Recent Advances in Data-Driven Wireless Communication Using Gaussian Processes: A Comprehensive Survey

Kai Chen²,³, Qinglei Kong²,³, Yijue Dai¹,², Yue Xu⁵, Feng Yin¹,²,*, Lexi Xu⁴,*, Shuguang Cui¹,²

¹ School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China
² Future Network of Intelligence Institute (FNii), The Chinese University of Hong Kong, Shenzhen 518172, China
³ School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China
⁴ Research Institute, China United Network Communications Corporation, Beijing 100048, China
⁵ Alibaba Group, Hangzhou 310052, China

* The corresponding author, email: yinfeng@cuhk.edu.cn and xulx29@chinaunicom.cn

Abstract: Data-driven paradigms are well-known and salient demands of future wireless communication. Empowered by big data and machine learning techniques, next-generation data-driven communication systems will be intelligent with unique characteristics of expressiveness, scalability, interpretability, and uncertainty awareness, which can confidently involve diversified latent demands and personalized services in the foreseeable future. In this paper, we review a promising family of nonparametric Bayesian machine learning models, i.e., Gaussian processes (GPs), and their applications in wireless communication. Since GP models demonstrate outstanding expressive and interpretable learning ability with uncertainty, they are particularly suitable for wireless communication. Moreover, they provide a natural framework for collaborating data and empirical models (DEM). Specifically, we first envision three-level motivations of data-driven wireless communication using GP models. Then, we present the background of the GPs in terms of covariance structure and model inference. The expressiveness of the GP model using various interpretable kernels, including stationary, non-stationary, deep and multi-task kernels, is showcased. Furthermore, we review the distributed GP models with promising scalability, which is suitable for applications in wireless networks with a large number of distributed edge devices. Finally, we list representative solutions and promising techniques that adopt GP models in various wireless communication applications.

Keywords: wireless communication; Gaussian process; machine learning; kernel; interpretability; uncertainty

I. INTRODUCTION

Recently, there has experienced an explosion of works in artificial intelligence (AI) for wireless communications [1–6]. Furthermore, traditional paradigms (TPs) based on mathematical modeling have greatly hindered the progress of future wireless communications and negatively affected its emerging applications, such as the Internet of Vehicles (IoV) [7–9], Internet of Things (IoT) [10–14], augmented/virtual reality (AR/VR) [15, 16], and energy efficient 5G [17–21]. Increasingly, many new breeds of smart connected sensors and AI-enabled applications heavily depend on intelligent real-time response and explainable decision making, e.g., emergency braking in self-driving vehicles, obstruction warning for drones, fault diagnosis for intelligent manufacturing, environmental perception for cooperative multirobot systems, and predictable human-computer interaction for AR/VR, to reduce response times and human-interventions. These application-driven requirements demand the next-generation communication systems [22–24] to be intelligent with the
following welcome features: flexibility, scalability, interpretability, and especially uncertainty modeling to confidently involve latent demands and personalized service in the future.

Compared with traditional paradigms in wireless communication, a significant advantage of machine learning is its capability of gaining knowledge and automatically extracting information without specific rules [25]. However, due to the insufficient interpretation of state prediction, machine learning methods with black-box decision making [26–28] always confuse the diagnosis and analysis of complex communication systems and lead to a passive understanding of its functioning mechanism. To promote interpretable machine learning for data-driven wireless communications, in this paper, we review the Gaussian process (GP) model, and present their applications in wireless communications due to their interpretable learning ability with uncertainty.

GP is a generalization of the Gaussian probability distribution, which means GP is any distribution over functions $f(x)$ such that any finite set of function values has a joint Gaussian distribution [25, 27]. The GP provides a model where a posterior distribution over the unknown function is maintained as evidence is accumulated. This allows GPs to learn the underlying functions of wireless communication systems when a large number of observations are collected. In contrast to the popular deep neural network (DNN) [31] and other learning models, GP model shows a unique property of uncertainty qualification with a closed-form mathematical expression of great value to data-driven wireless systems that demand controllable and understandable state prediction. In table 1, we compare GPs with DNN, reinforcement learning (RL), and TPs in terms of model expressiveness, interpretability, scalability, uncertainty modeling, sample efficiency, and collaboration of data and empirical models (DEM).

