Continuous-time Stochastic Modeling and Estimation of Electricity Load

Roohallah Khatami, *Student Member, IEEE*, Masood Parvania, *Member, IEEE*, Pramod Khargonekar, *Fellow, IEEE*, and Akil Narayan

***Abstract*—The current discrete-time (e.g., hourly) modeling and prediction methods fall short in capturing and anticipating the sub-interval variations of electricity load. This leads to inability of power system operators to appropriately utilize the available resources to follow and compensate the load variations. This paper takes a novel and different approach on modeling electricity load, and proposes a continuous-time model for characterizing the uncertainty and variability of load. More specifically, the electricity load is modeled as a continuous-time stochastic process that is projected on a reduced-order function space spanned by Bernstein polynomials, which ensures the continuity of the process over the estimation and forecasting horizons. We assume a Gaussian process (GP) prior on the load process and design a covariance function that reflects the periodicity and smoothness of electricity load. We develop a computationally efficient method for estimating the hyper-parameters of the model using the solution of a maximum likelihood estimation problem and form the posterior GP process. The proposed method is utilized to model and predict the load of California Independent System Operator (CAISO). The proposed model uniquely predicts the continuous-time mean value and uncertainty envelopes of future CAISO load, which inherently embeds information on the continuous-time variations and the associated ramping requirements of the load.**

I. Introduction

A. Background on Load Modeling and Prediction

Electricity load modeling and prediction is a fundamental stage of power systems operation and exploring mathematical load modeling techniques dates back to decades earlier [1]. Remarkable efforts in this realm have culminated in developing several load forecasting methods among them the linear regression, stochastic time series, exponential smoothing, artificial neural networks (ANN), and Gaussian process (GP) predictors [2], [3]. These methods cover a wide range of modeling time-scales from short-term to medium- and long-term, where our focus in this work is the short-term load model.

The linear regression methods offer more simplicity as compared to the other methods, and their flexibility is commensurate with the richness of opted functions [1], [4]. Richer functions, however, do not necessarily guarantee the accuracy of modeling and prediction, specifically in the presence of high noise levels. In fact, excessively complex functions may perform well in modeling the observed data, though they may suffer from overfitting [5] that brings about considerable prediction errors. The stochastic time series models, including autoregressive moving average and autoregressive integrated moving average [6]–[8], alleviate to some extent the overfitting problem and reflect implicitly the correlation of forecasted data through the feedback of time-lagged load values and the associated error terms. Exponential smoothing methods assign exponentially decreasing factors to the time-lagged components of time series, such that farther observations have less impact on the present forecast [9], [10].

The ANN and GP predictors follow the principles of supervised learning [11]–[14]. In [12], the uncertainty intervals of the predicted load points are also furnished as the outputs of the ANN model, yet the covariance matrix of the forecasted load points is not provided. The GP models assume a Gaussian prior on the stochastic load process [13], [14], and derive the Gaussian predictive distribution as the posterior distribution of future load points. The predictive distribution not only provides the pointwise load values, but also their covariance matrix. Further, in derivation of GP predictors a compromise between data-fit and complexity is inherently made through the maximum likelihood problem, which alleviates to a good extent the overfitting issue [13].

Several works have leveraged the appealing properties of GP predictors to forecast electric load or wind power [15]–[18]. In [15], the authors compare the performance of GP and ANN models and showcase the computational merit of the former over the latter. In [16], an ensemble prediction model composed of 52 ANN and 5 GP sub-models are used to forecast the 48 hour ahead wind power. The GP sub-models serve as auxiliary predictors to provide initial points for the main ANN models. In [17], a censored GP model is used along with data from numerical weather prediction for mapping wind speed values to wind turbines output power, taking into account the wind direction, temperature, air pressure, and humidity. The model in [18] utilizes a teaching learning based optimization to accelerate the GP learning process.

B. Continuous-time Load Modeling

The load of power systems varies continuously in time, and its variability and uncertainty is best characterized by a continuous-time stochastic process. Integrating a continuous-time load trajectory in power systems control and operation problems, however, would result in a continuous-time optimal control problem with infinite dimensional decision space that is computationally intractable to solve. The common practice to overcome this problem has been to divide the modeling time horizon using a finite number of sampling points and

This work was supported in part by the U.S. NSF grant ECCN-1549924, and in part by the U.S. DOE grant DE-OE0000882.

R. Khatami and M. Parvania are with the Department of Electrical and Computer Engineering, the University of Utah, Salt Lake City, UT 84108 USA (e-mails: roohallah.khatami@utah.edu, masood.parvania@utah.edu).

P. Khargonekar is with the Department of Electrical Engineering and Computer Science, the University of California, Irvine, CA 92697 USA (e-mail: pramod.khargonekar@uci.edu).

A. Narayan is with the Department of Mathematics, the University of Utah, Salt Lake City, UT 84108 USA (e-mail: akil@sci.utah.edu).

approximate the continuous-time load trajectory with a zero-order piecewise constant trajectory, as shown in Fig. 1-(a). This time-discretization approach, though, does not appropriately capture the load dynamics and variations in smaller time scales, and neglects a great deal of prior information about the load process. Increasing the accuracy of piecewise constant load approximation would require an increasing number of sampling points, which would in turn increase the dimensionality of the associated power system control problem that the load is considered as an input to.

In a series of recent works, we have developed an alternative approach for sampling the load trajectories and control decisions of the associated control problems in power systems operation [19]–[23]. The approach projects the realizations of load trajectory on a countable and finite-dimensional function space, as schematically shown in Fig. 1-(b). In [19], [20], we have shown that Bernstein polynomials represent a perfect choice for modeling the load trajectory, and leveraged multiple properties of Bernstein polynomials (e.g., convex hull property) for scalable and accurate solution of the associated optimal control problems. However, in our past works (and to the best of our knowledge in other works), the focus has been on deterministic continuous-time representation of load trajectories and the development of stochastic continuous-time load models that would account for the inherent uncertainty of load in a continuous-time fashion is remained unexplored.

