in only 25 observations leads in almost every setting to the worst results. But also, a high resolution of 200 observations influences the performance of the models badly. The number of observations of 50 and 100 achieves the most consistent results with 100 observations being the best setup for all households, except for HH3 with discretization type A.

Further, the optimal number of hidden states differs for each household. However, a main takeaway of this analysis is that models of discretization type A reach their optimal setting with a lower number of hidden states than models of discretization type B, which leads to more efficient models. This effect can be explained by the design of discretization. The discretization of equal-mass bins favors the stochastic characteristic of the HMM models because it leads to a balanced weighting of all observations and thus the underlying hidden states.

4.2. Sensitivity Analysis for T

The sensitivity analysis investigates the convergence of the prediction performance conditioned by increasing the historic window length T . The experiments were conducted for each household and discretization type with the optimal models hyperparameter setting summarized in 6. As the visualized results do not differ significantly between the households, only one representative figure 12 for HH1 is presented here, while the corresponding figures for HH2-HH5 are placed in the appendix B.

The results of the sensitivity analysis support the hypothesis that with increasing the historic window length and thus increasing available information, the performance of the forecasts improves. All households show a similar pattern of converging mean-CRPS to a lower limit. This limit is reached in most cases at $T = 10$, for HH2 already at $T = 5$ and only in the case of HH1 with discretization in equal-mass bins at $T = 25$. There are no

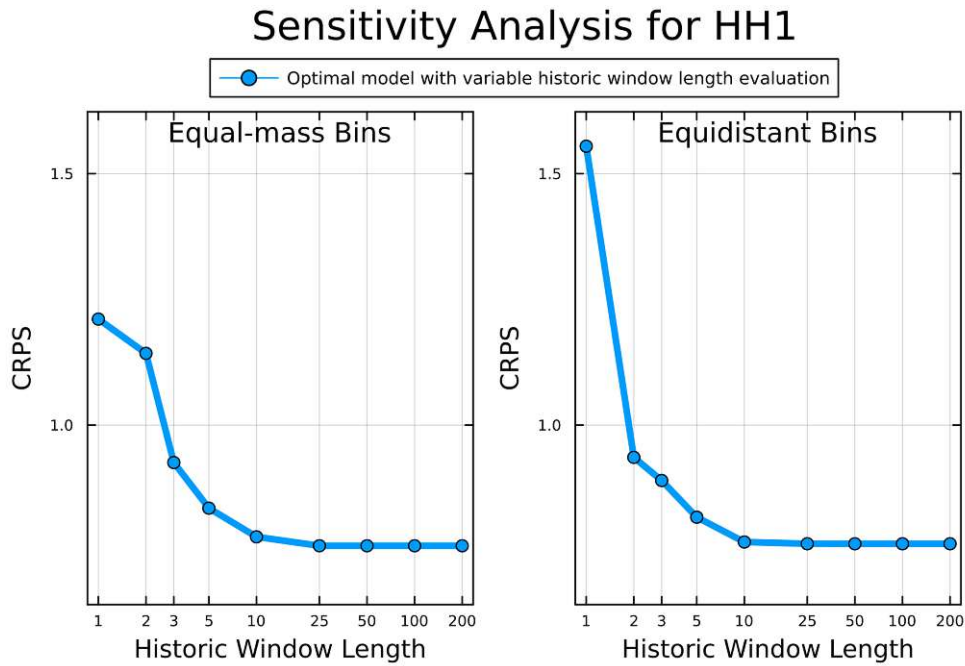


Figure 12.: Representative figure of the result of the sensitivity analysis for HH1.

significant differences between the discretization types, even though they differ greatly in the number of hidden state, like eg. in HH1 and HH5.

In addition, it verifies that the choice of T during the hyperparameter analysis was well taken. For further experiments in the future, the historic window length can be chosen even lower to improve computational evaluation times.

Lastly, considering the converging point to be around 10, it can be said that only historic data up to two and a half hours ago adds beneficial information to the HMM forecasting model. That also implies that the model does not detect recognizable behavior and patterns before this point, inferring that HMM forecasting models should only be applied for short-time pattern forecasting.

5. Results Model Analysis

This chapter presents the results of the model analysis and discusses their implications to better understand the model’s characteristics. For this, the optimal models resulting from the hyperparameter analysis in section 4.1 are tested on the test data set of their corresponding household. First, we evaluate the optimal model’s performance, then we investigate the calibration of the predicted distribution, and in the end, analyze the models behavior for multi-step-ahead predictions.

5.1. Evaluation of Prediction Accuracy

| Household | N | M | CRPS | Household | N | M | CRPS |
|-----------|----|-----|-------|-----------|----|-----|-------|
| HH1 | 40 | 100 | 0.802 | HH1 | 80 | 100 | 0.812 |
| HH2 | 30 | 100 | 0.720 | HH2 | 40 | 100 | 0.723 |
| HH3 | 50 | 50 | 1.363 | HH3 | 50 | 100 | 1.357 |
| HH4 | 60 | 100 | 0.956 | HH4 | 50 | 100 | 0.901 |
| HH5 | 40 | 100 | 0.819 | HH5 | 80 | 100 | 0.831 |

(a) Discretization Type A (b) Discretization Type B

Table 7.: The hyperparameter setting of the optimal models for each household

Table 7 shows the evaluation results of the optimal models together with their hyperparameter setting. The historical window length T for the sliding window of the forecasting method was fixed to be 30, which is sufficient as shown in section 4.2. Comparing the performance of the discretization methods, we observe for all households except HH4 that both discretization methods achieve similar mean-CRPS values not varying greater than 2%. Thus, no discretization method can be superior in evaluation of the mean-CRPS. HH3 stands out with a much greater mean-CRPS than the other households, which can be caused by either a greater volatility of the electricity demand or a conceptual drift in energy consumption for the test data compared to the training data. Also interestingly, in comparing the models for each household separately, we observe that the model with less hidden states performs better than the second model of the household.

5.2. Calibration via PIT Histograms

The PIT histograms are shown in figures 13 to 17. The underlying pit values are calculated with the same settings and distribution forecasts as in the previous section.

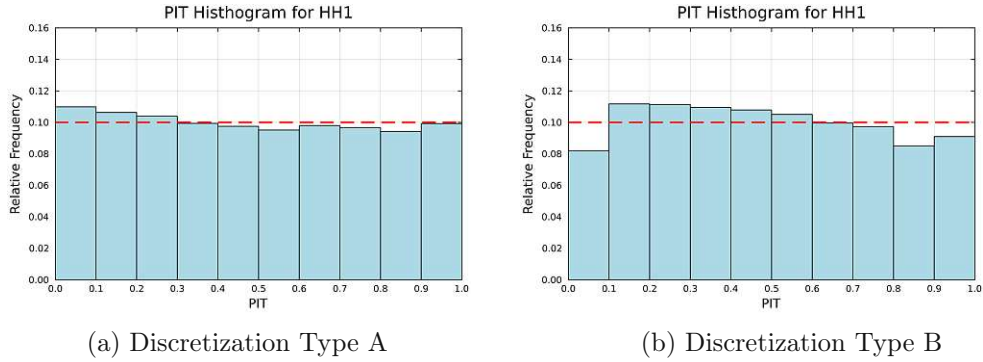


Figure 13.

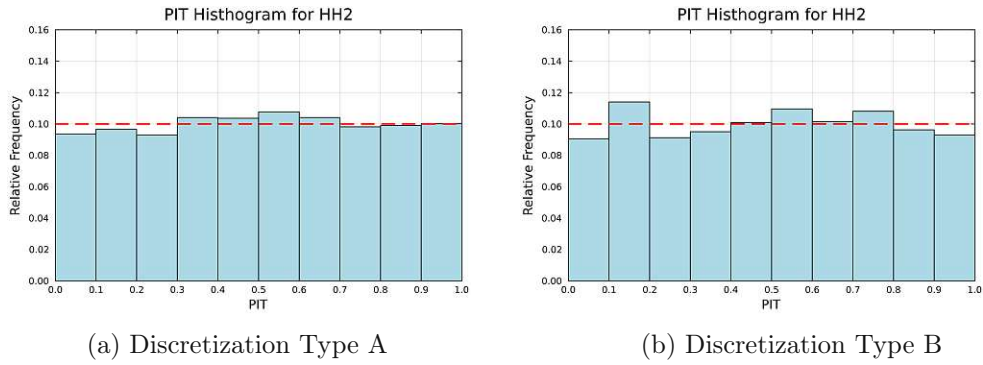


Figure 14.

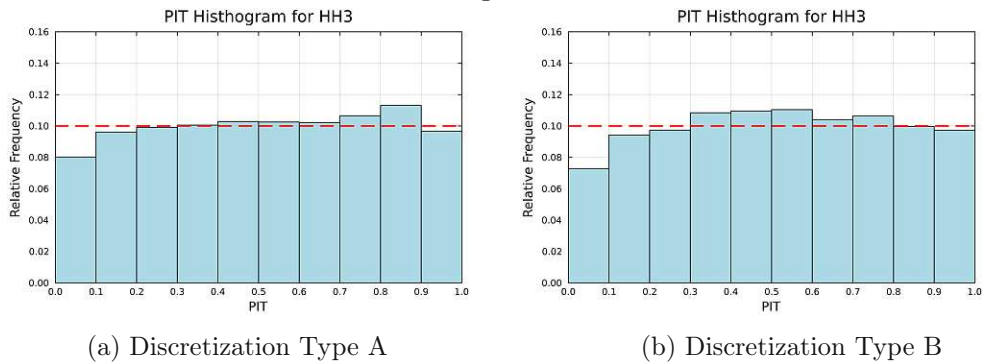


Figure 15.

