Grzegorz DUDEK

Czestochowa University of Technology

Short-Term Load Forecasting Based on Kernel Conditional Density Estimation

Abstract. *A short-term load forecasting model based on the kernel estimation of the conditional probability density distribution is proposed. The pattern vector of the load time series sequence can be treated as the multivariate random variable whose value determines the pattern component values of the next sequence, which is forecasted. Probability density functions are obtained from historical load time series by means of nonparametric density estimation. This approach uses the product kernel estimators. The kernel function smoothing parameters are determined using cross-validation procedure. The suitability of the proposed approach is illustrated through applications to real load data.*

Streszczenie. *Proponuje się model prognostyczny do sporządzania krótkoterminowych prognoz obciążeń systemów elektroenergetycznych w oparciu o estymację jądrową rozkładu warunkowej gęstości prawdopodobieństwa. Wektor obrazu sekwencji szeregu czasowego obciążeń może być traktowany jako wielowymiarowa zmienna losowa, która determinuje wartość składowych obrazu następnej, prognozowanej sekwencji. Funkcje gęstości prawdopodobieństwa utworzono na podstawie historycznych szeregów czasowych obciążeń za pomocą estymacji nieparametrycznej. To podejście używa produktowych estymatorów jądrowych. Parametry wygładzania funkcji jądrowych określa się w procedurze walidacji krzyżowej. Użyteczność proponowanego podejścia zilustrowano aplikacjami do rzeczywistych danych. (**Krótkoterminowe prognozowanie obciążeń elektroenergetycznych za pomocą jądrowej estymacji gęstości warunkowej**).*

Keywords: short-term load forecasting, kernel density estimation, nonparametric regression.
Słowa kluczowe: prognozowanie krótkoterminowe obciążeń, jądrowa estymacja gęstości, regresja nieparametryczna.

Introduction

The short-term load forecasting (STLF) is extremely important to balance the electricity generated and consumed at any moment. Precise load forecasts are necessary for electric companies to make important decisions connected with electric power production and transmission planning, such as unit commitment, generation dispatch, hydro scheduling, hydro-thermal coordination, spinning reserve allocation and interchange evaluation. Understanding the load behavior as the basic driver of electricity prices becomes more important in restructured power market.

Many STLF models have been designed. Conventional STLF models use smoothing techniques, regression methods and statistical analysis. ARIMA and related models are very popular, where the load is modeled by an autoregressive moving average difference equation. In recent years artificial intelligence methods have been widely applied to STLF: neural networks, fuzzy systems and expert systems.

In this article nonparametric regression method is used to STLF. The regression relationship can be modelled as [1]:

$$y = m(x) + \varepsilon\,, \tag{1}$$

where y is the response variable; x – the predictor; ε – the error, which is assumed to be normally and independently distributed with zero mean and constant variance; $m(x) = \mathrm{E}(Y \mid X = x)$ is a regression curve.

The aim of regression is to estimate the function m. This task can be done essentially in two ways. The first approach to analyze a regression relationship is called parametric since it is assumed that the mean curve m has some prespecified functional form and is fully described by a finite set of parameters (e.g. a polynomial regression equation). In the alternative nonparametric regression the regression curve does not take a predetermined form but is constructed according to information derived from the data. The regression function is estimated directly, rather than to estimate parameters. Most methods of nonparametric regression implicitly assume that m is a smooth and continuous function. The most popular nonparametric regression models are [1, 2]: kernel estimators, k-nearest neighbor estimators, orthogonal series estimators and spline smoothing.

Nonparametric conditional density estimation is a way to model the relationship between past and future loads of the power system. Using kernel estimates of density functions we get the regression function as a locally weighted average of the response variables. This nonparametric model, called Nadaraya-Watson estimator, in application to the STLF is presented in this article.

Load forecasting model

First step to build the STLF model is data analysis and preprocessing. The load time series are characterized by annual, weekly and daily cycles due to changes in industrial activities and climatic conditions. The goal of data preprocessing is to get rid of the time series trend and seasonality, and simplify the model.