C. Contribution and Paper Structure

In this paper, we expand the function space representation of load trajectories using Bernstein polynomials in [19]–[22], and develop a continuous-time stochastic process model for electricity load. More specifically, the electricity load is modeled in Section II as a continuous-time stochastic process that is projected on a function space spanned by Bernstein polynomials, where the projection coefficients are random variables and form a multivariate probability distribution. In Section III, we adapt the Bayesian inference method to estimate the proposed load process using the past observations of the load. We assume a GP prior on the stochastic load process and design a covariance function that includes a periodic squared exponential and a pure squared exponential kernel function. We then estimate the hyper-parameters of the model using the solution of a maximum likelihood estimation problem and form the posterior GP model. The posterior GP model is then utilized to develop the predictive process that predicts the mean and covariance of future load process. The proposed continuous-time GP load model is utilized in Section IV to develop a model for the real load data of California Independent System Operator (CAISO), and predict the future load values. The conclusions are drawn in Section V.

II. Function Space Representation of Stochastic Load Process

Let us assume that the load of power systems, $D(t)$, over $t \in \mathcal{T}$ is defined on a filtered probability space, $(\Omega, \mathcal{F}, \mathbb{P}, \mathfrak{F})$, with the continuous sample space Ω, the set of events $\mathcal{F}$, the probability measure $\mathbb{P}$, and the filtration $\mathfrak{F}$. The continuous-time trajectory $D_\omega(t)$ denotes a realization ω of the sample space Ω. Let us subdivide the modeling horizon $\mathcal{T}$ into I intervals $\mathcal{T}_i = [t_i, t_{i+1}), \rightarrow \mathcal{T} = \cup_{i=0}^{I-1} \mathcal{T}_i$, with lengths $T_i = t_{i+1} - t_i$, and construct a subset of basis functions formed by the Bernstein polynomials of degree Q in each interval $\mathcal{T}_i$, forming a spline function space $\mathbf{e}^{(Q)}(t) = (e_1^{(Q)}(t), \dots, e_P^{(Q)}(t))^T$ to represent the whole $\mathcal{T}$, which contains $P = (Q+1)I$ functions with components defined as:

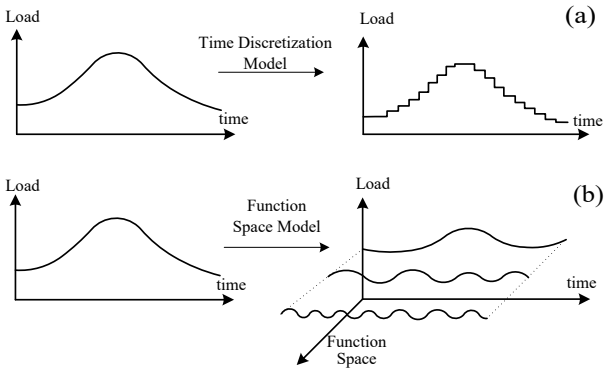

Fig. 1. Load trajectory: a) pointwise model, b) function space model

$$e^{(Q)}_{i(Q+1)+q}(t) = b_{q,Q}\left(\frac{t - t_i}{T_i}\right), t \in [t_i, t_{i+1}), \tag{1}$$

for $i = 0, \dots, I-1; q = 0, \dots, Q$, where $b_{q,Q}(t)$ represents the Bernstein polynomials of degree Q defined as [24]:

$$b_{q,Q}(t) = \binom{Q}{q} t^q (1-t)^{Q-q}, \quad t \in [0, 1). \tag{2}$$

The Bernstein function space $\mathbf{e}^{(Q)}(t)$ guarantees the continuity of desired order at the internal points of intervals, however, maintaining C^1 continuity of the load trajectory at connection points imposes constraints on the Bernstein coefficients of adjacent intervals, which forms a new reduced-order Bernstein function space $\mathbf{w}^{(Q)}(t) = \mathbf{M}\mathbf{e}^{(Q)}(t)$ with dimension $Z = (Q-1)I + 2$, where $\mathbf{M}$ is a $Z \times P$ linear mapping matrix.

Here, we propose to represent the stochastic load process $D(t)$ on the Bernstein function space $\mathbf{w}^{(Q)}(t)$ as follows:

$$D(t) = \mathbf{B}\mathbf{w}^{(Q)}(t), \quad t \in \mathcal{T}, \tag{3}$$

where $\mathbf{B} = (B_1, B_2, \dots, B_Z)$ is the Z-dimensional vector of random variables representing the Bernstein coefficients of projecting $D(t)$ in $\mathbf{w}^{(Q)}(t)$. The Bernstein coefficients in $\mathbf{B}$ together form a multivariate distribution system with mean and covariance functions $\boldsymbol{\mu}_\mathbf{B}$ and $\boldsymbol{\Sigma}_\mathbf{B}$ defined as:

$$\boldsymbol{\mu}_\mathbf{B} = \mathbf{E}[\mathbf{B}], \tag{4}$$

$$\boldsymbol{\Sigma}_\mathbf{B} = \mathbf{E}[(\mathbf{B} - \boldsymbol{\mu}_\mathbf{B})^T (\mathbf{B} - \boldsymbol{\mu}_\mathbf{B})], \tag{5}$$

where $\mathbf{E}[\cdot]$ is the expected value operator.

