Variational Inference for Infinite Mixtures of Gaussian Processes With Applications to Traffic Flow Prediction

Shiliang Sun, *Member, IEEE*, and Xin Xu, *Member, IEEE*

***Abstract*—This paper proposes a new variational approximation for infinite mixtures of Gaussian processes. As an extension of the single Gaussian process regression model, mixtures of Gaussian processes can characterize varying covariances or multimodal data and reduce the deficiency of the computationally cubic complexity of the single Gaussian process model. The infinite mixture of Gaussian processes further integrates a Dirichlet process prior to allowing the number of mixture components to automatically be determined from data. We use variational inference and a truncated stick-breaking representation of the Dirichlet process to approximate the posterior of hidden variables involved in the model. To fix the hyperparameters of the model, the variational EM algorithm and a greedy algorithm are employed. In addition to presenting the variational infinite-mixture model, we apply it to the problem of traffic flow prediction. Experiments with comparisons to other approaches show the effectiveness of the proposed model.**

***Index Terms*—Bayesian learning, Dirichlet process, Gaussian process, traffic flow prediction, variational inference.**

I. Introduction

GAUSSIAN processes have proven to be a successful tool for regression problems, e.g., modeling the robot arm inverse dynamics [1]. Formally, a Gaussian process is a collection of random variables, any finite number of which obeys a joint Gaussian prior distribution. For regression, the function to be estimated is assumed to be generated by an infinite-dimensional Gaussian distribution, and the observed outputs are contaminated by additive Gaussian noise.

A common representation for the dependency among outputs used in Gaussian processes is $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma_n^2\mathbf{I})$, where vector $\mathbf{y}$ consists of training outputs $\{y_1, \ldots, y_N\}$, $\mathbf{K}$ is the kernel matrix, and $\mathbf{I}$ is a unit matrix. The entries of $\mathbf{K}$ are given by a kernel function (also called covariance function) $\kappa(\cdot, \cdot)$ between pairs of inputs. Suppose that $\mathbf{x}_i$ and $\mathbf{x}_j$ are two d-dimensional input vectors. The kernel matrix with the squared exponential kernel function can be given as $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp[-\sum_{m=1}^{d}(x_{im} - x_{jm})^2/(2\sigma_m^2)]$. The hyperparameters $\sigma_f, \sigma_1, \ldots, \sigma_d, \sigma_n$ can be estimated from the training data.

However, Gaussian processes suffer from two major limitations. First, using a stationary covariance function, Gaussian processes cannot account for varying covariances or multimodal outputs. Second, parameter inference in Gaussian processes requires the inversion of matrix $\mathbf{K} + \sigma_n^2\mathbf{I}$, which is computationally cubic with respect to the number of the training data. Consequently, mixtures of Gaussian processes [2]–[4] inspired by the mixtures of experts [5], [6] have been proposed to address these problems, some of which used infinite mixtures that place a Dirichlet process prior on the mixture component to allow the number of components to be automatically determined from data. Mixtures of Gaussian processes naturally overcome the first limitation of the single Gaussian process model by using multiple Gaussian processes. In addition, the second limitation is resolved, because now, the inversion of a large matrix is replaced by inversions of multiple smaller matrices [4].

The following two kinds of probability models are considered in these mixtures of Gaussian processes: 1) a conditional model [2], [3] and 2) a full generative model [4], [7]. The first model does not model the distribution of the input, whereas the second model formulates the joint distribution between the input and the output and leads to a powerful consistent manner to designate the responsibility of each component for a certain input. We also adopt the full generative model in this paper.

Inference problems for these models involve fixing the values of hyperparameters and estimating the posterior distribution of hidden variables composed of parameters and latent variables. The complexity of the inference problem for mixtures of Gaussian processes necessitates approximate inference techniques. Although the stochastic Markov chain Monte Carlo sampling methods can be used for approximate inference, they are computationally demanding, and it is difficult to diagnose the convergence of the sampling process [8]. As an alternative, variational inference is a deterministic approximation technique that provides an analytical approximation to the true posterior. This condition is accomplished by exploiting a given factorization form or a specific parameter form such as Gaussian distributions, which makes integrations or summations involved feasible [8], [9]. Because variational inference is faster and more convenient for predicting the outputs of new inputs (e.g., considering the real-time requirement of traffic prediction in

Manuscript received August 19, 2010; revised November 5, 2010; accepted November 13, 2010. Date of publication December 10, 2010; date of current version June 6, 2011. This work is supported in part by the National Natural Science Foundation of China under Project 61075005. The Associate Editor for this paper was L. Li.

S. Sun is with the Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China (e-mail: shiliangsun@gmail.com; slsun@cs.ecnu.edu.cn).

X. Xu is with the Institute of Automation, College of Mechatronics and Automation, National University of Defense Technology, Changsha 410073, China (e-mail: xinxu@nudt.edu.cn; xuxin_mail@263.net).

Digital Object Identifier 10.1109/TITS.2010.2093575

intelligent transportation systems), we adopt variational inference techniques in this paper.

Most previous inference methods for a mixture of Gaussian processes are built on Markov chain Monte Carlo sampling methods. Recently, a variational mixture of Gaussian processes has been proposed [7]. However, it is not an infinite-mixture model, and thus, the number of components needs to be specified *a priori*, which could cause difficult model selection problems. Moreover, it was only validated on 2-D data sets and lacks evaluations on higher dimensional applications, as acknowledged by the authors themselves. In this paper, we propose a new variational inference approach to infinite mixtures of Gaussian processes (IMGP) using the Dirichlet process prior and apply it to the real high-dimensional application problem of traffic flow prediction.

The contribution of this paper is twofold. First, for IMGP, we propose a new variational approximation for estimating hidden variables and hyperparameters. To the best of our knowledge, this approach has not been attempted. Second, the infinite-mixture model and the corresponding variational approximation are, for the first time, applied to the traffic prediction problem and successfully outperform the state-of-the-art Bayesian network (BN) approach [10]. This case would be very interesting to intelligent transportation systems.

The remainder of this paper is organized as follows. Section II introduces the infinite-mixture model of Gaussian processes, including the adopted Gaussian and Dirichlet processes. Section III presents the detailed variational inference techniques for estimating the posterior distribution of hidden variables and the values of hyperparameters. Section IV reports experimental results on applying the proposed model to traffic flow prediction and compares it with some other methods, including the BN approach. Finally, Section V gives concluding remarks and future research directions.

II. Infinite Mixtures of Gaussian Processes

By introducing an additional random variable, the Gaussian process model can reach an equivalent representation that removes the dependency between training outputs. This condition will facilitate the variational approximation treatments. This section gives the formulation of this Gaussian process model and briefly reviews the Dirichlet process model, with an emphasis on its stick-breaking representation. These approaches would be instructive to the presentation of the subsequent variational inference techniques.

We also give the graphical model representation for the adopted IMGP. The local expert and gating network, by the terminology of mixtures of experts [5], can be characterized from the distributions represented by the graphical model.

A. Gaussian Processes

Suppose that the training set $\mathbf{D}$ has N examples $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$. For IMGP, each Gaussian process component is assumed to have a support set of M training examples $(M < N)$. The support sets $\mathbf{I}_k$ for the kth component with $k \in \{1, \ldots, \infty\}$ used to model the corresponding Gaussian process are selected from the original training set. The $M \times M$ kernel matrix $\mathbf{K}_k$, which is confined to the corresponding support set of the kth Gaussian process, is defined by the kernel function as

$$\kappa_k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_{kf}^2 \exp\left(- \sum_{m=1}^{d} \frac{1}{2\sigma_{km}^2} (x_{im} - x_{jm})^2 \right). \quad (1)$$

Suppose that, for input $\mathbf{x}$, the component indicator variable $z = k$. The Gaussian process is then specified by the following univariate Gaussian distribution model:

$$p(y|\mathbf{x}, z = k, \mathbf{w}_k, r_k) = \mathcal{N}(y|\mathbf{w}_k^\top \phi_k\left(\mathbf{x}\right), r_k^{-1}) \quad (2)$$

where weight vector $\mathbf{w}_k$ has a Gaussian distribution $\mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{U}_k^{-1})$, $\phi_k(\mathbf{x})$ is given by the kernel function values between $\mathbf{x}$ and the support set, i.e.,

$$\phi_k(\mathbf{x}) = \left[\kappa_k\left(\mathbf{x}, \mathbf{x}_1^{\mathbf{I}_k}\right), \kappa_k\left(\mathbf{x}, \mathbf{x}_2^{\mathbf{I}_k}\right), \ldots, \kappa_k\left(\mathbf{x}, \mathbf{x}_{|\mathbf{I}_k|}^{\mathbf{I}_k}\right)\right]^\top \quad (3)$$

and r_k^{-1} is the variance. The inverse covariance (also called precision) $\mathbf{U}_k$ is set to $\mathbf{K}_k + \sigma_{kb}^2\mathbf{I}$, where σ_{kb}^2 is used to avoid matrix singularity [7]. Define $\boldsymbol{\theta}_k = \{\sigma_{kf}, \sigma_{k1}, \sigma_{k2}, \ldots, \sigma_{kd}, \sigma_{kb}\}$. $\boldsymbol{\theta}_k$ and $\mathbf{I}_k$ are used to define the parameters of the Gaussian process model and are thereby called hyperparameters, which are omitted in (2). Parameter r_k has a Gamma prior distribution $\Gamma(r_k|a_0, b_0) \propto r_k^{a_0-1}e^{-b_0 r_k}$ with hyperparameters a_0 and b_0.