Let $\mathbf{L}_i = [L_{i,1}\ L_{i,2}\ \dots\ L_{i,24}]$ be a vector of hourly power system loads in the following hours of the day preceding the day of forecast, and let $\mathbf{L}_{i+\tau} = [L_{i+\tau,1}\ L_{\ i+\tau,2}\ \dots\ L_{i+\tau,24}]$ be a vector of hourly loads of the day of forecast, where $\tau > 0$ denotes the forecast horizon. The load patterns are defined: input $\mathbf{x}_i = [x_{i,1}\ x_{i,2}\ \dots\ x_{i,24}]$ and output $\mathbf{y}_i = [y_{i,1}\ y_{i,2}\ \dots\ y_{i,24}]$, which are vectors with components calculated as follows:

$$x_{i,h} = \frac{L_{i,h} - \overline{L}_i}{\sqrt{\sum_{l=1}^{24}(L_{i,l} - \overline{L}_i)^2}}\,, \tag{2}$$

$$y_{i,h} = \frac{L_{i+\tau,h} - \overline{L}_i}{\sqrt{\sum_{l=1}^{24}(L_{i+\tau,l} - \overline{L}_i)^2}}\,, \tag{3}$$

where: i – the day number; h = 1, 2, …, 24 – the hour of the day; τ – the forecast horizon, here τ = 1; $L_{i,h}$ – the load at hour h of day i; $\overline{L}_i$ – the mean load of day i.

Definition (2) expresses normalization of the original load vectors $\mathbf{L}_i$. After normalization they have the unity length, zero mean and the same variance.

Forecast patterns (3) are analogous to input patterns (2), but they are encoded using the current loads determined from the nearest past of the process history, what enables decoding of the forecasted vector $\mathbf{L}_{i+\tau}$ after the forecast of pattern $\mathbf{y}$ is determined.

Let $\mathbf{Z}$ be a d+1-dimensional random variable consisted of the forecast pattern component Y and the input pattern $\mathbf{X}$: $\mathbf{Z} = [Y\ \mathbf{X}]^T$. The pattern $\mathbf{X}$ is a d-dimensional conditioning random variable, its values condition the value of the random variable Y. The density distribution of the variable $\mathbf{Z}$ describes a function $f_{\mathbf{Z}}$: $\Re^{d+1} \rightarrow [0, \infty)$, and the density distribution of $\mathbf{X}$ describes a function $f_{\mathbf{X}}$: $\Re^d \rightarrow [0, \infty)$. The conditional density of Y given $\mathbf{X} = \mathbf{x}$ is defined as:

$$f_{Y|\mathbf{X}}(y \mid \mathbf{x}) = \frac{f_{\mathbf{Z}}(y, \mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}. \tag{4}$$

Given a random sample:

$$\begin{bmatrix} y_1 \\ \mathbf{x}_1 \end{bmatrix}, \begin{bmatrix} y_2 \\ \mathbf{x}_2 \end{bmatrix}, \ldots, \begin{bmatrix} y_n \\ \mathbf{x}_n \end{bmatrix}, \tag{5}$$

we can determine the kernel density estimator of $\mathbf{Z}$ – $\hat{f}_{\mathbf{Z}}$, of $\mathbf{X}$ – $\hat{f}_{\mathbf{X}}$ and, in consequence, the conditional kernel density estimator:

$$\hat{f}_{Y|\mathbf{X}}(y \mid \mathbf{x}) = \frac{\hat{f}_{\mathbf{Z}}(y, \mathbf{x})}{\hat{f}_{\mathbf{X}}(\mathbf{x})}. \tag{6}$$

In general the kernel density estimator of an n-element random sample $v_1, v_2, \ldots, v_n$ – $\hat{f}$: $\Re \rightarrow [0, \infty)$ has the form:

$$\hat{f}(v) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{v - v_j}{h}\right), \tag{7}$$

where $h \in \Re^+$ is a bandwidth (smoothing parameter).
The K is called a kernel, which is a continuous, bounded, and symmetric real function which integrates to 1: $\int K(t)dt = 1$. Several types of kernel functions are commonly used [1, 2]: uniform, triangle, Epanechnikov, biweight or Gaussian:

$$K(t) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right). \tag{8}$$

Performance of the kernel is measured by the mean integrated squared error (MISE). The Epanechnikov kernel is the most efficient in minimizing the error when approximating the true density by the kernel density, but the efficiencies of other kernels are only a little lower. Thus the choice of a kernel is not as important as the choice of a bandwidth.

