Project: Electricity Day-Ahead Price Forecasting

Ryan Tang[1]

[1]Duke University, Statistical Science

I. Introduction

The electricity market is a peculiar one among all commodities. It is economically non-storable. The power system requires a constant balance between production and consumption. The spot price varies wildly due to temperature, wind speed, sun exposure, and various business and residential activities that exhibit long-term and short-term seasonal patterns. Nevertheless, the world cannot run without electricity. After the deregulation in the 1990s, the introduction of competitive electricity markets reshaped the traditional monopolistic, government-controlled sector. With the wide participation range from public and private generators and distributors to many individual trading firms, the activities have been driving energy prices lower simultaneously, ensuring a stable electricity market. However, forecasting the future electricity price is not a simple endeavor. Utility generators need to do the forecast to plan the production the next day or even the next few years. Distributors need an accurate forecast to satisfy all retail demand in real-time with the lowest price possible. And trading firms need accurate predictions for their arbitrage strategies to drive profitability. Hence, the demand for accurate forecasts is paramount.

There are two types of forward markets, long-term, and day-ahead. Companies use the long-term markets to plan generation and consumption and hedge risks one to three years ahead. And most of the immediate planning is done in the day-ahead market for the preceding day's retail demand. If there are any differences between the planned load and the realized load, the grid operator often uses the spot market to mark the differences in real-time. The focus of this journal will be the day-ahead market, with its intricacy, where the forecast is delivered one day ahead at once for all 24 hours. In other words, the day-ahead market does not allow continuous trading. Before a specific cutoff time, all bids and offers must be submitted for the next day of each hour. When making the forecast, we have 24 outputs to make.

Hong's team hosted a global energy forecasting competition in 2014 with multiple tracks, GEFCom2014, and open-sourced all the data [2]. It has been utilized widely by other researchers as a simple benchmark dataset after that. We, too, will use the same dataset to conduct price forecasting to have a few benchmarks to compare results. The data shown in Figure 1 provided by Hong's team is minimalist, with only 1 price series and 2 co-variate series. In particular, the system load is the forward forecast of consumption or the total amount of energy that needs to be transferred instantaneously on the grid due to demand. Hence, the concept of congestion and grid failure, because when the concurrent amount of electricity on the grid is too high, it overheats the wires and causes congestion and failures. The difference between system and zonal is that zonal only counts the amount of power towards the target zone, and the system counts the total amount of power. Sometimes, a zone can have congestion, and power has to transmit through different wires toward the target zone; therefore, the differences between the two loads also provide signals. Note, here, the price is only the zonal price.

The real-world forecasting task is, in fact, much more complex and takes many other factors into account in the modeling process, such as temperatures, generator failures, transmission congestion, and the inter-dependency of the electricity grid network, which were not present in the dataset. Nevertheless, the data is enough for our purpose.

Not all predictions are treated equally. Due to the inherited volatility of electricity prices, point estimates are often insufficient for energy companies' decision-making and planning process because most transactions are conducted with the forward market. Therefore, coming up with probabilistic predictions is crucial for the task, and it seems to be a natural application for Bayesian analysis. Therefore, here we first formulate the problem as a time series forecasting under certainty by utilizing probabilistic modeling, applying Hidden Markov Model (HMM) to the problem, and comparing it with a few empirical results presented in Nowotarski's paper [4]. In the last section, we perform diagnostics on HMM and propose future improvements.

II. Previous Work

Despite its importance, electric price forecasting (EPF) is an underdeveloped and often overlooked research area [4]. The publication pace had only started to pick up attention since 2000, the California crisis. And the pace