This linear Gaussian process model has been used, e.g., in [7] and [11], and is an equivalent parametric representation of Gaussian processes, which can readily be shown. For the data set $\mathbf{D}$, we may assume that the output $\mathbf{y} = [y_1, \ldots, y_N]^\top$ is generated by $\mathbf{y} = \mathbf{K}\mathbf{w} + \xi$, where $\mathbf{K}$ is the kernel matrix, $\mathbf{w}$ has a Gaussian distribution $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{K}^{-1})$, and noise variable ξ has a distribution $\mathcal{N}(\xi|\mathbf{0}, \sigma_n^2\mathbf{I})$. The joint distribution of $\mathbf{y}$ is therefore Gaussian $\mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma_n^2\mathbf{I})$. We can recover this model by setting variables involved in (2) as $\mathbf{w}_k = \mathbf{w}$, $\mathbf{K}_k = \mathbf{K}$, $\sigma_{kb} = 0$, and $r_k^{-1} = \sigma_n^2$ and the support set as the whole data set $\mathbf{D}$.

B. Dirichlet Processes for Mixtures of Gaussian Processes

The Dirichlet process [12] is a prior model used for Bayesian data analysis. Each draw from a Dirichlet process is a discrete distribution, with marginal distributions being Dirichlet distributions.

Let Φ be a random variable, H be a distribution over Φ, and α_0 be a positive real scalar. Distribution G is said to be Dirichlet process distributed $G \sim DP(\alpha_0, H)$ if any k partitions $\{A_1, \ldots, A_k\}$ of the corresponding probability space obey a Dirichlet distribution, i.e.,

$$(G(A_1), \ldots, G(A_k)) \sim \mathrm{Dir}\left(\alpha_0 H(A_1), \ldots, \alpha_0 H(A_k)\right) \quad (4)$$

where k is a natural number [13], [14]. The Dirichlet process can be adopted to extend a usual finite mixture model to a mixture with a countably infinite number of components. This approach would be clear by considering the following stick-breaking construction of the Dirichlet process [15].

Consider two infinite collections of independently drawn random variables $\Phi_i \sim H$ and $\nu_i \sim \mathrm{Beta}(1, \alpha_0)$ for

$i = \{1, \ldots, \infty\}$, where the beta distribution has the form $\mathrm{Beta}(\nu|a, b) \propto \nu^{a-1}(1-\nu)^{b-1}$. By introducing a proportion variable $\pi_i = \nu_i \prod_{j=1}^{i-1}(1-\nu_j)$, we can reach the stick-breaking representation of G as

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\Phi_i} \tag{5}$$

where δ_{Φ_i} is a delta function whose value is 1 at location Φ_i and 0 elsewhere. The mixing proportions $\{\pi_i\}$ always sum to one and can imaginarily be given by breaking a unit length stick into a countably infinite number of pieces. The product $\prod_{j=1}^{i-1}(1-\nu_j)$ denotes the previous remaining length of stick, and multiplication by ν_i gives the length of the stick currently broken off.

The N observations $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ can be modeled by the associated latent parameters $\{\Phi_n\}_{n=1}^N$, which characterize the generation of these observations. Because distribution G is discrete, $\{\Phi_n\}_{n=1}^N$ should take no more than N different values. Consequently, the N examples can be partitioned into different groups, and the whole model serves as a mixture model. It is validated that the number of clusters only logarithmically grows in N [13].

Let z_n be a latent variable that assigns the index of the parameter associated with example $(\mathbf{x}_n, y_n)$. The distribution of z_n can be regarded as a multinomial distribution with parameters $\{\pi_1, \ldots, \pi_\infty\}$. With the values of $\{z_1, \ldots, z_N\}$ and the model assumption on the observations (independent or dependent), we can formulate the joint distribution of the training set with the associated parameters. Because the Gaussian process model adopted in this paper breaks the dependency among outputs, given all the parameters, the data would independently be drawn (the model for the inputs will be introduced in the next section).

C. Graphical Model Representation

The graphical model for our IMGP is shown in Fig. 1. The distribution over the input space for a mixture component is given by a Gaussian distribution with a full covariance, i.e.,

$$p(\mathbf{x}|z = k, \boldsymbol{\mu}_k, \mathbf{R}_k) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \mathbf{R}_k^{-1}\right) \tag{6}$$

where $\mathbf{R}_k$ is the inverse covariance. This input model is identical to the model used in [4] and is often flexible enough to provide a good performance, although we can consider to adopt mixtures of Gaussian distributions (e.g., [16]) to further finely model the input space. Parameters $\boldsymbol{\mu}_k$ and $\mathbf{R}_k$ are further specified by a Gaussian distribution prior and a Wishart distribution prior with additional hyperparameters, respectively, i.e.,

$$\boldsymbol{\mu}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \mathbf{R}_0^{-1}\right), \quad \mathbf{R}_k \sim \mathcal{W}(\mathbf{W}_0, \nu_0). \tag{7}$$

To relate our mixture model with the mixture of experts model, we calculate the responsibility of the mixture components for a new input $\mathbf{x}$ as

$$p(z|\mathbf{x}) = \frac{p(z)p(\mathbf{x}|z)}{\sum_z p(z)p(\mathbf{x}|z)} \tag{8}$$

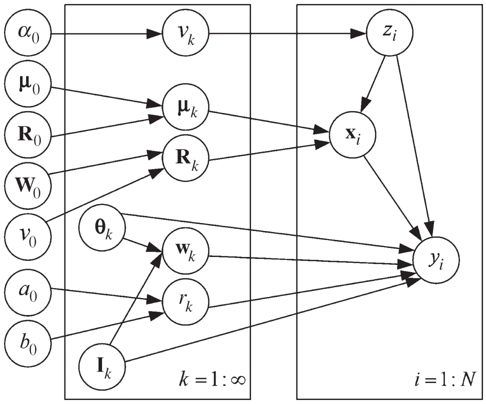

Fig. 1. Graphical model representation of IMGP.

where all the parameters and hyperparameters involved are already determined by optimizing certain objectives on the available data $\mathbf{D}$. The distribution of the corresponding output y can be formulated by the component responsibility and the Gaussian process model given in (2). For regression problems such as the traffic prediction considered in this paper, the prediction for y would be the weighted average of the Gaussian means, where the weights are the responsibilities, and the Gaussian means are provided by (2) evaluated at the new input $\mathbf{x}$. Therefore, by the terminology of mixtures of experts, (8) takes the role of the gating network, whereas (2) serves as the local Gaussian process expert.

III. Variational Inference for Infinite Mixtures of Gaussian Processes

Define parameter sets $\bar{\boldsymbol{\nu}} = \{\nu_1, \ldots, \nu_\infty\}$, $\bar{\boldsymbol{\mu}} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_\infty\}$, $\bar{\mathbf{R}} = \{\mathbf{R}_1, \ldots, \mathbf{R}_\infty\}$, $\bar{\mathbf{w}} = \{\mathbf{w}_1, \ldots, \mathbf{w}_\infty\}$, and $\bar{r} = \{r_1, \ldots, r_\infty\}$ and the latent variable set $\bar{z} = \{z_1, z_2, \ldots, z_N\}$. Thus, the hidden variable set is $\Omega = \{\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}, \bar{\mathbf{R}}, \bar{\mathbf{w}}, \bar{r}, \bar{z}\}$. The variables in the leftmost column in Fig. 1, $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\infty\}$ and $\{\mathbf{I}_1, \ldots, \mathbf{I}_\infty\}$ constitute the hyperparameter set. The role of inference would be to determine the values of the hyperparameters and the posterior distribution of the hidden variables.

