Non-parametric Regression Model for Continuous-time Day Ahead Load Forecasting with Bernstein Polynomial

Roya Nikjoo*, Abouzar Estebsari†, Mohammad Nazari*
*KTH Royal Institute of Technology, Stockholm, Sweden
†Department of Energy, Politecnico di Torino, Turin, Italy

***Abstract*—Growing perception of diverse generation resources and demand response operation of power system with high uncertainty has increased the attention to a more dynamic and accurate day-ahead load prediction. In this paper, we develop an stochastic model for short term load forecasting based on the Gaussian process, in which the non parametric estimator of the regression functions are obtained by using Bernstein polynomials. One of the major features of this model is its ability to predict a continuous load at any time of the day with a regression function. We use the historical data for training and the constrained marginal likelihood problem is optimized for finding the hyperparameters of the model. Real data sets from California ISO were used for training and testing the model. The results are compared to the day ahead piecewise constant load and the real time load. The common error measures are employed to infer the deviation of the load forecast from the real data.**

***Index Terms*—Bernstein polynomials, Load Profile, Regression Model, Non-parametric**

I. Introduction

In power system operation, the electricity demand must be well matched with the supply. This fact requires an accurate dynamic load forecasting that the generation must fulfill. Beside automatic generation control, other factors such as unit commitment and economic dispatch are also dependent on the load forecast. However, uncertainty in the load demand and power generation availability leads to uncertainty in the day-ahead scheduling. This characteristic encourages the use of stochastic process to define the probability of supply and demand for unit commitment which needs to be done several hours in advance. The current unit commitment provides hourly schedules for the generations and assumes the load demand to be constant at each hour. However, in reality neither generation nor demand has the stepwise profile at each hour. Several load forecasting approaches for short term (hour to week ahead) time scales are proposed. They can be categorized as artificial intelligence methods and statistical techniques. Artificial Neural Network (ANN) [1] and neuro-fuzzy network [2] are between those artificial intelligence methods that has been used for load forecasting. There are several works that have been done on hourly load forecasting in day-ahead scheduling with stochastic models [3]. Among statistical techniques, regression models and time series are more popular. Time series methods assume internal structures for data, such as autocorrelations and trends, which they try to detect and explore. ARMA (Autoregressive Moving Average) for stationary processes and ARIMA (Autoregressive Integrated Moving Average) for nonstationary processes are most common used time series method [4], [5]. Regression models [6] uses several features for the linearization of the function to interpret the relation of different factors in an easy way. Multivariate polynomial and exponential regression [7] has been proposed for both short term and medium term load forecasting. Support vector machine, based on statistical learning theory is another method that has been used for load forecasting [8]. none of the methods that has been applied for the short term load forecasting offers a continuous time load profile estimation. There are several factors which can affect the load forecasting including the time, weather and customer classes [9]. However, in this work, we mostly focus on continuous time modeling of hourly loads without specifically involving other parameters. This paper describes a stochastic model based on the Gaussian process with Bernstein polynomial for the continuous time load estimation. We develop a stochastic model which estimate a continuous time function for each hour in the load profile.

II. Gaussian Process

In a Gaussian process, every input as x is associated with a normally distributed random variable. It is often shown as a vector x as it can include several variables. The output or target y may either be continues or discrete. For the first case, the process is known as a regression and for the latter one as a classification. Sets of samples with n observation $D = (xi, yi)|i = 1, , n$ is using as a training data to define a function f that predicts the output for any possible input variable. Therefore, the characteristics of this function needs to be defined. Estimation of this function is based on the prior information and a finite set of observations. Wide range of functions can be assumed which Gaussian process can help in choosing between infinite set of possible functions. A Gaussian stochastic process governs the properties of the function. Prior is considered as our belief about the function in absence of knowledge about the data. Gaussian process specifies that the prior variance doesnt depend on the x variable and also it can be assumed that the mean of $f(x)$ for any x is zero independent of its value. Then the combination of the prior and data leads to the posterior distribution of

the function. Gaussian process is not a parametric model and doesnt intend to fit the data but its uncertainty reduces close to the observations. The specification of prior is important for inference of the function property. Smoothness and stationary properties can be specified by the covariance of the Gaussian process. Therefore finding a suitable covariance function is a challenge. In time series, the key aspect is how observations are related to each other in time. This concept is formalized through the covariance between elements which measures the degree of second order variation between two elements at two different times. If the statistical properties doesnt change over time, e.g. expectations, variances and other properties remain the same, the process is considered as stationary. If the mean or the variance of the time series changes over time, the process is non-stationary. The aim is to inference about the relationship between the inputs and the targets. Standard linear regression model with Gaussian noise can be defined as:

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

x is the input vector, f is the function value and y is the observed target data. The observed value differs from the function values by additive Gaussian noise with zero mean and variance σ_n^2, $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. The vector of the noise or error is distributed normally and is independent of the random function f. A Gaussian process is completely specified by its mean $m(x)$ and covariance $k(x, x')$ function.

$$f(x) \sim \mathcal{G}P(m(x), k(x, x')) \tag{2}$$

$$y \sim \mathcal{G}P(m(x), k(x, x') + \sigma_n^2 \delta_{ii'}) \tag{3}$$

where $\delta_{ii'}$ is the Kronecker's delta and $\delta_{ii'} = 1$ if $i = i'$, otherwise it is zero. If there wouldn't be any noise in the observations, the terms related to the Gaussian noise can be neglected. Mean function $m(x)$ and covariance function of $f(x)$, $k(x, x')$ are defined as:

$$\begin{aligned} m(x) &= \mathbb{E}[f(x)] \\ k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \end{aligned} \tag{4}$$

E stands for the expected value. To define $f(x)$ in the form of Gaussian distribution, it will be:

$$\mathbf{f} \sim \mathcal{N}(\mu, \Sigma) \tag{5}$$

where μ and Σ are vectors of the mean and the covariance. For multivariate normal distribution with positive definite covariance matrix, the Gaussian probability density function is expressed by:

$$p(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \, exp(-\frac{1}{2} \, (y - \mu)^T \Sigma^{-1} (y - \mu)) \tag{6}$$

The regression function $f(x)$ is supposed to satisfy certain shape restriction with nonparametric form for all x values and with realization of the observation points. In our problem, Gaussian process is defined over the time, where the index of set of variables is time. For a fixed time, our aim is to estimate:

$$\hat{f}(x_t) = E[f(x_t)|y(t_1), ..., y(t_n)] \tag{7}$$

where $\hat{f}(x_t)$ is the solution of the prediction problem which can be obtained by the maximum likelihood estimation. Maximum-likelihood estimation (MLE) is a method for estimating the parameters of a statistical model with available observation data. It is an interesting approach which can be used also for continuous time processes [9]. As our data assumed to have Gaussian distribution, logarithmic marginal likelihood function is defined as:

$$log(L) = -\frac{1}{2} \, log|\Sigma| - \frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu) - \frac{n}{2} \, log(2\pi) \tag{8}$$

We use the term marginal to emphasize that we are dealing with a non-parametric model. Assume Bayesian linear regression model for $f(x)$:

$$f(x) = \varphi(x)^T \beta \tag{9}$$

where β is a vector of weights and considered to be a prior parameter. $\varphi(x)$ is a function that maps the D dimensional input vector x to N dimensional feature space using a sets of basis function (e.g. Bernstein polynomials). As long as the projections are fixed functions, the model is linear in the parameters [10]. We need to define the prior parameters before look at the observation. Prior is not dependent on the training data but it has some properties of the function. We take prior on β to be Gaussian:

$$\beta \sim \mathcal{N}(b, \Sigma_b) \tag{10}$$

without loss of generality we assume for now that the mean is zero, $b = 0$. Then the covariance function of the $f(x)$ will be:

$$k(x, x') = \varphi(x)^T \Sigma_b \varphi(x) \tag{11}$$

As the Gaussian process uses priors, the smoothness of the prior is defined by the covariance function. Choosing a proper covariance matrix for our model is important. For example, if we expect that for close-by input variables, the output will be close (continuity assumption), the covariance function must satisfy this characteristic. The covariance between the outputs is written as a function of the inputs. There are a number of common covariance functions available. The squared exponential covariance is used in our model which corresponds to a Bayessian linear regression model with an infinite number of basic functions:

$$cov(f(x_p), f(x_q)) = k(x_p, x_q) = exp(-\frac{1}{2}|x_p - x_q|^2) \tag{12}$$

This function is infinitely differentiable. If the covariance wont be zero, we say that the errors of x_p and x_q are correlated. The covariance can be normalized to form a correlation coefficient:

$$r = \frac{k(x_p, x_q)}{\sqrt{k(x_p, x_p) k(x_q, x_q)}} \tag{13}$$

Usually the covariance functions have some free parameters called hyperparameters. In case of vague prior information, we use a hierarchical prior, where the mean and covariance functions are parameterized in terms of hyperparameters. The

squared exponential covariance function with hyperparameters in one dimension has the following form:

$$k_y(x_p, x_q) = \sigma_f^2 exp(-\frac{1}{2l^2}(x_p - x_q)^2) + \sigma_n^2 \delta_{pq} \quad (14)$$

l is called the length scale of the process which practically shows how close two points have to be to notably influence each other. σ_f^2 is the signal variance and δ_{pq} is a Kronecker delta which is one when p = q and zero otherwise. We can now find the values of the hyperparameters which optimizes the marginal likelihood based on its partial derivatives which are easily evaluated.

