ARTICLE IN PRESS

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Lasso estimation for GEFCom2014 probabilistic electric load forecasting

Florian Ziel [a,*], Bidong Liu [b]

[a] Europa-Universität Viadrina, Frankfurt (Oder), Germany

[b] University of North Carolina at Charlotte, Charlotte, NC, USA

ARTICLE INFO

Keywords:
Probabilistic forecasting
Threshold AR
Time-varying effects

ABSTRACT

We present a methodology for probabilistic load forecasting that is based on lasso (least absolute shrinkage and selection operator) estimation. The model considered can be regarded as a bivariate time-varying threshold autoregressive(AR) process for the hourly electric load and temperature. The joint modeling approach incorporates the temperature effects directly, and reflects daily, weekly, and annual seasonal patterns and public holiday effects. We provide two empirical studies, one based on the probabilistic load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014-L), and the other based on another recent probabilistic load forecasting competition that follows a setup similar to that of GEFCom2014-L. In both empirical case studies, the proposed methodology outperforms two multiple linear regression based benchmarks from among the top eight entries to GEFCom2014-L.

1. Introduction

We present a methodology for probabilistic load forecasting that is based on lasso (least absolute shrinkage and selection operator) estimation. The lasso estimator introduced by Tibshirani (1996) has the properties of automatically shrinking parameters and selecting variables. Thus, it enables us to estimate high-dimensional parameterizations. The procedure learns from the data in the sense that the parameters of less important variables will automatically be given low or even zero values. The time series model considered is a bivariate time-varying threshold autoregressive (AR) model for the hourly load and temperature. The model is specified so that it captures several stylized facts in load forecasting, such as the underlying daily, weekly, and annual seasonal patterns, the non-linear relationship between load and temperature, and holiday and long term effects.

In this paper, we illustrate the proposed methodology using two case studies from two recent forecasting competitions. The first is from the probabilistic load forecasting track of the Global Energy Forecasting Competition 2014, denoted GEFCom2014-L. The topic of GEFCom2014-L is month-ahead hourly probabilistic load forecasting using hourly temperature data from 25 weather stations. More details about GEFCom2014-L, such as rules and data, are provided by Hong et al. (2016). When implementing the proposed methodology, we create a new virtual temperature time series by averaging the temperatures of stations 3 and 9. These stations are chosen because they give the best in-sample fits to a cubic regression of the load against the temperature.

The second case study is from the year-ahead probabilistic load forecasting competition organized by Tao Hong from UNC Charlotte in fall 2015, which was an

* Corresponding author.
E-mail addresses: ziel@europa-uni.de (F. Ziel), bliu8@uncc.edu (B. Liu).

http://dx.doi.org/10.1016/j.ijforecast.2016.01.001
0169-2070/

Please cite this article in press as: Ziel, F., & Liu, B., Lasso estimation for GEFCom2014 probabilistic electric load forecasting. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2016.01.001

extended version of GEFCom2014-L. Here, we refer this competition as GEFCom2014-E. The competition included five tasks, in each of which the participants were asked to forecast the next year of hourly loads and submit the forecasts as 99 quantiles. The historical dataset for the first task was six years (2004–2009) of hourly temperature data and four years (2006–2009) of hourly load data. Each of the remaining four tasks then included an additional year of hourly load and temperature data for the period forecast as the previous task. The data for GEFCom2014-E are also provided by Hong et al. (2016). Florian Ziel joined this competition using the methodology proposed here, and ranked second out of 16 participating teams.

The structure of this paper is as follows: Section 2 introduces the time series model; Section 3 discusses the lasso estimation algorithm; Section 4 describes two benchmarks that are developed from the methodology used by Bidong Liu to win a place in the top eight in GEFCom2014-L; and Section 5 presents the empirical results. The paper is concluded in Section 6.

2. Time series model

Let $(\boldsymbol{Y}_t)_{t\in\mathbb{Z}}$, with $\boldsymbol{Y}_t = (Y_{\mathcal{L},t}, Y_{\mathcal{T},t})'$, be the $d = 2$-dimensional time series model of interest, and denote $\mathcal{D} = \{\mathcal{L}, \mathcal{T}\}$. Thus, $Y_{\mathcal{L},t}$ is the electric load and $Y_{\mathcal{T},t}$ is the temperature at time point t.

For $(\boldsymbol{Y}_t)_{t\in\mathbb{Z}}$, the joint multivariate time-varying threshold AR model (VAR) considered is given by

$$Y_{i,t} = \phi_{i,0}(t) + \sum_{j\in\mathcal{D}}\sum_{c\in C_{i,j}}\sum_{k\in I_{i,j,c}} \phi_{i,j,c,k}(t)\max\{Y_{j,t-k}, c\} + \varepsilon_{i,t} \quad (1)$$

for $i \in \mathcal{D}$, where $\phi_{i,0}$ are the time-varying intercepts and $\phi_{i,j,k,c}$ are time-varying autoregressive coefficients. Moreover, $C_{i,j}$ are the sets of all thresholds considered, $I_{i,j,c}$ are the index sets of the corresponding lags, and $\varepsilon_{i,t}$ is the error term. We assume that the error process is uncorrelated, with a zero mean and constant variance.

