Demand Forecasting at Low Aggregation Levels using Factored Conditional Restricted Boltzmann Machine

Elena Mocanu
Phuong H. Nguyen
Madeleine Gibescu
Department of Electrical Engineering
Eindhoven University of Technology, Netherlands
{e.mocanu, p.nguyen.hong, m.gibescu}@tue.nl

Emil Mahler Larsen
Pierre Pinson
Department of Electrical Engineering
Technical University of Denmark,
Lyngby, Denmark
{emlar, ppin}@elektro.dtu.dk

***Abstract*—The electrical demand forecasting problem can be regarded as a nonlinear time series prediction problem depending on many complex factors since it is required at various aggregation levels and at high temporal resolution. To solve this challenging problem, various time series and machine learning approaches have been proposed in the literature. As an evolution of neural network-based prediction methods, deep learning techniques are expected to increase the prediction accuracy by allowing stochastic formulations and bi-directional connections between neurons. In this paper, we investigate a newly developed deep learning model for time series prediction, namely Factored Conditional Restricted Boltzmann Machine (FCRBM), and extend it for electrical demand forecasting. The assessment is made on the EcoGrid dataset, originating from the Bornholm island experiment in Denmark, consisting of aggregated electric power consumption, local price and meteorological data collected from 1900 customers. The households are equipped with local generation and smart appliances capable of responding to real-time pricing signals. The results show that for the short-term (5 minute to 1 day ahead) prediction problems solved here, FCRBM outperforms the benchmark machine learning approach, i.e. Support Vector Machine.**

***Index Terms*—Demand Forecasting, Deep Learning, Factored Conditional Restricted Boltzmann Machine, Support Vector Machine.**

I. Introduction

The electrical demand forecasting problem, at various aggregation levels, can be regarded as a highly nonlinear time series prediction problem. The complexity of the consumers' energy producing and consuming technologies and the uncertainty in the influencing factors, yield frequent fluctuations. Traditionally, the short-term forecasting problem is referring to 1 hour and 15 minutes resolutions, but higher resolutions make the problem even more complicated. Moreover, urbanization and electrification trends show that the total energy demand will increase in the future, and the penetration of energy from renewable sources is increasing as well. Future smart grids need a system that can monitor, predict, schedule, learn and make decisions regarding local energy consumption and production in real time. Modeling and predicting energy consumption in smart buildings can provide valuable information to facilitate Demand Response (DR) or Demand Side Management(DSM) programs.

The short-term (electrical) energy demand forecasting problem was extensively pursued in the literature over decades by various traditional time series and machine learning methods. Some of these methods are used to predict consumption by correlating it with influencing variables, such as climate conditions or energy prices. Interested readers are referred to [1]–[4] for a more comprehensive discussion about building modeling with focus on electrical demand forecasting. Moreover, to account for the evolution of future building energy management systems, there are also some representative approaches which combine some of the above modeling methods to optimize predictive performance, such as semi-parametric regression models used to forecast the contribution of load from some non-linear variable [5], exponential smoothing [6], multivariate state-space models and seasonal time series models [7]–[9]. On the other hand, it is worth noting that some of the most widely used machine learning methods for energy prediction are Artificial Neural Networks (ANNs) [10] and Support Vector Machines (SVMs) [5], [11], [12].

This paper focuses on Deep Learning methods [13] for electrical energy demand prediction, with an application to the aggregated profiles collected from the Danish Island Bornholm within the EcoGrid project [14]. Due to the fact that energy consumption can be seen as a time series problem, it investigates the application of Factored Conditional Restricted Boltzmann Machines (FCRBM) [15], recently introduced stochastic machine learning methods which were used successfully until now to model highly non-linear time series (e.g. human motion style, structured output prediction) [15]–[17]. Consequently,

This research has been partly funded by AgentschapNL - TKI Switch2SmartGrids of Dutch Top Sector Energy. The authors would like to thank to our EcoGrid EU partners and DMI for providing the meteorological dataset.

we adapt the FCRBM architecture for demand forecasting problems by merging the style and feature labels into one, and by rewriting the equations and the derivatives of the learning rules according to the new configuration of the model. As a secondary contribution, we analyze how external factors (e.g. weather conditions, electricity prices) can be used to improve the forecasting accuracy and we propose the use of a Gaussian Restricted Boltzmann Machine to perform feature extraction in a fully automated manner and to reduce their dimensionality.

The remainder of this paper is organized as follows. Section II provides some background knowledge on unsupervised learning with Restricted Boltzmann Machines and Section III presents our proposed method using the FCRBM, including the adaptions necessary for demand prediction. Section IV describes the methodology and data description, followed by Section V where the experiment and results are detailed. Finally, SectionVI concludes the paper and presents directions for future research.

II. Background

Literature provides a wide range of techniques that can solve the demand forecasting problem. The electrical demand has a non-linear and non-stationary profile, which favours a probabilistic approach. In general, we attempt to model the probability of a data point, x using a function of the form $f(x;\theta)$, where θ is a vector of model parameters. Learning the model parameters, θ, can be done by maximizing the probability of a training set of data, or equivalently and often more convenient, by minimizing the negative log $p(x_i;\theta)$. This is not always a trivial task. In the context of our proposed method, we used another common method to learn the parameters of the model by minimizing the Kullback-Leibler (KL) divergence between the empirical and the approximated distributions of the model, as follows:

$$min_{\boldsymbol{\Theta}}[\mathrm{KL}(p_{model}(\mathbf{V}|\boldsymbol{\Gamma};\boldsymbol{\Theta})||p_{\text{empirical}}(\mathbf{V}|\boldsymbol{\Gamma}))] \tag{1}$$

where $\boldsymbol{\Gamma}$ represents the total input set and $\mathbf{V}$ is the total output set. The rest of this section presents the background knowledge useful to the reader to understand the remaining of the paper.