From these PIT histograms, it is visible that models with discretization type *B* have a worse calibration than models with discretization type *A* across all households. This effect can be seen especially strong for HH4 and HH5. While the HH4 model with equal-mass

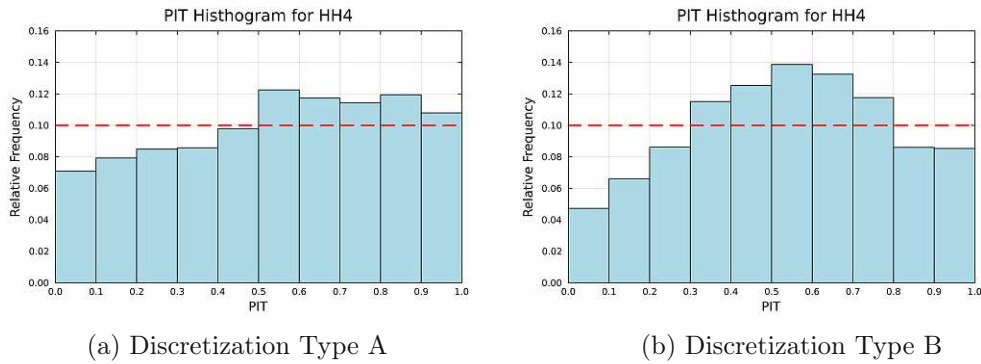


Figure 16.

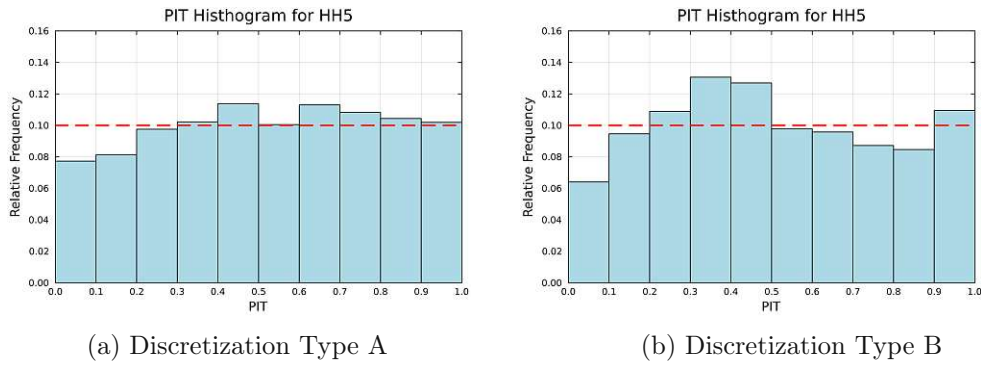


Figure 17.

bins shows a slight skewness towards higher quantiles, the forecast of the HH4 model with equidistant bins is strongly overdispersed. This implies that the reliability of the prediction increases if the discretization with equal-mass bins is carried out. Considering the similar values of the mean-CRPS in the households for different discretization methods, we can assume that discretization type *B* models compensate the lack of calibration with a greater sharpness.

In conclusion, we can state that the choice of discretization has an impact on the calibration of the models forecasts. Following the main principle of probabilistic forecasting (see section 2.2.2), models with discretization into equal-mass bins should be favored.

5.3. Convergence for Increasing Forecasting Horizon

Here, the optimal models are evaluated with increasing forecasting horizons instead of the one-step-ahead predictions previously used. Figure 18 shows the mean-CRPS values of all optimal models for increasing forecasting horizon. It is visible that with increasing forecasting horizon the predictive performance decreases, as one can expect. Interestingly, increasing the forecasting horizon from one to two steps into the future, all models succumb a great loss of accuracy observable by the big jumps in the mean-CRPS. After a

forecasting horizon greater than 2 the model performances slowly converges moderately against a stable mean-CRPS value (only exception being HH5). Comparing the models with different discretization types for each household individually, no significant differences exist. The results are valuable in giving first information about the models performance for future research since they suggest a usage of the proposed HMM forecasting method only for very short-term forecasting tasks.

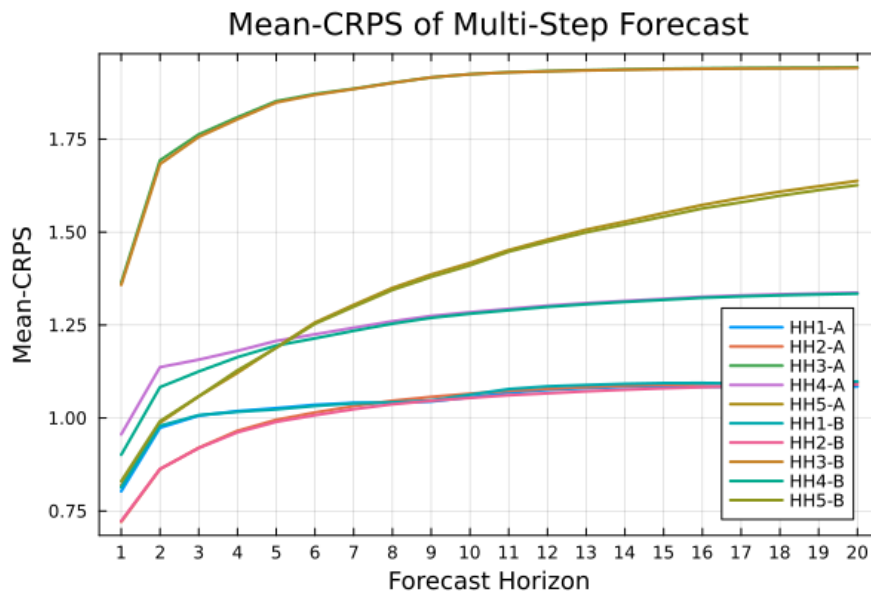


Figure 18.: Trajectories of the mean-CRPS value for increasing forecasting horizons of the optimal models.

The next graphics display the histograms for all convergence points evaluated over every validation data instance and each optimal model. The figures 19-23 show the convergence points histograms for a tolerance of $\epsilon = 0.1$ and figures 24-28 the convergence points histogram of $\epsilon = 0.01$. Trivially, the convergence to the 10% epsilon-neighborhood is reached earlier than in the 1% epsilon-neighborhood. Referring to the rougher convergence, a lot of forecasts converge around 5 steps into the future, most forecasts some when before 10 steps, and only a few forecasts do change significantly their shape later than 10 (like eg. in HH1.). With a finer convergence of $\epsilon = 0.01$, these boundaries shift to the right. In this case, there are also great differences between the households. Especially HH1 and HH5 require more steps to reach their stationary distributions. These converging points serve as an absolute limit for T for these HMM forecasting models to be applied in a reasonable way, since they converge to the empirical distribution and do not utilize beneficial information anymore.

5. Results Model Analysis

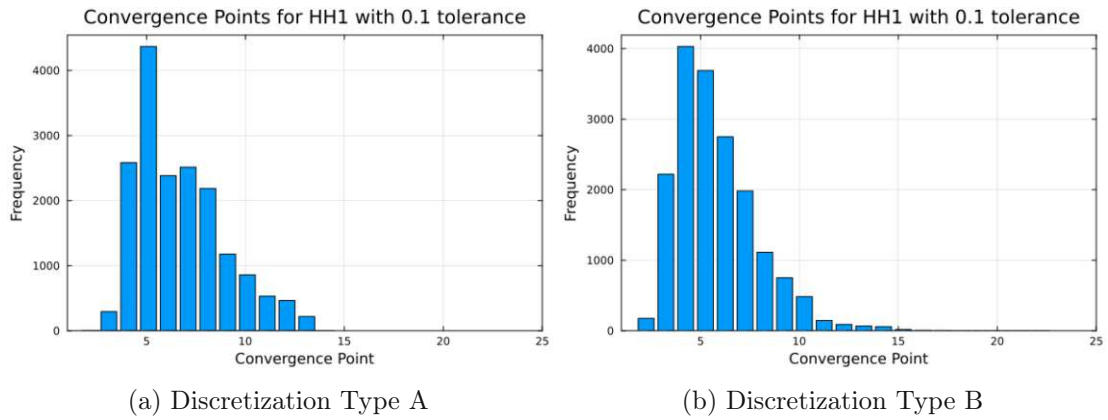


Figure 19.