Given the function space representation (3), we model the probability measure $\mathbb{P}$ of the stochastic load process $D(t)$ with

the mean and covariance functions $\mathbb{D}(t)$ and $cov(D(t), D(t'))$ respectively calculated as follows:

$$\mathbb{D}(t) = \mathbf{E}[D(t)] = \mathbf{E}[\mathbf{B}]\mathbf{w}^{(Q)}(t) = \boldsymbol{\mu}_{\mathbf{B}}\mathbf{w}^{(Q)}(t), \quad t \in \mathcal{T}, \quad (6)$$

$$\begin{aligned} cov(D(t), D(t')) &= \mathbf{E}[(D(t) - \mathbb{D}(t))(D(t') - \mathbb{D}(t'))] \\ &= \mathbf{w}^{(Q)^T}(t)\ \mathbf{E}[(\mathbf{B} - \boldsymbol{\mu}_{\mathbf{B}})^T(\mathbf{B} - \boldsymbol{\mu}_{\mathbf{B}})]\ \mathbf{w}^{(Q)}(t') \\ &= \mathbf{w}^{(Q)^T}(t)\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{w}^{(Q)}(t'), \quad (t, t') \in (\mathcal{T}, \mathcal{T}), \end{aligned} \quad (7)$$

We wish to learn the stochastic load process (3) through learning the mean and covariance functions $\mathbb{D}(t)$ and $cov(D(t), D(t'))$. As apparent in (6) and (7), this boils down to learning the mean and covariance functions $\boldsymbol{\mu}_{\mathbf{B}}$ and $\boldsymbol{\Sigma}_{\mathbf{B}}$ of the Bernstein coefficients $\mathbf{B}$, which is discussed next.

III. Learning the Stochastic Load Process

In this section, we aim to adapt the Bayesian inference method to learn the parameters $\boldsymbol{\mu}_{\mathbf{B}}$ and $\boldsymbol{\Sigma}_{\mathbf{B}}$ of the stochastic load process $D(t)$ given the past load observations as realizations of the process. In Bayesian inference, we first assume a prior distribution over the parameters and then calculate the posterior process, which is a conditional probability distribution conditioned on a set of noisy observations. The posterior not only estimates the structure of the observed load data, but is also used to estimate future load realizations.

Our primary task here is to estimate the mean and covariance of the Bernstein coefficients $\mathbf{B}$. We may assume a prior on $\mathbf{B}$, e.g., multivariate normal distribution, and formulate a maximum likelihood estimation problem to estimate the posterior distribution, given the past observations of the load process. The main drawback of this approach is that the Z^2 components of the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{B}}$ would be variables of the maximum likelihood problem, which would be computationally burdensome (e.g., Z equals 50 for Bernstein coefficients of degree 3 and 24 hourly spline model). In the next section, we develop a Gaussian Process model, which enables using the kernel trick [13] to formulate a computationally efficient maximum likelihood problem with less variables.

A. Gaussian Process Prior on Stochastic Load Process

Gaussian Process (GP) is generalized multivariate Gaussian distribution where the observations occur in continuous time and any finite subset of the characterizing domain also follows a Gaussian distribution. Here, we assume a GP prior on stochastic load process $D(t)$ over $\mathcal{T}$, i.e., $D(t) \sim \mathcal{GP}\left(\mathbb{D}(t), cov(D(t), D(t'))\right)$. Given the function space representation of $D(t)$ in (3), Bernstein coefficients $\mathbf{B}$ would inherit the Gaussian properties of $D(t)$, and therefore form a multivariate Gaussian distribution, i.e., $\mathbf{B} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{B}}, \boldsymbol{\Sigma}_{\mathbf{B}})$.

The electricity load demonstrates recurring (e.g., daily) patterns that should be appropriately taken into account by the proposed GP load model. Covariance function is the crucial part of the GP model, as it embeds features of the load process (e.g., continuity, smoothness, and periodicity) that we wish to learn. In addition, covariance function encodes the nearness and similarity of the historical load data and supports predicting the future values of load that are likely to follow similar characteristics of the historical data. Therefore, the choice of covariance function is important to reveal and learn the underlying characteristics of the load process. We propose the following kernel function $k(t, t')$ as the covariance function of the GP load model:

$$\begin{aligned} cov(D(t), D(t')) &= k(t, t') \\ &= \sigma_{d_1}^2 \exp\left(-\frac{1}{{l_1}^2}\sin^2\frac{\pi(t - t')}{\tau}\right) \\ &\quad + \sigma_{d_2}^2 \exp\left(-\frac{|t - t'|^2}{{l_2}^2}\right), \ (t, t') \in (\mathcal{T}, \mathcal{T}). \end{aligned} \quad (8)$$

The covariance function (8) includes two kernel functions that embed the periodicity of load, smoothness and continuity, as well as the increasing uncertainty and vanishing dependence between the load realizations over time. More specifically, both terms in (8) are squared exponential functions, which ensure the continuity and smoothness of any order, where l_1, l_2, $\sigma_{d_1}^2$ and $\sigma_{d_2}^2$ respectively represent the characteristic length-scales and the variances for the two terms. In addition, the first term in (8) is a periodic squared exponential that incorporates the periodicity of the load, where τ indicates the period of the load process. As the main focus of the present work is short-term modeling of load (e.g., day-ahead), the first term would only capture the daily pattern of load, though the implementation of weekly and seasonal periodicity is not more challenging and could be achieved by enhancing the covariance function in (8) with additional periodic terms. The second term in (8) is a pure squared exponential kernel function that tends to zero for the future points far from the observation points, i.e., $|t - t'| \gg 0$, implying the increasing vagueness and uncertainty of load prediction for fairly distant future. The periodic squared exponential term in (8), though, is periodic and repeats the same mean and covariance for each period.

B. Estimating the Posterior Gaussian Process

Here we aim to obtain the posterior GP load model by estimating the parameters of the mean and covariance functions of the GP prior, given the past observations of the load process.