$$\frac{\partial L}{\partial \theta_m} = -(y-\mu)^T \Sigma^{-1} \frac{\partial m}{\partial \theta_m} \quad (15)$$

$$\frac{\partial L}{\partial \theta_k} = \frac{1}{2} trace(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_k}) - \frac{1}{2}(y-\mu)^T \frac{\partial \Sigma}{\partial \theta_k} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_k}(y-\mu) \quad (16)$$

θ_m includes the hyperparameters of mean function $\mu = m(x)$ (which can be the coefficient of polynomial function) and θ_k the hyperparameters of the covariance function, $\Sigma = k(x, x')$.

III. Predictive Gaussian Process

By combining the likelihood and the prior, the posterior gather everything we know about the parameters.The major goal of the posterior is to be used for the prediction of the future cases.

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma_* \\ \Sigma_*^T & \Sigma_{**} \end{bmatrix}) \quad (17)$$

IV. Bernstein Polynomial Estimator

Polynomials are among popular mathematical tools for approximating and forming spline curves of any function with any desired accuracy. They have a variety of basis functions which can fit different applications. They can be easily differentiated and integrated and they have other useful properties. Bernstein basis is a commonly used base for the space of polynomial which can suitably be used for load profile modeling. The Bernstein polynomial estimator can be used for nonparametric curve estimations. It has many useful properties in preserving the different shapes of the regression functions. Using Bernstein as an estimator is computationally efficient as its coefficients can be computed easily as a solution of quadratic programming problem [11].

The Bernstein basis polynomial functions of order N are defined as:

$$b_k(x, N) = \binom{N}{k} x^k (1-x)^{N-k}$$

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

$$for \quad k = 0, 1, ..., N \quad and \quad x \in [0 \quad 1] \quad (18)$$

And the Bernstein polynomial of degree N can be expressed by:

$$B_N(x,) = \sum_{k=0}^{N} \beta_k . b_k(x, N) \quad (19)$$

Where β_k is a coefficient of Bernstein polynomial. Load profile at each hour (segment) can be addressed in the interval of $[01]$ which can have several data points (e.g. the load value every 5 minutes) and each segment can be modeled by one Bernstein polynomial. However, if each segment will be defined in the closed interval of $[a, b]$, then the corresponding polynomial is written as:

$$B_N(x, \beta) = \frac{1}{(b-a)^N} \sum_{k=0}^{N} \beta_k . \binom{N}{k} (x-a)^k (b-x)^{N-k} \quad (20)$$

Assume to have a set of observations $(X_i, Y_i) \in \{(X_1, Y_1), ..., (X_n, Y_n)\}$, to obtain the spline approximation of the load curve from the observations, the Bernstein coefficients must be calculated. Least Squares method is a standard approach to approximate the solution by minimizing the sum of the squares of the errors made in the results.

$$min \quad ||\mathbf{b_k}(X_i)^T \beta_{\mathbf{N}} - Y_i||^2 \quad (21)$$

where $\mathbf{b_k}(x) = (b_0(x, N), ..., b_N(x, N))'$ and $\beta_{\mathbf{N}} = (\beta_{0,N}, ..., \beta_{N,N})'$

V. Continuity at the joints

As the goal is the piecewise approximation of the load profile, series interconnection of hourly curves linked together at joints must be considered. This requires the continuity conditions at the joints or control points. Different level of continuity can be defined. Parametric continuity of C0 refers to the continuity of two consecutive segments at the joint point:

$$B_N^s(x_j, \beta) = B_N^{s+1}(x_j, \beta) \quad (22)$$

Where s is the number of segment and j is the index of a joint point of the segments. To consider the continuity condition in the Bernstein modeling problem, the constrained least square problem must be solved. The C^0 continuity condition is added as an equality constrained to the optimization problem and least square formula will be minimized subjected to that. However, we might require other orders of continuity to satisfy the smoothness and accuracy of the approximation. The curve segments should have the same slopes when they join together. Therefore, C^1 continuity refers to the slope matches where the curves join.

$$B'_N{}^s(x_j, \beta) = B'_N{}^{s+1}(x_j, \beta) \quad (23)$$

Derivative of Bernstein polynomials of degree N are polynomials of degree $N-1$ and this derivative can be written as a linear combination of Bernstein polynomials:

$$\frac{d}{dx} b_k(x, N) = N(b_{k-1}(x, N-1) - b_k(x, N-1)) \quad (24)$$

The C^1 continuity condition adds more equality constraints to the optimization problem.