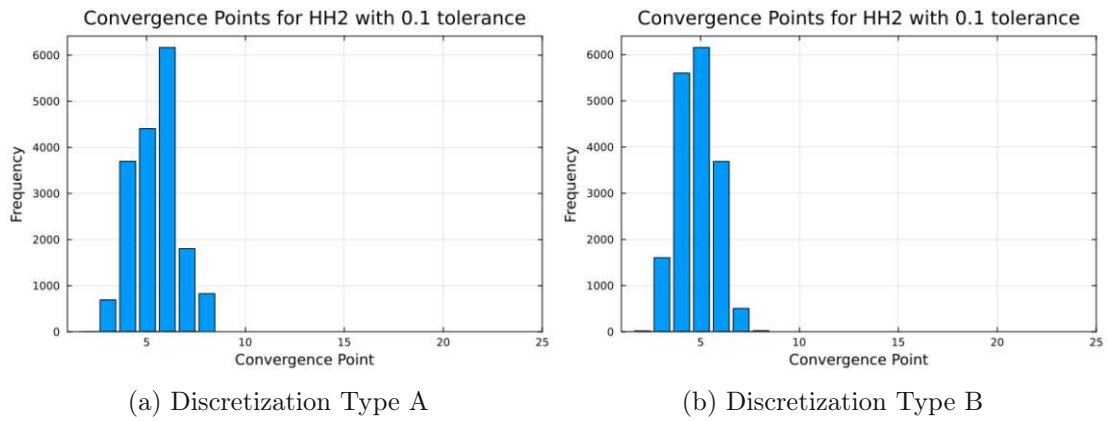


Figure 20.

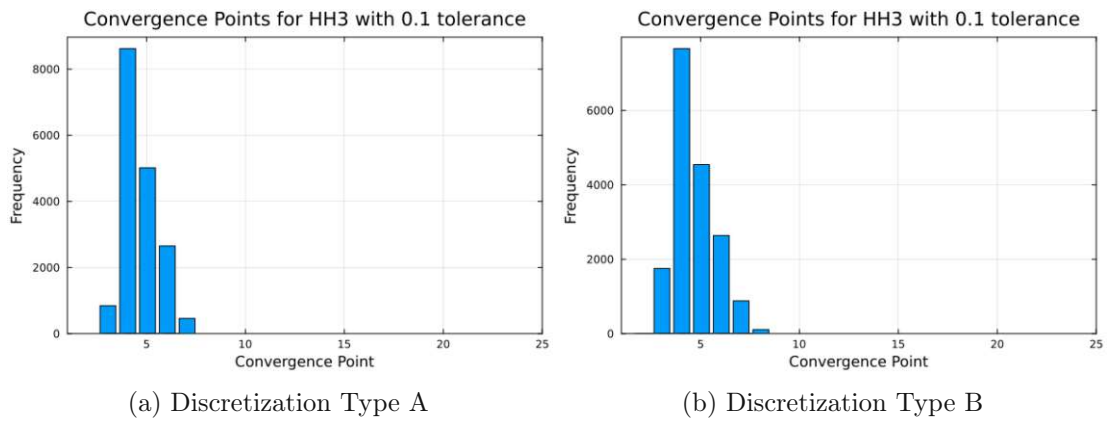


Figure 21.

5. Results Model Analysis

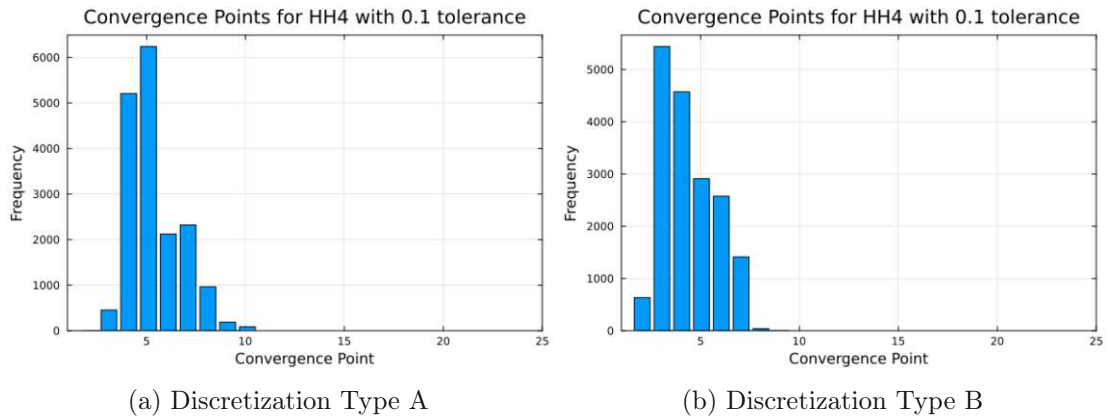


Figure 22.

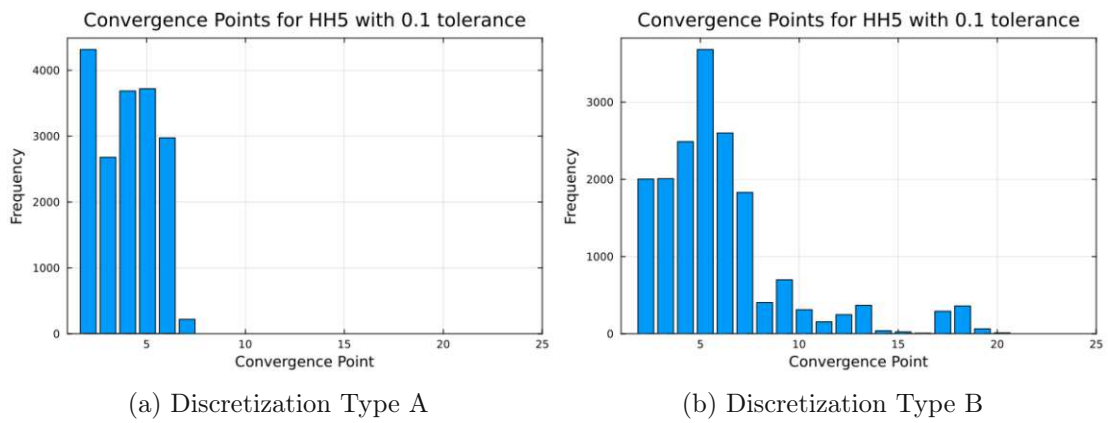


Figure 23.

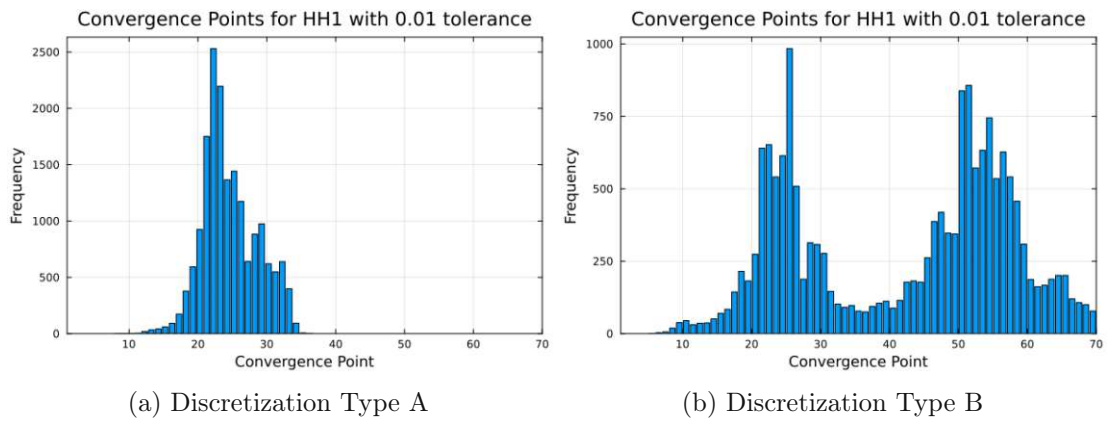


Figure 24.

5. Results Model Analysis

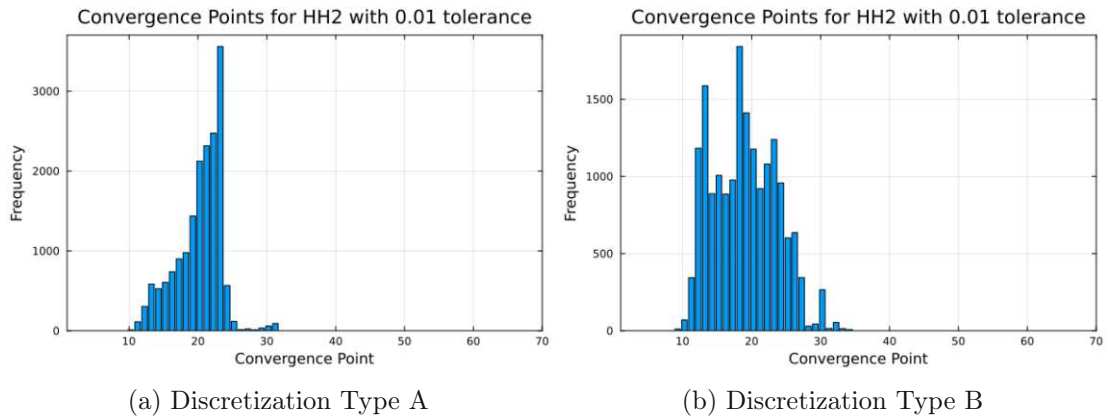


Figure 25.