Let $\hat{\Omega}$ be the set of past realizations of load during comparable periods of the modeling time horizon $\mathcal{T}$ (e.g., specific day of week). The discrete sample space $\hat{\Omega}$ represents an approximation of the actual continuous sample space Ω of the proposed load model. Let the continuous-time trajectory $D_\omega(t)$ denote a realization $\omega \in \hat{\Omega}$ of the load process, the N-dimensional vector $\mathbf{D}_\omega(t_n) = (D_\omega(t_1), \ldots, D_\omega(t_N))$ represent the noisy discrete-time samples of such realization at discrete times $t_n, n \in \{1, \ldots, N\}$ over $\mathcal{T}$, and the $(N \times |\hat{\Omega}|)$-dimensional vector $\mathbf{D}_\omega = (\mathbf{D}_1(t_n), \ldots, \mathbf{D}_{|\hat{\Omega}|}(t_n))$ represent the vector of all noisy samples of all realizations, that is:

$$\mathbf{D}_\omega = \mathbf{D} + \boldsymbol{\epsilon}, \quad (9)$$

where $\mathbf{D} = (\mathbf{D}(t_n), \ldots, \mathbf{D}(t_n))$ is a $N \times |\hat{\Omega}|$-dimensional vector in which $\mathbf{D}(t_n) = (D(t_1), \ldots, D(t_N))$ is value of the proposed load model evaluated at N sampling points; and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1(t_n), \ldots, \boldsymbol{\epsilon}_{|\hat{\Omega}|}(t_n))$ is a $N \times |\hat{\Omega}|$-dimensional vector in which $\boldsymbol{\epsilon}_\omega(t_n) = (\epsilon_\omega(t_1), \ldots, \epsilon_\omega(t_N))$ is the vector of independent identically distributed Gaussian noises with zero mean and σ^2 variance, i.e., $\epsilon_\omega(t_n) \sim \mathcal{N}(0, \sigma^2)$. Given that

every finite subset of any GP has a Gaussian distribution, $N \times |\hat{\Omega}|$ samples in the realization sample vector $\mathbf{D}_\omega$ form a multivariate Gaussian probability distribution with mean value $\boldsymbol{\mu}_{\mathbf{D}_\omega}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{D}_\omega}$ defined as:

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{D}_\omega} &= \mathbf{E}[\mathbf{D}_\omega] = \mathbf{E}[\mathbf{D}] + \mathbf{E}[\boldsymbol{\epsilon}] = \boldsymbol{\mu}_{\mathbf{D}}, \quad (10)\\ \Sigma_{\mathbf{D}_\omega} &= \mathbf{E}[(\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}_\omega})^T(\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}_\omega})]\\ &= \mathbf{E}[(\mathbf{D} + \boldsymbol{\epsilon} - \boldsymbol{\mu}_{\mathbf{D}})^T(\mathbf{D} + \boldsymbol{\epsilon} - \boldsymbol{\mu}_{\mathbf{D}})]\\ &= \mathbf{E}[(\mathbf{D} - \boldsymbol{\mu}_{\mathbf{D}})^T(\mathbf{D} - \boldsymbol{\mu}_{\mathbf{D}})] + \mathbf{E}[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}] +\\ &\quad \mathbf{E}[\boldsymbol{\epsilon}^T]\mathbf{E}[(\mathbf{D} - \boldsymbol{\mu}_{\mathbf{D}})] + \mathbf{E}[(\mathbf{D} - \boldsymbol{\mu}_{\mathbf{D}})^T]\mathbf{E}[\boldsymbol{\epsilon}]\\ &= \mathbf{K}(\boldsymbol{\theta}) + \sigma^2\mathbf{I}, \quad (11)\end{aligned}$$

where $\boldsymbol{\mu}_{\mathbf{D}}$ is a $N \times |\hat{\Omega}|$-dimensional vector of mean values, and $\mathbf{K}(\boldsymbol{\theta})$ is the covariance function of the load process $D(t)$ evaluated at the sampling points:

$$\boldsymbol{\mu}_{\mathbf{D}} = (\mathbb{D}(t_n), \ldots, \mathbb{D}(t_n)), \quad (12)$$

$$\mathbf{K}(\boldsymbol{\theta}) = [k(t_n, t_{n'})],\ \forall n, n' \in \{1, \ldots, N \times |\hat{\Omega}|\}, \quad (13)$$

where $\boldsymbol{\theta} = \{l_1, \sigma_{d_1}^2, l_2, \sigma_{d_2}^2, \tau\}$ is the vector of hyper-parameters associated with the kernel function (8).

We formulate a maximum likelihood estimation problem to estimate the hyper-parameters of the GP model, i.e., $\boldsymbol{\theta}$, $\boldsymbol{\mu}_{\mathbf{D}}$, σ^2. In order to further reduce computational burden of the problem and without loss of generality, we assume that the mean vector $\boldsymbol{\mu}_{\mathbf{D}}$ is constant over the samples, i.e., $\boldsymbol{\mu}_{\mathbf{D}} = \mu\mathbf{1}$, where μ is the mean value and $\mathbf{1}$ is a $N \times |\hat{\Omega}|$-dimensional vector of ones. The maximum likelihood estimation problem, formulated in (14), maximizes the log marginal likelihood with respect to the hyper-parameters given the load realizations $\mathbf{D}_\omega$:

$$\begin{aligned}\max_{\boldsymbol{\theta}, \boldsymbol{\mu}_{\mathbf{D}}, \sigma^2} \; & -\frac{1}{2}(\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}})\big(\mathbf{K}(\boldsymbol{\theta}) + \sigma^2\mathbf{I}\big)^{-1}(\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}})^T\\ & -\frac{1}{2}\log|\mathbf{K}(\boldsymbol{\theta}) + \sigma^2\mathbf{I}| - \frac{N \times |\hat{\Omega}|}{2}\log(2\pi), \quad (14)\end{aligned}$$

where $\mathbf{I}$ is the $(N \times |\hat{\Omega}|) \times (N \times |\hat{\Omega}|)$ identity matrix. The load realizations $\mathbf{D}_\omega$ only appear in the first term of (14) that optimizes the data fit. The second term, however, is independent of the realizations and represents the complexity penalty. The first and second terms together make a trade-off between fitting the realizations and complexity of solution, obviating the overfitting problem and providing the fittest and simplest GP model. The last term in (14) is merely a scaling factor. The solution of the maximum likelihood estimation problem in (14) would provide the optimal estimates of the hyper-parameters given the load realizations, which could then be used to form the posterior GP load model in terms of the mean vector, and covariance function $k(t, t')$ in (8).