By the graphical model given in Fig. 1, the joint distribution of all the random variables (the hyperparameters are omitted) is given by

$$\begin{aligned} p(\mathbf{D}, \Omega) &= p(\bar{\boldsymbol{\nu}})p(\bar{\boldsymbol{\mu}})p(\bar{\mathbf{R}})p(\bar{\mathbf{w}})p(\bar{r}) \\ &\quad \times \prod_{i=1}^{N} p(z_i|\bar{\boldsymbol{\nu}})p(\mathbf{x}_i|z_i, \bar{\boldsymbol{\mu}}, \bar{\mathbf{R}})p(y_i|\mathbf{x}_i, z_i, \bar{\mathbf{w}}, \bar{r}) \\ &= \prod_{k=1}^{\infty} p(\nu_k)p(\boldsymbol{\mu}_k)p(\mathbf{R}_k)p(\mathbf{w}_k)p(r_k) \\ &\quad \times \prod_{i=1}^{N} p(z_i|\bar{\boldsymbol{\nu}})p(\mathbf{x}_i|z_i, \bar{\boldsymbol{\mu}}, \bar{\mathbf{R}})p(y_i|\mathbf{x}_i, z_i, \bar{\mathbf{w}}, \bar{r}) \end{aligned} \tag{9}$$

where we have used the equalities $p(\bar{\boldsymbol{\nu}}) = \prod_k p(\nu_k)$, $p(\bar{\boldsymbol{\mu}}) = \prod_k p(\boldsymbol{\mu}_k)$, $p(\bar{\mathbf{R}}) = \prod_k p(\mathbf{R}_k)$, $p(\bar{\mathbf{w}}) = \prod_k p(\mathbf{w}_k)$, and $p(\bar{r}) = \prod_k p(r_k)$. Although we can formulate the posterior $p(\Omega|\mathbf{D})$ by the Bayes theorem [17] $p(\Omega|\mathbf{D}) = p(\mathbf{D}, \Omega)/p(\mathbf{D})$, the evaluation of $p(\mathbf{D})$ from $p(\mathbf{D},\Omega)$, which involves integration and summation over the hidden variables, is infeasible even for a very small number of mixture components due to the coupling among these variables. Consequently, this paper uses variational inference to approximate the posterior distribution.

Variational methods have their root in the early work of calculus of variations, which concerns functional derivatives about how the value of a functional changes with respect to infinitesimal changes of the input function [8], [18]. We can optimize a certain functional to explore desirable input functions that correspond to the optimal value of the functional. By restricting the range of input functions to be explored, variational methods naturally lead to approximations, although they are not intrinsically approximate [8]. For example, possible restrictions include specific parametric forms such as a Gaussian or particular factorization assumptions.

Suppose that $q(\Omega)$ is an approximation for the true posterior $p(\Omega|\mathbf{D})$. A useful decomposition for variational inference is

$$\ln p(\mathbf{D}) = \mathcal{L}(q) + \mathrm{KL}(q\|p) \tag{10}$$

where $\mathcal{L}(q) = \int q(\Omega)\ln\{p(\mathbf{D},\Omega)/q(\Omega)\}d\Omega$, and $\mathrm{KL}(q\|p) = \int q(\Omega)\ln\{q(\Omega)/p(\Omega|\mathbf{D})\}d\Omega$. The Kullback–Leibler divergence $\mathrm{KL}(q\|p)$ is nonnegative and is zero if and only if $q(\Omega) = p(\Omega|\mathbf{D})$ [19]. $\mathcal{L}(q)$ is the lower bound of $\ln p(\mathbf{D})$. Maximizing this lower bound is equivalent to minimizing the Kullback–Leibler divergence. However, minimizing $\mathrm{KL}(q\|p)$ to solve $q(\Omega)$ is infeasible, because $p(\Omega|\mathbf{D})$ is unknown. Therefore, optimizing the lower bound is widely used in variational inference to reach a good approximation distribution.

A. Truncated Representation of the Dirichlet Process

To formulate the variational posterior $q(\Omega)$, we approximate the posterior Dirichlet process by a truncated stick-breaking representation, as shown in [14] and [20]. We fix a value T and let $q(\nu_T = 1) = 1$, which implies that the mixture proportions π_i are zero for $i > T$, and use the following factorized variational distribution to approximate $p(\Omega|\mathbf{D})$:

$$q(\Omega) = \prod_{t=1}^{T-1} q(\nu_t)\prod_{k=1}^{T} q(\boldsymbol{\mu}_k)q(\mathbf{R}_k)q(\mathbf{w}_k)q(r_k)\prod_{n=1}^{N} q(z_n). \tag{11}$$

Note that the hidden variables in Ω do not share the same variational parameters, e.g., distributions $q(\nu_1)$ and $q(\nu_2)$ usually have different parameters, and there have been no restrictions placed on the functional forms of the individual factor distributions [8]. Variational inference in terms of this factorized form is also called mean field variational inference [21].

The truncation level T is not a part of our prior infinite-mixture model. It is a variational parameter for pursuing an approximation to the true posterior. Although it can freely be set or selected by maximizing $\mathcal{L}(q)$ without fear of overfitting, in this paper, we just fix a single value as done in [14].

With this full factorization formulation, we can solve for the variational distribution by maximizing the lower bound $\mathcal{L}(q)$ in (10). The solution is quite simple. That is, to compute the variational distribution for a hidden variable $\boldsymbol{\omega} \in \Omega$, we need to compute the posterior mean of $\ln p(\mathbf{D},\Omega)$ over the variational distributions of all the other latent variables as

$$\ln q(\boldsymbol{\omega}) = \mathbb{E}_{\Omega\backslash\boldsymbol{\omega}}\left[\ln p(\mathbf{D},\Omega)\right] + \mathrm{const} \tag{12}$$

where "const" denotes a constant that is independent of $\boldsymbol{\omega}$ and is used to normalize the corresponding distribution [8], [14].

B. Variational Distribution

This section details how we can make use of (12) to calculate the variational factors. Note that the variational inference is essentially iterative, because it represents a distribution factor using knowledge about other factors.

1) $q(\nu_t)$: Any terms that are independent of $\boldsymbol{\nu}_t$ will be absorbed into the additive constant. Thus, we have

$$\ln q(\nu_t) = \ln p(\nu_t) + \sum_{n=1}^{N}\mathbb{E}_{\Omega\backslash\nu_t}\left[\ln p(z_n|\bar{\boldsymbol{\nu}})\right] + \mathrm{const}. \tag{13}$$

To solve the term $\mathbb{E}_{\Omega\backslash\nu_t}[\ln p(z_n|\bar{\boldsymbol{\nu}})]$, we employ the following expression [14]:

$$\begin{aligned}&\mathbb{E}_q[\ln p(z_n|\bar{\boldsymbol{\nu}})]\\ &\quad=\mathbb{E}_q\left[\ln\left(\prod_{i=1}^{\infty}(1-\nu_i)^{1[z_n>i]}\nu_i^{1[z_n=i]}\right)\right]\\ &\quad=\sum_{i=1}^{\infty}\{q(z_n>i)\mathbb{E}_q[\ln(1-\nu_i)]+q(z_n=i)\mathbb{E}_q[\ln\nu_i]\}.\end{aligned} \tag{14}$$

Notice that $q(z_n > T) = 0$. Therefore, we can truncate the aforementioned summation at $i = T$ [14]. This approach yields

$$\begin{aligned}&\mathbb{E}_q\left[\ln p(z_n|\bar{\boldsymbol{\nu}})\right]\\ &\quad=\sum_{i=1}^{T}\{q(z_n > i)\mathbb{E}_q\left[\ln(1-\nu_i)\right] + q(z_n = i)\mathbb{E}_q[\ln\nu_i]\}.\end{aligned}$$

Consequently, for the variational distribution $q(\nu_t)$, we have

$$\begin{aligned}&\ln q(\nu_t) + \mathrm{const}\\ &\quad= \ln p(\nu_t) + \sum_{n=1}^{N}\left[q(z_n > t)\ln(1-\nu_t) + q(z_n = t)\ln\nu_t\right]\\ &\quad= \ln p(\nu_t) + \left[\sum_{n=1}^{N} q(z_n > t)\right]\ln(1-\nu_t)\\ &\qquad+ \left[\sum_{n=1}^{N} q(z_n = t)\right]\ln\nu_t.\end{aligned} \tag{15}$$

Because $\nu_i \sim \mathrm{Beta}(1,\alpha_0)$, we have $p(\nu_t) \propto (1-\nu_t)^{\alpha_0-1}$. Define $\nu_{t1} = \sum_{n=1}^{N} q(z_n > t)$ and $\nu_{t2} = \sum_{n=1}^{N} q(z_n = t)$. According to (15), we get $q(\nu_t) \propto \nu_t^{\nu_{t2}}(1-\nu_t)^{(\nu_{t1}+\alpha_0)-1}$. That is, the variational distribution $q(\nu_t)$ is also a beta distribution, with $\nu_t \sim \mathrm{Beta}(\nu_{t2}+1, \nu_{t1}+\alpha_0)$.