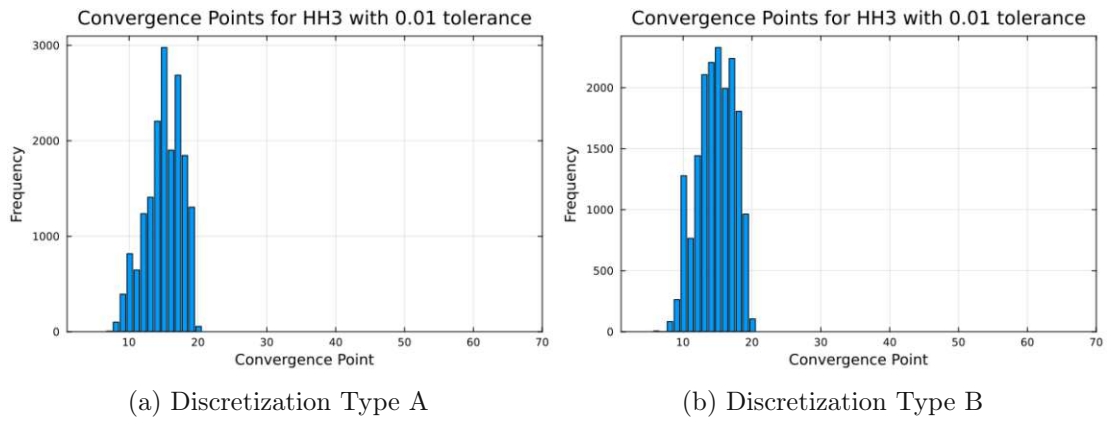


Figure 26.

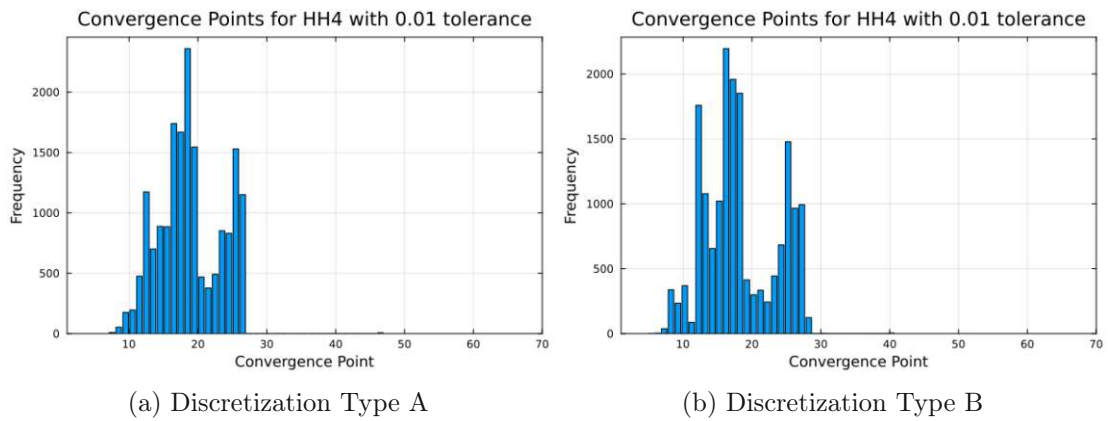


Figure 27.

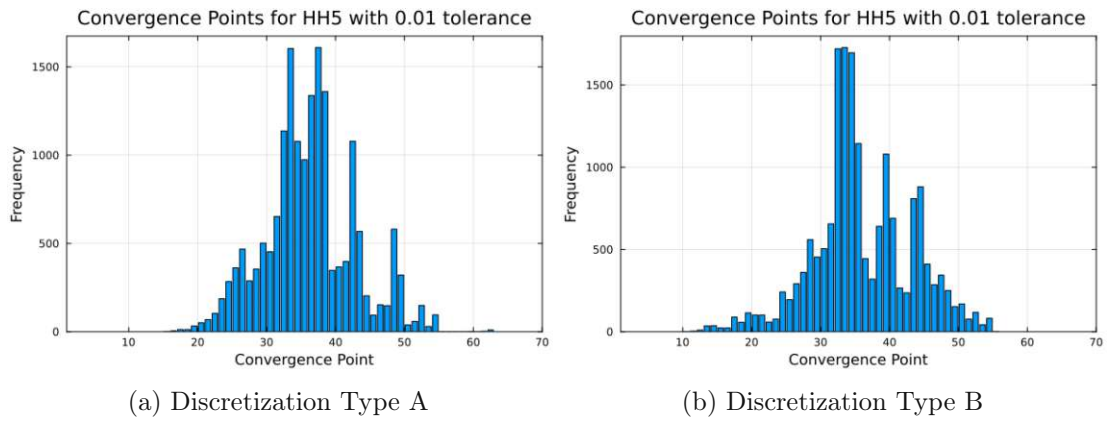


Figure 28.

6. Results of Benchmarks

This chapter addresses the final open question: How does the HMM model perform compared to other state-of-the-art probabilistic short-term power load forecasting methods for households?

The results of the benchmarks are visualized in the figures 29-33. The models HMM-A and HMM-B correspond to the optimal HMM models for the different discretization types. These bar charts show the mean-CRPS of all models sorted in ascending order by best performance for each household separately, while the numerical results are summarized in table 9 in appendix C.

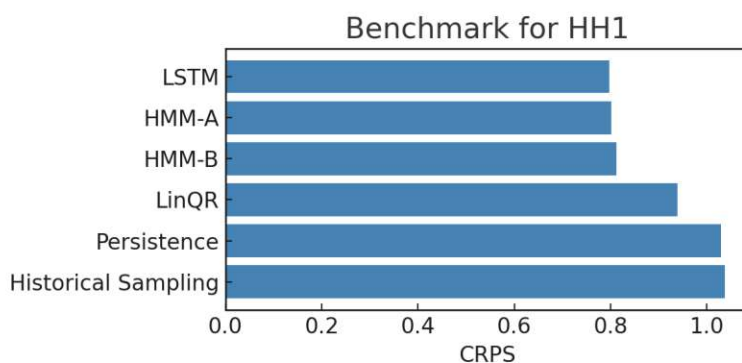


Figure 29.

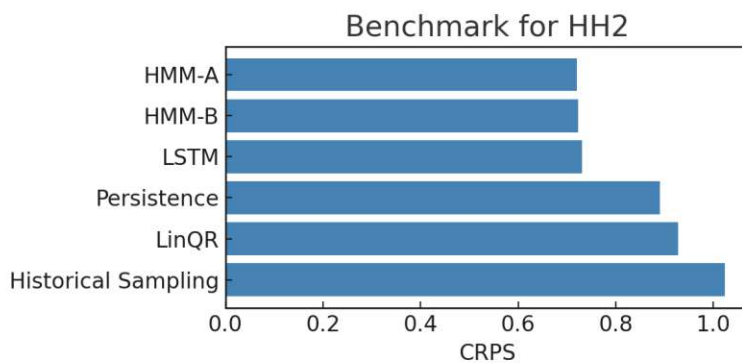


Figure 30.

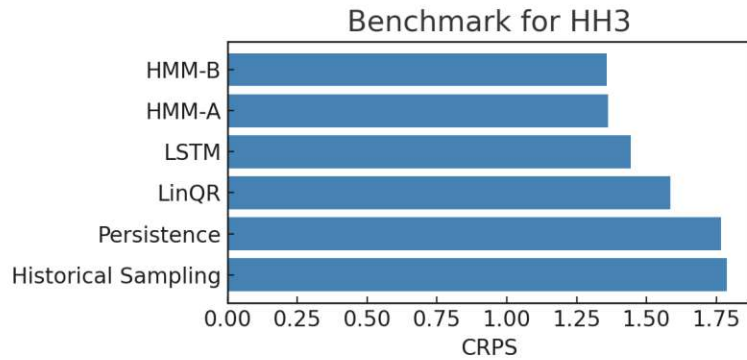


Figure 31.

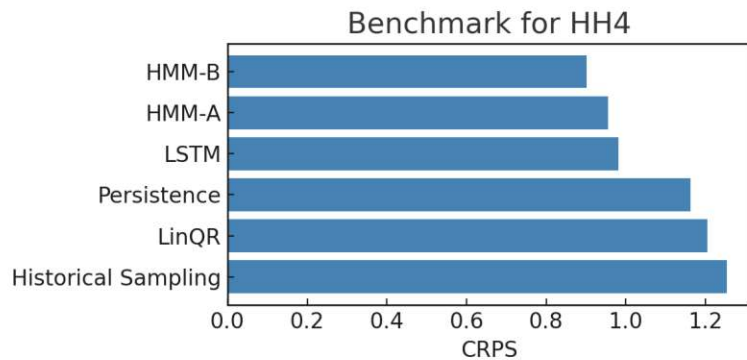


Figure 32.