A. Restricted Boltzmann machine

Restricted Boltzmann Machines (RBMs) [18] have been applied in different machine learning fields including, multi-class classification [19], collaborative filtering [20], among others. They are energy-based models for unsupervised learning. These models have stochastic nodes and layers, making them less vulnerable to local minima [15]. Further, due to their multiple layers and neural configurations, RBMs possess excellent generalisation capabilities [13]. Formally, an RBM consists of visible and hidden binary layers. The visible layer represents the data, while the hidden one increases the learning capacity by enlarging the class of distributions that can be represented to an arbitrary complexity [15]. This paper follows a standard notation where i represents the indices of the visible layer, j those of the hidden layer, and $w_{i,j}$ denotes the weight connection between the i^{th} visible and j^{th} hidden unit. Further, v_i and h_j denote the state of the i^{th} visible and j^{th} hidden unit, respectively. According to the above definitions, the energy function[1] of an RBM is given by:

$$E(v,h) = -\sum_{i,j} v_i h_j w_{ij} - \sum_i v_i a_i - \sum_j h_j b_j \tag{2}$$

where, a_i and b_j represent the biases of the visible and hidden layers, respectively. The joint probability of a state of the hidden and visible layers is defined as: $P(v,h) = \frac{\exp(-E(v,h))}{Z}$ with $Z = \sum_{x,y}\exp(-E(x,y))$. To determine the probability of a data point represented by a state v, the marginal probability is used. This is determined by summing out the state of the hidden layer as: $p(v) = \sum_h P(v,h) = \frac{\sum_h\left(\exp\left(-\sum_{i,j} v_i h_j w_{ij} - \sum_i v_i a_i - \sum_j h_j b_j\right)\right)}{Z}$. Parameters are fitted by maximising the likelihood function. In order to maximise the likelihood of the model, the gradients of the energy function with respect to the weights have to be calculated. Usually, in RBMs maximum likelihood can not be simply applied due to intractability problems. To deal with these problems, Contrastive Divergence, explained next, was introduced.

B. Contrastive Divergence

In Contrastive Divergence (CD) [21], learning follows the gradient of:

$$CD_n = D_{KL}(p_0(\mathbf{x})||p_\infty(\mathbf{x})) - D_{KL}(p_n(\mathbf{x})||p_\infty(\mathbf{x})) \tag{3}$$

where, $p_n(.)$ is the distribution of a Markov chain running for n steps. Since the visible units are conditionally independent given the hidden units and vice versa, learning can be performed using one step Gibbs sampling, which is carried in two half-steps: (1) update all the hidden units, and (2) update all the visible units. Thus, in CD_n the weight updates are done as follows: $w_{ij}^{\tau+1} = w_{ij}^{\tau} + \alpha\left(\left\langle\langle h_j v_i\rangle_{p(\mathbf{h}|\mathbf{v};\mathbf{W})}\right\rangle_0 - \langle h_j v_i\rangle_n\right)$ where τ is the iteration, α is the learning rate, $\left\langle\langle h_j v_i\rangle_{p(\mathbf{h}|\mathbf{v};\mathbf{W})}\right\rangle_0 = \frac{1}{N}\sum_{k=1}^{N} v_i^{(k)} P(h_j^{(k)} = 1|\mathbf{v}^{(k)};\mathbf{W})$ and $\langle h_j v_i\rangle_n = \frac{1}{N}\sum_{k=1}^{N} v_i^{(k)(n)} P(h_j^{(k)(n)} = 1|\mathbf{v}^{(k)(n)};\mathbf{W})$ with N being the total number of input instances, and the superscript $^{(n)}$ indicates that the states are obtained after n iterations of Gibbs sampling from the Markov chain starting at $p_0(\cdot)$.

III. Factored conditional restricted Boltzmann machine

This section presents the adapted mathematical details of the proposed method, namely Factored Conditional Restricted Boltzmann Machine (FCRBM) [15], to achieve an accurate and robust prediction at the low aggregation level of electrical energy demand profiles.

[1]Please note that the energy function of RBM should not be confused with the aggregated electrical energy demand.

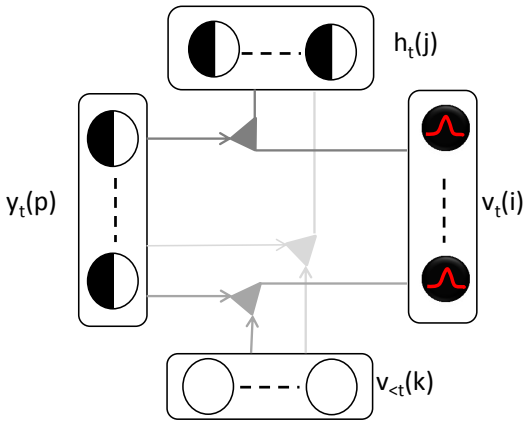

Figure 1. The general architecture of FCRBM, where $v_{<t}$ is the conditional history layer (input), h is the hidden layer, y is the style layer and v is the visible layer (output). Where ◐ denotes binary neurons, ○ represent the real values and the others are Gaussian value.