2) $q(\boldsymbol{\mu}_k)$: Likewise, any terms that are independent of $\boldsymbol{\mu}_k$ will be absorbed into the additive constant as

$$\ln q(\boldsymbol{\mu}_k)=\ln p(\boldsymbol{\mu}_k)+\sum_{n=1}^{N}\mathbb{E}_{\Omega\backslash\boldsymbol{\mu}_k}\left[\ln p(\mathbf{x}_n|z_n,\bar{\boldsymbol{\mu}},\bar{\mathbf{R}})\right]+\text{const.} \tag{16}$$

For the term $\mathbb{E}_{\Omega\backslash\boldsymbol{\mu}_k}[\ln p(\mathbf{x}_n|z_n,\bar{\boldsymbol{\mu}},\bar{\mathbf{R}})]$, we employ the following expression:

$$\begin{aligned}\mathbb{E}_q\left[\ln p(\mathbf{x}_n|z_n,\bar{\boldsymbol{\mu}},\bar{\mathbf{R}})\right]&=\mathbb{E}_q\left[\ln\left(\prod_{i=1}^{\infty}p(\mathbf{x}_n|\boldsymbol{\mu}_i,\mathbf{R}_i)^{1[z_n=i]}\right)\right]\\&=\sum_{i=1}^{\infty}\{q(z_n=i)\mathbb{E}_q[\ln p(\mathbf{x}_n|\boldsymbol{\mu}_i,\mathbf{R}_i)]\}\\&=\sum_{i=1}^{T}\{q(z_n=i)\mathbb{E}_q[\ln p(\mathbf{x}_n|\boldsymbol{\mu}_i,\mathbf{R}_i)]\}.\end{aligned} \tag{17}$$

Consequently, for the variational distribution $q(\boldsymbol{\mu}_k)$, we have

$$\begin{aligned}&\ln q(\boldsymbol{\mu}_k)+\text{const}\\&=\ln p(\boldsymbol{\mu}_k)+\sum_{n=1}^{N}\{q(z_n=k)\mathbb{E}_{\mathbf{R}_k}[\ln p(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{R}_k)]\}\\&=\ln p(\boldsymbol{\mu}_k)-\frac{1}{2}\sum_{n=1}^{N}\left\{q(z_n=k)\left(\mathbf{x}_n-\boldsymbol{\mu}_k\right)^{\top}\mathbb{E}\mathbf{R}_k(\mathbf{x}_n-\boldsymbol{\mu}_k)\right\}\\&=-\frac{1}{2}(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)^{\top}\mathbf{R}_0(\boldsymbol{\mu}_k-\boldsymbol{\mu}_0)\\&\quad-\frac{1}{2}\sum_{n=1}^{N}\left\{q(z_n=k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\top}(\mathbb{E}\mathbf{R}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)\right\}.\end{aligned} \tag{18}$$

Define $\mathbf{R}_{k1}=\mathbb{E}\mathbf{R}_k$, $\mathbf{R}_{k2}=\sum_{n=1}^{N}q(z_n=k)\mathbf{R}_{k1}\mathbf{x}_n$, and $\mathbf{R}_{k3}=\sum_{n=1}^{N}q(z_n=k)\mathbf{R}_{k1}$. Plugging them into the aforementioned equation results in

$$\begin{aligned}-2\ln q(\boldsymbol{\mu}_k)+\text{const}&=\boldsymbol{\mu}_k^{\top}\mathbf{R}_0\boldsymbol{\mu}_k-2\boldsymbol{\mu}_k^{\top}\mathbf{R}_0\boldsymbol{\mu}_0\\&\quad+\boldsymbol{\mu}_k^{\top}\mathbf{R}_{k3}\boldsymbol{\mu}_k-2\boldsymbol{\mu}_k^{\top}\mathbf{R}_{k2}.\end{aligned} \tag{19}$$

Therefore, the variational distribution $q(\boldsymbol{\mu}_k)$ is a Gaussian distribution, with

$$\boldsymbol{\mu}_k\sim\mathcal{N}\left((\mathbf{R}_0+\mathbf{R}_{k3})^{-1}(\mathbf{R}_0\boldsymbol{\mu}_0+\mathbf{R}_{k2}),(\mathbf{R}_0+\mathbf{R}_{k3})^{-1}\right).$$

3) $q(\mathbf{R}_k)$: The central equation is

$$\ln q(\mathbf{R}_k)=\ln p(\mathbf{R}_k)+\sum_{n=1}^{N}\mathbb{E}_{\Omega\backslash\mathbf{R}_k}\left[\ln p(\mathbf{x}_n|z_n,\bar{\boldsymbol{\mu}},\bar{\mathbf{R}})\right]+\text{const.} \tag{20}$$

By (17), we have

$$\begin{aligned}&\ln q(\mathbf{R}_k)+\text{const}\\&=\ln p(\mathbf{R}_k)+\sum_{n=1}^{N}\left\{q(z_n=k)\mathbb{E}_{\boldsymbol{\mu}_k}[\ln p(\mathbf{x}_n|\boldsymbol{\mu}_k,\mathbf{R}_k)]\right\}\\&=\left[\frac{\nu_0-d-1}{2}\ln|\mathbf{R}_k|-\frac{1}{2}\mathrm{tr}\left(\mathbf{R}_k\mathbf{W}_0^{-1}\right)\right]\\&\quad+\frac{1}{2}\sum_{n=1}^{N}\{q(z_n=k)[\ln|\mathbf{R}_k|\\&\qquad-\mathrm{tr}\left[\mathbf{R}_k\mathbb{E}\left((\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\top}\right)\right]]\}.\end{aligned} \tag{21}$$

Define $\boldsymbol{\mu}_{k1}=\sum_{n=1}^{N}q(z_n=k)$ and $\boldsymbol{\mu}_{k2}=\sum_{n=1}^{N}q(z_n=k)\mathbb{E}_{\boldsymbol{\mu}_k}[(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\top}]$. The variational distribution $q(\mathbf{R}_k)$ is thus a Wishart distribution, with

$$\mathbf{R}_k\sim\mathcal{W}(\mathbf{W}^k,\nu^k) \tag{22}$$

where $(\mathbf{W}^k)^{-1}=(\mathbf{W}_0)^{-1}+\boldsymbol{\mu}_{k2}$, and $\nu^k=\nu_0+\boldsymbol{\mu}_{k1}$.

4) $q(\mathbf{w}_k)$: Absorbing any terms independent of $\mathbf{w}_k$ into the additive constant results in

$$\begin{aligned}\ln q(\mathbf{w}_k)&=\ln p(\mathbf{w}_k)\\&\quad+\sum_{n=1}^{N}\mathbb{E}_{\Omega\backslash\mathbf{w}_k}\left[\ln p(y_n|\mathbf{x}_n,z_n,\bar{\mathbf{w}},\bar{r})\right]+\text{const.}\end{aligned} \tag{23}$$

We have

$$\begin{aligned}&\mathbb{E}_q\left[\ln p(y_n|\mathbf{x}_n,z_n,\bar{\mathbf{w}},\bar{r})\right]\\&\quad=\mathbb{E}_q\left[\ln\left(\prod_{i=1}^{\infty}p(y_n|\mathbf{x}_n,\mathbf{w}_i,r_i)^{1[z_n=i]}\right)\right]\\&\quad=\sum_{i=1}^{\infty}\left\{q(z_n=i)\mathbb{E}_q\left[\ln p(y_n|\mathbf{x}_n,\mathbf{w}_i,r_i)\right]\right\}\\&\quad=\sum_{i=1}^{T}\left\{q(z_n=i)\mathbb{E}_q\left[\ln p(y_n|\mathbf{x}_n,\mathbf{w}_i,r_i)\right]\right\}.\end{aligned} \tag{24}$$

Consequently, the variational distribution $q(\mathbf{w}_k)$ can be reformulated as

$$\begin{aligned}&\ln q(\mathbf{w}_k)+\text{const}\\&\quad=\ln p(\mathbf{w}_k)+\sum_{n=1}^{N}\left\{q(z_n=k)\mathbb{E}_{r_k}\left[\ln p(y_n|\mathbf{x}_n,\mathbf{w}_k,r_k)\right]\right\}\\&\quad=-\frac{1}{2}\mathbf{w}_k^{\top}\mathbf{U}_k\mathbf{w}_k\\&\qquad-\frac{1}{2}\sum_{n=1}^{N}\left\{q(z_n=k)\left(y_n-\mathbf{w}_k^{\top}\phi_k(\mathbf{x}_n)\right)^2\mathbb{E}r_k\right\}.\end{aligned} \tag{25}$$

Define $r_{k1}=\mathbb{E}r_k$, $r_{k2}=\sum_{n=1}^{N}q(z_n=k)r_{k1}y_n\phi_k(\mathbf{x}_n)$, and $r_{k3}=\sum_{n=1}^{N}q(z_n=k)r_{k1}\phi_k(\mathbf{x}_n)\phi_k(\mathbf{x}_n)^{\top}$. Plugging them into the aforementioned equation, we get

$$-2\ln q(\mathbf{w}_k)+\text{const}=\mathbf{w}_k^{\top}\mathbf{U}_k\mathbf{w}_k-2\mathbf{w}_k^{\top}r_{k2}+\mathbf{w}_k^{\top}r_{k3}\mathbf{w}_k. \tag{26}$$

Therefore, the variational distribution $q(\mathbf{w}_k)$ is a Gaussian distribution, with

$$\mathbf{w}_k\sim\mathcal{N}\left((\mathbf{U}_k+r_{k3})^{-1}r_{k2},(\mathbf{U}_k+r_{k3})^{-1}\right).$$