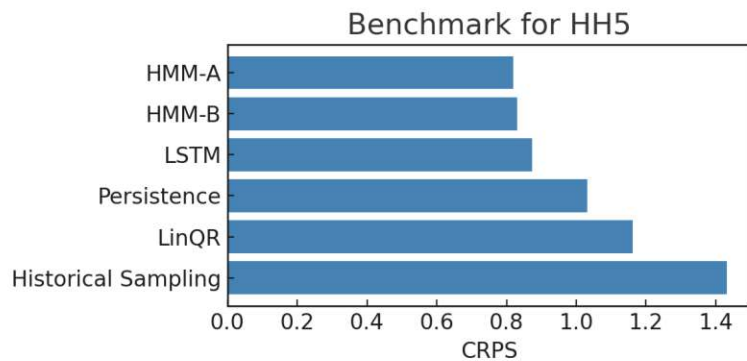


Figure 33.

The charts show that the models with the best performance are HMM-A, HMM-B and LSTM which are always ranked above the other models. The persistence model and LinQR are located in the midfield, while historical sampling is consistently the worst model. Both HMM models generally perform slightly better than LSTM with exception of HH1 in which LSTM achieves the best mean-CRPS value by a small margin.

A standard method of probabilistic forecasting evaluation is to calculate the relative improvements of the mean-CRPS in comparison to the baseline models. This allows a fast subsumptions of the results and comparisons between literature. Same as in Botman,

Lago, Becker, Vanthournout, and Moor [5], our chosen baseline model is the persistence model to computed the relative improvements, which are presented in table 8.

| HH | Hist. Sampling | LinQR | LSTM | HMM-A | HMM-B |
|-----|----------------|---------|--------|--------|--------|
| HH1 | -0.78% | 8.75% | 22.45% | 22.06% | 21.09% |
| HH2 | -14.94% | -4.27% | 17.87% | 19.10% | 18.76% |
| HH3 | -1.19% | 10.25% | 18.18% | 22.82% | 23.16% |
| HH4 | -7.83% | -3.70% | 15.58% | 17.73% | 22.46% |
| HH5 | -38.66% | -12.60% | 15.41% | 20.64% | 19.48% |

Table 8.: The relative improvement of the mean-CRPS values of the models compared to the persistence model

The relative improvements of the HMM-A and HMM-B models lie between 17% to 22% and 18% to 23% respectively. The LSTM model has a range of 15% to 22% and lacks in some cases up to 7% of improvement behind the HMM models. Based on these results, it can be confidently concluded that the HMM models are able to achieve similar and often better results in predicting the one-step-ahead distribution than the LSTM model and other benchmark models. However, to verify this result, further experiments should be repeated on a bigger number of households. For example in Botman, Lago, Becker, Vanthournout, and Moor [5] 100 households for multiple data sets were investigated. In addition, it should be considered that no data set specific hyperparameter tuning was conducted for the benchmark models, which would have exceeded the scope of this thesis. Thus, we expect that the benchmark models can improve slightly with a more detailed study. However, also the HMM models could be potentially improved, since the here presented design does not make use of temporal information, which is the standard for load forecasting models. This can be achieved by either adding temporal data as an additional feature to the observation state, or to train multiple HMM models for different day times (like eg. night vs day) or seasons. The here inferred results are in any case promising that HMM can be considered a competitive alternative in probabilistic load forecasting. Nevertheless, future studies into this topic are recommended.

Lastly, we want to put the results of this thesis in relation to existing literature. In Botman, Lago, Becker, Vanthournout, and Moor [5], similar benchmark models and results are obtained in a detailed study for a global approach of forecasting individual households load. In difference to this thesis, a multi-step forecasting method to predict the distributions for each quarter hour of the next 24 hours is conducted. Compared to the same baseline model, the evaluated forecasting models achieve relative improvements of 20% to 34%. This general better performance compared to the presented results here is explained by the difference of forecasting horizon. For the one-step-ahead prediction, the persistence model is more accurate than for a longer forecasting horizon which implies that the relative improvements to the persistence model are lower for the one-step-ahead model. Considering this argument, this comparison shows that the results of this thesis have a similar improvement level and thus, are in line with existing literature. Another

interesting observation of this literature is that the historical sampling model performs for the 24h-multiple-step-ahead prediction equally good as the other advanced forecasting methods while in the one-step-ahead prediction performs very poorly. This effect is most probably caused by the fact that the historical sampling model does not take recent historic data into account but relies purely on temporal information. This implies that for very short-time forecasting, temporal data does not have a great significance as recent historic data.

7. Conclusion

This thesis contributes to the area of forecasting by presenting a novel method for probabilistic forecasting based on HMM. Although HMM has been a popular modeling method for decoding and classification, it has been applied only scarcely for prediction purposes with little to no general theoretical background. For this reason, we developed a simple theoretical framework for HMM based forecasting in alignment with the existing standard HMM literature and derived an efficient HMM forecasting algorithm.

To test the predictive capability of HMM, we selected the use-case of probabilistic forecasting the electrical power load of households. The electrical demand of household suits as a case study for the HMM forecasting method because of its strong volatile characteristics and its current focus of research in the era of energy transition. The proposed HMM forecasting method offers here a great alternative to existing probabilistic forecasting methods due to its efficiency, closed mathematical design and tractability.

The dedicated goal was to forecast the distribution of the electric power load 15 minutes into the future. We tested the model on 5 different households from the WPUQ data set [37]. A detailed methodology was applied including an encompassing hyperparameter tuning which was often neglected in previous HMM based forecasting studies. A novelty of this thesis is a discretization method based on occurrences, which leads to smaller and thus more efficient optimal models. While the number of hidden states differ for the optimal models and lie between 40 and 80, the optimal number of observations is almost consistently 100 with only one model being the exception with 50 observations. These results show that for future studies 100 observations are a good choice for the number of discretizations, while for the number of hidden states a hyperparameter tuning is recommended and should focus on the area between 40 to 80 states. From the sensitivity analysis for the historic window length during forecasting, we conclude that a relative short historic window length of two and a half hours is sufficient which also reveals limitations of the proposed forecasting method. To wit, it shows that the HMM model does not consider information prior to this historic window and implying that it cannot detect long term patterns and behavior.

The model analysis revealed that the novel discretization method with equal-mass bins induces significantly better calibrated forecasting models than the discretization into equidistant bins, even though there are no major differences in performance measured with the mean-CRPS. As a last study, the predicted distributions for multi-step-ahead forecasts were investigated and showed that for most cases the predicted distribution converges to the empirical distribution at a time horizon of around 3-5 hours into the future. We consider this time horizon as the upper limit for a potential application of the HMM forecasting method.

We compared the HMM forecasting models with other probabilistic forecasting methods. For the benchmark experiments, two heuristic models serve as baseline, while one statistical and one machine learning model represent the state-of-the-art methods. The results show that the HMM models outperform generally the competing models. The best benchmark model, the LSTM, accomplishes a relative improvement to the baseline reaching from 15% to 22%, while the relative improvement of the HMM models ranges from 17% to 23% which improves the prediction accuracy for some households up to 7 percent points.

These are promising results; however, a few limitations have to be considered. This work served as a first study of the potential usage of HMM as a forecasting method further experiments are needed to verify the results on a large scale. Due to the scope of this thesis, no household specific hyperparameter tuning was conducted for the benchmark models. Thus, there are capacities to compare the proposed HMM forecasting method with other competitive models. The study focused mainly on the one-step-ahead predictions, while first experiments indicated that the predictive accuracy quickly declines for longer forecasting horizons. This is possibly caused by the simple design of the proposed HMM forecasting method that was purposely chosen for this fundamental study. These limitations are the choice of a discrete state and observation HMM and also excluding additional features like temporal information.

The results show that HMM provides a great opportunity to probabilistic forecasting which should be pursued in further research. Future studies should focus on extending the here proposed HMM forecasting method with the following points. First, the here presented results should be verified on a larger scale including different and bigger data sets. These data sets are ideally open source and already popular in the energy community to allow comparability to historical and future literature. Also, the predictive performance for longer time horizons should be compared with existing state-of-the-art probabilistic load forecasting methods. Furthermore, the HMM forecasting model itself can also be extended in various ways. One possible adaptation is the inclusion of temporal data and model typical daytime patterns by changes in design and additional input features. Here, results of already existing HMM research can be beneficial to include. This can potentially increase the forecasting horizon drastically. Another option is to consider a continuous observation HMM instead of the discrete version. For this case, the pre- and postprocessing step can be reduced, but certain assumptions of the distribution have to be made. At last, this thesis did not exploit the tractability of HMM, one of the great advantages of the model. Here, research can study the internal behavior of the HMM forecasting method and potentially improve the predictive performance.

In conclusion, this thesis shows that HMM has the potential to become a competitive alternative to existing probabilistic forecasting methods and provides the basis for future research in this field.