A. Total energy for FCRBM

The total energy function, $\mathbf{E}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}, \mathbf{y}_t)$ for FCRBM, is computed as the sum of the first and third order energy terms as follows:

$$\mathbf{E}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}, \mathbf{y}_t) = \mathbf{E}_I + \mathbf{E}_{III} \tag{4}$$

where $\mathbf{E}_I$ and $\mathbf{E}_{III}$ are defined as:

$$\mathbf{E}_I = \frac{1}{2}\sum_{i=1}^{n_1}(v_{i,t} - \hat{a}_{i,t})^2 - \sum_{j=1}^{n_2}\hat{b}_{j,t}h_{j,t}$$

$$\mathbf{E}_{III} = -\sum_{f=1}^{F}\left[\sum_{i=1}^{n_1}W_{if}^v v_{i,t}\sum_{j=1}^{n_2}W_{jf}^h h_{j,t}\sum_{p=1}^{n_3}W_{pf}^y y_{p,t}\right]$$

$$= -\sum_{f=1}^{F}\left[\sum_{i=1}^{n_1}\left[\sum_{j=1}^{n_2}\left[\sum_{p=1}^{n_3}W_{if}^v W_{jf}^h W_{pf}^y v_{i,t}h_{j,t}y_{p,t}\right]\right]\right]$$

where F, n_1, n_2, and n_3, represent the total number of factors, and the number of units in each of the visible, hidden, and label layers, respectively. The terms $\hat{a}_{i,t}$ and $\hat{b}_{j,t}$ are called dynamic biases, which are defined as:

$$\hat{a}_{i,t} = a_i + \sum_m A_{i,m}^v \sum_k A_{k,m}^{v<t} v_{k,<t} \sum_p A_{p,m}^y y_{p,t} \tag{6a}$$

$$\hat{b}_{j,t} = b_j + \sum_n B_{j,n}^h \sum_k B_{k,n}^{v<t} v_{k,<t} \sum_p B_{p,n}^y y_{p,t} \tag{6b}$$

with $A_{i,m}^v$, $A_{k,m}^{v<t}$, $A_{l,m}^y$, $B_{j,n}^h$, $B_{k,n}^{v<t}$, $B_{l,n}^y$, are dynamic biases of each of the layers.

B. Probabilistic inference in FCRBM

Inference in FCRBM is conducted in parallel, since there are no connections between the neurons in the same layer. Specifically, this means determining two conditional distributions. Firstly, the conditional probability distribution of the hidden neurons, $p(h_{j,t} = 1 | \mathbf{v}_t, \mathbf{v}_{<t}, \mathbf{y}_t)$, is given by a sigmoidal function evaluated on the total input to each hidden unit,$h_{j,t}^* = \sum_f W_{if}^h \sum_i W_{if}^v v_{i,t} \sum_p W_{pf}^y y_{p,t}$, via the factors. Secondly, the probability of the visible neurons, $p(v_{i,t} | \mathbf{h}_t, \mathbf{v}_{<t}, \mathbf{y}_t)$, is given by a Gaussian distribution over the total input, $v_{i,j}^* = \sum_f W_{if}^v \sum_j W_{jf}^h h_{j,t} \sum_p W_{pf}^y y_{p,t}$, to each visible unit via the factors. Therefore, for each of the jth hidden and ith visible unit, inference is performed using:

$$p(h_{j,t} = 1 | \mathbf{v}_t, \mathbf{v}_{<t}, \mathbf{y}_t) = \text{sigmoid}(\hat{b}_{j,t} + h_{j,t}^*) \tag{7}$$

$$p(v_{i,t} = x | \mathbf{h}_t, \mathbf{v}_{<t}, \mathbf{y}_t) = \mathcal{N}(\hat{a}_{i,t} + v_{i,j}^*, \sigma_i^2) \tag{8}$$

where $\mathcal{N}(\mu, \sigma_i^2)$ denotes the Gaussian probability density function with mean μ and variance σ_i^2.

C. Learning & Update Rules in FCRBM

The general update rule for all the hyper-parameters θ is given by:

$$\theta_{\tau+1} = \theta\tau + \rho\Delta\theta_\tau + \alpha(\Delta\theta_{\tau+1} - \gamma\theta_\tau) \tag{9}$$

where τ, ρ, α and γ represent the update number, momentum, learning rate, and weights decay, respectively. More details regarding the the choice of this parameters are described in [22]. The update rules for each of the weights matrices and biases can be computed by deriving the energy function from (2) with respect to each of these variables (i.e., the factored visible weights, factored label weights, factored hidden weights, and the biases of each of the layers), yelding:

1) Weights update: Three update rules corresponding to each of $\mathbf{W}^v$, $\mathbf{W}^h$, $\mathbf{W}^y$ need to be derived. Firstly, the factored visible weights W_{if}^v is computed by derivaiting the total energy function, provide in (4), with respect to W_{if}^v is:

$$\frac{\partial\, \mathbf{E}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}, \mathbf{y}_t)}{\partial W_{if}^v} = -v_{i,t}\sum_{j=1}^{n_2}W_{jf}^h h_{j,t}\sum_{p=1}^{n_3}W_{pf}^y y_{p,t} \tag{10}$$

