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/© 2016 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

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

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

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

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

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.

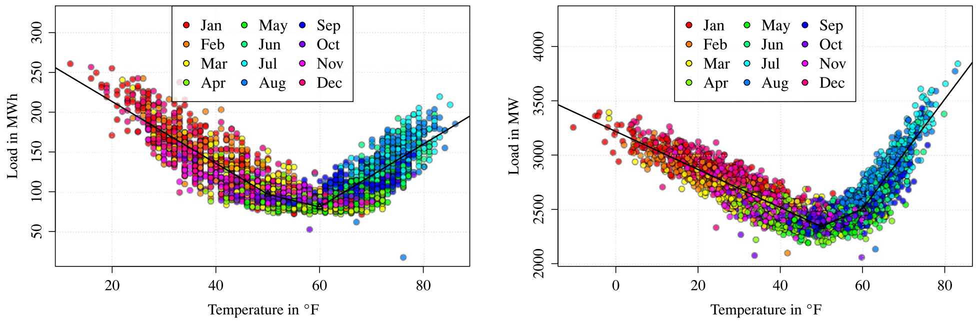
We can also introduce many other thresholds into the model in order to increase the flexibility. However, this enlarges the parameter space, which results in longer computation times and raises the concern of over-fitting. The lasso estimation algorithm can help to ease these two concerns. Even better,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



(a) GEFCom2014-L data.

(b) GEFCom2014-E data.

Fig. 1. Temperature against load for all days at 00:00, with fitted values of Eq. (2) for both data sets.

Table 1
Considered lags of the required index sets.

Index sets	Contained lags
$I_{\mathcal{L},\mathcal{L},-\infty}$	$1, \dots, 1200$
$I_{\mathcal{L},\mathcal{L},c}$ (with $c \neq -\infty$), $I_{\mathcal{L},\mathcal{T},c}$	$1, \dots, 200$
$I_{\mathcal{T},\mathcal{T},-\infty}$	$1, \dots, 360$
$I_{\mathcal{T},\mathcal{L},-\infty}$	–

Table 2
List of all groups $\mathcal{G}_1, \dots, \mathcal{G}_8$ of basis functions considered.

Group	Description
$\mathcal{G}_1$	Hourly impacts on the seasonal daily pattern
$\mathcal{G}_2$	Hourly impacts on the seasonal weekly pattern
$\mathcal{G}_3$	Daily impacts on the seasonal annual pattern
$\mathcal{G}_4$	Smooth annual impacts
$\mathcal{G}_5$	Long term trend effects
$\mathcal{G}_6$	Fixed date public holidays effects
$\mathcal{G}_7$	Varying date public holidays effects
$\mathcal{G}_8$	Interaction effects between $\mathcal{G}_1$ and $\mathcal{G}_4$

2.3. The time-varying coefficients

The assumed structure of the time-varying coefficients is of substantial importance as well, as they have big impacts not only on the seasonality and public holiday effects, but also on the long term trend behavior. However, we keep most of the coefficients constant, allowing only the important ones to vary over time. The intercepts $\phi_{i,0}$ in Eq. (1) are important and are allowed to vary over time for both the load and the temperature. For the load, we additionally allow $\phi_{\mathcal{L},\mathcal{L},-\infty,k}$ with $k \in \{1, 2, 24, 25\}$ to vary over time, and for the temperature, $\phi_{\mathcal{T},\mathcal{T},-\infty,k}$ with $k \in \{1, 2\}$. Thus, in total, the two intercepts and the four autoregressive load and two autoregressive temperature coefficients are allowed to vary over time. Obviously, this choice can be modified based on a knowledge of the important parameters. Again, the more the parameters vary over time, the larger the parameter space, which increases both the computation time and the risk of over-fitting.

For the time varying coefficients, we assume a structure similar to that of Ziel et al. (2015). For a time-varying parameter of interest ξ (e.g., $\phi_{i,0}$ or $\phi_{i,j,c,k}$), we assume that

$$\xi(t) = \xi_0 + \boldsymbol{\xi}'\boldsymbol{B}^{\xi}(t) = \xi_0 + \sum_{l=1}^{N_\xi} \xi_l B_l^{\xi}(t), \tag{3}$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{N_\xi})'$ is the vector of coefficients that applies to the basis functions $\boldsymbol{B}^{\xi} = (B_1^{\xi}, \dots, B_{N_\xi}^{\xi})'$. Obviously, the sum in Eq. (3) is empty for constant parameters.