In order to calculate the components of $\boldsymbol{\Sigma}_{\mathbf{B}}$, i.e., the covariance matrix of the Bernstein coefficients in (3), we select an arbitrary set of evaluation times $t_s, s \in \{1, \ldots, S\}$ over $\mathcal{T}$ and evaluate the covariance function $k(t, t')$ in (8) at these times. We then enforce the equality of the covariance function values in (8) with the right-hand-side covariance representation in (7), and form a set of linear equations where the unknowns are the components of $\boldsymbol{\Sigma}_{\mathbf{B}}$:

$$\mathbf{K}^s(\boldsymbol{\theta}^*) = \mathbf{W}^{sT}\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W}^s, \quad (15)$$

where $\boldsymbol{\theta}^*$ is the optimum estimate of $\boldsymbol{\theta}$, $\mathbf{K}^s(\boldsymbol{\theta}^*) = [k(t_s, t_{s'})]$, $\forall s, s' \in \{1, \ldots, S\}$, $\mathbf{W}^s \equiv \mathbf{W}(t_s) = \big(\mathbf{w}^{(Q)}(t_1), \ldots, \mathbf{w}^{(Q)}(t_S)\big)$. In order to have a full-rank set of equations in (15), the choice of the evaluation times τ_s for forming the equations should meet the following two conditions: 1) number of evaluation times S should be equal to Z, 2) the evaluation times need to be distributed over $\mathcal{T}$ such that $Q-1$ samples for the first and last intervals (i.e., $i = 0$ and $i = I-1$), and $Q-2$ samples for the rest of intervals be allocated. Solving the full-rank set of linear equations (15), one can calculate the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{B}}$ as follows:

$$\boldsymbol{\Sigma}_{\mathbf{B}} = \left(\mathbf{W}^{sT}\right)^{-1}\mathbf{K}^s\left(\mathbf{W}^s\right)^{-1}. \quad (16)$$

Note that in the proposed estimation method in this section, there are only seven hyper-parameters to optimize in the maximum likelihood estimation problem (14) and finally calculate $\boldsymbol{\Sigma}_{\mathbf{B}}$ in (16), while in the direct approach we would have all the $Z^2 + 2$ components of $\boldsymbol{\Sigma}_{\mathbf{B}}$ as variables of the maximum likelihood estimation problem. This considerably reduces the computation burden of estimating the proposed load process.

C. Derivation of Predictive Gaussian Process

After solving the maximum likelihood estimation problem in (14) and forming the posterior GP load model using the optimal estimation of hyper-parameters, we can utilize the posterior GP to form the joint distribution of the load realizations $\mathbf{D}_\omega$ and the load values at future test times $t_m^*,\ m \in \{1, \ldots, M\}$, $\mathbf{D}^* = (D(t_1^*), \ldots, D(t_M^*))$ as follows:

$$\begin{bmatrix}\mathbf{D}_\omega^T\\ \mathbf{D}^{*T}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_{\mathbf{D}}^T\\ \boldsymbol{\mu}_{\mathbf{D}}^{*T}\end{bmatrix}, \begin{bmatrix}\mathbf{K}(\boldsymbol{\theta}^*) + \sigma_n^2\mathbf{I} & \mathbf{K}^*(\boldsymbol{\theta}^*)\\ \mathbf{K}^{*T}(\boldsymbol{\theta}^*) & \mathbf{K}^{**}(\boldsymbol{\theta}^*)\end{bmatrix}\right), \quad (17)$$

where $\boldsymbol{\mu}_{\mathbf{D}}^* = \mu\mathbf{1}_M$; $\mathbf{K}^*(\boldsymbol{\theta}^*) = [k(t_n, t_m^*)]$, $\forall n \in \{1, \ldots, N \times |\hat{\Omega}|\}$, $\forall m \in \{1, \ldots, M\}$; and $\mathbf{K}^{**}(\boldsymbol{\theta}^*) = [k(t_m^*, t_{m'}^*)]$, $\forall m, m' \in \{1, \ldots, M\}$. Then we derive the posterior distribution of $\mathbf{D}^*$, i.e., $p(\mathbf{D}^*|t_1, \ldots, t_{N\times|\hat{\Omega}|}, t_1^*, \ldots, t_M^*, \mathbf{D}_\omega, (\boldsymbol{\theta}, \mu, \sigma^2))$, which is a Gaussian with a M-dimensional mean vector $\boldsymbol{\mu}^*$ and a $M \times M$ covariance matrix $\boldsymbol{\Sigma}^*$ defined as follows:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_D^* + (\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}})\left(\mathbf{K}(\boldsymbol{\theta}^*) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{K}^*(\boldsymbol{\theta}^*), \quad (18)$$

$$\boldsymbol{\Sigma}^* = \mathbf{K}^{**}(\boldsymbol{\theta}^*) - \mathbf{K}^{*T}(\boldsymbol{\theta}^*)\left(\mathbf{K}(\boldsymbol{\theta}^*) + \sigma^2\mathbf{I}\right)^{-1}\mathbf{K}^*(\boldsymbol{\theta}^*). \quad (19)$$