1.1 Related Work

A GP can model a large and complex wireless communication system through the design of its covariance function (also called kernel function), which encodes one’s assumption about the auto-covariance of an unknown function. Therefore, a kernel is crucial in a GP model, as it implies the characteristics of distribution over functions. In addition, scalable inference is another core aspect in GP because the computational complexity of GP is cubic $O(n^3)$ for large scale wireless communication systems. This usually prevents the GP from learning a big data problem. Thus, the most important advances in wireless communication using GPs are related to both the kernel function design and scalable inference, which are extensively studied [2, 35, 36, 34]. There are a number of tutorials and survey works reviewing GPs in wireless communication, such as [2, 37]. In a recent magazine paper [2], GP model was introduced as a replacement of the traditional deep learning model to address critical uncertainty issues of the next-generation data-driven wireless communication system, however, technical details of the GP models were kept brief. In the magazine paper [37], GP model was introduced for nonlinear signal processing applications. Therein, a number of illustrative applications of signal processing for wireless communication, such as probabilistic channel equalization and tracking of nonlinear channel variations, were presented. The main attention of [37] lies in demonstrating the superior nonlinear mapping ability of GP models, while not on model expressiveness and interpretability aspects. In contrast, our survey paper gives both a complete overview of the mathematical advances developed in recent years and a comprehensive survey of wireless communication applications benefited from using GP models.

Generally, for kernel function designation, there are broadly four categories of covariance design in the existing works for GP, including (1) compositional kernel design [38, 39], where kernels are constructed compositionally from several existing base kernels; (2) spectral kernel learning, where kernels are derived by modeling the kernel spectral density as a mixture of distributions [40–43]; (3) deep kernel representation [44, 45], where DNN plays a role in nonlinear mapping between input space and feature space; and (4) multi-task kernel [46, 47], where adjacent devices (tasks) share knowledge and interact with each other to obtain collective intelligence. In the next sections, we review related works in detail.

To overcome the computational complexity issue of GP [2, 27, 34, 48] used in large scale wireless communication systems, scalable inference can be achieved by exploring (1) low-rank covariance matrix approximation [49, 50], (2) special structures of the kernel matrix [51, 52], (3) Bayesian committee machine (BCM),
which distributes computations to a big number of computing units [53, 54], (4) variational Bayesian inference [55, 56], and (5) special optimization [57, 43]. Notably, these scalable methods are not exclusive, and we can combine some of them to get a better method, for instance, stochastic variational inference (SVI) [55, 56] combining the strength of inducing points for low-rank approximation and variational inference.

1.2 Outline

The main contributions of this survey are summarized below:

- We extensively discuss the generally desired AI features of next-generation data-driven wireless communication systems, namely, expressiveness, scalability, interpretability, and uncertainty modeling. Regarding these aspects, we compare GP with other machine learning methods and then conclude that GP can cover these qualities better (as shown in Table 1).

- We broadly analyze and explain four categories of covariance design in terms of mathematical theorem and GP kernel expression, including (1) stationary kernel, (2) non-stationary kernel, (3) deep kernel, and (4) multi-task kernel. These kernels leverage both the expressiveness and interpretability of the GP model in wireless communication.

- Due to the scalability demand and distributed deployment of wireless communication systems, we review and evaluate the advances of distributed GP with scalable inference for big data of cloud intelligence as well as AI-enabled edge devices.

- We show an exemplary case by extrapolating the number of online 5G users collected from a real world 5G wireless base station. The results show the expressiveness, interpretability, and uncertainty modeling of GPs for data-driven wireless communication.

- We exhibit some representative wireless communication scenarios for applying the GP models and further envision some open issues and challenges of using GPs for future data-driven wireless communication.