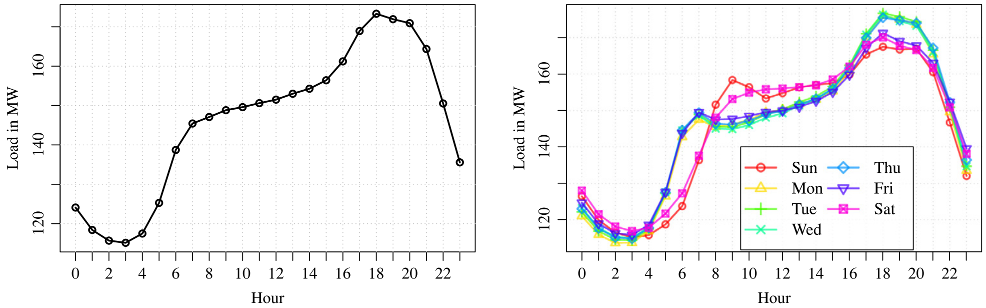
The basis functions of the time-varying coefficients have to be chosen accurately. The selection is modular, meaning that several effects can be added and merged easily. We consider a selection of several groups of regressors, as listed in Table 2.

Below, we explain the groups $\mathcal{G}_1, \dots, \mathcal{G}_8$ one by one. The daily and weekly mean electric loads of the GEFCom2014-L data are given in Fig. 2. Fig. 2(a) shows the clear, distinct seasonal daily pattern, with low values at night and high values during the day. The group $\mathcal{G}_1$ will cover this effect. Obviously, this requires 24 parameters. However, Fig. 2(b) shows that the Saturdays and Sundays show behaviors that differ from those of the typical working days from Monday to Friday, for which we see basically the same behaviors every day. Nevertheless, Monday morning and Friday evening do show transition effects towards and from the weekend. $\mathcal{G}_2$ will cover the full weekly structure, with 168 parameters being required. As has been mentioned, there is redundancy in the pattern; e.g., the Tuesdays, Wednesdays and Thursdays generally exhibit similar behaviors. This structure is taken into account automatically when using the regressors $\mathcal{G}_1$ and $\mathcal{G}_2$ in combination with the lasso estimation technique. The basis functions of groups $\mathcal{G}_1$ and $\mathcal{G}_2$ are defined as

$$B_k^{\mathcal{G}_1}(t) = \begin{cases} 1, & k \leq \text{HoD}(t) \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad B_k^{\mathcal{G}_2}(t) = \begin{cases} 1, & k \leq \text{HoW}(t) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\text{HoD}(t)$ and $\text{HoW}(t)$ give the hour-of-the-day $(1, 2, \dots, 24)$ and hour-of-the-week $(1, 2, \dots, 168$, start counting at 0:00 on Sunday) of time point t. Note that the parametrization in Eq. (4) is done using cumulative components. Therefore, the "$\leq$" relation is used instead of the commonly used "$=$" relation. As an example, $B_2^{\mathcal{G}_1}$ models the additional impact of hour 1:00 on hour 0:00 (which is modeled by $B_1^{\mathcal{G}_1}$), instead of modeling the direct impact of hour 1:00, which would be associated with the "$=$" relation in Eq. (4). In other words, we are modeling the changes in the impacts associated with an hour, instead of the absolute effects. Our estimation method will mean

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



(a) Hourly mean load during a day.

(b) Hourly mean load during a week.

Fig. 2. Hourly mean load during a day (a) and week (b) of the GEFCom2014-L data.

that a parameter is included in the model only if the corresponding change is significant.

Similarly to the daily and weekly patterns, there is also an annual seasonal pattern. To capture this, we introduce

$$B_k^{\mathcal{G}_3}(t) = \begin{cases} 1, & k \leq \text{DoY}(t) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\text{DoY}(t)$ gives the day-of-the-year $(1, 2, \dots, 365)$ of time point t in a common year with 365 days. In a leap year, $\text{DoY}(t)$ also takes values from $(1, 2, \dots, 365)$, but the 29th February has the same value (namely 59) as the 28th February. Similarly to above, we model the changes in the annual pattern, not the direct impact.

The next group of basis functions concerns smooth annual impacts. This will capture effects similar to those in $B_k^{\mathcal{G}_3}$, but in a smoother manner. We consider periodic B-splines, which results in a local modeling approach. Specifically, we use cubic B-splines with a periodicity of $8765.76 = 24 \times 365.24$ on an equidistant grid with six basis functions. In Fig. 3(a), we see these basis functions on a time range of three years. We clearly observe the local impact. Thus, for example, the dashed yellowish function $(k = 2)$ covers only effects in the summer, but has no impact in the winter.

The most tricky basis function group relates to the long term effects. The challenging part is the distinction between spurious effects and real long term changes in the load behavior. The spurious effect problem is crucial for long term forecasting, whereas it is negligible for shorter time horizons. To make the problem clear, suppose that the available time series ends in 31th December, and the last two months, November and December, had low load values for some unknown reason. Now, the question is whether this was a random effect (just a realization of rare or outlier events) or a structural change in the load level (induced, for example, by an increase in energy efficiency that is not captured by external regressors). Conservatively, statistical modeling would typically suggest that it is a random effect unless the structural change is significant enough to be detected by the modeling approach.