The final goal here is to find the Bernstein coefficients of the predictive mean denoted as $\boldsymbol{\mu}_{\mathbf{B}}^*$, and the associated covariance matrix denoted as $\boldsymbol{\Sigma}_{\mathbf{B}}^*$. In order to derive $\boldsymbol{\mu}_{\mathbf{B}}^*$ and $\boldsymbol{\Sigma}_{\mathbf{B}}^*$, we need to substitute in (18) and (19) the values of $\mathbf{K}(\boldsymbol{\theta}^*)$, $\mathbf{K}^*(\boldsymbol{\theta}^*)$, and $\mathbf{K}^{**}(\boldsymbol{\theta}^*)$ in terms of $\boldsymbol{\Sigma}_{\mathbf{B}}$ from (7) as follows:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_{\mathbf{D}}^* + (\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}})\left(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W} + \sigma^2\mathbf{I}\right)^{-1}\left(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W}^*\right), \quad (20)$$

$$\begin{aligned}\boldsymbol{\Sigma}^* = {} & \mathbf{W}^{*T}\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W}^*\\ & -\left(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W}^*\right)^T\left(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W} + \sigma^2\mathbf{I}\right)^{-1}\left(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W}^*\right), \quad (21)\end{aligned}$$

where $\mathbf{W} = \left(\mathbf{w}^{(Q)}(t_1), \ldots, \mathbf{w}^{(Q)}(t_{N\times|\hat{\Omega}|})\right)$ and $\mathbf{W}^* = \left(\mathbf{w}^{(Q)}(t_1^*), \ldots, \mathbf{w}^{(Q)}(t_M^*)\right)$ are respectively $Z \times (N \times |\hat{\Omega}|)$ and $Z \times M$ matrices. Now, factoring $\mathbf{W}^*$, we recast (20) as:

$$\begin{aligned}\boldsymbol{\mu}^* &= \left(\mu\mathbf{1}_Z + (\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}})\left(\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{B}}\mathbf{W} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{W}^T\boldsymbol{\Sigma}_{\mathbf{B}}\right)\mathbf{W}^*\\ &= \boldsymbol{\mu}_{\mathbf{B}}^*\mathbf{W}^*, \quad (22)\end{aligned}$$

and derive the predictive mean of Bernstein coefficients as:

$$\boldsymbol{\mu}^*_{\mathbf{B}} = \mu \mathbf{1}_Z + (\mathbf{D}_\omega - \boldsymbol{\mu}_{\mathbf{D}}) \big(\mathbf{W}^T \boldsymbol{\Sigma}_{\mathbf{B}} \mathbf{W} + \sigma^2 \mathbf{I}\big)^{-1} \mathbf{W}^T \boldsymbol{\Sigma}_{\mathbf{B}}. \quad (23)$$

Similarly, factoring $\mathbf{W}^{*T}$ and $\mathbf{W}^*$, we recast (21) as:

$$\begin{aligned} \boldsymbol{\Sigma}^* &= \mathbf{W}^{*T} \big(\boldsymbol{\Sigma}_{\mathbf{B}} - \boldsymbol{\Sigma}_{\mathbf{B}} \mathbf{W} \big(\mathbf{W}^T \boldsymbol{\Sigma}_{\mathbf{B}} \mathbf{W} + \sigma^2 \mathbf{I}\big)^{-1} \mathbf{W}^T \boldsymbol{\Sigma}_{\mathbf{B}}\big) \mathbf{W}^* \\ &= \mathbf{W}^{*T} \boldsymbol{\Sigma}^*_{\mathbf{B}} \mathbf{W}^*, \end{aligned} \quad (24)$$

and derive the predictive covariance of the Bernstein coefficients as follows:

$$\boldsymbol{\Sigma}^*_{\mathbf{B}} = \boldsymbol{\Sigma}_{\mathbf{B}} - \boldsymbol{\Sigma}_{\mathbf{B}} \mathbf{W} \big(\mathbf{W}^T \boldsymbol{\Sigma}_{\mathbf{B}} \mathbf{W} + \sigma^2 \mathbf{I}\big)^{-1} \mathbf{W}^T \boldsymbol{\Sigma}_{\mathbf{B}}. \quad (25)$$

The mean vector (23) and covariance matrix (25) together form the predictive distribution of the Bernstein coefficients of the proposed GP load model, i.e., $\mathbf{D}^* \sim \mathcal{N}(\boldsymbol{\mu}^*_{\mathbf{B}}, \boldsymbol{\Sigma}^*_{\mathbf{B}})$.

IV. Numerical Results

In this section, we use the hourly electricity load data of CAISO for 10 consecutive Tuesdays, from Dec. 5, 2017 to Feb. 13, 2018 [25] as the load realizations, for estimating the proposed load process projected on the function space spanned by Bernstein polynomials of degree 3 ($Q = 3$). The estimated model is then used to predict the electricity load of Feb. 20, 2018. The GPML toolbox [26] is utilized to implement the kernel function in (8), solve the maximum likelihood problem in (14), and derive the optimal hyper-parameters. The GPML uses the iterative conjugate gradient method to solve the maximum likelihood problem, as a fast and computationally efficient numerical solution method, which alleviates the heavy matrix calculations for a large number of observations.

Choosing appropriate initial values for the hyper-parameters is important for solution of the maximum likelihood estimation. We set the initial value of τ at 24, since the load process approximately experiences daily periods. We also set the initial value of μ to $24,000$, which is the approximate average of the realizations. Further, in order to initialize the rest of the hyper-parameters, we first evaluate the log marginal likelihood in (14) at a limited set of hyper-parameter samples, determine the optimal subset of hyper-parameter samples that maximize the function, and use this subset as the initial guess for the conjugate gradient method. The initial values of the hyper-parameters, and the resulting estimated values using the solution of the maximum likelihood estimation are shown in Table I. As the optimal estimate of σ_{d_1} is almost twice the optimal estimate of σ_{d_2}, the periodic term contributes more in the kernel function formation compared to the exponential term. Also, as expected, the optimal load process period is very close to its initial value of 24. In Fig. 2, we aim to investigate the sensitivity of log marginal likelihood to each of the hyperparameters l_1, l_2, σ_{d_1}, and σ_{d_2} separately, where the rest are set to their initial values. In Fig. 2, the log marginal likelihood is more sensitive to l_1 and σ_{d_1}, compared to l_2 and σ_{d_2}. Besides, among all the hyper-parameters, l_2 has the least impact on the log marginal likelihood.