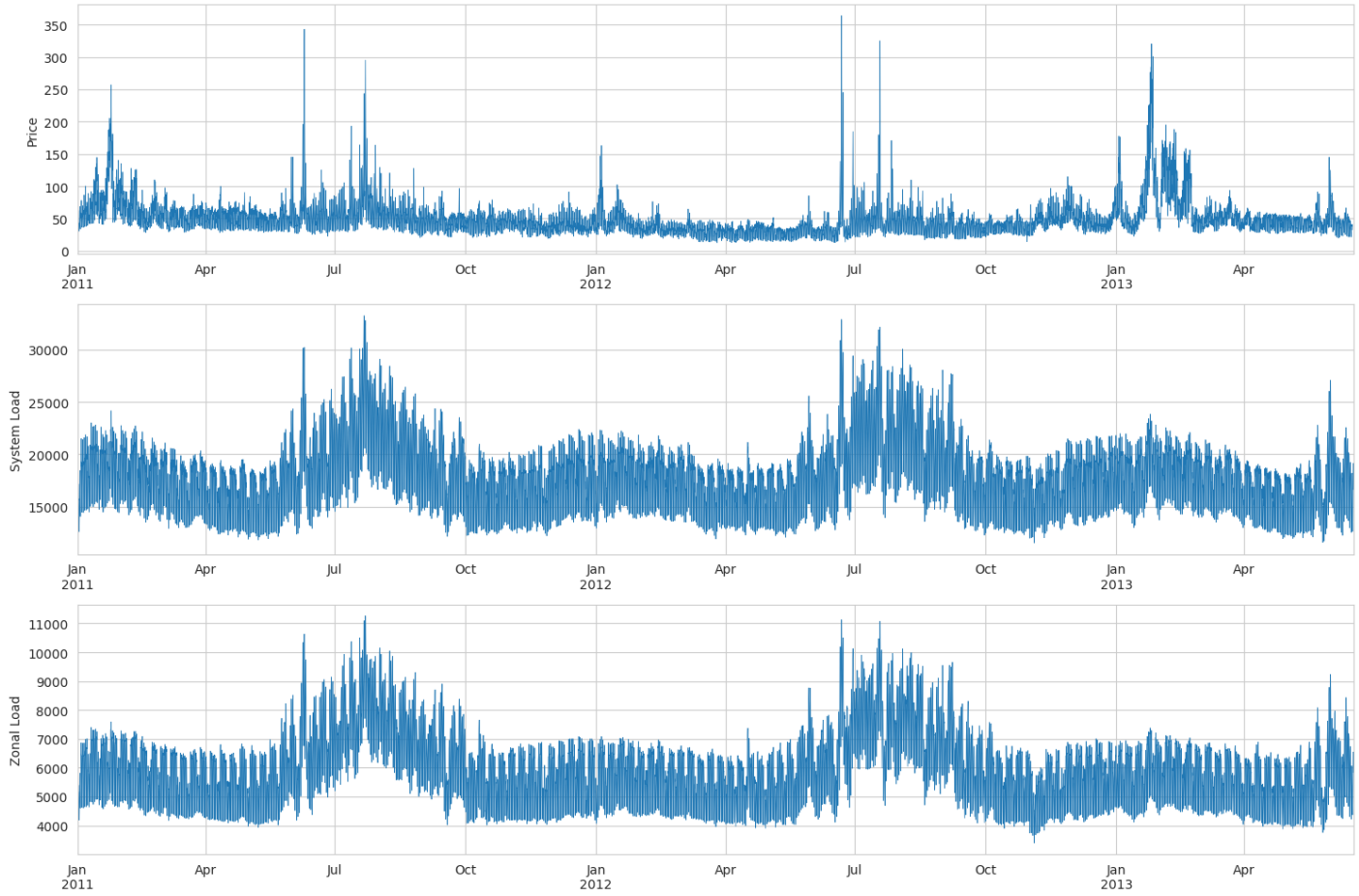

Fig. 1: Plots of the electricity price and the two covariates, load estimates.

had only started to pick up since 2005. Many publications focus on using neural networks in prediction, some utilize classic time series methodologies, but the majority concentrates on point forecasting. As we highlighted the importance of probabilistic electric price forecasting forecast (PEPF) previously, it is, even more, an undeveloped regime and barely picks up the pace since 2011.