We model long term effects using monotonically increasing basis functions. These are constant in the past, then strictly monotonically increasing in a certain time range in which the long term transition effect might have taken place, then constant after this possible transition. The time range in which the basis function is monotonically increasing should be larger than a year, in order to reduce the probability of including spurious effects. Furthermore, the distances between these basis functions should be relatively large as well. We consider a distance of one year between the basis functions, with a support of two years for the transition effect. Specifically, we use cumulative quadratic B-splines as basis functions for the long term effects. We consider only basis functions where the in-sample basis functions take a smallest value of at least 10% of the overall maximum and at most 90% of the overall maximum. This will reduce the danger of modeling a spurious effect. We end up with only a few basis functions. An illustrative example for an in-sample period of 12 years (2001–2012), with 2013 as the out-of-sample year, is given in Fig. 3(b). Note that the number of long term basis functions in group $\mathcal{G}_5$ depends on the data range.

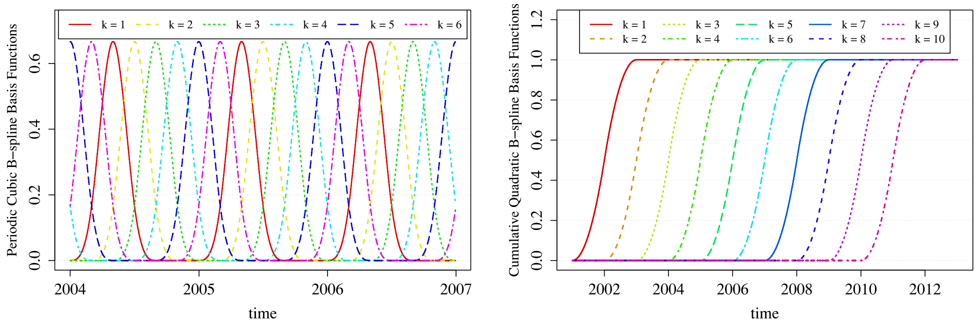
The next two groups, $\mathcal{G}_6$ and $\mathcal{G}_7$, contain the public holiday information. In general, the electric load exhibits special behaviors on public holidays, which eventually disturb the standard weekly pattern. For modeling purpose, we group the public holidays into two classes: those with fixed dates, such as New Year's Day (January 1), and those with flexible dates, such as Thanksgiving Day (fourth Thursday in November). We consider all United States federal public holidays, and denote the sets of public holidays with fixed and flexible dates as $\mathbb{F}$ix and $\mathbb{F}$lex.

As days in $\mathbb{F}$lex are always on a specific weekday, we can expect the same behavior on these public holidays each year. If a week includes a public holiday, then the typical weekly structure in Fig. 2(b) changes. Not only is the structure of the public holiday affected, the hours before and after the public holiday are also affected, due to transition effects. Therefore, for each flexible public holiday $F \in \mathbb{F}\text{lex}$, we define a basis of $6+24+6 = 36$ h (six hours before F, 24 h at F, and six hours after F). Specifically, it is given by

$$B_k^{F}(t) = \begin{cases} 1, & k \leq \text{Ho}F(t) \\ 0, & \text{otherwise} \end{cases}$$

where $\text{Ho}F(t)$ gives the hours from $1, 2, \dots, 36$ at time point t around the public holiday, starting counting from 18:00.

The impact of the days in $\mathbb{F}$ix is complex, because it depends on the weekday of incidence. Some research has found that it is usually similar to that of a Sunday (see



(a) Periodic cubic B-spline basis within three years.

(b) Cumulated quadratic B-spline basis on 12 years.

Fig. 3. Illustration of basis functions for $\mathcal{G}_4$ and $\mathcal{G}_5$.

e.g. Ziel et al., 2015). We will introduce an effective coefficient $C(t)$ for each hour of the week. Using $C(t)$, we can define the basis functions for $H \in \mathbb{F}\text{ix}$

$$B_k^H(t) = \begin{cases} C(t), & k \leq \text{Ho}H(t) \\ 0, & \text{otherwise}, \end{cases}$$

where $\text{Ho}H(t)$ gives the hours around the public holidays from $1, 2, \ldots, 36$ at time point t, starting counting from 18:00. The coefficients $C(t)$ are defined as follows. If the public holiday is on a Sunday, then the effective coefficient is zero, assuming that there is no additional impact of the public holiday on a Sunday. Thus, we refer to these 24 hourly mean load values as the low level load target. If such a public holiday occurs during the core working days, such as Tuesday, Wednesday or Thursday, we expect a full impact, with an effective coefficient of one. We refer to the 24 hourly mean load values of these three days as the high level load target. If the holiday happens on a Monday, Friday or Saturday, the impact should be between the two situations above, and the effective coefficient is usually between zero and one. If we denote the hourly mean load of the week from Fig. 2(b) by the actual load target, then we define the coefficients as $C(t) = \max\{1 - \frac{\text{high level load target}(t) - \text{actual load target}(t)}{\text{high level load target}(t) - \text{low level load target}(t)}, 1\}$.