5) $q(r_k)$: The central equation for solving $q(r_k)$ is

$$\ln q(r_k)=\ln p(r_k)+\sum_{n=1}^{N}\mathbb{E}_{\Omega\backslash r_k}[\ln p(y_n|\mathbf{x}_n,z_n,\bar{\mathbf{w}},\bar{\mathbf{r}})]+\text{const}. \tag{27}$$

By (24), we rewrite the aforementioned equation as

$$\begin{aligned}
&\ln q(r_k)+\text{const}\\
&\quad=\ln p(r_k)+\sum_{n=1}^{N}\{q(z_n=k)\mathbb{E}_{\mathbf{w}_k}[\ln p(y_n|\mathbf{x}_n,\mathbf{w}_k,r_k)]\}\\
&\quad=[(a_0-1)\ln r_k-b_0 r_k]\\
&\qquad+\frac{1}{2}\sum_{n=1}^{N}\Big\{q(z_n=k)\Big[\ln r_k-r_k\mathbb{E}\Big[\big(y_n-\mathbf{w}_k^\top\phi_k(\mathbf{x}_n)\big)^2\Big]\Big]\Big\}.
\end{aligned}$$

Define $\mathbf{w}_{k1}=(1/2)\sum_{n=1}^{N}q(z_n=k)$ and $\mathbf{w}_{k2}=(1/2)\sum_{n=1}^{N}q(z_n=k)\mathbb{E}_{\mathbf{w}_k}[(y_n-\mathbf{w}_k^\top\phi_k(\mathbf{x}_n))^2]$. The variational distribution $q(r_k)$ is a Gamma distribution, with

$$r_k\sim\Gamma(a_0+\mathbf{w}_{k1},b_0+\mathbf{w}_{k2}). \tag{28}$$

6) $q(z_n)$: The equation for solving $q(z_n)$ is slightly more complex, because z_n is involved in multiple variational factors, i.e.,

$$\begin{aligned}
\ln q(z_n)+\text{const}=\mathbb{E}_{\Omega\backslash z_n}\big[&\ln p(z_n|\bar{\boldsymbol{\nu}})+\ln p(\mathbf{x}_n|z_n,\bar{\boldsymbol{\mu}},\bar{\mathbf{R}})\\
&+\ln p(y_n|\mathbf{x}_n,z_n,\bar{\mathbf{w}},\bar{r})\big].
\end{aligned} \tag{29}$$

The right-hand side of (29) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{\infty}\{1[z_n>i]\mathbb{E}\left[\ln(1-\nu_i)\right]+1[z_n=i]\mathbb{E}[\ln\nu_i]\\
&\qquad+1[z_n=i]\mathbb{E}\left[\ln p(\mathbf{x}_n|\boldsymbol{\mu}_i,\mathbf{R}_i)\right]\\
&\qquad+1[z_n=i]\mathbb{E}\left[\ln p(y_n|\mathbf{x}_n,\mathbf{w}_i,r_i)\right]\}\\
&=\sum_{i=1}^{T}\{1[z_n>i]\mathbb{E}\left[\ln(1-\nu_i)\right]+1[z_n=i]\mathbb{E}[\ln\nu_i]\\
&\qquad+1[z_n=i]\mathbb{E}\left[\ln p(\mathbf{x}_n|\boldsymbol{\mu}_i,\mathbf{R}_i)\right]\\
&\qquad+1[z_n=i]\mathbb{E}\left[\ln p(y_n|\mathbf{x}_n,\mathbf{w}_i,r_i)\right]\}\\
&=\sum_{i=1}^{T}\Big\{1[z_n>i]\mathbb{E}\left[\ln(1-\nu_i)\right]+1[z_n=i]\mathbb{E}[\ln\nu_i]\\
&\qquad+1[z_n=i]\frac{1}{2}\left[\mathbb{E}\ln|\mathbf{R}_i|-d\ln(2\pi)\right.\\
&\qquad\qquad\left.-\mathbb{E}\left((\mathbf{x}_n-\boldsymbol{\mu}_i)^\top\mathbf{R}_i(\mathbf{x}_n-\boldsymbol{\mu}_i)\right)\right]\\
&\qquad+1[z_n=i]\frac{1}{2}\Big[\mathbb{E}\ln r_i-\ln(2\pi)\\
&\qquad\qquad-\mathbb{E}\left(r_i\left(y_n-\mathbf{w}_i^\top\phi_i(\mathbf{x}_n)\right)^2\right)\Big]\Big\}.
\end{aligned} \tag{30}$$

Define $\ln\rho_{nt}=\mathbb{E}[\ln\nu_t]+\sum_{i=1}^{t-1}\mathbb{E}[\ln(1-\nu_i)]+(1/2)[\mathbb{E}\ln|\mathbf{R}_t|+\mathbb{E}\ln r_t-(d+1)\ln(2\pi)-\mathbb{E}((\mathbf{x}_n-\boldsymbol{\mu}_t)^\top\mathbf{R}_t(\mathbf{x}_n-\boldsymbol{\mu}_t))-\mathbb{E}(r_t(y_n-\mathbf{w}_t^\top\phi_t(\mathbf{x}_n))^2)]$ and $\tilde{\rho}_{nt}=\rho_{nt}/\sum_{i=1}^{T}\rho_{ni}$. Then, we have $q(z_n=t)=\tilde{\rho}_{nt}$, which means that z_n is chosen according to a multinomial probability distribution.

C. Inferring the Hyperparameters

Some hyperparameters are quite generic without the need for further estimation and are thus fixed. $\boldsymbol{\mu}_0$ and $\mathbf{R}_0$ are set to the mean $\boldsymbol{\mu}_x$ and inverse covariance $\mathbf{R}_x$ of the training data, respectively. Parameter ν_0, which is the number of degrees of freedom under a Wishart distribution, is set to the dimensionality d of the inputs. $\mathbf{W}_0$ is set to $\mathbf{R}_x/d$ such that the mean of $\mathbf{R}_k$ under the Wishart distribution is $\mathbf{R}_x$. Parameters a_0 and b_0 in the gamma distribution $\Gamma(r_k|a_0,b_0)$ are set with $a_0=10^{-2}$ and $b_0=10^{-4}$, respectively, to give broad priors following [7] and [22]. The concentration parameter α_0 in defining $\nu_k\sim\text{Beta}(1,\alpha_0)$ is set to 1, as done in [14]. This parameter can also be predetermined by calculating the expected number of mixture components according to [23]. For example, if we have 2000 training examples and $\alpha_0=1$, the expected number will be around 8.

The only hyperparameters to be estimated are $\Theta=\{\boldsymbol{\theta}_{1:T},\mathbf{I}_{1:T}\}$ under the truncated stick-breaking representation. The variational expectation–maximization (EM) algorithm [8] and a greedy algorithm are, respectively, used to estimate the covariance-function-related hyperparameters $\boldsymbol{\theta}_{1:T}$ and support sets $\mathbf{I}_{1:T}$ following [7].

1) Fixing $\boldsymbol{\theta}_{1:T}$: The variational EM algorithm iteratively maximizes $\mathbb{E}_{\Omega}\ln p(\mathbf{D},\Omega|\boldsymbol{\theta}_{1:T})$ with the E- and M-steps to update hyperparameters, where the expectation is taken over the variational distribution. For the standard EM algorithm [8], [10], the expectation would be calculated over a true posterior distribution.

The factors in $p(\mathbf{D},\Omega)$ related to these hyperparameters should be retained for maximization, whereas other factors can be regarded as being absorbed into a constant. Therefore, we can maximize the following expression:

$$\begin{aligned}
&\mathbb{E}\left\{\sum_{k=1}^{T}\ln p(\mathbf{w}_k)+\sum_{n=1}^{N}\ln(y_n|\mathbf{x}_n,z_n,\bar{\mathbf{w}},\bar{r})\right\}\\
&\quad=\frac{1}{2}\sum_{k=1}^{T}\left[\ln|\mathbf{U}_k|-\text{tr}\left(\mathbf{U}_k\mathbb{E}\left(\mathbf{w}_k\mathbf{w}_k^\top\right)\right)\right]\\
&\qquad+\sum_{n=1}^{N}\sum_{k=1}^{T}\left[q(z_n=k)\mathbb{E}\ln p(y_n|\mathbf{x}_n,\mathbf{w}_k,r_k)\right]\\
&\quad=\frac{1}{2}\sum_{k=1}^{T}\Bigg[\ln|\mathbf{U}_k|-\text{tr}\left(\mathbf{U}_k\mathbb{E}\left(\mathbf{w}_k\mathbf{w}_k^\top\right)\right)\\
&\qquad\qquad+2\sum_{n=1}^{N}q(z_n=k)\mathbb{E}\ln p(y_n|\mathbf{x}_n,\mathbf{w}_k,r_k)\Bigg].
\end{aligned} \tag{31}$$

The term $\mathbb{E}(\mathbf{w}_k\mathbf{w}_k^\top)$ can easily be computed by exploiting the formulation of $q(\mathbf{w}_k)$. The computation of $\mathbb{E}\ln p(y_n|\mathbf{x}_n, \mathbf{w}_k, r_k)$ uses the following result:

$$\begin{aligned}&\mathbb{E}\ln p(y_n|\mathbf{x}_n, \mathbf{w}_k, r_k) + \text{const}\\ &\quad= \frac{1}{2}\left\{\mathbb{E}\ln r_k - (\mathbb{E}r_k)\mathbb{E}\left[\left(y_n - \mathbf{w}_k^\top\phi_k(\mathbf{x}_n)\right)^2\right]\right\} \qquad (32)\end{aligned}$$

where $\mathbb{E}\ln r_k$ and $\mathbb{E}r_k$ are readily computed using properties of the Gamma distribution.