In the case of the multidimensional random variables $\mathbf{v} \in \Re^d$, often the product kernel is used:

$$\hat{f}(\mathbf{v}) = \frac{1}{nh_1 \ldots h_d}\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{v_k - v_{j,k}}{h_k}\right), \tag{9}$$

with different smoothing parameter h_k in the k-th direction.
The kernel estimator of the conditional density in this case has the form:

$$\hat{f}_{Y|\mathbf{X}}(y \mid \mathbf{x}) = \frac{\frac{1}{g}\sum_{j=1}^{n} K\left(\frac{y - y_j}{g}\right)\prod_{k=1}^{d} K\left(\frac{x_k - x_{j,k}}{h_k}\right)}{\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{x_k - x_{j,k}}{h_k}\right)}, \tag{10}$$

where g and h_k are bandwidths.
The expected value of random variable y with respect to a conditional probability distribution $f_{Y|\mathbf{X}}(y \mid \mathbf{x})$ is [1]:

$$m(\mathbf{X}) = E(Y \mid \mathbf{X}) = \int y f_{Y|\mathbf{X}}(y \mid \mathbf{x})dy = \frac{\int y f_{\mathbf{Z}}(y, \mathbf{x})dy}{f_{\mathbf{X}}(\mathbf{x})}. \tag{11}$$

If we replace the density function $f_{\mathbf{Z}}(y, \mathbf{x})$ by its kernel estimate, the numerator of (11) has the form [1]:

$$\begin{aligned} &\int y \hat{f}_{\mathbf{Z}}(y, \mathbf{x})dy = \\ &= \frac{1}{nh_1 \ldots h_d}\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{x_k - x_{j,k}}{h_k}\right)\int \frac{y}{g} K\left(\frac{y - y_j}{g}\right)dy = \\ &= \frac{1}{nh_1 \ldots h_d}\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{x_k - x_{j,k}}{h_k}\right)\int (sg + y_j)K(s)ds = \\ &= \frac{1}{nh_1 \ldots h_d}\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{x_k - x_{j,k}}{h_k}\right)y_j, \end{aligned} \tag{12}$$

where we used the facts that kernel functions integrate to 1 and are symmetric around zero.

After replacing the density function $f_{\mathbf{X}}(\mathbf{x})$ in (11) by its kernel estimate we get the multivariate generalization of the Nadaraya-Watson estimator:

$$\hat{m}(\mathbf{x}) = \frac{\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{x_k - x_{j,k}}{h_k}\right)y_j}{\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{x_k - x_{j,k}}{h_k}\right)}, \tag{13}$$

which is a weighted sum of the observed responses. For observations which are closer to input $\mathbf{x}$ the weights are larger.

The choice of bandwidths is made so that some global error criterion is minimized (traditionally in STLF this criterion is MAPE). The bandwidth values decide about the bias-variance tradeoff of the estimator. The small bandwidth value results in undersmoothing, whereas the large value results in oversmoothing.

The bandwidths were determined in leave-one-out cross-validation procedure. This method is based on regression smoothers (13), in which one, ith observation is left out. The value of y_i is predicted across the subsamples $\{[y_j\ \mathbf{x}_j]^{\mathrm{T}}\}_{j \neq i}$, and the error criterion value is calculated. Repeating this for i = 1, 2, …, n we determine the mean value of the error at the specific vector of bandwidths $\mathbf{h} = [h_1, \ldots, h_d]$. We are searching for the optimal bandwidth values in the iterating process, trying a sequence of bandwidth vectors $\mathbf{h}_l$, l = 1, 2, …, which elements $h_{l,k}$ are calculated according to the formula:

$$h_{l,k} = a_l h_k^* \quad l = 1,2,\ldots, \quad k = 1,2,\ldots,d, \tag{14}$$

where $a_l = 0.5 + 0.05(l - 1)$ and h_k^* is calculated using the Scott's rule [3]:

$$h_k^* = \hat{\sigma}_k n^{-1/(d+4)}, \quad (15)$$

where $\hat{\sigma}_k$ is the sample standard deviation of x_k and n is the sample size.