Weron has extensively reviewed various models used in EPF literature and their relative strengths and weaknesses [1]. He segmented all modeling approaches into 5 categories, but we think 3 are more appropriate because of their significant overlap. The multi-agent model is the first segment that has its root in game theory. It tries to model market competition, supply and demand, and many other fundamental factors by formulating equilibrium and a strategic game. Such formulation is neat for theoretical pursuit and creating beautiful theorem but lacks a general solution except for some toy examples. And often, such a method relies on fundamental data and only updates weekly, monthly, or even yearly, making them unsuitable for day-ahead prediction. The second major segment consists of many traditional models, including many variants of neural networks, support vector machines, random forests, gradient boosting machines, etc. These model are great point predictor that works well with non-linearity. However, there has not been a thorough investigation of these methods, according to Weron, because it is difficult to establish a benchmark due to the temporal nature of the task. The last segment consists of many statistical and probabilistic models, including the classic time series models, AR, ARX, GARCH, regression, and Markovian models like Jump Diffusion and Markov-Regime Switching. Weron summarized the former tends to perform poorly on sudden spikes, and the latter tends to capture spikes and volatility quite well but lacks forward predictability. We believe the statement is biased because they never considered Bayesian analysis in the context.

We introduce two classic benchmarks for comparison purposes, which we will briefly discuss in this section. Although most methods discussed by Weron to date only offer point estimations by design, many auxiliary techniques can be utilized to construct a prediction interval. We also introduce one of the most straightforward methods, historical simulation, too, in this section. So we can make a comparison between the benchmarks and our method coherently.

A. Benchmark

We called this the naive model because it simply uses the last day or week's hourly price for the current estimates based on the day of the week. Namely,

$$\hat{P}_{d,h} = P_{d-1,h}$$

on Tuesday, Wednesday, Thursday or Friday, and

$$\hat{P}_{d,h} = P_{d-7,h}$$

on Monday, Saturday, or Sunday. We used the conventional electricity price notation here that $P_{d,h}$ stands for the electricity price at day d hour h; hence $P_{d-1,h}$ is the h-th hourly price from yesterday. Obviously, there is no training or parameter to tune except that we need a week's data for initial calibration.

B. Autoregressive Model (ARX)

Another popular model that is adopted widely in the community uses an autoregressive (AR) model with extra factors (ARX), X stands for exogenous variables. Simply put, it is a linear regression on AR. Misiorek et al. introduced it originally, and the model uses a centered log-price $p_{d,h}$, few autoregressive lag terms, log zonal load $z_{d,h}$, and day of week effects D_{day} which follow the following equation:

$$\begin{aligned} p_{d,h} &= \log(P_{d,h}) - \frac{1}{T}\sum_{t=1}^{T}\log(P_{t,h}) \\ \hat{p}_{d,h} &= \beta_h^{\intercal} x_{d,h} \\ x_{d,h} &= (p_{d-1,h}, p_{d-2,h}, p_{d-7,h}, p_{d-1}^{min}, \\ &\qquad z_{d,h}, D_{sat}, D_{sun}, D_{mon})^{\intercal} \end{aligned}$$

Here T is the model calibration, or training, period which was chosen as 365 days. In our case, the day-of-week effects were handled using one-hot-encoding for only Monday, Sunday, and Saturday.

C. Prediction Interval PI Construction

The interval bounds are usually determined simply using historical residual analysis. Hence, the prediction interval, $[\hat{L}_t, \hat{U}_t]$, lower and upper bounds are estimated directly using the observed prediction errors ϵ_t with the respective quantile in the training sets.

$$\begin{aligned} \epsilon_t &= P_t - \hat{P}_t \\ \hat{L}_t &= \hat{P}_t + \hat{F}_{\epsilon,t}^{-1}(1-q) \\ \hat{U}_t &= \hat{P}_t + \hat{F}_{\epsilon,t}^{-1}(q) \end{aligned}$$

where $\hat{F}_{\epsilon,t}^{-1}(\cdot)$ is the inverse CDF function of the empirical prediction residual at time t and q is the quantile in interest, and $\hat{P}_t$ is the predicted electricity price at time t. We used this methodology to construct the PI for all the benchmark models.