Consequently, the following expression should be maximized to fix the hyperparameters:

$$\begin{aligned}\sum_{k=1}^{T}\Bigg\{&\ln|\mathbf{U}_k| - \operatorname{tr}\left(\mathbf{U}_k\mathbb{E}\left(\mathbf{w}_k\mathbf{w}_k^\top\right)\right)\\ &- (\mathbb{E}r_k)\sum_{n=1}^{N} q(z_n = k)\mathbb{E}\left[\left(y_n - \mathbf{w}_k^\top\phi_k(\mathbf{x}_n)\right)^2\right]\Bigg\} \qquad (33)\end{aligned}$$

where $\mathbf{U}_k$ and $\phi_k(\mathbf{x}_n)$ are related to the hyperparameters $\boldsymbol{\theta}_{1:T}$. In particular, to optimize $\boldsymbol{\theta}_k$, the following expression can be maximized:

$$\begin{aligned}&\ln|\mathbf{U}_k| - \operatorname{tr}\left(\mathbf{U}_k\mathbb{E}\left(\mathbf{w}_k\mathbf{w}_k^\top\right)\right)\\ &\quad- (\mathbb{E}r_k)\sum_{n=1}^{N} q(z_n = k)\mathbb{E}\left(\left(y_n - \mathbf{w}_k^\top\phi_k(\mathbf{x}_n)\right)^2\right). \qquad (34)\end{aligned}$$

Because the variational posterior of $\mathbf{w}_k$ is a Gaussian (suppose that the mean is μ and the covariance is Σ), variable $\phi_k(\mathbf{x}_n)^\top\mathbf{w}_k - y_n$ is also Gaussian distributed with mean $\phi_k(\mathbf{x}_n)^\top\mu - y_n$ and variance $\phi_k(\mathbf{x}_n)^\top\Sigma\phi_k(\mathbf{x}_n)$. We have

$$\begin{aligned}&\mathbb{E}\left(\left(y_n - \mathbf{w}_k^\top\phi_k(\mathbf{x}_n)\right)^2\right)\\ &\quad= \phi_k(\mathbf{x}_n)^\top(\Sigma + \mu\mu^\top)\phi_k(\mathbf{x}_n) + y_n^2 - 2y_n\phi_k(\mathbf{x}_n)^\top\mu. \qquad (35)\end{aligned}$$

Now, the formulation for optimizing $\boldsymbol{\theta}_k$ can be simplified as

$$\begin{aligned}\ln|\mathbf{U}_k| - \operatorname{tr}(\mathbf{U}_kA) - b\sum_{n=1}^{N} q(z_n = k)\,&\left[\phi_k(\mathbf{x}_n)^\top A\phi_k(\mathbf{x}_n)\right.\\ &\left.- 2y_n\phi_k(\mathbf{x}_n)^\top\mu\right] \qquad (36)\end{aligned}$$

where we have defined symmetric matrix $A = \mathbb{E}(\mathbf{w}_k\mathbf{w}_k^\top) = \Sigma + \mu\mu^\top$ and scalar $b = (\mathbb{E}r_k)$.

To maximize the objective in (36), the conjugate gradient ascent method [24] is used, which is outlined as follows for completeness.

For an objective function denoted as $\ell(\mathbf{w})$ with input $\mathbf{w}$, the gradient of the objective is $\mathbf{g} = \nabla_{\mathbf{w}}\ell(\mathbf{w})$, and the Hessian is

$$\mathbf{H} = \frac{d^2\ell(\mathbf{w})}{d\mathbf{w}d\mathbf{w}^\top}.$$

The Newton step along a direction $\mathbf{u}$ is

$$\mathbf{w}' \leftarrow \mathbf{w} - \frac{\mathbf{g}^\top\mathbf{u}}{\mathbf{u}^\top\mathbf{H}\mathbf{u}}\mathbf{u} \qquad (37)$$

where the values of $\mathbf{g}$ and $\mathbf{H}$ are taken at $\mathbf{w}$. For conjugate gradient ascent, $\mathbf{u} = \mathbf{g} - \beta\mathbf{u}^{last}$, and β can be given by the popular Hestenes–Stiefel formula [24] as

$$\beta = \frac{\mathbf{g}^\top(\mathbf{g} - \mathbf{g}^{last})}{(\mathbf{u}^{last})^\top(\mathbf{g} - \mathbf{g}^{last})}. \qquad (38)$$

The initial value of $\mathbf{u}$ can be $\mathbf{g}$ at some initial guess for $\mathbf{w}$.

To apply the conjugate gradient ascent method to the optimization of (36), the gradient and Hessian should be calculated. Now, we provide the main elements in formulating the gradient and Hessian. Suppose that θ_i and θ_j are entries of $\boldsymbol{\theta}_k$. We have

$$\begin{aligned}\frac{\partial\ln|\mathbf{U}_k|}{\partial\theta_i} &= \operatorname{tr}\left(\mathbf{U}_k^{-1}\frac{\partial\mathbf{U}_k}{\partial\theta_i}\right)\\ \frac{\partial\operatorname{tr}(\mathbf{U}_kA)}{\partial\theta_i} &= \operatorname{tr}\left(\frac{\partial\mathbf{U}_k}{\partial\theta_i}A\right)\\ \frac{\partial\left(\phi_k(\mathbf{x}_n)^\top A\phi_k(\mathbf{x}_n)\right)}{\partial\theta_i} &= \frac{\partial\phi_k(\mathbf{x}_n)^\top}{\partial\theta_i}\frac{\partial\left(\phi_k(\mathbf{x}_n)^\top A\phi_k(\mathbf{x}_n)\right)}{\partial\phi_k(\mathbf{x}_n)}\\ &= 2\frac{\partial\phi_k(\mathbf{x}_n)^\top}{\partial\theta_i}A\phi_k(\mathbf{x}_n)\\ \frac{\partial(\phi_k(\mathbf{x}_n)^\top\mu)}{\partial\theta_i} &= \frac{\partial\phi_k(\mathbf{x}_n)^\top}{\partial\theta_i}\mu \qquad (39)\\ \frac{\partial^2\ln|\mathbf{U}_k|}{\partial\theta_i\partial\theta_j} &= \partial\left(\frac{\partial\ln|\mathbf{U}_k|}{\partial\theta_i}\right)/\partial\theta_j\\ &= \operatorname{tr}\left(\mathbf{U}_k^{-1}\frac{\partial^2\mathbf{U}_k}{\partial\theta_i\partial\theta_j}\right)\\ &\quad- \operatorname{tr}\left(\mathbf{U}_k^{-1}\frac{\partial\mathbf{U}_k}{\partial\theta_j}\mathbf{U}_k^{-1}\frac{\partial\mathbf{U}_k}{\partial\theta_i}\right)\\ \frac{\partial^2\operatorname{tr}(\mathbf{U}_kA)}{\partial\theta_i\partial\theta_j} &= \operatorname{tr}\left(\frac{\partial^2\mathbf{U}_k}{\partial\theta_i\partial\theta_j}A\right)\\ \frac{\partial^2\left(\phi_k(\mathbf{x}_n)^\top A\phi_k(\mathbf{x}_n)\right)}{\partial\theta_i\partial\theta_j} &= 2\left[\frac{\partial^2\phi_k(\mathbf{x}_n)^\top}{\partial\theta_i\partial\theta_j}A\phi_k(\mathbf{x}_n)\right.\\ &\quad\left.+\frac{\partial\phi_k(\mathbf{x}_n)^\top}{\partial\theta_j}A\frac{\partial\phi_k(\mathbf{x}_n)}{\partial\theta_i}\right]\\ \frac{\partial^2\left(\phi_k(\mathbf{x}_n)^\top\mu\right)}{\partial\theta_i\partial\theta_j} &= \frac{\partial^2\phi_k(\mathbf{x}_n)^\top}{\partial\theta_i\partial\theta_j}\mu. \qquad (40)\end{aligned}$$

2) Fixing $\mathbf{I}_{1:T}$: Although the support sets can be identified from the likelihood functions similar to what variational EM uses, it would computationally be very difficult, considering the complex relationship between the support sets and the likelihood. As an alternative, we use the method proposed in [7] and [11] to fix support sets.

To find support set $\mathbf{I}_k$ for the kth component, the idea is to maximize the probability density of $q(\mathbf{w}_k)$ at its mean while holding the distributions of other latent variables fixed. The objective turns out to be maximizing the determinant of the inverse covariance [7]. The rationality lies at the assumption that a good support set should make the posterior distribution highly peaked, particularly when we have a large training set.