The searching process stops when L successive iterations do not bring better results.

In this procedure, due to the size of the problem and many variants of the models (see next section), the whole vector of h_k is optimized, not the individual bandwidths.

The input variables x_k are strongly correlated what means that the same input information is repeated many times. A removal of the redundant or unpredictive variables can improve the model quality. Moreover it causes the model simplification and improvement in generalization. In order to select the optimal subset of input variables the sequential forward and sequential backward selection methods (SFS, SBS) [4] are used.

The SFS, based on the simple greedy deterministic heuristics, starts with an empty variable subset and adds one new variable to the current set of selected variables in each step. To determine which variable to add, the algorithm tentatively adds to the candidate variable subset one variable that is not already selected and tests the accuracy of a forecasting model built and optimized on the tentative variable subset. The variable that results in the highest accuracy is definitely added to the variable subset. The process stops after an iteration where no variable additions result in an improvement in accuracy. Similar method is the SBS which starts with all the possible features and discards one at the time.

Application examples

The model based on the Nadaraya-Watson estimator (N-WE) described above was used to five practical problems of the next day load curve forecasting. Data are described in Table 1.

Table 1. Description of data used in experiments

Data symbol	Data description
A	Time series of the hourly loads of the Polish power system from the period 2002-2006, mean load of the system ~16 GW
B	Time series of the hourly loads of the Polish power system from the period 1997-2000, mean load of the system ~15,5 GW
C	Time series of the hourly loads of the local power system from the period July 2001-January 2003, mean load of the system ~1,2 GW
D	Time series of the hourly loads of the local power system from the period June 1998-July 2002, mean load of the system ~300 MW
E	Time series of the hourly load demands of the chemical plant from the period 1999-2001, mean load demand of the plant ~80 MW

For each day type and each hour of the day a separate forecasting model was built. MAPE (Mean Absolute Percentage Error) was used as a model quality criterion.

Datasets were divided into two subsets – training one and test one. The first sequences of the time series (two thirds of the whole time series) were included in the training set and the latest sequences were included in the test set. The models were optimized on training sets and then tested on test sets.

Table 2 contains forecasting errors for the test sets with and without variable selection. For comparison, forecast using the simple nearest neighbour (NN) method, multilayer perceptron (MLP) and the model based on the fuzzy estimators (FE) [5, 6] were calculated. More results in [6] can be found, where many other models were tested (including models based on artificial immune systems, neural gas, self-organizing maps and k-NN estimators).

The NN method applies the following rule: the forecasted y-pattern paired with the input x-pattern is the same as the y-pattern paired with nearest neighbour of the input x-pattern found in the reference set.

The MLP model consisted of only one linear neuron and was trained using the Bayesian regularization. For each day type and hour of the day a separate net (24 inputs and 1 output) was created and trained. That simple net structure was one of the best comparing to other structures tested in [5] because of good generalization properties.

The forecast results for these methods are presented in Table 2.

Table 2. Forecast errors for the test sets

Data symbol	N-WE			NN	MLP	FE
	Without selection	SFS	SBS			
A	1.73	1.84	1.77	1.94	2.02	1.76
B	2.05	2.19	2.06	2.55	2.24	2.14
C	4.12	4.47	4.43	5.12	4.89	4.08
D	3.63	3.64	3.60	3.98	3.71	3.63
E	8.32	8.56	8.57	9.18	8.32	8.24

More detailed results for the test part of A time series are presented in Fig. 1.