The last group of basis functions focuses on interaction effects, which are important for the temperature modeling. As the length of the night changes over the year, the daily seasonal pattern changes over the year as well. We create the interaction group by multiplying the basis function of each group by that of another group. Thus, the interaction groups tend to require many parameters. For that reason, for the last group $\mathcal{G}_8$ we consider only the multiplication of the daily seasonal component $\mathcal{G}_1$ with the smooth annual basis functions $\mathcal{G}_4$. Specifically, $\mathcal{G}_8$ contains the basis functions $B_{24(j-1)+i}^{\mathcal{G}_8}(t) = B_i^{\mathcal{G}_1}(t)B_j^{\mathcal{G}_4}(t)$ for $i \in \{1, \ldots, 24\}$ and $j \in \{1, \ldots, 6\}$.

For all basis function groups, we can define the full basis function vector $\boldsymbol{B}^{\xi}$ for a parameter ξ. Hence, the basis functions for a time-varying parameter $\xi_{\mathcal{L}}$ associated with the load are given by $\boldsymbol{B}^{\xi_{\mathcal{L}}} = (\boldsymbol{B}^{\mathcal{G}_1}, \boldsymbol{B}^{\mathcal{G}_2}, \boldsymbol{B}^{\mathcal{G}_3}, \ldots, \boldsymbol{B}^{\mathcal{G}_8})$, where $\boldsymbol{B}^{\mathcal{G}_1} = (B_1^{\mathcal{G}_1}, \ldots, B_{24}^{\mathcal{G}_1})$, $\boldsymbol{B}^{\mathcal{G}_2} = (B_1^{\mathcal{G}_2}, \ldots, B_{168}^{\mathcal{G}_2})$, $\boldsymbol{B}^{\mathcal{G}_3} = (B_1^{\mathcal{G}_3}, \ldots, B_{365}^{\mathcal{G}_3})$, $\boldsymbol{B}^{\mathcal{G}_4} = (B_1^{\mathcal{G}_4}, \ldots, B_6^{\mathcal{G}_4})$, ... define the vectors of the basis functions. For the time-varying parameters $\xi_{\mathcal{T}}$ of the temperature modeling process, we define $\boldsymbol{B}^{\xi_{\mathcal{T}}} = (\boldsymbol{B}^{\mathcal{G}_1}, \boldsymbol{B}^{\mathcal{G}_4}, \boldsymbol{B}^{\mathcal{G}_8})$. Thus, only daily and smooth annual effects and their interactions are allowed. In particular, we do not include any weekly, public holiday or long term effects when modeling the temperature.

3. Estimation and forecasting method

The introduction mentions that we use a lasso estimation technique, which is a penalized ordinary least squares regression estimator. The ordinary least squares (OLS) representation of Eq. (1) is given by

$$\mathcal{Y}_i = \mathcal{X}_i\boldsymbol{\beta}_i + \mathcal{E}_i. \tag{6}$$

Here, $\mathcal{Y}_i = (Y_{i,1}, \ldots, Y_{i,n})'$, $\mathcal{X}_i$ is the $n \times p_i$-dimensional regressor matrix that corresponds to Eq. (1), $\boldsymbol{\beta}_i$ is the full parameter vector of length p_i, $\mathcal{E}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n})'$ is the residual vector, and n is the number of observations. However, we perform a lasso estimation not on Eq. (6) directly, but on its standardized version. Therefore, we standardize Eq. (6) so that the regressors and the regressand all have a variance of one and a mean of zero. Thus, we obtain the standardized version of Eq. (6):

$$\widetilde{\mathcal{Y}}_i = \widetilde{\mathcal{X}}_i\widetilde{\boldsymbol{\beta}}_i + \widetilde{\mathcal{E}}_i. \tag{7}$$

We can easily compute $\boldsymbol{\beta}_i$ by rescaling, if $\widetilde{\boldsymbol{\beta}}_i$ is determined. The lasso optimization problem of Eq. (7) is given by

$$\widehat{\widetilde{\boldsymbol{\beta}}}_i = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p_i}} \|\widetilde{\mathcal{Y}}_i - \widetilde{\mathcal{X}}_i\boldsymbol{\beta}\|_2^2 + \lambda_i\|\boldsymbol{\beta}\|_1, \tag{8}$$