Because $\mathbf{w}_k \sim \mathcal{N}((\mathbf{U}_k + r_{k3})^{-1} r_{k2}, (\mathbf{U}_k + r_{k3})^{-1})$, the objective function is then

$$|\mathbf{U}_k + r_{k3}| = \left|\mathbf{U}_k + (\mathbb{E}r_k) \sum_{n=1}^{N} q(z_n = k)\phi_k(\mathbf{x}_n)\phi_k(\mathbf{x}_n)^\top\right|. \tag{41}$$

The procedure is explained as follows. First, we initialize the support sets at random. Then, we run the variational EM algorithm to obtain the variational posterior $q(\Omega)$ and the hyperparameters $\boldsymbol{\theta}_{1:T}$. Finally, a greedy algorithm is used to select the support sets from randomly chosen candidate sets to alleviate the computational burden [7]. The greedy algorithm incrementally selects examples to maximize the objective in (41). The whole process is repeated until convergence or a prefixed number of iterations are reached.

D. Prediction

For a new input $\mathbf{x}$, the predictive distribution is

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{D}, \Theta) &= \int p(y|\mathbf{x}, \Omega, \Theta) p(\Omega|\mathbf{D}, \Theta) d\Omega \\ &\simeq \int p(y|\mathbf{x}, \Omega, \Theta) q(\Omega) d\Omega \\ &\simeq p(y|\mathbf{x}, \hat{\Omega}, \Theta) \end{aligned} \tag{42}$$

where two approximations are used to make the computation feasible [7]. The first approximation replaces the true posterior by the variational posterior, and the second approximation replaces the average with respect to the distribution $q(\Omega)$ by a single value $\hat{\Omega}$. $\hat{\Omega}$ represents the posterior means of all the hidden variables involved. This approach is reasonable, because given sufficient data, posterior distributions are usually highly peaked [7], [8].

The predictive distribution can further be formulated as

$$\begin{aligned} p(y|\mathbf{x}, \hat{\Omega}, \Theta) &= \sum_{k=1}^{T} p(z = k, y|\mathbf{x}, \hat{\Omega}, \Theta) \\ &= \sum_{k=1}^{T} p(z = k|\mathbf{x}, \hat{\Omega}) p(y|\mathbf{x}, z = k, \hat{\Omega}, \Theta) \\ &= \sum_{k=1}^{T} \frac{p(z = k|\hat{\Omega}) p(\mathbf{x}|z = k, \hat{\Omega})}{\sum_{i=1}^{T} p(z = i|\hat{\Omega}) p(\mathbf{x}|z = i, \hat{\Omega})} \\ &\quad \times p(y|\mathbf{x}, z = k, \hat{\Omega}, \Theta). \end{aligned}$$

Because $p(y|\mathbf{x}, z = k, \hat{\Omega}, \Theta)$ is Gaussian distributed, the prediction $\hat{y}$ for the new input $\mathbf{x}$ would be the weighted average of the T Gaussian means, and the weights are given by $p(z = k|\hat{\Omega}) p(\mathbf{x}|z = k, \hat{\Omega}) / \sum_{i=1}^{T} p(z = i|\hat{\Omega}) p(\mathbf{x}|z = i, \hat{\Omega})$ $(k = 1, \ldots, T)$.

IV. Experiments on Traffic Prediction

Traffic flow prediction, which is defined to be predicting future traffic flows of a certain road segment, is an important direction in the research of intelligent transportation systems [10], [25]–[29]. Short-term traffic flow prediction, which is one of the most important and difficult tasks, determines the traffic volume in the next time interval, usually in the range of 5–30 min.

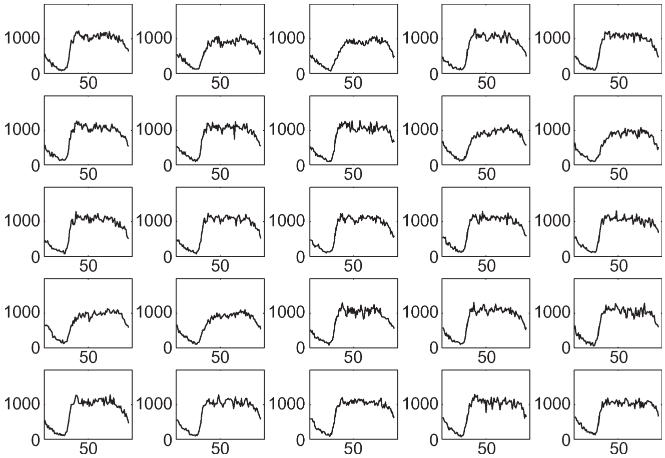

Fig. 2. Traffic flows of Ka during 25 days. The first row corresponds to the first five days, and so on, for a total of 25 days.

The focus of this section is on applying the proposed IMGP model and variational inference techniques to predicting the short-term traffic flows of road links based on their own historical traffic volumes. We will also compare this new approach with some other traffic prediction methods, including one state-of-the-art method BNs.

A. Data Description

The data analyzed are the vehicle flow rates recorded at an interval of 15 min with the Urban Traffic Control/Split Cycle and Offset Optimization Technique system by the Traffic Management Bureau of Beijing [25]. The unit of the data is vehicles per hour (veh/hr). As done in [10] and [25], we carry out a one-step prediction, and the prediction time horizon is 15 min. That is, some historical data are used to forecast the traffic flow rate for the next recording interval. Nine road links from the urban traffic map of Beijing City are used for the experiments, which demonstrate a wide spatial spread in the transportation network [25]. The road links are denoted as follows: 1) Bb; 2) Ch; 3) Dd; 4) Fe; 5) Gd; 6) Hi; 7) Ia; 8) Jf; 9) Ka.

The raw data for use are of 25 days, which include 2400 recording points for each road link. The starting time for data recording on a day is 00:00 at midnight. To illustrate the patterns of traffic flows, Fig. 2 depicts the raw data of road link Ka over 25 days.

B. Model Training

For experiments on each road link, the first 2112 points (from 22 days) are employed as the training set to infer the variational posterior distribution of hidden variables and the values of hyperparameters. The remaining data serve as the test set for model evaluation and comparison. For each example constructed from the training and test sets, the input includes four continuous traffic volumes, and the corresponding output is the very successive traffic volume.

The variational inference algorithm for model training is given in Table I. For the current traffic prediction problem, the parameters in the algorithm input are fixed as $T = 5$, $C = 100$, $S = 50$, $M_s = 10$, $M_{em} = 50$, $M_e = 50$, and $M_m = 50$. The candidate support sets and hyperparameters $\boldsymbol{\theta}_{1:T}$ are randomly

TABLE I
VARIATIONAL INFERENCE ALGORITHM FOR IMGP

```
Input
  $\mathbf{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$: training set
  $T$: truncated level in the Dirichlet process
  $C$: size of candidate support sets
  $S$: size of support sets
  $M_s$: number of iterations for updating support sets
  $M_{em}$: number of iterations in the EM algorithm
  $M_e$: number of iterations for updating variational posterior (E-step)
  $M_m$: number of iterations for updating hyperparameters (M-step)
Procedure
  Initialize $T$ candidate support sets, hyperparameters $\boldsymbol{\theta}_{1:T}$ and $\mathbf{I}_{1:T}$
  for $i = 1 : M_s$
      for $j = 1 : M_{em}$ % EM algorithm
          for $k = 1 : M_e$ % E-step
              Update the variational posterior distribution $q(\Omega)$ for all
              hidden variables $\Omega$
          end for
          for $l = 1 : M_m$ % M-step
              Update $\boldsymbol{\theta}_{1:T}$ using conjugate gradient ascent
          end for
      end for
      Update support sets $\mathbf{I}_{1:T}$
  end for
Output Variational posterior $q(\Omega)$; hyperparameters $\boldsymbol{\theta}_{1:T}$ and $\mathbf{I}_{1:T}$
```

initialized, whereas hyperparameters $\mathbf{I}_{1:T}$ are initially specified by running the k-means clustering algorithm [17].

To illustrate the result of variational posterior estimation, we give the partition results of the training data indicated by $q(z_n)$ for road link Ka. The proportions of the numbers of training data that belong to each of the five mixture components are 34.09%, 17.12%, 26.57%, 14.13%, and 8.08%, respectively.

C. Prediction Results

After model training, we can carry out traffic prediction on the test set using the method given in Section III-D. As aforementioned, to make prediction feasible, we use the posterior means of the hidden variables to represent their distributions. To justify this approach, we show the proportions of the five mixture components calculated in terms of the variational posterior means of $\nu_{1:T}$ for road link Ka. The resulting proportions for the five mixture components are 34.11%, 17.14%, 26.54%, 14.11%, and 8.09%, respectively, which are quite similar to the proportions accounted for by the training set. This case indicates the rationality of using posterior means to approximate the predictive distribution.