Bibliography

- [1] S. Arnold, E.-M. Walz, J. Ziegel, and T. Gneiting. *Decompositions of the mean continuous ranked probability score*. 2023. arXiv: [2311.14122](https://arxiv.org/abs/2311.14122) [stat.ME]. URL: <https://arxiv.org/abs/2311.14122>.
- [2] L. E. Baum. „An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Process“. In: *Inequalities* 3 (1972), pp. 1–8.
- [3] L. E. Baum and T. Petrie. „Statistical Inference for Probabilistic Functions of Finite State Markov Chains“. In: *The Annals of Mathematical Statistics* 37.6 (1966), pp. 1554–1563.
- [4] C. M. Bishop. *Pattern recognition and machine learning*. corr. print. New York, NY: Springer, 2007.
- [5] L. Botman, J. Lago, T. Becker, K. Vanthournout, and B. D. Moor. „A global probabilistic approach for short-term forecasting of individual households electricity consumption“. In: *Applied Energy* 382 (2025), p. 125168.
- [6] M. Brand, N. Oliver, and A. Pentland. „Coupled hidden Markov models for complex action recognition“. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1997. DOI: [10.1109/CVPR.1997.609450](https://doi.org/10.1109/CVPR.1997.609450).
- [7] A. Brockwell. „Universal Residuals: A Multivariate Transformation“. In: *Statistics and probability letters* 77 (2007), pp. 1473–1478. DOI: [10.1016/j.spl.2007.02.008](https://doi.org/10.1016/j.spl.2007.02.008).
- [8] K. Chen, K. Chen, Q. Wang, Z. He, J. Hu, and J. He. „Short-Term Load Forecasting With Deep Residual Networks“. In: *IEEE Transactions on Smart Grid* 10.4 (2019), pp. 3943–3952. DOI: [10.1109/TSG.2018.2844307](https://doi.org/10.1109/TSG.2018.2844307).
- [9] H. Cloke and F. Pappenberger. „Ensemble Flood Forecasting: A Review“. In: *Journal of Hydrology* 375 (2009), pp. 613–626. DOI: [10.1016/j.jhydrol.2009.06.005](https://doi.org/10.1016/j.jhydrol.2009.06.005).
- [10] A. P. Dempster, N. M. Laird, and D. B. Rubin. „Maximum Likelihood from Incomplete Data Via the EM Algorithm“. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (2018), pp. 1–22. DOI: [10.1111/j.2517-6161.1977.tb01600.x](https://doi.org/10.1111/j.2517-6161.1977.tb01600.x). eprint: https://academic.oup.com/jrsss/article-pdf/39/1/1/49117094/jrsss_39_1_1.pdf.
- [11] Y. Eren and İ. Küçükdemiral. „A comprehensive review on deep learning approaches for short-term load forecasting“. In: *Renewable and Sustainable Energy Reviews* 189 (2024), p. 114031. DOI: <https://doi.org/10.1016/j.rser.2023.114031>.

- [12] C. Erlwein, F. E. Benth, and R. Mamon. „HMM filtering and parameter estimation of an electricity spot price model“. In: *Energy Economics* 32.5 (2010), pp. 1034–1043. DOI: <https://doi.org/10.1016/j.eneco.2010.01.005>.
- [13] K. Ghasvarian Jahromi, D. Gharavian, and H. R. Mahdiani. „Wind power prediction based on wind speed forecast using hidden Markov model“. In: *Journal of Forecasting* 42.1 (2023), pp. 101–123. DOI: <https://doi.org/10.1002/for.2889>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2889>.
- [14] T. Gneiting and M. Katzfuss. „Probabilistic Forecasting“. In: *Annual Review of Statistics and Its Application* 1. Volume 1, 2014 (2014), pp. 125–151. DOI: <https://doi.org/10.1146/annurev-statistics-062713-085831>.
- [15] T. Gneiting, L. Stanberry, E. Grimit, L. Held, and N. Johnson. „Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds“. In: *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* 17 (2008), pp. 211–235. DOI: [10.1007/s11749-008-0114-x](https://doi.org/10.1007/s11749-008-0114-x).
- [16] J. J. J. Groen, R. Paap, and F. Ravazzolo. „Real-Time Inflation Forecasting in a Changing World“. In: *Journal of Business & Economic Statistics* 31.1 (2013), pp. 29–44. DOI: [10.1080/07350015.2012.727718](https://doi.org/10.1080/07350015.2012.727718).
- [17] A. Gupta and B. Dhingra. „Stock market prediction using Hidden Markov Models“. In: *2012 Students Conference on Engineering and Systems*. 2012. DOI: [10.1109/SCES.2012.6199099](https://doi.org/10.1109/SCES.2012.6199099).
- [18] S. Haben, S. Arora, G. Giasemidis, M. Voss, and D. Vukadinović Greetham. „Review of low voltage load forecasting: Methods, applications, and recommendations“. In: *Applied Energy* 304 (2021), p. 117798. DOI: <https://doi.org/10.1016/j.apenergy.2021.117798>.
- [19] S. Haben, M. Voß, and W. Holderbaum. *Core Concepts and Methods in Load Forecasting: With Applications in Distribution Networks*. 2023. DOI: [10.1007/978-3-031-27852-5](https://doi.org/10.1007/978-3-031-27852-5).
- [20] M. R. Hassan and B. Nath. „Stock market forecasting using Hidden Markov Model: A new approach“. In: vol. 2005. 2005. DOI: [10.1109/ISDA.2005.85](https://doi.org/10.1109/ISDA.2005.85). URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-33846963860&doi=10.1109/2fISDA.2005.85&partnerID=40&md5=8eac7c94e8828f9e1a4d05997cca4c85>.
- [21] T. Hong and S. Fan. „Probabilistic electric load forecasting: A tutorial review“. In: *International Journal of Forecasting* 32.3 (2016), pp. 914–938. DOI: <https://doi.org/10.1016/j.ijforecast.2015.11.011>.
- [22] T. Hong, P. Pinson, and S. Fan. „Global Energy Forecasting Competition 2012“. In: *International Journal of Forecasting* 30.2 (2014), pp. 357–363. DOI: <https://doi.org/10.1016/j.ijforecast.2013.07.001>.

- [23] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman. „Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond“. In: *International Journal of Forecasting* 32.3 (2016), pp. 896–913. DOI: <https://doi.org/10.1016/j.ijforecast.2016.02.001>.
- [24] T. Hong, J. Xie, and J. Black. „Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting“. In: *International Journal of Forecasting* 35.4 (2019), pp. 1389–1399. DOI: <https://doi.org/10.1016/j.ijforecast.2019.02.006>.
- [25] Q. Huangfu and J. A. J. Hall. *Parallelizing the dual revised simplex method*. 2015. arXiv: 1503.01889 [math.OC]. URL: <https://arxiv.org/abs/1503.01889>.
- [26] T. H. Jordan. „The Value, Protocols, and Scientific Ethics of Earthquake Forecasting“. In: *EGU General Assembly Conference Abstracts*. 2013, EGU2013-12789.
- [27] M. Khadr. „Forecasting of meteorological drought using Hidden Markov Model (case study: The upper Blue Nile river basin, Ethiopia)“. In: *Ain Shams Engineering Journal* 7.1 (2016), pp. 47–56. DOI: [10.1016/j.asej.2015.11.005](https://doi.org/10.1016/j.asej.2015.11.005).
- [28] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: <https://arxiv.org/abs/1412.6980>.
- [29] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. „A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA“. In: 1996. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0030333286&partnerID=40&md5=831af18283d8d9dcdd265177d2bbe2e4>.
- [30] B. Liu, J. Nowotarski, T. Hong, and R. Weron. „Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts“. In: *IEEE Transactions on Smart Grid* 8.2 (2017), pp. 730–737. DOI: [10.1109/TSG.2015.2437877](https://doi.org/10.1109/TSG.2015.2437877).
- [31] J. Luo, Y. Zheng, T. Hong, A. Luo, and X. Yang. „Fuzzy support vector regressions for short-term load forecasting“. In: *Fuzzy Optimization and Decision Making* 23.3 (2024), pp. 363–385. DOI: [10.1007/s10700-024-09425-x](https://doi.org/10.1007/s10700-024-09425-x).
- [32] S. Makonin, F. Popowich, I. V. Bajić, B. Gill, and L. Bartram. „Exploiting HMM Sparsity to Perform Online Real-Time Nonintrusive Load Monitoring“. In: *IEEE Transactions on Smart Grid* 7.6 (2016), pp. 2575–2585. DOI: [10.1109/TSG.2015.2494592](https://doi.org/10.1109/TSG.2015.2494592).
- [33] A. H. Murphy and R. L. Winkler. „Probability Forecasting in Meteorology“. In: *Journal of the American Statistical Association* 79.387 (1984), pp. 489–500.
- [34] J. Nowotarski and R. Weron. „Recent advances in electricity price forecasting: A review of probabilistic forecasting“. In: *Renewable and Sustainable Energy Reviews* 81 (2018), pp. 1548–1568. DOI: <https://doi.org/10.1016/j.rser.2017.05.234>.
- [35] L. Rabiner. „A tutorial on hidden Markov models and selected applications in speech recognition“. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.
- [36] Z. Saffer. *Stochastische Prozesse für Informatik*. Wien: TU-Verlag, 2022.