Furthermore, it is important that we are using the whole dataset with all hours to model the hourly load and temperature, instead of using a dataset that is sliced by hour to model the loads of specific hours, as is often done in literature. Forecasting algorithms applied to the whole dataset can learn about those events better, since the full dataset is more informative than the small hourly datasets.

The modeling process has three crucial components: the choice of the thresholds sets $C_{i,j}$, the choice of the lag sets $I_{i,j,k}$ and the time-varying structure of the coefficient. We describe these issues in the following three subsections.

2.1. Choice of the threshold sets

The choice of the threshold sets $C_{i,j}$ will characterize the potential non-linear impacts in the model. Note that if we choose $C_{i,j} = \{-\infty\}$, the model in Eq. (1) will turn into a standard multivariate time-varying AR process.

For load data, the temperature typically has a non-linear effect on the electric load. Fig. 1 shows the temperature at 00:00 of every day in the sample against the corresponding load. In general, we observe a decreasing relationship for lower temperatures and an increasing one for higher temperatures. To emphasize the non-linear relationship, we added the fitted line of the toy example regression

$$Y_{\mathcal{L},t} = c_0 + c_1 Y_{\mathcal{T},t} + c_2\max\{Y_{\mathcal{T},t}, 50\} + c_3\max\{Y_{\mathcal{T},t}, 60\} + \epsilon_t. \quad (2)$$

This is a simple threshold model, with thresholds at 50 °F and 60 °F.

In Fig. 1, we see that the threshold model in Eq. (2) captures the relationship using piecewise linear functions. Even though this is just an illustrative example, we see that this type of model is able to approximate all non-linear relationships between the load and temperature.

We can also introduce many other thresholds into the model in order to increase the flexibility. However, this enlarges the parameter space, which results in longer computation times and raises the concern of over-fitting. The lasso estimation algorithm can help to ease these two concerns. Even better, it will keep only significant non-linear impacts.

We choose the threshold sets manually for both data sets. For the GEFCom2014-L data, we consider $C_{\mathcal{L},\mathcal{T}} = \{-\infty, 20, 30, 40, 45, 50, 55, 60, 65, 70, 80\}$ as thresholds of the temperature to electric load impact, and $C_{\mathcal{L},\mathcal{L}} = \{-\infty, 100, 125, 150, 175, 200, 225\}$ for the load to load effects. Remember that the thresholds corresponding to $-\infty$ model the linear effects. For the other sets, we assume no non-linear effects, so $C_{\mathcal{T},\mathcal{L}} = C_{\mathcal{T},\mathcal{T}} = \{-\infty\}$. For the GEFCom2014-E data, we are use different thresholds, as the scale is different. In detail, we use $C_{\mathcal{L},\mathcal{T}} = \{-\infty, 10, 20, 30, 40, 45, 50, 60, 70, 80\}$, $C_{\mathcal{L},\mathcal{L}} = \{-\infty, 2500, 3000, 3500, 4000, 4500\}$ and $C_{\mathcal{T},\mathcal{L}} = C_{\mathcal{T},\mathcal{T}} = \{-\infty\}$ for the thresholds sets. Note that, in general, a data-driven threshold set selection is plausible as well, e.g., using a selected set of quantiles.

2.2. Choice of the relevant lag sets

The lag sets $I_{i,j,c}$ are essential for a good model, as they characterize the causal structure of the processes and the potential memory of the process. The lags in $I_{i,j,c}$ describe a potential lagged impact of the regressor j at threshold c to the process i. It is widely known that the load at time t is related to both its past and the temperature. Therefore, we choose $I_{\mathcal{L},\mathcal{L},c}$ and $I_{\mathcal{L},\mathcal{T},c}$ to be non-empty for all c. For the temperature, the situation is slightly different. Here, we assume that the temperature depends on its past, so $I_{\mathcal{T},\mathcal{T},-\infty}$ is non-empty as well. However, it is clear that the electric load does not effect the temperature, so $I_{\mathcal{T},\mathcal{L},-\infty}$ is empty.

The selected index sets are given in Table 1. Here, similarly as for the threshold sets, larger sets increase the parameter space, thus increasing the computational burden. However, they have to be chosen to be large enough to capture the relevant information. $I_{\mathcal{L},\mathcal{L},-\infty}$ contains all lags up to 1200, so the maximal memory is the preceding 1200 h, which is slightly more than seven weeks. The most essential part is that the important lags of orders, such as 1, 24, 48 and 168, are included. A detailed discussion of the choice of the index sets is provided by Ziel, Steinert, and Husmann (2015).

Please cite this article in press as: Ziel, F., & Liu, B., Lasso estimation for GEFCom2014 probabilistic electric load forecasting. International Journal of Forecasting (2016), http://dx.doi.org/10.1016/j.ijforecast.2016.01.001