III. Evaluation Metrics

We are spitting out full probability distribution rather than a point estimate. However, the realization is a scalar. We need a specific way of assessing the performance of the probabilistic prediction in this context. The goal is to maximize sharpness while maintaining calibration. Calibration is crucial because we like to ensure the outputted distribution reflects the true, observed density. It is often time called unbiasedness. On the other hand, sharpness concerns the prediction variances, often called precision; the tighter the distribution, the better. Below, we introduce two widely used metrics for performance evaluations that will be used in our performance evaluations.

A. Unconditional coverage (UC)

The intuition behind this metric is that the ideal empirical coverage of the quantile should resemble the nominal quantile coverage. Here, we first introduce the hits and misses indicator $\mathbf{1}_t$ for every prediction at time t. Then, the unconditional coverage of our series of probabilistic predictions, UC, is defined as below.

$$\mathbf{1}_t = \begin{cases} 1 & \text{if } \hat{P}_t \in [\hat{L}_t, \hat{U}_t] \\ 0 & \text{otherwise} \end{cases}$$

$$UC = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_t$$

B. Continuous Ranked Probability Score (CRPS)

In a nutshell, CRPS measures how closely a predicted distribution is to a delta function that centers at the observed instance; the smaller the value, the better, and is defined as

$$CRPS(\hat{F}_{P_t}, P_t) = \int_{-\infty}^{\infty} (\hat{F}_{P_t} - \mathbf{1}_{\{P_t \le x\}})^2 dx$$

However, the integral is hard to estimate. Often, we have to rely on approximation. One useful way is to use a series of Pinball losses at different quantiles ranging from 1% to 99%, then average them up to arrive at the approximation. Pinball loss is an asymmetric linear loss version of L1-loss. Instead of calculating the loss at the median, 50-th quantile, it does on any arbitrary point. Hence, it is often called quantile loss.

$$\begin{aligned} &Pinball(\hat{Q}_{P_t}(q), P_t, q) \\ &\qquad = \begin{cases} (1-q)|\hat{Q}_{P_t}(q) - P_t| & \text{if } P_t < \hat{Q}_{P_t}(q) \\ q|\hat{Q}_{P_t}(q) - P_t| & \text{if } P_t \ge \hat{Q}_{P_t}(q) \end{cases} \end{aligned}$$

$$CRPS(\hat{F}_{P_t}, P_t) = \int_0^1 Pinball(\hat{Q}_{P_t}(q), P_t, q) dq$$

$\hat{Q}_{P_t}(q)$ is the price forecast at the q-th quantile and P_t is the observed price. We can see when $q = 0.5$, the Pinball loss reduces to the L1-loss.

IV. Problem Formulation (ARX-HMM)

In this section, we define the exact formulation of the proposed HMM model. Rather than relying on the historical simulation method for constructing the prediction interval, we use a forward sampling directly through the posterior predictive distribution using Markov Chain Monte Carlo (MCMC), which comes naturally through our graphical model and is one of the biggest advantages on using Bayesian method to quantify uncertainty directly in terms of posterior. Like all, HMM, the model consists of a discrete hidden state denoted $Z_t \in \{0, 1\}$, a conditional observation model, $p(y_t|Z = z_k, x_t)$, and a transition matrix A. Here we assume time-invariant A but parameterized the observation model using Gaussian distribution with $\theta_k = (\beta_k, \sigma_k^2)$ conditionally independent of each hidden state k. Lastly, we placed priors on top of all of them for Bayesian methods. Mathematically, the following are the model specs. Figure 2 has the formulation in a neatly hand-drawn graphical model diagram.

$$
\begin{aligned}
p(z_0) &= p(\pi) \\
p(z_t|z_{t-1}) &= A_{ji} \\
p(y_t|x_i, Z_t = z_k) &= \mathcal{N}(y_t|x_i^\intercal \beta_k, \sigma_k^2) \\
A_{j\cdot} &\sim Dir(\{\alpha_1, ..., \alpha_k\} = \mathbf{1}_k) \\
\beta_k &\sim \mathcal{N}(\mu_{ko} = 0, \sigma_{ko}^2 = 100) \\
1/\sigma_k^2 &\sim Gamma(1, 20)
\end{aligned}
$$