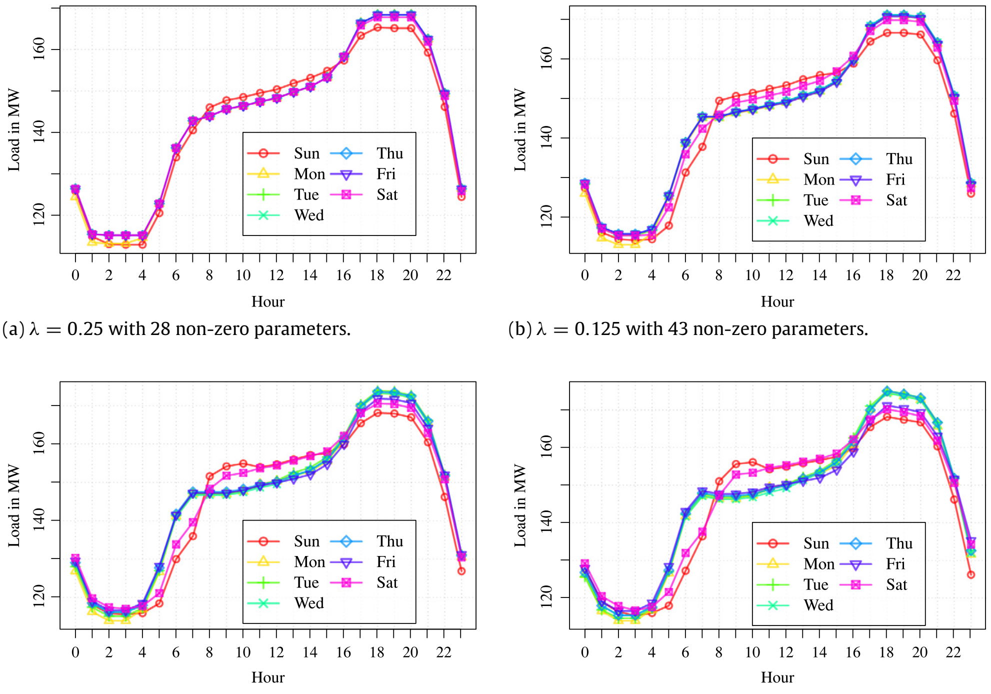
with tuning parameters of λ_i, and $\|\cdot\|_1$ and $\|\cdot\|_2$ as the ℓ_1- and ℓ_2-norm. For $\lambda_i = 0$, Eq. (8) is the standard OLS problem. For huge λ_i values, we have a huge penalty on the parameters and receive the estimator $\widehat{\widetilde{\boldsymbol{\beta}}}_i = \boldsymbol{0} = (0, \ldots, 0)'$, meaning that no parameters are included in the model. In a moderate range of λ_i values, we get different solutions. Thus, it holds that the larger λ_i is, the fewer parameters are included in the estimated model.

To obtain a better understanding of this feature, we consider a simple lasso problem, given by

$$\|\mathcal{Y}_i - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{9}$$

where $\mathbb{X}$ is the regressor matrix that contains the 24 basis functions of $\mathcal{G}_1$ and the 168 basis functions of $\mathcal{G}_2$. We remember that the OLS solution of this problem corresponds to Fig. 2(b) and requires 168 parameters to capture all of the effects fully. In Fig. 4, we plot the fitted

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



(a) λ = 0.25 with 28 non-zero parameters.

(b) λ = 0.125 with 43 non-zero parameters.

(c) λ = 0.0625 with 80 non-zero parameters.

(d) λ = 0.03125 with 101 non-zero parameters.

Fig. 4. Fitted model for Eq. (9) for selected λ values and the corresponding numbers of non-zero parameters.

values of the solution of Eq. (9) for four different λ values. As has been mentioned, we see that the smaller λ is, the more parameters are included in the model, and therefore the closer the solution gets to Fig. 2(b). For example, in Fig. 4(c), we observe a pattern that does not differ from that in Fig. 2(b) in any way that is easy to observe by eye-balling, even though only 80 parameters are required to capture the structure, instead of 168. In contrast, Fig. 4(a), with only 28 parameters, does not cover the pattern well; thus, for example, the seasonal pattern is the same during the morning and noon hours for all days except Sundays. This indicates that the 28-parameter solution does not include enough parameters for appropriate modeling.

Note that the shrinkage property of the lasso is relevant as well as the selection property. For example, if we have the lasso solution in Fig. 2(b) with 80 non-zero parameters, then this is different from the OLS solution of the corresponding 80 regressors. In general, the lasso solution tends to have smaller estimated parameters (in terms of absolute values) than the OLS solution, due to the shrinkage towards **0**. Specifically, the in-sample residual sums of squares (RSS) are always larger for the lasso solution than for the OLS solution. Thus, even though there might be many non-zero parameters in the final estimated model, their contributions to the model are small. This shrinkage property reduces the parameter uncertainty, and may give a better out-of-sample performance.

In general, the tuning parameters λ_i should be chosen using a selection algorithm. Usually, the optimal λ_i will be chosen from a given grid Λ_i by minimizing an information criterion. We select the tuning parameter using the minimum Bayesian information criterion (BIC), which is a conservative information criterion that avoids over-fitting. For the grid Λ_i, we choose an exponential grid, as was suggested by Friedman, Hastie, and Tibshirani (2010).

As the computation algorithm, we consider the fast coordinate descent algorithm and the corresponding R package functions of the `glmnet` package; see e.g. Friedman et al. (2010) for more details. The asymptotic computational complexity of the coordinate descent algorithm is only $\mathcal{O}(np_i)$. This is optimal, as np_i is the number of elements in the regression matrix. Thus, we can estimate the model efficiently and can carry out model selection easily. Another positive feature is that we do not require the data set to be divided into training and test data sets, as we can tune the model based on statistical theory (like the BIC).

For each forecasting task, we use all available data for the lasso estimation procedure. Given the estimated model, we can use a residual-based bootstrap to simulate future scenario sample paths, as per (Ziel et al., 2015). We consider a total of $N = 10\,000$ sample paths here. The corresponding empirical percentiles are used as estimates for the target quantiles.