Secondly, the factored hidden weights W_{jf}^h are update. Following the same reasoning we obtain:

$$\frac{\partial\, \mathbf{E}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}, \mathbf{y}_t)}{\partial W_{jf}^h} = -h_{j,t}\sum_{i=1}^{n_1}W_{if}^v v_{i,t}\sum_{p=1}^{n_3}W_{pf}^y y_{p,t} \tag{11}$$

Thirdly, by deriving the total energy function with respect to W_{pf}^y, we obtain the update rule for the factored label weights:

$$\frac{\partial\, \mathbf{E}(\mathbf{v}_t, \mathbf{h}_t | \mathbf{v}_{<t}, \mathbf{y}_t)}{\partial W_{pf}^y} = -y_{p,t}\sum_{i=1}^{n_1}W_{if}^v v_{i,t}\sum_{j=1}^{n_2}W_{jf}^h h_{j,t} \tag{12}$$

2) Biases update: The derivatives to find the update rules for the parameters which compose the dynamic biases of the present layer (i.e. A_{im}^v, $A_{k,m}^{v<t}$, $A_{l,m}^y$) are:

$$\frac{\partial\, \mathbf{E}}{\partial A_{im}^v} = v_{i,t}\sum_k A_{k,m}^{v<t} v_{k,<t}\sum_p A_{p,m}^y y_{p,t} \tag{13a}$$

$$\frac{\partial\, \mathbf{E}}{\partial A_{k,m}^{v<t}} = v_{k,<t}\sum_i A_{i,m}^v v_{i,t}\sum_p A_{p,m}^y y_{p,t} \tag{13b}$$

$$\frac{\partial\, \mathbf{E}}{\partial A_{l,m}^y} = y_{p,t}\sum_i A_{i,m}^v v_{i,t}\sum_k A_{k,m}^{v<t} v_{k,<t} \tag{13c}$$

Further, the derivatives to find the update rules for the parameters which compose the dynamic biases of the hidden layer (i.e. $B^h_{j,n}$, $B^{v<t}_{k,n}$, $B^y_{l,n}$) are presented as:

$$\frac{\partial\ \mathbf{E}}{\partial B^h_{j,n}} = -h_{j,t}\sum_k B^{v<t}_{k,n} v_{k,<t}\sum_p B^y_{p,n} y_{p,t} \tag{14a}$$

$$\frac{\partial\ \mathbf{E}}{\partial B^{v<t}_{k,n}} = -h_{j,t} v_{k,<t}\sum_j B^h_{j,n}\sum_p B^y_{p,n} y_{p,t} \tag{14b}$$

$$\frac{\partial\ \mathbf{E}}{\partial B^y_{l,n}} = -h_{j,t} y_{p,t}\sum_j B^h_{j,n}\sum_k B^{v<t}_{k,n} v_{k,<t} \tag{14c}$$

Using the energy derivative of the hyper parameters and the Contrastive Divergence expression shown in (3), we can calculate the $delta$ rule leading to:

$$\Delta W \propto \langle\frac{\partial\ \mathbf{E}}{\partial W}\rangle_0 - \langle\frac{\partial\ \mathbf{E}}{\partial W}\rangle_k, \text{using eq. (10), (11), (12)} \tag{15a}$$

$$\Delta A \propto \langle\frac{\partial\ \mathbf{E}}{\partial A}\rangle_0 - \langle\frac{\partial\ \mathbf{E}}{\partial A}\rangle_k, \text{using eq. (13)} \tag{15b}$$

$$\Delta B \propto \langle\frac{\partial\ \mathbf{E}}{\partial B}\rangle_0 - \langle\frac{\partial\ \mathbf{E}}{\partial B}\rangle_k, \text{using eq. (14)} \tag{15c}$$

with k being a Markov chain step running for a total number of K steps and starting at the original data distribution.

IV. Methodology and Data Description

In this section, we describe the metrics chosen for prediction accuracy assessment, followed by a brief description of the dataset. Finally, we propose and describe an automatic method for feature extraction enforced by our dataset characteristics.

A. Metrics for prediction assessment

To quantify the performance of the prediction methods, we used a variety of standard metrics. Firstly, the prediction accuracy is evaluated using three popular metrics capable to put a different penalty on the same error, namely the root mean square error, $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}$, the normalized root-mean-square error, $NRMSE[\%] = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}/(v_{max} - v_{min}) \cdot 100$, and the mean absolute percentage error, $MAPE = \frac{1}{n}\sum_{i=1}^{n}|v_i - \hat{v}_i|/\max(v_i) \cdot 100$, where n represents the total number of predicted steps, v_i represents the true values for the time-step i and $\hat{v}_i$ represents the value predicted by the model at the same time-step. Secondly, the Pearson Correlation Coefficient (PCC) is used to indicate the degree of the linear dependence between the real and the predicted values, as follows:

$$PCC(v, \hat{v}) = \frac{\mathbb{E}[(v - \mu_v)(\hat{v} - \mu_{\hat{v}})]}{\sigma_v \sigma_{\hat{v}}}$$

where $\mathbb{E}[\cdot]$ is the expected value operator with means μ_v and $\mu_{\hat{v}}$, and standard deviations σ_v and $\sigma_{\hat{v}}$, for the true and estimated values, respectively. The PCC value is within the range [-1,1]. The sign of the correlation coefficient defines the direction of the relationship, either positive or negative. Besides using PCC in the demand forecast evaluation process, in the second part of the experiments, the PCC values were used to highlight the most influential factors for the electrical energy demand profiles.