VI. Constrained least square problem

Our optimization problem is maximizing the marginal likelihood with respect to the hyper parameters. In addition to that the optimization must also satisfy continuity criteria. The load profile estimation must have the C^0 continuity at joints, and that the slope of the curve at joints, must have C^1 continuity. This is a linearly-constrained quadratic minimization, an ideal problem for Lagrange multipliers. If the covariance considered to be constant, the marginal likelihood must be optimized with respect to the Bernstein coefficients and satisfies the continuity conditions. By substituting the mean function $m(x) = \mu$ with Bernstein polynomial function $\mathbf{b_k}(t)^T\beta_\mathbf{N}$, the augmented objective function (the Lagrangian) is then:

$$\begin{aligned}\mathcal{L}(\beta_\mathbf{N}, \lambda) =& y^T\Sigma^{-1}y - y^T\Sigma^{-1}\mathbf{b_N^T}\beta_\mathbf{N} - {\beta_\mathbf{N}}^T\beta_\mathbf{N}\Sigma^{-1}y + \\ & {\beta_\mathbf{N}}^T\beta_\mathbf{N}\Sigma^{-1}\mathbf{b_N^T}\beta_\mathbf{N} + \lambda^T(C\beta_\mathbf{N} - d)\end{aligned} \tag{25}$$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\beta_\mathbf{N}} =& -y^T\Sigma^{-1}\mathbf{b_N^T} - (\mathbf{b_N}\Sigma^{-1}y)^T + \\ & (\mathbf{b_N}\Sigma^{-1}\mathbf{b_N^T}\beta_\mathbf{N})^T + {\beta_\mathbf{N}}^T\mathbf{b_N}\Sigma^{-1}\mathbf{b_N^T} + \lambda^T C = 0\end{aligned} \tag{26}$$

C and d are the matrices with continuity parameters. Minimizing $\mathcal{L}$ with respect to $\beta_\mathbf{N}$ and λ results in a system of linear equations for the optimum coefficients ${\beta_\mathbf{N}}^*$ and Lagrange multipliers λ^*.

$$\begin{bmatrix}{\beta_\mathbf{N}}^* \\ \lambda^*\end{bmatrix} = \begin{bmatrix}2\mathbf{b_N}\Sigma^{-1}\mathbf{b_N^T} & C^T \\ C & 0\end{bmatrix}\begin{bmatrix}2\mathbf{b_N}\Sigma^{-1}y \\ d\end{bmatrix} \tag{27}$$

If C has independent rows and $\begin{bmatrix}\beta_\mathbf{N} \\ C\end{bmatrix}$ has independent column, the KKT matrix is invertible.

VII. Model Accuracy Evaluation

To evaluate our model, two error types have been considered and employed. The Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE) are common types of error estimations. While the MAE gives the same weight to all errors, the RMSE gives errors with larger absolute values more weight than errors with smaller absolute values. Mean Average Percentage Error (MAPE) also has been commonly used to measure the forecasting performance. Root mean squared residual error at each point provides the squared difference between the observation values y and the predicted function m(x). For an unbiased estimator, the RMSD is the square root of the variance, known as the standard error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m(x_i) - y_i)^2} \tag{28}$$

Mean absolute error is used to know how close predictions are to the actual observations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|m(x_i) - y_i| \tag{29}$$

$$MAPE = 100.\frac{1}{n}\sum_{i=1}^{n}|\frac{m(x_i) - y_i}{y_i}| \tag{30}$$

The root mean squared prediction error is computed on out-of-sample data:

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m(x_i^*) - y_i^*)^2} \tag{31}$$

$m(x_i^*)$ is the estimated function at test values and y_i^* is the true values for the test data points.

VIII. Result and Discussion

To analyze our proposed model, 26 sets of five-minute net-load data from California Independent System Operator CAISO, is used for training the process. CAISO is the largest Independent System Operator ISOs in the world managing about 80 percent of California's electric flow. These data are selected for a specific day of a week from six consecutive months (May-Oct), Fig. 1. The load has been normalized with respect to its maximum. The results are compared to the piece wise constant loads to show the effectiveness of continuous load forecasting for each hour. As the load for each hour interval can be modeled with a continuous Bernstein function, it will provide a more realistic load profile which can be used for the real-time economic dispatch. Bernstein polynomial of any degree can be used for the modeling but there is an optimal degree which can be a better fit to the load profile. In order to choose a proper order of Bernstein polynomials, the prediction error is estimated for different orders from N=1,...,10.