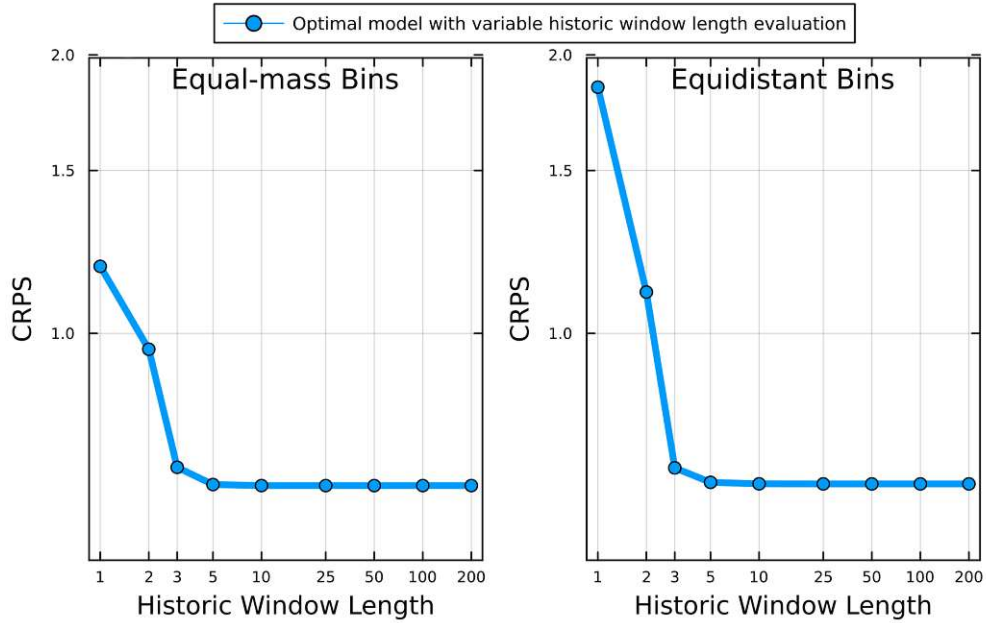
- [37] M. Schlemminger, T. Ohrdes, E. Schneider, and M. Knoop. „Dataset on electrical single-family house and heat pump load profiles in Germany“. In: *Scientific Data* 9 (2022), p. 56. DOI: [10.1038/s41597-022-01156-1](https://doi.org/10.1038/s41597-022-01156-1).
- [38] C. Strähl and J. Ziegel. „Cross-calibration of probabilistic forecasts“. In: 11.1 (2017), pp. 608–639. DOI: [10.1214/17-EJS1244](https://doi.org/10.1214/17-EJS1244).
- [39] V. Ter-Hovhannisyanyan, A. Lomsadze, Y. O. Chernoff, and M. Borodovsky. „Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training“. In: *Genome Research* 18.12 (2008), pp. 1979–1990. DOI: [10.1101/gr.081612.108](https://doi.org/10.1101/gr.081612.108).
- [40] R. Tian and G. Shen. „Predictive power of Markovian models: Evidence from US recession forecasting“. In: *Journal of Forecasting* 38.6 (2019), pp. 525–551. DOI: <https://doi.org/10.1002/for.2579>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/for.2579>.
- [41] D. van der Meer, J. Widén, and J. Munkhammar. „Review on probabilistic forecasting of photovoltaic power production and electricity consumption“. In: *Renewable and Sustainable Energy Reviews* 81 (2018), pp. 1484–1512. DOI: <https://doi.org/10.1016/j.rser.2017.05.212>.
- [42] Y. Wang, D. Gan, M. Sun, N. Zhang, Z. Lu, and C. Kang. „Probabilistic individual load forecasting using pinball loss guided LSTM“. In: *Applied Energy* 235 (2019), pp. 10–20. DOI: <https://doi.org/10.1016/j.apenergy.2018.10.078>.
- [43] L. Xu, M. Hu, and C. Fan. „Probabilistic electrical load forecasting for buildings using Bayesian deep neural networks“. In: *Journal of Building Engineering* 46 (2022), p. 103853. DOI: <https://doi.org/10.1016/j.jobe.2021.103853>.
- [44] M. Zimmermann and F. Ziel. „Spatial Meteorological, Socio-Economic, and Political Risks in Probabilistic Electricity Demand Forecasting“. In: (2024). DOI: [10.2139/ssrn.5063126](https://doi.org/10.2139/ssrn.5063126).

A. Results of Hyperparameter Analysis

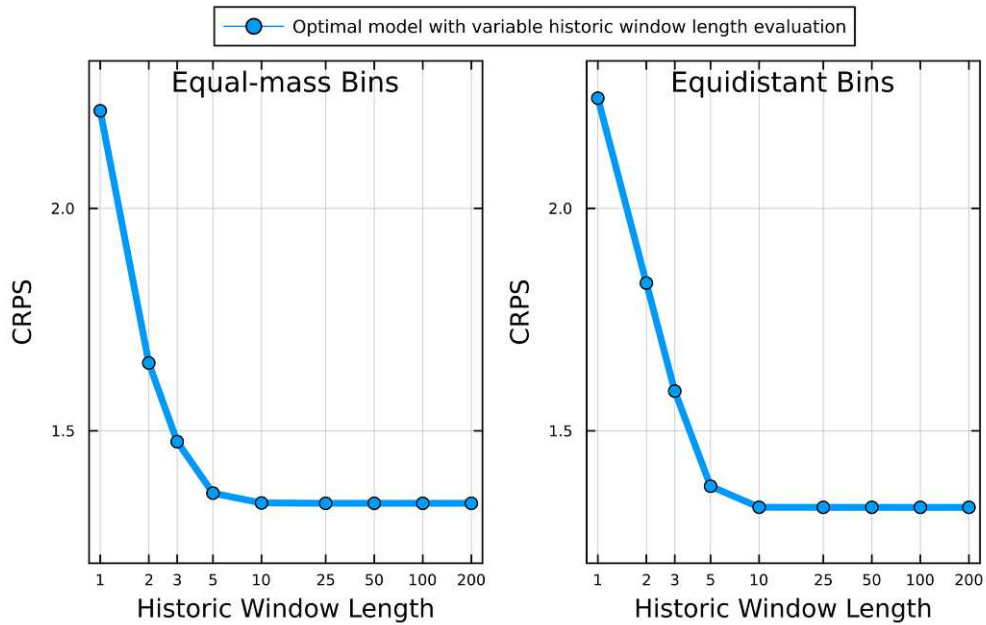
| HH | Discr. Type | M \ N | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 100 |
|-----|-------------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| HH1 | A | 25 | 0,8896 | 0,8853 | 0,8796 | 0,8815 | 0,8922 | 0,8799 | 0,8899 | 0,9003 | 0,8915 |
| | | 50 | 0,8638 | 0,8550 | 0,8402 | 0,8338 | 0,8354 | 0,8390 | 0,8442 | 0,8444 | 0,8432 |
| | | 100 | 0,8545 | 0,8500 | 0,8294 | 0,8232 | 0,8415 | 0,8359 | 0,8442 | 0,8352 | 0,8405 |
| | | 200 | 0,8574 | 0,8415 | 0,8505 | 0,8461 | 0,8587 | 0,8524 | 0,8472 | 0,8463 | 0,8766 |
| | B | 25 | 0,9196 | 0,9039 | 0,9055 | 0,9032 | 0,9063 | 0,9049 | 0,9023 | 0,9051 | 0,9000 |
| | | 50 | 0,8740 | 0,8387 | 0,8435 | 0,8434 | 0,8436 | 0,8413 | 0,8339 | 0,8396 | 0,8415 |
| | | 100 | 0,8455 | 0,8505 | 0,8304 | 0,8316 | 0,8298 | 0,8280 | 0,8415 | 0,8259 | 0,8390 |
| | | 200 | 0,8591 | 0,8349 | 0,8444 | 0,8363 | 0,8472 | 0,8450 | 0,8520 | 0,8347 | 0,8599 |
| HH2 | A | 25 | 0,7354 | 0,7381 | 0,7503 | 0,7539 | 0,7567 | 0,7569 | 0,76 | 0,7696 | 0,778 |
| | | 50 | 0,7279 | 0,6987 | 0,6934 | 0,7013 | 0,6995 | 0,7101 | 0,708 | 0,7191 | 0,7165 |
| | | 100 | 0,7014 | 0,7078 | 0,6846 | 0,6883 | 0,6903 | 0,6857 | 0,692 | 0,6973 | 0,6966 |
| | | 200 | 0,7075 | 0,6855 | 0,6882 | 0,6897 | 0,6946 | 0,7035 | 0,6988 | 0,7013 | 0,711 |
| | B | 25 | 0,8042 | 0,7995 | 0,803 | 0,7897 | 0,7912 | 0,799 | 0,7899 | 0,7872 | 0,7871 |
| | | 50 | 0,7088 | 0,7048 | 0,7089 | 0,6966 | 0,7019 | 0,7016 | 0,7051 | 0,7026 | 0,7031 |
| | | 100 | 0,6889 | 0,6939 | 0,6938 | 0,6874 | 0,6894 | 0,6917 | 0,7003 | 0,6946 | 0,6971 |
| | | 200 | 0,6947 | 0,699 | 0,6938 | 0,6932 | 0,6948 | 0,6996 | 0,7021 | 0,6993 | 0,6989 |
| HH3 | A | 25 | 1,4811 | 1,4117 | 1,4017 | 1,4062 | 1,4035 | 1,408 | 1,4057 | 1,4133 | 1,4228 |
| | | 50 | 1,4248 | 1,3885 | 1,3741 | 1,3697 | 1,3658 | 1,3767 | 1,3745 | 1,387 | 1,391 |
| | | 100 | 1,423 | 1,4065 | 1,3803 | 1,3815 | 1,3883 | 1,3876 | 1,3967 | 1,402 | 1,4123 |
| | | 200 | 1,4279 | 1,3965 | 1,4036 | 1,4022 | 1,4095 | 1,4215 | 1,4094 | 1,4269 | 1,4508 |
| | B | 25 | 1,4666 | 1,4525 | 1,4317 | 1,4269 | 1,43 | 1,4221 | 1,4195 | 1,4133 | 1,4201 |
| | | 50 | 1,4026 | 1,3866 | 1,3768 | 1,3776 | 1,3668 | 1,3687 | 1,37 | 1,3669 | 1,3826 |
| | | 100 | 1,4257 | 1,3744 | 1,3684 | 1,3712 | 1,3587 | 1,3768 | 1,3739 | 1,3689 | 1,3832 |
| | | 200 | 1,396 | 1,3854 | 1,3892 | 1,3915 | 1,3945 | 1,4115 | 1,4067 | 1,4289 | 1,4382 |
| HH4 | A | 25 | 1,0409 | 1,0328 | 1,0234 | 1,0291 | 1,0106 | 1,0356 | 1,0232 | 1,0494 | 1,0348 |
| | | 50 | 0,9857 | 0,9812 | 0,9506 | 0,9595 | 0,9578 | 0,9672 | 0,9753 | 0,9747 | 0,9688 |
| | | 100 | 1,0053 | 0,9771 | 0,9661 | 0,9542 | 0,9684 | 0,9505 | 0,9643 | 0,9734 | 0,9597 |
| | | 200 | 0,9845 | 0,9779 | 1,0166 | 0,9853 | 0,9986 | 0,9869 | 0,985 | 0,994 | 1,0403 |
| | B | 25 | 0,9993 | 0,9831 | 0,9897 | 0,9701 | 0,9765 | 0,9715 | 0,9764 | 0,9705 | 0,9661 |
| | | 50 | 0,9918 | 0,9595 | 0,9529 | 0,9464 | 0,9418 | 0,9414 | 0,9476 | 0,9493 | 0,9502 |
| | | 100 | 0,969 | 0,9455 | 0,9365 | 0,9242 | 0,9157 | 0,9236 | 0,9282 | 0,9229 | 0,9261 |
| | | 200 | 0,9739 | 0,9579 | 0,947 | 0,959 | 0,9769 | 0,9632 | 0,9716 | 0,9553 | 0,984 |
| HH5 | A | 25 | 0,9637 | 0,9145 | 0,92 | 0,9115 | 0,9022 | 0,8931 | 0,9158 | 0,904 | 0,9095 |
| | | 50 | 0,9345 | 0,8685 | 0,8421 | 0,8509 | 0,8361 | 0,8342 | 0,8357 | 0,8336 | 0,8414 |
| | | 100 | 0,8911 | 0,8364 | 0,8303 | 0,8219 | 0,835 | 0,8243 | 0,8297 | 0,8371 | 0,8388 |
| | | 200 | 0,8955 | 0,8398 | 0,8457 | 0,873 | 0,8461 | 0,8487 | 0,8492 | 0,8558 | 0,8549 |
| | B | 25 | 0,8934 | 0,8816 | 0,8897 | 0,8646 | 0,8745 | 0,8606 | 0,8556 | 0,8558 | 0,8583 |
| | | 50 | 0,8785 | 0,8622 | 0,8543 | 0,8427 | 0,8514 | 0,8528 | 0,8382 | 0,8293 | 0,8443 |
| | | 100 | 0,9075 | 0,8312 | 0,8314 | 0,8269 | 0,824 | 0,8291 | 0,8326 | 0,8213 | 0,835 |
| | | 200 | 0,8698 | 0,8347 | 0,8333 | 0,84 | 0,8388 | 0,839 | 0,8389 | 0,8498 | 0,8398 |