B. Dataset description

In this work we have used the EcoGrid dataset collected from the Danish Island Bornholm in the first seven months of 2014. The dataset includes the aggregated energy consumption, the real-time price (RTP), forecast prices (DA, HA) and meteorological data [23]. Altogether, this dataset has 50677 records at 5 minutes resolution, each record containing 16 different features, leading to more than 800000 data points. Table I summarizes some basic statistical information about the entire dataset used in the experiments, such as mean and standard deviation for each feature. Furthermore, the last column of Table I shows the correlation coefficient between the electrical energy demand values and the additional information available in the EcoGrid database. Figure 2 shows the

TABLE I. Summary of the meteorological and price data correlated with the aggregated electrical energy demand.

	Mean (μ)	Std.dev. (Σ)	PCC w.r.t energy
Price RTP	236.17	92.04	-0.0319
Price DA	233.10	98.34	0.0113
Price HA	234.84	86.67	-0.0231
cloud base height	2117.7	2832	-0.1635
water vapor	0.0053	0.0018	-0.7451
relative humidity	0.8413	0.1181	0.3486
temperature	6.9421	5.5756	-0.8384
global irradiance	521.40	806.29	-0.5832
diffuse irradiance	214.65	365.56	0.4031
wind speed	5.6202	2.8643	0.2598
cloud cover	0.6583	0.3869	0.3109
rain	0.0507	0.2035	0.0724
wind gust	9.5449	4.5317	0.2105
atmospheric pressure	1010.8	8.3686	-0.1970

aggregated electrical demand composition with a 5 minutes sampling rate which is the basis of our analysis. This aggregated electrical demand involves 1900 customers equipped with local generation. The decreasing trend of the demand with time, observed in Figure 2 is mainly due to the negative correlation with the temperature, i.e. PCC=-0.83, but also suggests that a load shifting may have occurred in the presence of a large renewable energy penetration.

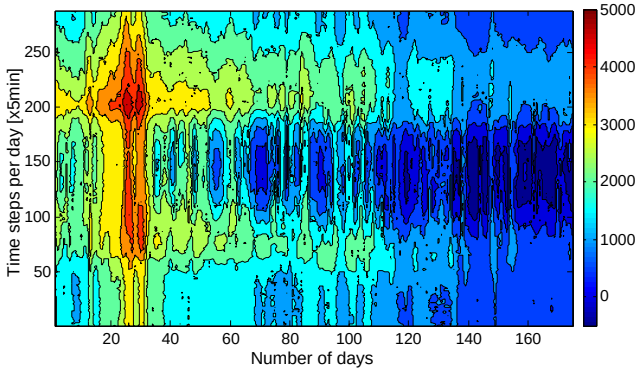

Figure 2. Electrical demand profile at low aggregation level between 1 January 2014 until 25 June 2014

C. Feature extraction

The growth of distributed energy resources (DERs), together with smart appliances, capable of responding to real-time pricing signals yields poor correlations with energy consumption at low time resolutions, as it can be observed in Table I. Thus, constructing a proper combination of the additional information needed in order to improve the forecast accuracy is not a trivial task. Moreover, the extracted information aims to be a non-redundant generalization of the price and weather data. Besides that, from a computational perspective we want to have a lower dimensional data set. One traditional way to perform feature extraction is based on statistical hypothesis testing in order to determine if the distributions of values of a feature for two different classes are distinct. Still this solution creates results which are hard to interpret. A simpler solution is to use Principal Component Analysis (PCA), which is part of a wide area of clustering methods. However, PCA loses its information-theoretic optimality as soon as the data becomes dependent [24].

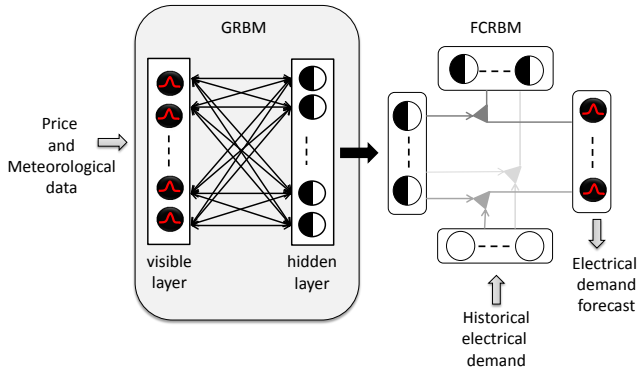

Figure 3. General architecture of a Gaussian restricted Boltzmann machine (GRBM) as input to the FCRBM.

More recently, it was shown that restricted Boltzmann machines are capable to learn low-dimensional codes that work much better than PCA as a tool to reduce the dimensionality of the data [25]. Consequently, we propose a combination between a Gaussian Restricted Boltzmann Machine (GRBM) and a FCRBM to perform dimensionality reduction and time series prediction, as depicted in Figure 3. The RBM mathematical details were previously described in Section II-A. This is further enhanced with Gaussian neurons in the visible layer in order to transform the RBM into an GRBM [25].