4. Benchmarks

The scenario-based probabilistic forecasting methodology proposed by Hong, Wilson, and Xie (2014b) was used by two teams from among the top eight (Jingrui Xie, third; Bidong Liu, eighth) in GEFCom2014-L. In this paper, we develop two benchmarks using this method, with two underlying models. The first one is Tao's Vanilla Benchmark model, as used in GEFCom2012 (Hong, Pinson, & Fan, 2014a), abbreviated as *Vanilla* in this paper. The second one is a recency effect model proposed by Wang, Liu, and

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

Hong (2016), abbreviated as *Recency* in this paper. In the GEFCom2014-L case study, instead of performing weather station selection as was discussed by Hong, Wang, and White (2015), we create a temperature series by averaging the 25 weather stations, in order to keep the benchmarks simple and easily reproducible. Note that this is different from the way in which the temperature series is created when implementing the lasso-based methodology as discussed in Section 1.

4.1. Vanilla model

The *Vanilla* model for the load $Y_{\mathcal{L},t}$ is given as:

$$\begin{aligned} Y_{\mathcal{L},t} &= \beta_0 + \beta_1 \text{MoY}(t) + \beta_2 \text{DoW}(t) + \beta_3 \text{HoD}(t) \\ &\quad + \beta_4 \text{DoW}(t)\text{HoD}(t) + f(Y_{\mathcal{T},t}) + \epsilon_t, \end{aligned} \tag{10}$$

where β_i are the regression coefficients, $\text{MoY}(t)$ gives the month-of-the-year $(1, \ldots, 12)$ of time t, $\text{DoW}(t)$ gives the day-of-the-week $(1, \ldots, 7$, with Sunday $= 1$, Monday $= 2, \ldots)$ of time t, $\text{HoD}(t)$ gives the hour-of-the-day $(1, \ldots, 24)$ of time t as for Eq. (4), and

$$\begin{aligned} f(Y_{\mathcal{T},t}) = {} & \beta_5 Y_{\mathcal{T},t} + \beta_6 Y^2_{\mathcal{T},t} + \beta_7 Y^3_{\mathcal{T},t} + \beta_8 Y_{\mathcal{T},t}\text{MoY}(t) \\ & + \beta_9 Y^2_{\mathcal{T},t}\text{MoY}(t) + \beta_{10} Y^3_{\mathcal{T},t}\text{MoY}(t) \\ & + \beta_{11} Y_{\mathcal{T},t}\text{HoD}(t) + \beta_{12} Y^2_{\mathcal{T},t}\text{HoD}(t) \\ & + \beta_{13} Y^3_{\mathcal{T},t}\text{HoD}(t). \end{aligned} \tag{11}$$

Here, for task 1, we are using the model specified in Eq. (10) as the underlying model, with the parameters estimated using the most recent 24 months (from 01/2009 to 12/2010) of hourly loads and temperatures. The 10 years of weather history (2001–2010) are used to generate 10 weather scenarios. In total, we have 10 load forecasts for each hour in 01/2011. We then compute the 99 quantiles required based on these 10 forecasts using the empirical distribution function. The 99 quantiles for the other 11 months of 2011 are generated similarly. For instance, when forecasting the load for 05/2011, the 24 months of hourly loads and temperatures from 05/2009 to 04/2011 are used for parameter estimation.

4.2. Recency model

The underlying model for the second benchmark is given as:

$$\begin{aligned} Y_{\mathcal{L},t} = {} & \beta_0 + \beta_1 \text{MoY}(t) + \beta_2 \text{DoW}(t) + \beta_3 \text{HoD}(t) \\ & + \beta_4 \text{DoW}(t)\text{HoD}(t) + f(Y_{\mathcal{T},t}) + \sum_{j \in \mathcal{J}} f(\widetilde{Y}_{\mathcal{T},t,j}) \\ & + \sum_{k \in \mathcal{K}} f(Y_{\mathcal{T},t-k}) + \epsilon_t, \end{aligned} \tag{12}$$

where f is as in Eq. (11) and the daily moving average temperature of the jth day $\widetilde{Y}_{\mathcal{T},t,j}$ is defined as

$$\widetilde{Y}_{\mathcal{T},t,j} = \frac{1}{24} \sum_{h=24j-23}^{24j} \widetilde{Y}_{\mathcal{T},t-h}. \tag{13}$$

The sets $\mathcal{J}$ and $\mathcal{K}$ in Eq. (12) are given by $\mathcal{J} = \{1, \ldots, J\}$ and $\mathcal{K} = \{1, \ldots, K\}$ for $J > 0$ and $K > 0$; they are empty if $J = 0$ and $K = 0$. Note that for $(J, K) = (0, 0)$ we receive the *Vanilla* model in Eq. (10). The 'average-lag' pair (J, K) needs to be identified before the *Recency* model can be used to generate forecasts for the target month. Since the pattern of load against temperature varies each year, the optimal pair selected changes every year correspondingly. To identify the optimal pair for the year i, we use the data from years $(i - 3)$ and $(i - 2)$ for training, and those from year $(i-1)$ for validation. The pair that results in the lowest mean absolute percentage error (MAPE) in the validation period will be selected, and the corresponding *Recency* model will then be used to forecast the year i. We search for the optimal (J, K) on the grid $\{0, \ldots, 7\} \times \{0, \ldots, 48\}$. The optimal pair identified for the year 2011 using this method for the GEFCom2014-L data is (2, 10).