For the rest of this paper, we begin by introducing the motivation of using GPs for data-driven wireless communication in section II and then give the mathematical background of GPs in section III. In section IV and section V, we present the advances of GPs. A demonstration for wireless communication using GPs is given in section VI. In section VII and section VIII, we give existing GP applications and future researches on wireless communications, respectively.

II. DATA-DRIVEN WIRELESS COMMUNICATION: MOTIVATION WITH UNIQUE FEATURES

In this section, we present the unique features for the next-generation data-driven wireless communication using machine learning methods with expressiveness, scalability, uncertainty modeling, and interpretability (see Table 1).

Due to the inherent intelligence requirements in data-driven wireless communication systems, there are three levels of motivations to apply GPs. First, the low-level motivation is based on the demands of smart, efficient, and flexible decision making, planning, and prediction in future wireless communication systems [3], which cannot be achieved by applying traditional paradigms. Then, the comparison between GP and other machine learning methods brings the middle-level motivation and comprehensively explains why we tend to choose the GP model for data-driven wireless communication systems [25, 27]. As shown in Section VI, the high-level motivation is derived from the competitive applications empowered by GPs in wireless communication. Specifically, the motivations can be summarized as fol-
For future wireless communication systems, it is expected that there are many latent demands and personalized services driven by diversified applications. These latent demands and personalized services can be further modeled and improved by using machine learning methods, with the growth of historical data, and ever-increasing computing power. There are many features describing future wireless communication: (a) expressiveness correlated to model complexity which results from diversified application scenarios [4, 58]; (b) scalability on big data due to the ever-growing network size with network densification and an increasing number of connected intelligent devices [3, 59]; (c) uncertainty resulting from a dynamical communication environment [60, 61]; and (d) interpretable knowledge discovery and representation for understanding the mechanism of complex systems [62, 63]. In particular, uncertainty modeling is critical for state prediction in wireless networks since there are always multiple noises and dynamic factors intervening the status of the system and the mobile users’ experience.

As a class of Bayesian nonparametric model, GP provides a principled, practical, probabilistic approach for learning the patterns encoded by kernel structure [27]. Among all machine learning models, the GP has a tight connection with various learning models [25-27], including spline models, support vector machines (SVMs), regularized least-squares models, relevance vector machines (RVMs), autoregressive moving averages (ARMAs), and deep neural networks (DNNs). In particular, GPs have advantages with respect to the interpretation of model learning, model selection, and uncertainty prediction from a Bayesian point of view. Using an appropriate kernel structure and computational approximation, GP can model any function of wireless communication systems with flexibility and scalability. Owing to the Bayesian rules, GP with a measure of uncertainty is more robust to overfitting problems. In comparison with other machine learning models, the GP model can simultaneously meet the requirements of expressiveness, scalability, uncertainty modeling, and interpretability [25, 27] for data-driven wireless communications.

Thanks to the Bayesian properties, GP model has eye-catching interpretations in terms of model construction, selection, and hyper-parameter adaptation (see section III). Such interpretation strengths promote a large number of GP models to empower diversified wireless communication applications. There are five popular GP models using different kernels to support various wireless communication tasks, such as the GP models with stationary spectral mixture (SM) [40, 64] and compositional kernels [65] (see section 4.1), non-stationary (NS) kernels [66-69] (see section 4.2), deep kernels [44, 70] (see section 4.3), and multi-task kernels [34, 46] (see section 4.4). Furthermore, GPs have scalability variations with distributed inference to scale large data on a big number of edge devices (see section V). The distributed GPs can make full use of the computational resources of local edge devices in wireless networks to gain efficiency improvement as well as privacy protection [35, 71].