TABLE I
Errors for different Bernstein polynomial orders

N	RMSE	MAE
1	4.1254	28.4927
2	1.5036	15.6126
3	1.5036	15.6128
4	1.5036	15.6142
5	1.5036	15.6128
6	33.0849	37.3396
7	392.2946	90.8602
8	6.1328e+03	414.6701
9	6.4669e+12	1.0410e+07
10	6.0388e+13	2.6205e+07

As it can be seen in the table I, Bernstein order from 2 to 5 can fit the load profile with reasonable accuracy but with orders above 5, it cannot follow the shape well anymore and will show large oscillations for certain parts.

As we haven't considered the weather variables such as temperature in the model, the collected historical data needs to be in the same weather condition to cover the close load profiles with the least diversity. Forecasting a day-ahead load profile also requires the use of the trained model with the same seasonal condition. Load profile of each hour of the day get modeled by continuous Bernstein polynomials. Scheduling of the generation with their real ramping will benefit from the continuous load forecasting and leads to economical benefits.

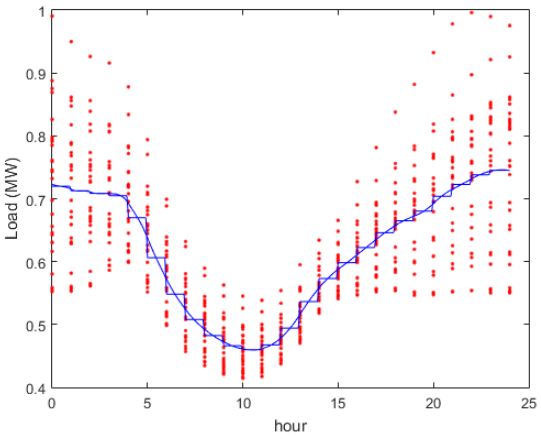

Fig. 1. Load forecasting with Bernstein model of order 5

IX. Conclusion

In this paper, a stochastic model based on the Gaussian process was presented for a short-term load forecasting. In this model, the non-parametric estimator of the regression functions is obtained by using Bernstein polynomials. Real data sets from California ISO were used as the historical data for training and the constrained marginal likelihood problem was optimized for finding the hyperparameters of the model. The common error measures were employed to infer the deviation of the load forecast from the real data. Bernstein polynomial of any degree could be used for the modeling but there are optimal degrees which can be a better fit to the load profile. As the load for each hour interval can be modeled with a continuous Bernstein function, it will provide a more realistic load profile which can be used for the real-time economic dispatch.

References

[1] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: a review and evaluation," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44–55, Feb 2001.

[2] Y. Bodyanskiy, S. Popov , and T. Rybalchenko, "Multilayer neuro-fuzzy network for short term electric load forecasting," *Lecture Notes in Computer Science*, vol. 5010, pp. 339–348, Feb 2008.

[3] E. A. Feinberg and D. Genethliou, "Chapter 12 load forecasting."

[4] M. T. Hagan and S. M. Behr, "The time series approach to short term load forecasting," *IEEE Transactions on Power Systems*, vol. 2, no. 3, pp. 785–791, Aug 1987.

[5] N. Amjady, "Short-term hourly load forecasting using time-series modeling with peak load estimation capability," *IEEE Transactions on Power Systems*, vol. 16, no. 3, pp. 498–505, Aug 2001.

[6] W. Charytoniuk, M. S. Chen, and P. Van Olinda, "Nonparametric regression based short-term load forecasting," *IEEE Transactions on Power Systems*, vol. 13, no. 3, pp. 725–730, Aug 1998.

[7] A. Nazib, F. Elkarmi, and O. Aloquili, "Medium-term electric load forecasting using multivariable linear and non-linear regression," *Smart Grid and Renewable Energy*, vol. 2, no. 2, pp. 126–135, 2011.

[8] E. E. Elattar, J. Goulermas, and Q. H. Wu, "Electric load forecasting based on locally weighted support vector regression," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 4, pp. 438–447, July 2010.

[9] Y. Feng, D. Gade, S. Ryan, J.-P. Watson, R. J.-B Wets, and D. Woodruff, "A new approximation method for generating day-ahead load scenarios," 01 2013, pp. 1–5.

[10] C. E. Rasmussen, "Gaussian processes for machine learning." MIT Press, 2006.

[11] P. D. Feigin, "Maximum likelihood estimation for continuous-time stochastic processes," *Advances in Applied Probability*, vol. 8, no. 4, pp. 712–736, 1976. [Online]. Available: http://www.jstor.org/stable/1425931