V. Numerical Results

To assess the performance of the proposed method we have conducted five sets of experiments, over a wide range of time horizons, as summarized in Table II. The experimental validation was done in two steps. Specifically, we have looked to the forecast problem in a traditional versus a price-responsive environment.

A. Implementation details

We made the implementation of FCRBM in Matlab® using the mathematical details described in Section III. The number of hidden neurons and the number of factors were set to 50. The learning rate was set to 10^{-4}, the momentum to 0.9, and the weight decay to 0.0002. These parameters were chosen carefully by performing a small cross-validation experiment and they were kept constant in all the experiments for a fair comparison. In the first set of experiments, the "traditional" forecast problem, we used in the class layer 10 neurons with the default value 1 and the number of history neurons was set to 864, corresponding to a historical time window of 3 days. In the second set of experiments, which includes the price-responsive environment, we used in the class layer of FCRBM the features extracted by GRBM from the price and meteorological data corresponding to each specific time window. More exactly, these features were a binary vector of 10 values. Besides that, we set the number of history neurons to 72, corresponding to a historical time window of 6 hours.

TABLE II. Summary of the experiments.

	Time horizon	Resolution
Scenario 1	5 minutes	5 minutes
Scenario 2	15 minutes	5 minutes
Scenario 3	1 hour	5 minutes
Scenario 4	6 hours	5 minutes
Scenario 5	1 day	5 minutes

Additionally, we made use of LibSVM library [26] to conduct a comparison of the FCRBM performance with a benchmark machine learning algorithm, namely the support vector machine with radial kernel function (SVM). To train both models, FCRBM and SVM, in the general forecast problem we have used the data from 1 January 2014 to 21 May 2014, while to test them we used the data from 21 May up to 25 June 2014. In the case of the price-responsive environment, we utilized 66% of the available data to train the models, and the rest of 34% to test them.

B. Electrical energy demand forecast

To quantify the performance of the proposed method, we used the four metrics described in Section IV-A. The results obtained with FCRBM have been further compared with other forecasting methods, such as SVM and persistence. Traditionally, the persistence method is recommended especially for very short-term forecasting [27], as it simply assumes that a constant value occurs over the forecast horizon.

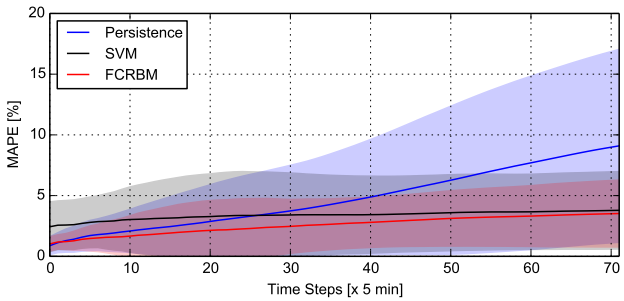

Figure 4. The prediction error of the aggregated demand with mean (straight line) and standard deviation (shaded area) for 6 hours, using FCRBM, SVM and persistence methods.

Figure 5 shows an example of aggregated energy demand prediction error for a max lead time of 72 time steps (6 hours ahead) averaged over the five weeks of the testing period, from 21 May 2014 to 25 June 2014. Therein, a slightly better accuracy with a lower variation is visible for FCRBM versus SVM in terms of the MAPE metric. Furthermore, Table III shows the performance of the proposed models for all five scenarios. Overall, FCRBM outperforms the other methods in

TABLE III. AGGREGATED ELECTRICAL DEMAND FORECASTING USING SUPPORT VECTOR MACHINE, FACTORED CONDITIONAL RESTRICTED BOLTZMANN MACHINE AND THE PERSISTENCE METHODS.

	Methods	NRMSE [%]	RMSE	MAPE	PCC
Scenario 1	Persistence	0.73	27.31	0.85	0
	SVM	1.69	62.90	1.95	0.13
	FCRBM	0.71	25.21	0.84	0.15
Scenario 2	Persistence	1.27	47.24	1.34	0
	SVM	2.07	77.30	2.28	0.29
	FCRBM	1.23	45.95	1.31	0.31
Scenario 3	Persistence	3.02	112.38	3.03	0.01
	SVM	2.91	108.44	3.07	0.45
	FCRBM	2.50	93.06	2.59	0.46
Scenario 4	Persistence	12.26	456.10	11.66	-0.01
	SVM	4.48	166.55	4.43	0.87
	FCRBM	4.30	160.10	4.18	0.88
Scenario 5	Persistence	11.76	437.35	10.39	0.01
	SVM	5.64	209.90	4.70	0.91
	FCRBM	5.19	193.14	4.49	0.91

all metrics, while SVM performs better than persistence for longer time horizons, and persistence perform better than SVM for the short-term scenarios 1 and 2. For a pictorial view of the short-therm forecast accuracy of the three methods we depict in Figure 5 an example of the true and forecast aggregated electrical energy demand over 6 hours horizon, with 5 minute resolution.