Table 3
The optimal pairs of (J, K) for the years from 2010 to 2014 in GEFCom2014-E.

Year	2010	2011	2012	2013	2014
J	1	1	1	1	0
K	9	0	8	13	13

In the GEFCom2014-E case study, the target years are from 2010 to 2014. The optimal pairs identified are listed in Table 3. After identifying the optimal pairs of (J, K), we follow the same steps as for the first benchmark discussed in Section 4.1, including two years of hourly loads and temperatures for parameter estimation and an empirical distribution function for extrapolating the 99 quantiles. But we use a recency model as the underlying model to do forecasting, instead of the vanilla model. When creating weather scenarios, we use 6 years (2004–2009) weather data for the target year of 2010, 7 years (2004–2010) for 2011, 8 years (2004–2011) for 2012, 9 years (2004–2012) for 2013 and 10 years (2004–2013) for 2014.

In order to keep the benchmarks simple and easy to reproduce, neither of the underlying models incorporate any other special treatments such as weather station selection, data cleansing, weekend and holiday effect modeling, or forecast combination.

5. Empirical results and discussion

We evaluate the forecasting performances based on the overall mean pinball loss function of the 99 percentiles. For more details on the pinball loss function and the evaluation methods used in GEFCom2014-L, see Hong et al. (2016).

5.1. GEFCom2014-L results

As an illustrative example, the 99 quantiles predicted for the April 2011 task are given in Fig. 5. We observe that the daily and weekly seasonal behaviors are captured well. Furthermore, the prediction intervals get wider with the increasing forecasting horizon, as expected.

The pinball scores of the proposed model (Lasso) and the two benchmarks are given in Table 4. We also list Bidong Liu's original GEFCom2014-L scores in the last column under BL. The main factors causing the differences between the two benchmarks and BL include the length of the training data and the extrapolation method. In

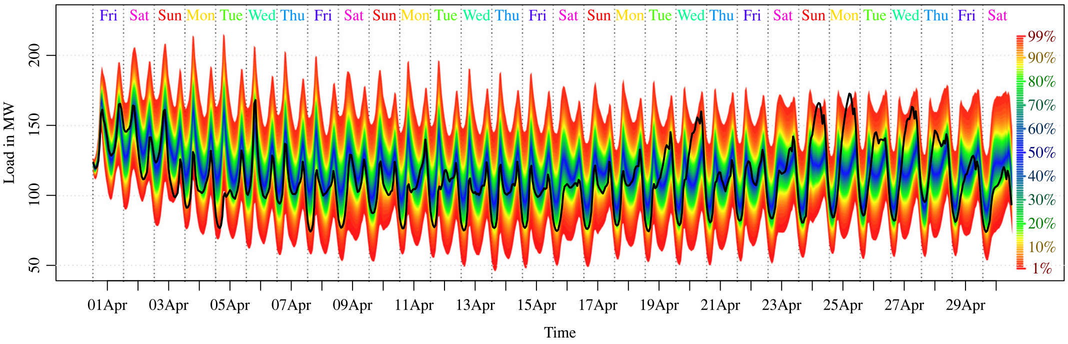


Fig. 5. April forecast of the GEFCom2014-L data with the corresponding legend and observed values (black line).

Table 4
Overall pinball scores for the GEFCom2014-L data.

Month	Lasso	*Vanilla*	*Recency*	BL	*Vanilla*-5Y
1	**9.88**	11.94	12.13	16.42	11.78
2	**9.54**	10.95	10.57	11.87	11.24
3	**7.97**	8.57	8.38	9.37	8.70
4	4.89	5.05	**4.80**	5.62	5.67
5	**5.96**	7.37	7.11	7.74	7.98
6	**5.86**	6.75	7.35	6.55	6.48
7	**7.66**	9.60	9.38	9.14	9.08
8	**10.70**	11.21	11.30	11.35	11.36
9	6.28	5.81	**5.65**	6.51	6.19
10	5.20	3.53	**3.40**	4.80	4.53
11	6.38	6.06	**5.93**	6.97	6.50
12	**8.99**	9.74	9.45	10.89	10.29
Average	**7.44**	8.05	7.95	8.94	8.32

Table 5
Overall pinball scores for the GEFCom2014-E data.