The optimal estimates of the hyper-parameters are used to form the posterior and predictive GP models. The estimated GP model is then used to forecast the electricity load of the next Tuesday after the training data (Feb. 20, 2018). The

TABLE I
Initial and Optimal Hyper-paramaters

Hyper-parameters	σ_{d_1}	σ_{d_2}	σ	l_1	l_2	τ	μ
Initial Values	1500	200	50	0.5	4.5	24.00	24000.00
Optimal Values	1699.65	790.83	226.96	0.42	3.25	24.03	24000.00

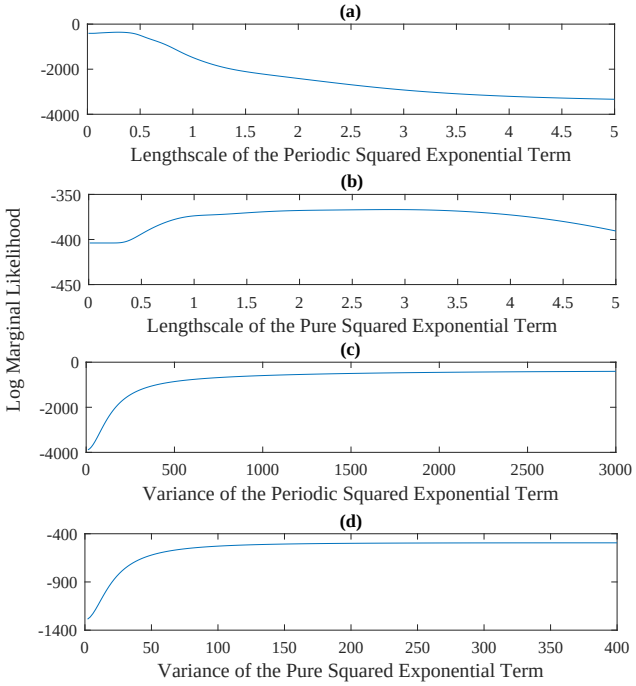

Fig. 2. Log marginal likelihood variations with respect to: (a) length-scale of the periodic term l_1, (b) length-scale of the exponential term l_2, (c) variance of the periodic term σ_{d_1}, (d) variance of the exponential term σ_{d_2}.

hourly training data, the continuous-time load forecast, the actual 5-min load of the forecast day, as well as the $\%99$ uncertainty envelope around the forecasted load trajectory are shown in Fig. 3. The continuous-time load forecast trajectory is drawn using Bernstein coefficients of the predictive distribution derived in (23), which smoothly varies over time, where the associated mean absolute percentage error (MAPE) is equal to $\%3.5$. The uncertainty envelope engulfs a good share of the hourly load realizations and the actual 5-min load of the forecast day is barely out of the uncertainty envelope.

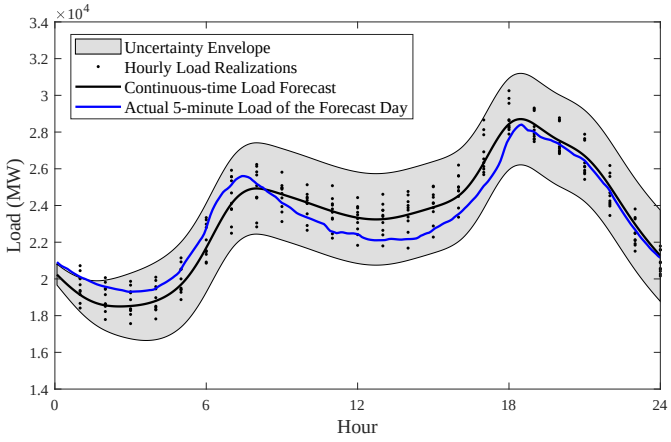

Fig. 3. Load process realizations and the predictive model outputs

The posterior covariance of the Bernstein coefficients are drawn in Fig. 4-(a) and (b), respectively, for one training day and the forecast day. Using Bernstein polynomials of degree 3, the reduced-order Bernstein function space that embeds the C^1 continuity has $(3-1)*24+2=50$ coefficients and the posterior covariance matrix of the Bernstein coefficients derived in (25) is of order 50×50. The magnitude of the covariance for the training day is obviously much lower than that of the forecast day. The periodic term of the kernel function (8) is clearly reflected in Fig. (25)-(a), where the covariance tends to repeat the same shape at the both ends (the previous and next realizations). Also, diagonal components of the covariance matrix monotonically increase with the increasing order of Bernstein coefficients, which reflects the increasing uncertainty of the data over time that is captured by the pure exponential term in the kernel function. In Fig. (25)-(b), since no observations are available in the forecast day, the magnitude of covariance increases considerably, impact of the exponential term of kernel function becomes dominant, and impact of the periodic term is less discernible.