III. BACKGROUND OF GAUSSIAN PROCESS FOR MACHINE LEARNING

There are multiple uncertainty issues in the modeling of wireless communication: (1) functional uncertainty describing the gap between the true function and learned model; (2) prediction uncertainty with a fuzzy range caused by the number of observed evidences; (3) input uncertainty due to the noise generated during the wireless propagation; and (4) output uncertainty due to unstable wireless propagation and poor precision of measuring sensors. Theoretically, these uncertainties, as well as interpretability, can be well represented by a GP model. In this section, we briefly describe the background of the Gaussian process for machine learning in terms of its mathematical definition, kernel function and model inference.

3.1 Definition of Gaussian Process

From the function-space view, a Gaussian process [26, 27] defines a distribution $p(f(x_1), f(x_2), \ldots, f(x_n)) \sim \mathcal{N}(m(x), K(x, x'))$ over functions, completely specified by its first and second-order statistics, namely, the mean $m(x)$ and the covariance $k(x, x')$ functions [72]. For a given input location $x \in \mathbb{R}^p$ of a real stochastic
process $f(x)$, the mean $m(x)$ and covariance function $k(x, x')$ are defined as:

$$m(x) = \mathbb{E}[f(x)],$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))].$$

Thus, a GP is expressed as $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$. Without loss of generality, the mean of a GP is often assumed to be zero anywhere because we usually do not have any prior knowledge about the mean. The covariance function (also called the kernel) between function values is applied to construct a positive definite covariance matrix on input points $X$ for the joint Gaussian distribution, here denoted by Gram matrix $K = K(X, X)$. By using a GP prior over functions in the kernel designation and parameter initialization, from the training data $X$, we can predict the unknown function value $\tilde{y}_*$ and its variance $\mathbb{V}[y_*]$ (that is, its uncertainty) for a test point $x_*$. Specifically, we have the following predictive equations for GP regression [27, 48]:

$$\tilde{y}_* = k_*^\top (K + \sigma_n^2 I)^{-1}y, \quad \mathbb{V}[y_*] = k(x_*, x_*) - k_*^\top (K + \sigma_n^2 I)^{-1}k_*,$$  

where $k_*^\top$ is the covariance vector between $x_*$ and $X$, $\sigma_n^2$ is the variance of the noise, and $y$ is the vector of observations corresponding to $X$.

### 3.2 Gaussian Process Kernel

Basically, the smoothness and generalization properties of GP depend on the kernel function and its hyper-parameters $\Theta$. Choosing an appropriate kernel function and the corresponding initial hyper-parameters are crucial to GP design since the posterior distribution can vary significantly for different kernels. The most extensively used covariance function is stationary. We
introduce a generalized theory of both stationary and non-stationary covariance functions in the later sections. For the underlying function to be modeled by a GP, there are many characteristics, such as exponentially decayed dependency and periodic dependency, which can be encoded by specific covariance functions.

To make the GP model applicable for practical applications, the inference of the GP model is also very important. During the inference phase of the GP model, the freedom of model selection is considerable even though an appropriate covariance was specified in advance. Typically, GPs contain hyper-parameters $\Theta$ describing the properties of the kernel and noise of the GP. Suppose we have chosen a covariance function $k(x, x')$ with hyper-parameters $\Theta_k$. The inference of the GP means Bayesian model selection with the possible best values of $\Theta = \{\Theta_k, \sigma_n^2\}$. Such selection can be accomplished by minimizing the negative log marginal likelihood (NLML), $L_{\text{NLML}} = -\log p(y|\mathcal{X}, \Theta)$. The inference and posterior sampling of a GP model is illustrated in Figure 1.

The NLML can be used for assessing the goodness of fit of the GP model. For the evaluation of GP model, we usually apply the mean squared error (MSE) and mean absolute error (MAE) to measure prediction performance. Specifically, the predictive uncertainty described in Eq. (2b) scores the confidence of the prediction.