Note we decided to only assume two hidden states. One because it already captures the spiky phenomenon quite well, we have a limited amount of data, and too many hidden states quickly exacerbate the sampling speed. Hence, the transition matrix is $A \in \mathbf{R}^{2\times 2}$, and we model each row with its Dirichlet prior. We decided to use the same features from the ARX model for the design matrix, which contains 8 expert-selected features, $\beta_k \in \mathbf{R}^8$. And, following the typical full-conditional of a Bayesian linear regression model, we pick the normal prior for β_k and gamma prior for the precision $1/\sigma_k^2$. We fit the model individually for each hourly time series, but we can still compactly write it still using a multivariate-normal with a diagonalized covariance matrix $\sigma^2\mathbf{I}_k$. Therefore, the resulting model consists of one for each hour, containing 528 parameters, 22 per hour. β_k, σ_k^2, A are time-invariant and are the point of interest in posterior analysis.

The model has no no-close-form solution, even if we used the normal conjugates. We have to rely on the Baum-Welch algorithm, especially the forward filtering backward sampling (FFBS) algorithm in MCMC, to sample the hidden state sequence, $Z_{1:T}^{(s)}$, first then utilize No-U-Term Hamiltonian Sampler (NUTS) for the rest of parameters at each sequential step. However, we group the learning and inference tasks by specifying all the parameters through Bayesian formulation. It helps alleviate over-fitting with Bayesian shrinkage and provides robust uncertainty estimation. Luckily, PYMC3 [3] already has a minimal HMM model in place and provides the advanced MCMC samplers out-of-box; all we need to do is adjust it to accommodate Bayesian Linear Regression at each observation step. For the entire implementation of the model in PYMC3, please refer to my Github repository.

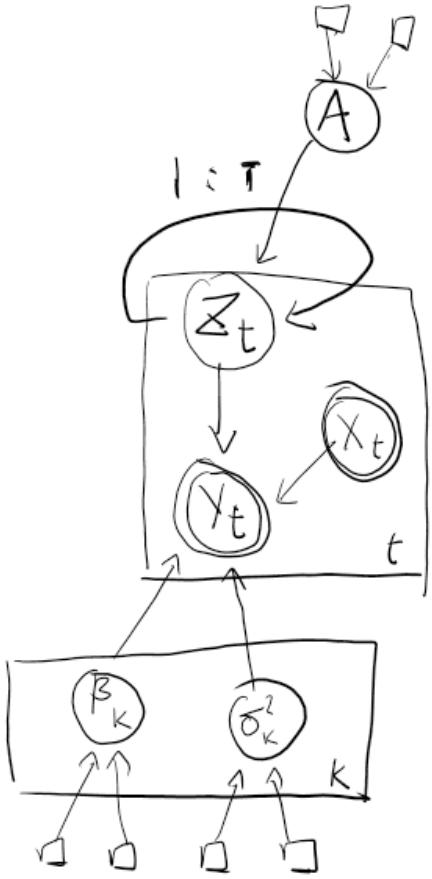

Fig. 2: The HMM graphical model with Gaussian regression observation parameterizations. The nodes with two square boxes pointed to are the ones with priors or hyper-priors. The nodes with two squares are observed in the data.

Model	UC	CRPS
ARX-HMM	0.900*	2.213
ARX	0.839	2.856
Naive	0.879	3.02

TABLE I: The test results for the 3 models. *ARX-HMM's UC estimation has a larger uncertainty because it is estimated with much fewer samples than the other two models. One major season is because ARX-HMM is much more expansive to run.