The prediction performance is measured by the criterion of root mean square error (RMSE). For a time series $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and its estimation $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$, the performance measure RMSE is given by the following formula:

$$\text{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)^{1/2}. \tag{43}$$

The BN approach [10] is one of the state-of-the-art methods for traffic flow prediction. It models traffic flows among adjacent road links in a transportation network by a BN. The joint distribution between the data used for predicting and the data to be predicted is described as a finite Gaussian mixture model whose parameters are estimated by an EM algorithm. This approach explicitly integrates information from adjacent road links to analyze the trend of the current link. When data used are restricted to recordings from a single link, it degenerates to the Markov chain model [25], [30]. It makes sense if we could compare the proposed method with this version of BNs for the current traffic prediction problem.

Table II gives the prediction results on the nine road links using our proposed variational inference model for IMGP. For comparison, we conducted experiments with the random walk (RW) approach (predicting a variable using its previous value) and ridge regression (RR) [8], in addition to the aforementioned BN approach. In Table II, we see that IMGP got the best results on seven of the nine road links, whereas on the other two links, its performance is very similar to the best result. The effectiveness of our proposed model is validated.

D. Discussion

Now, we discuss the differences between the previous BN model and the proposed IMGP model, with the aim to interpret their different behaviors in performance.

The major distinction lies at the capability to accommodate more related traffic flows. The BN approach can only consider a very limited number of related traffic flows as the input, because it directly learns a joint distribution between the input and output. Given limited training data, we cannot estimate a high-dimensional distribution well, because the number of unknown parameters will usually be high, which degrades the accuracy of parameter estimation. This case is a typical challenge of the curse of dimensionality [17].

However, the proposed mixture model of Gaussian processes effectively avoids this problem, because the maximum dimensionality considered for the involved flexible distribution is very limited, i.e., the dimensionality of the input in the training set $\mathbf{D}$. Although we would like to consider using many related traffic flows to regress the desired traffic flow such as enlarging the size of the support sets, the number of parameters to be estimated will not increase. For the current traffic prediction experiments, we used all the examples in a support set to regress a future traffic flow.

V. CONCLUSION

In this paper, we have presented a variational approximation for IMGP. In the mixture model, the input distribution is modeled by a multivariate Gaussian distribution with a full covariance, whereas the output distribution conditional on the input is a linear Gaussian process model. This linear Gaussian process makes effective variational inference possible. To make variational inference feasible, a truncated representation of the Dirichlet process and a factorization assumption for the posterior distribution are further used. Important techniques involved in our variational inference include the variational EM algorithm, conjugate gradient ascent algorithm, and a greedy algorithm for adapting support sets. To the best of our knowledge, the proposed approach is the first variational inference method for infinite Gaussian process mixture models.

To validate the proposed method, we applied it to the traffic flow prediction problem. Experiments and comparisons with other methods, including the BN approach, showed its

TABLE II
TRAFFIC PREDICTION RESULTS (IN RMSE) OF RW, RR, BN, AND IMGP

Method	Bb	Ch	Dd	Fe	Gd	Hi	Ia	Jf	Ka
RW	89.60	79.85	70.99	157.60	177.57	108.34	96.63	154.75	99.20
RR	87.04	75.61	67.25	152.33	171.06	107.52	95.04	147.50	94.19
BN	72.43	68.51	66.15	122.65	**151.31**	92.16	**84.20**	124.04	80.46
IMGP	**70.86**	**66.38**	**61.14**	**120.04**	152.35	**89.22**	84.22	**120.85**	**77.71**

effectiveness. This approach is an important attempt on using IMGP to the intelligent transportation field.

For future work, the proposed method can be extended in several ways, e.g., considering Gaussian mixture models (finite or infinite) to characterize the input distribution and integrating traffic flows from correlated road links in a transportation network to perform network-scale prediction.

REFERENCES

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[2] V. Tresp, "Mixtures of Gaussian processes," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 654–660, 2001.

[3] C. E. Rasmussen and Z. Ghahramani, "Infinite mixtures of Gaussian process experts," *Adv. Neural Inf. Process. Syst.*, vol. 14, pp. 881–888, 2002.

[4] E. Meeds and S. Osindero, "An alternative infinite mixture of Gaussian process experts," *Adv. Neural Inf. Process. Syst.*, vol. 18, pp. 883–890, 2006.

[5] R. A. Jacobs, M. I. Jordan, and G. E. Hinton, "Adaptive mixture of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, Spring 1991.

[6] L. Xu, M. I. Jordan, and G. E. Hinton, "An alternative model for mixtures of experts," *Adv. Neural Inf. Process. Syst.*, vol. 7, pp. 633–640, 1995.

[7] C. Yuan and C. Neubauer, "Variational mixture of Gaussian process experts," *Adv. Neural Inf. Process. Syst.*, vol. 21, pp. 1897–1904, 2009.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[9] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1/2, pp. 1–305, 2008.

[10] S. Sun, C. Zhang, and G. Yu, "A Bayesian network approach to traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 124–132, Mar. 2006.

[11] A. J. Smola and P. Bartlett, "Sparse greedy Gaussian process regression," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 619–625, 2001.

[12] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, Mar. 1973.

[13] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[14] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Anal.*, vol. 1, no. 1, pp. 121–144, 2006.

[15] J. Sethuraman, "A constructive definition of Dirichlet priors," *Stat. Sinica*, vol. 4, pp. 639–650, 1994.

[16] C. E. Rasmussen, "The infinite Gaussian mixture model," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 554–560, 2000.

[17] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2001.

[18] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures of Physics*. Reading, MA: Addison-Wesley, 1964.

[19] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[20] J. Ishwaran and L. James, "Gibbs sampling methods for stick-breaking priors," *J. Amer. Stat. Assoc.*, vol. 96, no. 453, pp. 161–174, Mar. 2001.

[21] G. Parisi, *Statistical Field Theory*. Reading, MA: Addison-Wesley, 1988.

[22] C. M. Bishop and M. Svensen, "Bayesian hierarchical mixtures of experts," in *Proc. 19th Conf. Uncertainty Artif. Intell.*, 2003, pp. 57–64.

[23] Y. W. Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning*. Berlin, Germany: Springer-Verlag, 2010.

[24] T. P. Minka, "A comparison of numerical optimizers for logistic regression," Mar. 26, 2007. [Online]. Available: http://research.microsoft.com/en-us/um/people/minka/papers/logreg/minka-logreg.pdf

[25] S. Sun and C. Zhang, "The selective random subspace predictor for traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 367–373, Jun. 2007.

[26] F.-Y. Wang, "Building an intellectual highway for ITS research and development," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 2–3, Mar. 2010.

[27] T. Thomas, W. Weijermars, and E. van Berkum, "Predictions of urban volumes in single time series," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 71–80, Mar. 2010.

[28] G. Vigos and M. Papageorgiou, "A simplified estimation scheme for the number of vehicles in signalized links," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 312–321, Jun. 2010.

[29] F.-Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 630–638, Sep. 2010.

[30] G. Yu, J. Hu, C. Zhang, L. Zhuang, and J. Song, "Short-term traffic flow forecasting based on Markov chain model," in *Proc. IEEE Intell. Veh. Symp.*, 2003, pp. 208–212.

Shiliang Sun (M'07) received the B.E. degree (with honors) in automatic control from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2002 and the Ph.D. degree (with honors) in pattern recognition and intelligent systems from Tsinghua University, Beijing, in 2007.

In 2004, he was a Microsoft Fellow. From 2009 to 2010, he was a Visiting Researcher with the Department of Computer Science, University College London, London, U.K., working within the Centre for Computational Statistics and Machine Learning. He is currently an Associate Professor with the Department of Computer Science and Technology and the Founding Director of the Pattern Recognition and Machine Learning Research Group, East China Normal University, Shanghai, China. His research interests include machine learning, pattern recognition, computer vision, intelligent transportation systems, and brain–computer interfaces.

Dr. Sun is a member of the Pattern Analysis, Statistical Modeling, and Computational Learning (PASCAL) Network of Excellence and an Associate Editor for the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Xin Xu (M'09) received the B.S. degree in electrical engineering in 1996 and the Ph.D. degree in control science and engineering from the National University of Defense Technology (NUDT), Changsha, China.

In August 2006 and from September to October 2007, he was a Visiting Scholar for Cooperational Research with the Hong Kong Polytechnic University, Kowloon, Hong Kong and the University of Strathclyde, Glasgow, U.K., respectively. He is currently a Full Professor with the Institute of Automation, College of Mechatronics and Automation, NUDT. He is a coauthor of four books and has published more than 50 papers in international journals and conference proceedings, including the *Journal of AI Research*. His research interests include reinforcement learning, learning control, robotics, data mining, autonomic computing, and computer security.

Dr. Xu is a member of the IEEE Technical Committee on Approximate Dynamic Programming and Reinforcement Learning and the IEEE Technical Committee on Robot Learning. He has published more than 50 papers in international journals and conference proceedings, including the IEEE TRANSACTIONS ON NEURAL NETWORKS.