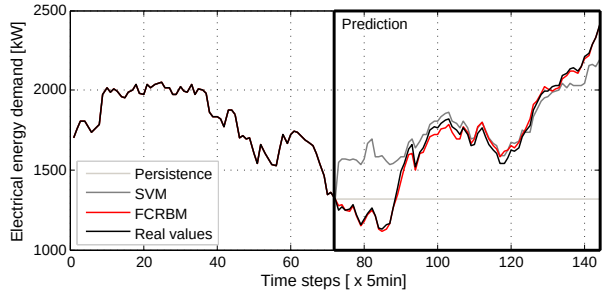

Figure 5. An example of true predicted aggregated electricity demand for six hours ahead, with five minutes resolution, using FCRBM, SVM and persistence methods.

C. Demand forecast in a price-responsive context

This second set of experiments investigates the possibility to increase the forecast accuracy by using fusion data from the environment in a price-responsive context, over a short period of time (approximately three days).

The price-responsiveness of the electricity demand is observed during a visual analysis of the entire dataset, and highlighted by the differences in the correlations observed in Table I and Table IV in terms of price values. So, we

TABLE IV. SUMMARY OF THE METEOROLOGICAL AND PRICE DATA CORRELATED WITH AGGREGATED ENERGY CONSUMPTION FOR A SHORT PERIOD

	Mean(μ)	Std.dev.(Σ)	PCC w.r.t energy
RTP	289.17	368.92	0.4912
DA	289.19	406.55	0.5144
HA	289.15	261.04	0.3051
cloud base height	904.42	1067	0.3089
water vapor	0.0049	3.5e-04	0.0597
relative humidity	0.8839	0.06	0.2300
temperature	5.6133	0.99	-0.1794
global irradiance	359.25	714.71	-0.2987
diffuse irradiance	115.31	162.28	-0.2432
wind speed	5.3913	1.8962	-0.4091
cloud cover	0.6494	0.40	0.4199
rain	0.1259	0.1756	0.14
wind gust	8.8503	3.0534	-0.4187
atmospheric pressure	1007.21	8.4877	0.1244

enhanced the FCRBM model with additional information and we analyzed the accuracy of the predictor. More exactly, following the method proposed in Section IV-C, we performed a fully automatic feature extraction computation using a GRBM model from the day-ahead price and cloud cover data. Then this encoded information is placed into the class layer of the FCRBM model.

TABLE V. IMPROVED ACCURACY OF THE AGGREGATED ELECTRICAL DEMAND FORECASTING USING PRICE AND METEOROLOGICAL DATA

Methods	NRMSE	RMSE	MAPE	PCC
Scenario 1				
Persistence	3.08	46.10	1.75	0.01
SVM (energy)	7.31	109.19	4.15	0.09
FCRBM (energy)	2.42	36.26	1.38	0.09
FCRBM (energy+weather)	2.60	38.95	1.48	0.08
FCRBM (energy+price)	2.28	34.08	1.30	0.12
FCRBM (energy+weather+price)	2.52	37.62	1.43	0.14
Scenario 2				
Persistence	3.74	55.78	1.89	0.01
SVM (energy)	8.09	120.82	4.48	0.34
FCRBM (energy)	3.37	50.36	1.75	0.33
FCRBM (energy+weather)	3.29	49.19	1.68	0.38
FCRBM (energy+price)	3.09	46.19	1.57	0.41
FCRBM (energy+weather+price)	2.79	41.74	1.39	0.47
Scenario 3				
Persistence	7.05	105.33	3.38	0.01
SVM (energy)	11.65	174.93	6.21	0.36
FCRBM (energy)	6.21	92.81	3.10	0.39
FCRBM (energy+weather)	5.88	87.84	2.92	0.32
FCRBM (energy+price)	5.76	86.08	2.80	0.45
FCRBM (energy+weather+price)	5.49	81.98	2.71	0.30
Scenario 4				
Persistence	23.08	344.63	10.73	-0.01
SVM (energy)	24.96	372.79	12.42	0.37
FCRBM (energy)	11.24	167.85	5.28	0.66
FCRBM (energy+weather)	11.09	165.73	5.50	0.80
FCRBM (energy+price)	8.42	125.76	4.45	0.95
FCRBM (energy+weather+price)	5.51	82.40	2.67	0.96

Figure 6 shows the best performer from Scenario 4 (6 hour ahead with 5 minute resolution) together with the corresponding FCRBM forecasting without any additional information, benchmarked by the persistence method, in terms of MAPE metric. The overall results are presented in Table V. There is

presented the performance of the forecasting methods analyzed for various combinations of input data which include, next to historical values for aggregated electrical demand, also prices and weather conditions. Although our data is multidimensional

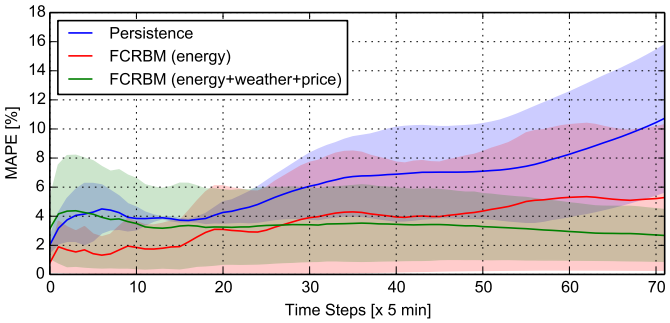

Figure 6. An example of aggregated electrical demand forecasting for six hours in therms of MAPE, with five minute resolution, using persistence, FCRBM (energy) and FCRBM with energy, weather and price data.

we have avoided the scalability problems by performing the feature extraction procedure. This leads to a light approach able to generalize over any other time series. It is worth highlighting, that we observed that by adding more external information we can slightly improve the overall accuracy for this real dataset.