V. Results & Inferences

The test results, out of sample predictions, are shown in Table I. We can see ARX-HMM significantly outperforms the two benchmarks. The unconditional coverage is nearly perfect; it is one of the advantages of using Bayesian

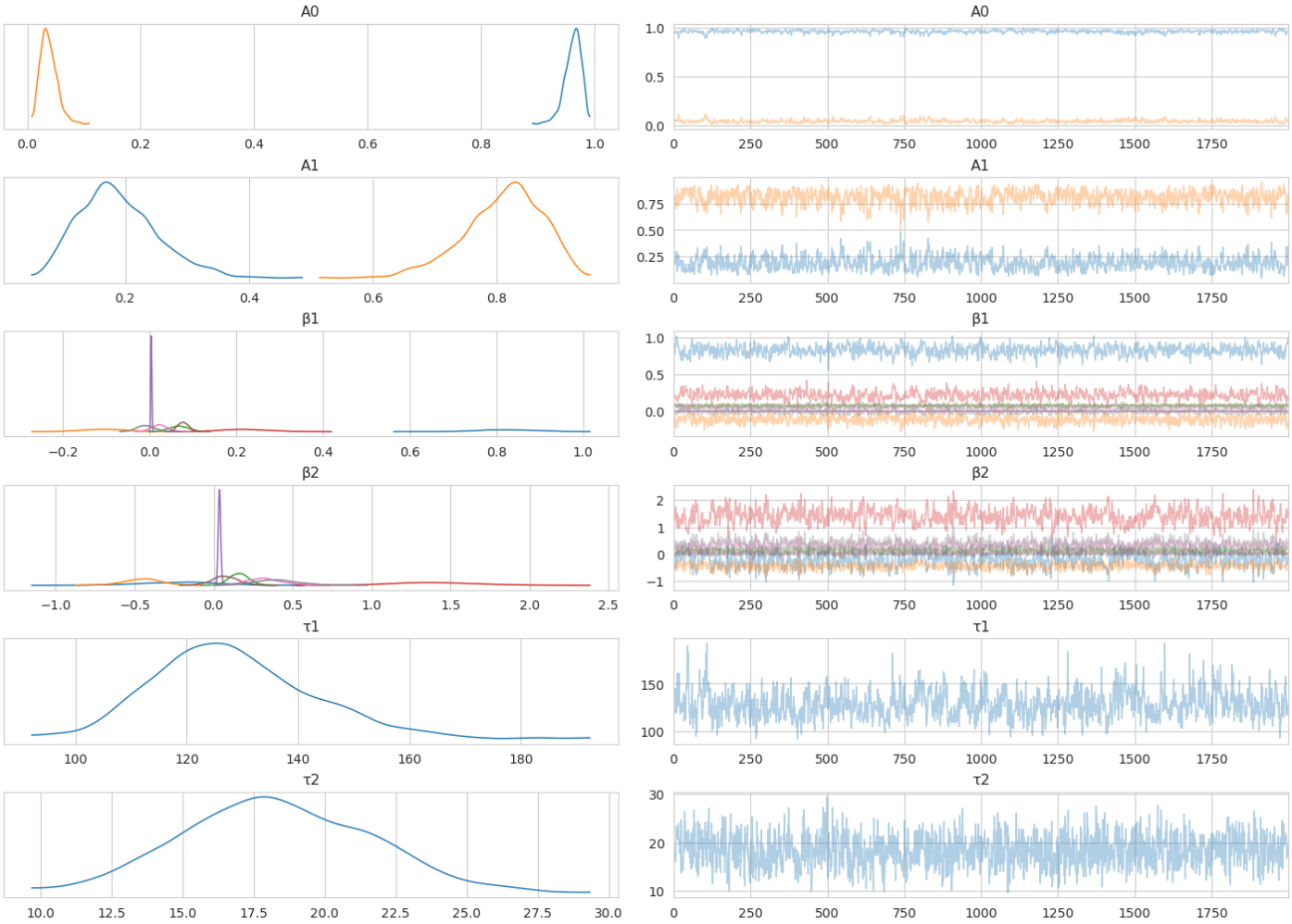

Fig. 3: MCMC Posterior Diagnostic Plots for 16-th hour.

methods. And CRPS is 22% less than the classic ARX model and 27% less than the naive model. Note we used exactly the same features from the ARX model. In other words, ARX-HMM had done a much better job of learning from the same dataset.