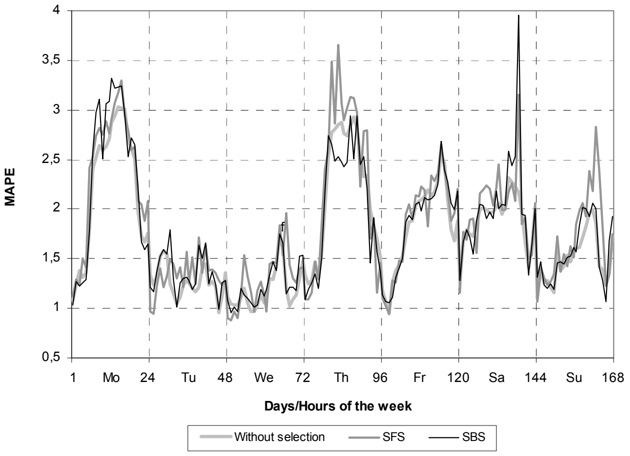

Fig.1. The forecast errors for data A

The variable selection methods caused the average reduction in the input numbers from 24 to 4.8 using SFS and to 13 using SBS. The training error decreased in all cases when SBS was used, but not always when SFS was used. This is because the variable selection methods are nonresistant to the local minima traps. The test error stayed unreduced in almost all cases after using variable selection methods or even was significantly higher (peaks observed on Fig. 1). The reason for this can be that the relation between selected input variables and output variable, which is determined on training set, is not valid on test set. Other, especially stochastic variable selection methods (genetic algorithms, simulated annealing, tournament searching [7]), not susceptible to the local minima traps, can improve results.

Conclusions

Nonparametric regression based on the kernel estimators in application to STLF, presented in this article, is a simple and theoretically well-founded method. The number of parameters, equals the number of the input variables, stays on the reasonable level, what makes good generalisation properties of the model. The parameter estimation can be done using rough (like in this work), gradient or stochastic methods. There is no problem to improve additional explanatory variables to the model, e.g. weather conditions [8].

Unlike the parametric models, where information contained in dataset is compressed into a set of equations, nonparametric models use all historical data and search through them for similar cases each time a forecast is made. For datasets used in STLF (hundreds or thousands of observations) the searching process is not time consuming.

In comparison to other STLF models, the proposed model gives very good results. Similar accuracy has been achieved by the model based on fuzzy estimators, which can be seen as a simplified version of the model based on kernel estimators. This type of models shows very important and desirable features – limited sensitivity to the parameter values (bandwidths) and to incomplete input information (indefinite values of some input variables) [5, 6].

The study was supported by the Research Project N N516 415338 financed by the Polish Ministry of Science and Higher Education.

REFERENCES

[1] Härdle W.K., Müller M., Sperlich S., Werwatz A., Nonparametric and Semiparametric Models, Springer 2004

[2] Kulczycki P,05). Kernel Estimators in Systems Analysis, WNT, Warsaw 2005 (in Polish)

[3] Scott D.W., Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, New York, Chichester 1992

[4] Theodoridis S., Koutroumbas K., Pattern Recognition, Elsevier Academic Press 2003

[5] Dudek G., Short-Term Load Forecasting using Fuzzy Clustering and Genetic Algorithms, *Final report of the Polish State Committee for Scientific Research founded grant no. 3T10B02329.* Dept. Elect. Eng., Częstochowa University of Technology 2006 (unpublished, in Polish)

[6] Dudek G., Similarity-Based Approaches to Short-Term Load Forecasting, in *Forecasting Models: Methods and Applications*, pp. 161-178, iConcept Press 2010

[7] Dudek G., Tournament Searching Method to Feature Selection Problem, in: Rutkowski L., Tadeusiewicz R., Zadeh L., Zurada J. (eds): *Lecture Notes in Artificial Intelligence*, Springer, Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing ICAISC 2010 (in print).

[8] Charytoniuk W., Chen M.S., Van Olinda P., Nonparametric Regression Based Short-Term Load Forecasting, *IEEE Transactions on Power Systems*, 13 (1998), n.3, 725-730

***Author**: Grzegorz Dudek PhD, Czestochowa University of Technology, Institute of Power Engineering, al. Armii Krajowej 17, 42-200 Czestochowa, Poland, E-mail: Dudek@el.pcz.czest.pl.*

.