VI. Conclusion

This paper proposes a powerful stochastic machine learning method to forecast the electricity demand at low aggregation levels, namely the Factored Conditional Restricted Boltzmann Machine. FCRBM has good generalization capabilities and it can be used to accommodate large sets of data, while its exploitation time in real-world settings is on the order of few milliseconds. Secondary, we propose the use of GRBM to extract features from external information and to reduce the dimensionality for the FCRBM. We validate our approach on a real dataset, consisting of 1900 households originating from the Danish island of Bornholm, collected within the EcoGrid project. In order to compare alternative approaches we used four different metrics and two benchmark forecasting methods, namely Support Vector Machine and persistence. On the one hand, the results show that FCRBM outperforms the other two methods, and on the other hand, they suggest that by adding more weather and price information to the FCRBM, its performance may be improved further. This promising method can in the future be applied to a fully automatic real-time prediction and optimal control of electrical energy consumption via demand response in a smart grid context.

References

[1] M. Krarti, *Energy Audit of Building Systems: An Engineering Approach, Second Edition*, ser. Mechanical and Aerospace Engineering Series. Taylor & Francis, 2012.

[2] A. Foucquier, S. Robert, F. Suard, L. Stphan, and A. Jay, "State of the art in building modelling and energy performances prediction: A review," *Renewable and Sustainable Energy Reviews*, vol. 23, no. 0, pp. 272 – 288, 2013.

[3] H. xiang Zhao and F. Magouls, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586 – 3592, 2012.

[4] A. I. Dounis, "Artificial intelligence for energy conservation in buildings," *Advances in Building Energy Research*, vol. 4, no. 1, pp. 267–299, 2010.

[5] S. Fan and R. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 134–141, Feb 2012.

[6] J. W. Taylor, "Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles," *International Journal of Forecasting*, vol. 26, no. 4, pp. 627 – 646, 2010.

[7] A. M. D. Livera, R. J. Hyndman, and R. D. Snyder, "Forecasting time series with complex seasonal patterns using exponential smoothing," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1513–1527, 2011.

[8] M. Aydinalp-Koksal and V. I. Ugursal, "Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector," *Applied Energy*, vol. 85, no. 4, pp. 271 – 296, 2008.

[9] L. Xuemei, D. Lixing, L. Jinhu, X. Gang, and L. Jibin, "A novel hybrid approach of kpca and svm for building cooling load prediction," in *Knowledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on*, 2010, pp. –.

[10] E. Mocanu, P. Nguyen, M. Gibescu, and W. Kling, "Comparison of machine learning methods for estimating energy consumption in buildings," in *International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, July 2014, pp. 1–6.

[11] S. A. Kalogirou, "Artificial neural networks in energy applications in buildings," *International Journal of Low-Carbon Technologies*, vol. 1, no. 3, pp. 201–216, 2006.

[12] S. Wong, K. K. Wan, and T. N. Lam, "Artificial neural networks for energy analysis of office buildings with daylighting," *Applied Energy*, vol. 87, no. 2, pp. 551–557, 2010.

[13] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009, also published as a book. Now Publishers, 2009.

[14] "Ecogrid eu project." [Online]. Available: http://www.eu-ecogrid.net/

[15] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Two distributed-state models for generating high-dimensional time series," *Journal of Machine Learning Research*, vol. 12, pp. 1025–1068, 2011.

[16] E. Mocanu, D. Mocanu, H. Ammar, Z. Zivkovic, A. Liotta, and E. Smirnov, "Inexpensive user tracking using boltzmann machines," in *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Oct 2014, pp. 1–6.

[17] V. Mnih, H. Larochelle, and G. Hinton, "Conditional restricted boltzmann machines for structured output prediction," in *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2011.

[18] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing: Volume 1: Foundations*, D. E. Rumelhart, J. L. McClelland *et al.*, Eds. Cambridge: MIT Press, 1987, pp. 194–281.

[19] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," 2008, pp. 536–543.

[20] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *In Machine Learning, Proceedings of the Twenty-fourth International Conference (ICML 2004). ACM*. AAAI Press, 2007, pp. 791–798.

[21] G. E. Hinton, "Training Products of Experts by Minimizing Contrastive Divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.

[22] G. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science, 2012, vol. 7700, pp. 599–619.

[23] E. Larsen, P. Pinson, G. Le Ray, and G. Giannopoulos, "Demonstration of market-based real-time electricity pricing on a congested feeder," in *12th International Conference on the European Energy Market (EEM)*, May 2015, pp. 1–5.

[24] B. C. Geiger and G. Kubin, "Signal enhancement as minimization of relevant information loss," *CoRR*, vol. abs/1205.6935, 2012.

[25] G. E. Hinton and R. R. Salakhutdinov.

[26] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[27] A. Foley, P. Leahy, A. Marvuglia, and E. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renewable Energy*, vol. 37, no. 1, pp. 1–8, 2012.