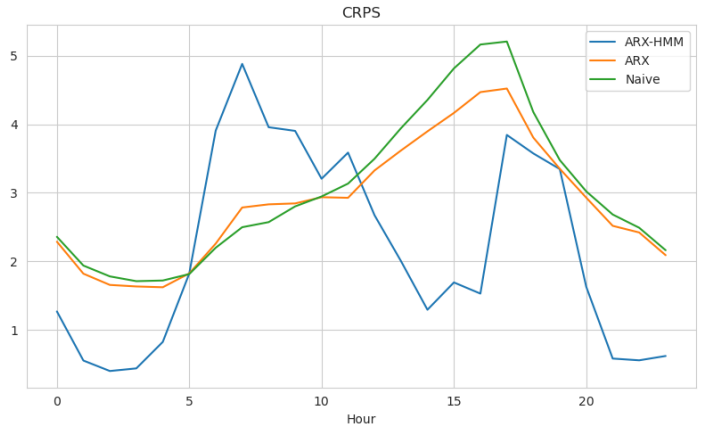

Fig. 4: Chart shows the CRPS hourly comparison of all three models

To dig in deeper about why ARX-HMM performs better, here we plot the hourly CRPS scores instead of the grand average in Figure 3, which sheds some light on the modeling differences. We can see ARX-HMM performs exceptionally well during the early, mid-day, and late night hours, 100%+ better, but underperforms during the morning rush hours, 6 Am to 10 AM. One reason for this artifact might be because morning rush hours have an inherently different process, and the current 2 hidden state assumption is not enough to capture it. Or, we might not have the relevant features, so ARX-HMM is being extra cautious around that region with higher CRPS scores. However, the performance of ARX-HMM has been surprisingly well.

Lastly, we can gain more insights directly from the posterior distributions' diagnostic and individual KDE plots, Figure 4. The transition probabilities from state 0, the state of low volatility, tends to stay in state 0 for about 95% of the time. And once we switched to the high volatility state, state 1, it stayed here with 80% probability and 20% of switch back to the low volatility state. Hence, the spiky price movement. And the spike tends to stay for

a bit before it goes away. The second interesting finding is that τ_2, which is the precision of the mean price, is about 7 times less than τ_1; in other words, the high volatility state has 7 times more volatility than the low volatility state.

VI. Conclusions

Besides some light review of the research area about electricity price forecasting and its history and deregulation, we introduced a novel, full Bayesian probabilistic modeling approach to the forecasting task, which we called ARX-HMM. Although much literature claimed that the HMM models have poor forward predictability, we proved the claim wrong through empirical results using a simple dataset and two widely adopted benchmarks. The success is mainly due to the robustness of Bayesian methods and their capability to directly quantify uncertainty through the posterior predictive distribution. By combining the ARX model with HMM model, we gave the model the ability to capture both short-term and long-term dependencies. Short-term through the direct autoregressive lag terms and long-term through the marginal of the hidden states. Bayes theorem ensures the model is well calibrated, and MCMC provided a way to sample directly from the posterior distributions. With the proposed model, we achieved around 25% less loss in terms of CRPS from the benchmarks.

References

[1] Rafał Weron. "Electricity price forecasting: A review of the state-of-the-art with a look into the future". In: *International Journal of Forecasting* 30.4 (2014), pp. 1030–1081. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2014.08.008. URL: https://www.sciencedirect.com/science/article/pii/S0169207014001083.

[2] Tao Hong et al. "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond". In: *International Journal of Forecasting* 32.3 (2016), pp. 896–913. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2016.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207016000133.

[3] John Salvatier, Thomas Wiecki, and Christopher Fonnesbeck. "Probabilistic programming in Python using PyMC3". In: (Jan. 2016). DOI: 10.7287/PEERJ.PREPRINTS.1686V1.

[4] Jakub Nowotarski and Rafał Weron. "Recent advances in electricity price forecasting: A review of probabilistic forecasting". In: *Renewable and Sustainable Energy Reviews* 81 (2018), pp. 1548–1568. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2017.05.234. URL: https://www.sciencedirect.com/science/article/pii/S1364032117308808.