Year	Lasso	FZ	*Vanilla*	*Recency*
2010	59.01	**58.02**	85.03	80.76
2011	**49.74**	54.50	59.54	56.77
2012	47.08	**46.51**	57.58	55.37
2013	62.53	63.71	62.59	**60.62**
2014	55.00	**52.25**	59.16	56.82
Average	**54.69**	55.00	64.78	62.07

GEFCom2014-L, Bidong Liu implemented the scenario-based method as described in Section 4 for months 2–12, but not for month 1. For parameter estimation, Bidong Liu used five years of historical data for most of the tasks during GEFCom2014-L. In addition, the required quantiles were generated by linear extrapolation. For illustrative purposes, we also list the pinball scores from the Vanilla benchmark estimated using five years of data in Table 4 under *Vanilla*-5Y.

We observe that the proposed lasso estimation method outperforms the two benchmarks, i.e., *Vanilla* and *Recency*, in 9 and 8 months out of 12, respectively. The reductions in the 12-month average pinball score are 6.4% and 7.6% relative to the *Recency* and *Vanilla* models, respectively. Although BL ranked among the top eight in GEFCom2014-L, its average pinball score is higher than those of the other four methods. The average pinball score of *Vanilla*-5Y (8.32) is higher than that of *Vanilla* (8.05), which reveals the necessity of selecting the right length for the training data.

5.2. GEFCom2014-E results

The pinball scores of the proposed method (Lasso) and the two benchmarks in the GEFCom2014-E case study are given in Table 5. We also provide the original scores of Florian Ziel (FZ) in the GEFCom2014-E. The FZ scores differ from those of the Lasso slightly, because the long term trend components ($\mathcal{G}_5$) were added to the time-varying parameters of Lasso. No long term modeling was considered for FZ, but a manual long-term effect adjustment was done for the years 2012 and 2013. In addition, the list of holidays considered was extended by including some bridging holidays, such as Christmas Eve (24 December), Boxing Day (26 December) and New Year's Eve (31 December).

Similarly to the GEFCom2014-L results, the lasso outperforms the two benchmarks in four of the five years. The average reductions in the pinball score relative to the *Recency* and *Vanilla* models are 11.9% and 15.6%, respectively.

5.3. Discussion

Even though the proposed methodology outperforms two credible benchmarks, we may be able to improve it in several ways. One model assumption is the homoscedasticity of the residuals, but in practice the residuals are heteroscedastic. Usually, we observe lower levels of variation at night and during low load seasons. Thus, the heteroscedasticity of the residuals should be taken into account when designing the model. Ziel (in press) and Ziel et al. (2015) suggest the use of an iteratively reweighted lasso approach, incorporating the volatility of the residuals. Their results suggest that a significant improvement in the forecasting results can be achieved. It might also help to apply a normality assumption with group analysis, as was discussed by Xie, Hong, Laing, and Kang (in press), or a block bootstrap method, as was used by Fan and Hyndman (2012), in order to incorporate the remaining dependency structure in the residuals. Another issue is the tuning of the lasso itself. We simply considered the Bayesian information criterion, but other special cases of the generalized information criterion (GIC) might yield better forecasting performances. Lastly, for the GEFCom2014-L data, the treatment of the available temperature information might be improved. For instance, the weather station selection methodology proposed by Hong et al. (2015) might provide a better incorporation of the temperature data.

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

6. Summary and conclusion

We introduce a methodology based on lasso estimation that can estimate parameters for a large pool of candidate variables in order to capture several distinct and well-known stylized facts in load forecasting. The proposed methodology ranked second in GEFCom2014-E. Two empirical studies based on two recent probabilistic load forecasting competitions (GEFCom2014-L and GEFCom2013-E) demonstrate the superiority of the proposed method to two credible benchmarks.

References

Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, *27*(1), 134–141.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Hong, T., Pinson, P., & Fan, S. (2014a). Global energy forecasting competition 2012. *International Journal of Forecasting*, *30*(2), 357–363.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2015 and beyond. *International Journal of Forecasting* http://dx.doi.org/10.1016/j.ijforecast.2016.02.001 (this issue).

Hong, T., Wang, P., & White, L. (2015). Weather station selection for electric load forecasting. *International Journal of Forecasting*, *31*(2), 286–295.

Hong, T., Wilson, J., & Xie, J. (2014b). Long term probabilistic load forecasting and normalization with hourly information. *IEEE Transactions on Smart Grid*, *5*(1), 456–462.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 267–288.

Wang, P., Liu, B., & Hong, T. (2016). Electric load forecasting with recency effect: a big data approach. *International Journal of Forecasting* http://dx.doi.org/10.1016/j.ijforecast.2015.09.006 (in press).

Xie, J., Hong, T., Laing, D. T., & Kang, C. (2015). On normality assumption in residual simulation for probabilistic load forecasting. *IEEE Transactions on Smart Grid*, http://dx.doi.org/10.1109/TSG.2015.2447007 (in press).

Ziel, F. (2015). Iteratively reweighted adaptive lasso for conditional heteroscedastic time series with applications to ar-arch type processes. *Computational Statistics and Data Analysis*, http://dx.doi.org/10.1016/j.csda.2015.11.016 (in press).

Ziel, F., Steinert, R., & Husmann, S. (2015). Efficient modeling and forecasting of electricity spot prices. *Energy Economics*, *47*, 98–111.

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