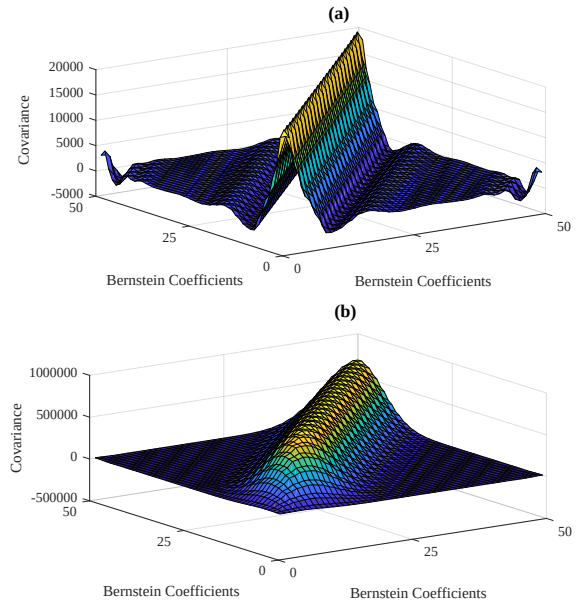

Fig. 4. Posterior covariance of the Bernstein coefficients (a) calculated at a training day (b) calculated at the prediction day

V. Conclusion

In this paper, a continuous-time stochastic process model is proposed to characterize the variability and uncertainty of electricity load. The load process is projected on a reduced-order Bernstein function space with embedded C^1 continuity, a GP prior is imposed on the associated coefficients, and the posterior distribution of the coefficients are derived using past observations of the load process. The kernel trick is applied to reduce the computational burden of estimating the model, where a covariance function is designed to capture the periodicity and smoothness of the load as well as the increasing uncertainty over time. The numerical studies show promising performance of the proposed model to learn the behavior and estimate the future process of the CAISO load. The proposed model uniquely predicts the continuous-time mean value and uncertainty envelopes of future loads, which inherently embeds information on continuous-time variations and the associated ramping requirements of the load.

References

[1] G. Heinemann, D. Nordmian, and E. Plant, "The relationship between summer weather and summer loads-a regression analysis," *IEEE Transactions on Power Apparatus and Systems*, no. 11, pp. 1144–1154, 1966.

[2] A. Muñoz, E. F. Sánchez-Úbeda, A. Cruz, and J. Marín, "Short-term forecasting in power systems: a guided tour," *Handbook of power systems II*, pp. 129–160, 2010.

[3] M. Shah and R. Agrawal, "A review on classical and modern techniques with decision making tools for load forecasting," *Int. Journal of Emerging Trends in Engineering and Development*, no. 3, pp. 174–184, 2013.

[4] G. Gross and F. D. Galiana, "Short-term load forecasting," *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1558–1573, 1987.

[5] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[6] J. W. Taylor and P. E. McSharry, "Short-term load forecasting methods: An evaluation based on european data," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 2213–2219, 2007.

[7] Y. Chakhchoukh, "A new robust estimation method for arma models," *IEEE Trans. Signal Processing*, vol. 58, no. 7, pp. 3512–3522, 2010.

[8] Y. Chakhchoukh, P. Panciatici, and L. Mili, "Electric load forecasting based on statistical robust methods," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 982–991, 2011.

[9] W. Christiaanse, "Short-term load forecasting using general exponential smoothing," *IEEE Trans. Power Systems*, no. 2, pp. 900–911, 1971.

[10] J. W. Taylor, "Short-term load forecasting with exponentially weighted methods," *IEEE Trans. Power Systems*, vol. 27, no. 1, pp. 458–464, 2012.

[11] 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, 2001.

[12] H. Quan, D. Srinivasan, and A. Khosravi, "Short-term load and wind power forecasting using neural network-based prediction intervals," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 2, pp. 303–315, 2014.

[13] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[14] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain, "Gaussian processes for time-series modelling," *Phil. Trans. R. Soc. A*, vol. 371, no. 1984, p. 20110550, 2013.

[15] P. Lauret, M. David, and D. Calogine, "Nonlinear models for short-time load forecasting," *Energy Procedia*, vol. 14, pp. 1404–1409, 2012.

[16] D. Lee and R. Baldick, "Short-term wind power ensemble prediction based on gaussian processes and neural networks," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 501–510, 2014.

[17] N. Chen, Z. Qian, I. T. Nabney, and X. Meng, "Wind power forecasts using gaussian processes and numerical weather prediction," *IEEE Transactions on Power Systems*, vol. 29, no. 2, pp. 656–665, 2014.

[18] J. Yan, K. Li, E. Bai, Z. Yang, and A. Foley, "Time series wind power forecasting based on variant gaussian process and tlbo," *Neurocomputing*, vol. 189, pp. 135–144, 2016.

[19] M. Parvania and A. Scaglione, "Unit commitment with continuous-time generation and ramping trajectory models," *IEEE Trans. Power Systems*, vol. 31, no. 4, pp. 3169–3178, 2016.

[20] ——, "Generation ramping valuation in day-ahead electricity markets," in *Proc. 49th Hawaii International Conference on System Sciences (HICSS)*, 2016, pp. 2335–2344.

[21] M. Parvania and R. Khatami, "Continuous-time marginal pricing of electricity," *IEEE Trans. Power Systems*, vol. 32, no. 3, pp. 1960–1969, 2017.

[22] R. Khatami, M. Parvania, and P. Khargonekar, "Scheduling and pricing of energy generation and storage in power systems," *IEEE Trans. Power Systems*, vol. 33, no. 4, pp. 4308–4322, 2018.

[23] R. Khatami, M. Heidarifar, M. Parvania, and P. Khargonekar, "Scheduling and pricing of load flexibility in power systems," *IEEE Journal of Selected Topics in Signal Processing*, 2018.

[24] P. M. Prenter *et al.*, *Splines and variational methods*. Courier Corporation, 2008.

[25] California ISO Operan Access Same-Time Information System, Feb. 2017. [Online]. Available: http://oasis.caiso.com

[26] C. E. Rasmussen and H. Nickisch, gaussian Process Machine Learning (GPML) Toolbox. [Online]. Available: http://www.jmlr.org/papers/v11/rasmussen10a.html