B. Results of Sensitivity Analysis

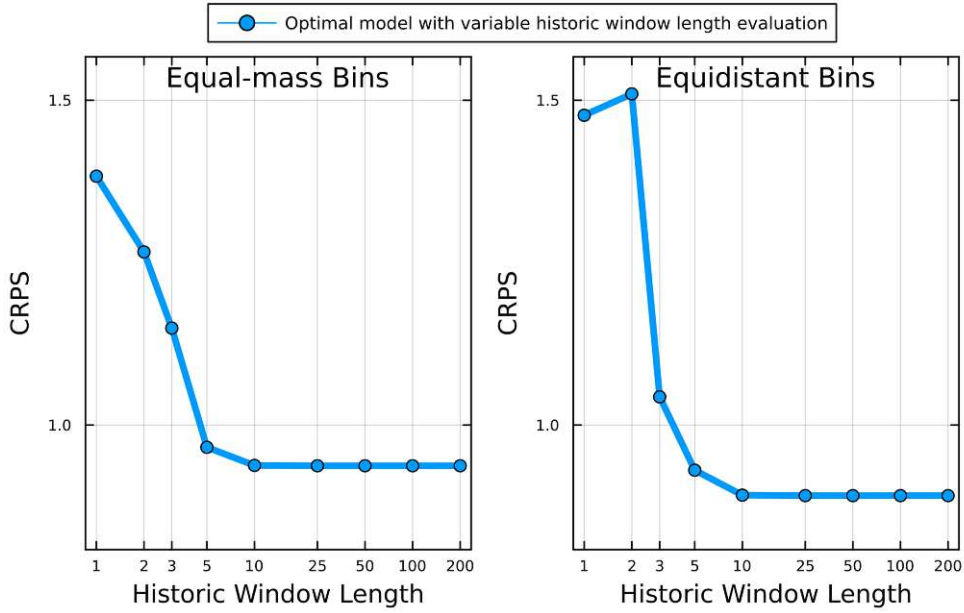
Sensitivity Analysis for HH2



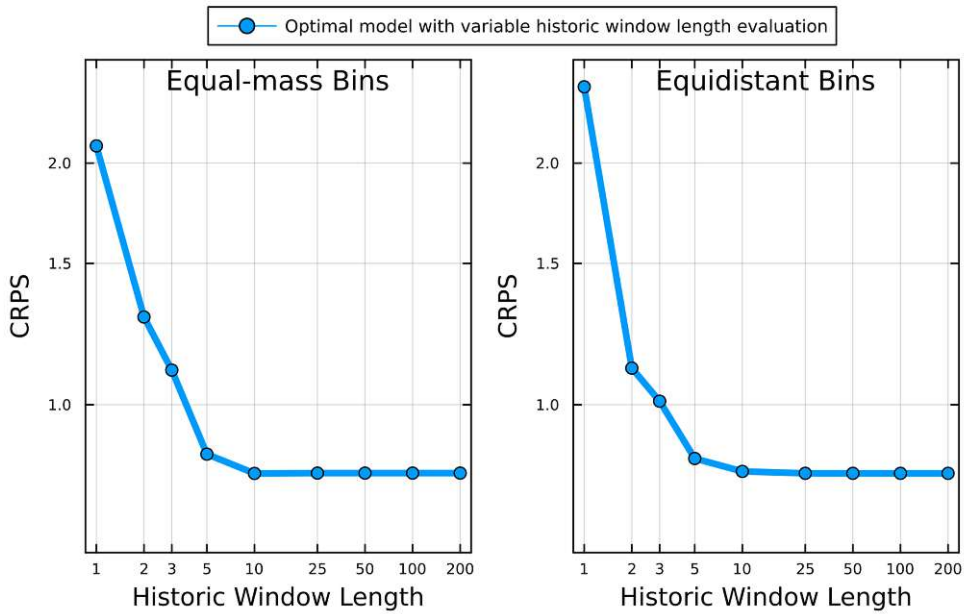
Sensitivity Analysis for HH3



Sensitivity Analysis for HH4



Sensitivity Analysis for HH5



C. Results of Benchmark

| HH | Persistence | Historical Sampling | LinQR | LSTM | HMM-A | HMM-B |
|-----|-------------|---------------------|-------|-------|-------|-------|
| HH1 | 1.029 | 1.037 | 0.939 | 0.798 | 0.802 | 0.812 |
| HH2 | 0.890 | 1.023 | 0.928 | 0.731 | 0.720 | 0.723 |
| HH3 | 1.766 | 1.787 | 1.585 | 1.445 | 1.363 | 1.357 |
| HH4 | 1.162 | 1.253 | 1.205 | 0.981 | 0.956 | 0.901 |
| HH5 | 1.032 | 1.431 | 1.162 | 0.873 | 0.819 | 0.831 |

Table 9.: Mean-CRPS of all models per household.