IV. ADVANCES IN GP KERNELS

4.1 Stationary Spectral Mixture Kernel

Data generated in wireless communication systems often demonstrate the following patterns: (1) weekly periodic trends on weekdays and weekends, (2) daily periodic trends in working hours and spare time, (3) decayed deviations in terms of small-scale variation, and (4) some noise introducing disorder fluctuations. These patterns are generally stationary and can be captured by the GP with a flexible kernel structure (see section VI). However, without tangible prior information, the number of patterns and their signal features are not clear for the definition and construction of a GP model. Alternatively, we can apply a universal representation of stationary kernels and then automatically infer the latent patterns through optimization, which can simplify the practice of machine learning in wire-

![Figure 2. Spectral densities (left) with a mixture of Gaussians and corresponding covariance functions (right) in the SM kernel. For SM, the location (black dot) of each component denotes the period of underfunction.](image)

less communication systems and enhance the efficiency of interpretable knowledge discovery.

In this section, we review the theoretical foundation of stationary covariance functions and recent GP works. Stationary covariance is regarded as a function of $\tau = x - x'$ other than input location $x$, which is invariant to translations in the input space [27]. For each covariance function of a stationary process, there is a corresponding representation, the Fourier transform of a positive finite measure $\psi$, in the frequency domain. Referring to [73, 74], Bochner’s theorem indicates the connection between the covariance function and its spectral density.

**Theorem 1** (Bochner’s Theorem [73, 74]). A complex-valued function $k$ on $\mathbb{R}^P$ is the covariance function of a weakly stationary mean square continuous complex-valued function $r(x)$ if and only if there exists a finite Borel measure $\psi$ on $\mathbb{R}^P$ such that $k(x, x') = \int_{\mathbb{R}^P} e^{2\pi i \tau \cdot (x - x')} d\psi$. Conversely, the Fourier transform of a measure $\psi$ defines a bounded linear functional on $C_0^\infty(\mathbb{R}^P)$.
valued random process on $\mathbb{R}^P$ if and only if it can be represented as
\[
k(\tau) = \int_{\mathbb{R}^P} e^{2\pi \mathbf{p}^\top \tau} \psi(d\mathbf{s}),
\]
where $\psi$ is a positive finite measure and $\mathbf{j}$ denotes the imaginary unit.

If $\psi$ has a density $\hat{k}(\mathbf{s})$ called the spectral density or power spectrum of $k$, Theorem (1) implies the following Fourier dual.

\[
\begin{align*}
k(\tau) &= \int \hat{k}(\mathbf{s}) e^{2\pi \mathbf{p}^\top \tau} d\mathbf{s}, \\
\hat{k}(\mathbf{s}) &= \int k(\tau) e^{2\pi \mathbf{p}^\top \tau} d\tau.
\end{align*}
\]

Based on Bochner’s theorem, a large number of expressible stationary kernels are proposed, including the spectral mixture kernels (SMs) and compositional kernels. Compositional kernels [75, 65] have advanced kernel structures constructed from a combination of normal kernels by using a series of kernel operations, such as plus, wrap, and product operations. Furthermore, one of the most representative stationary kernels is the SM kernel [40, 76, 41, 77, 78], as SM can approximate any stationary kernels with a sufficient number of components. Here, we mainly introduce the SM kernel. An SM kernel $k_{SM}$ is derived by representing its spectral density (the Fourier transform of a kernel) with a Gaussian mixture model (GMM) (see Figure 2), $\hat{k}_{SM}(\mathbf{s}) = \sum_{i=1}^{Q} w_i \varphi_{SM,i}(\mathbf{s})$, where $\hat{k}_{SM,i}(\mathbf{s}) = w_i [\varphi_{SM,i}(\mathbf{s}) + \varphi_{SM,i}(-\mathbf{s})]/2$, $Q$ is the number of Gaussians, $w_i$ is the weight of the $i$-th Gaussian, and $\varphi_{SM,i}(\mathbf{s}) = N(\mathbf{s}; \mu_i, \Sigma_i)$ is a scale-location Gaussian with mean $\mu_i$ and variance $\Sigma_i$. The symmetrization makes $\hat{k}_{SM,i}(\mathbf{s})$ even, that is, $\hat{k}_{SM,i}(\mathbf{s}) = \hat{k}_{SM,i}(-\mathbf{s})$ for all $\mathbf{s}$. Then, applying the inverse Fourier transform, we can obtain the SM kernel as follows:

\[
k_{SM}(\tau) = \mathcal{F}_{\omega \rightarrow \tau}^{-1} \left[ \hat{k}_{SM}(\mathbf{s}) \right](\tau) = \sum_{i=1}^{Q} w_i \cos \left( 2\pi \tau^\top \mu_i \right) \exp \left( -2\pi^2 \tau^\top \Sigma_i \tau^\top \right),
\]

where $\mathcal{F}_{\omega \rightarrow \tau}^{-1}$ denotes the inverse Fourier transform operator from the frequency domain to the time domain. For the SM kernel, we can interpret $w_i$, $\mu_i$, and $\Sigma_i$ as the signal variance, inverse period, and inverse length scale of the $i$-th covariance component, respectively. In summary, the SM kernel can be seen as a generalization of existing stationary kernels. Note that the GP model with an SM kernel has been used for wireless traffic prediction [34] and is trusted by the application of wireless communication. In section VI we predict the number of online 5G users by using a GP model with an SM kernel.

4.2 GP with Non-stationary Kernel

In addition to stationary patterns, there are also a few complex non-stationary patterns with time-varying characteristics for wireless communication, for instance, mmWave massive MIMO channel modeling [79], 5G wireless channel modeling [80], wireless control systems [81], 3D non-stationary unmanned aerial vehicle (UAV) MIMO channels [82], non-stationary mobile-to-mobile channels allowing for velocity and trajectory variations in mobile stations [83], and non-stationary channel modeling for vehicle-to-vehicle communications [84]. In contrast to the stationary kernel depending only on the distance $\tau$, the signal characteristics of non-stationary GP, such as frequencies, amplitudes, and spectral densities, have direct dependences on the input locations $\mathbf{x}$. The extension of Bochner’s theorem (see Theorem 1) to the non-stationary domain has a generalized spectral representation on the $P \times P$ surface

\[
k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^P} \int_{\mathbb{R}^P} e^{2\pi \mathbf{p}^\top (\mathbf{x}-\mathbf{x}'^\prime)} \mathbf{u}_S(d\mathbf{s}, d\mathbf{s}'),
\]

where $\mathbf{u}_S$ is a positive finite measure on spectral surface $P \times P$.

Arguably, the dot product kernel is the simplest non-stationary kernel [27]. The well-known and extensively used non-stationary kernels are linear and polynomial kernels [27], which are less parameterized for representing complex patterns. Since the introduction of the neural network (NN) kernel [28], GPs can approximate both DNN and one hidden layer neural network model (known for universal approximator and nonlinear property) with infinity neurons. After that, Gibbs [85] developed the non-stationary covariance function shown in Eq. (6) by considering a grid of exponential
Then, Higdon [86] proposed a non-stationary spatially evolving GP using a process convolution to model toxic waste remediation. Based on [86], Paciorek [87] generalized the Gibbs kernel using non-stationary quadratic form $Q_{x,x'} = (x-x')(\Sigma_x + \Sigma_{x'})^{-1}(x-x')$ instead of $\tau$ in any stationary kernel, where $\Sigma_x$ is the positive length scale function of input $x$. After proposing the SM kernel (see Eq. (4)), in [66], a non-stationary SM (NSM, see Figure 3) kernel was introduced by modeling the spectral surface as a two-dimensional GMM.

$$k_{NSM}(x, x') = \sum_{i=1}^{Q} w_i^2 \exp(-2\pi^2 \tilde{x}\Sigma_{x}\tilde{x}) \times \Psi_{\mu_i,\mu'_i}(x) \Psi_{\mu_i,\mu'_i}(x'),$$

(7)

where $\tilde{x} = (x, -x')^\top$ and

$$\Psi_{\mu_i,\mu'_i}(x) = \begin{pmatrix} \cos 2\pi \mu x + \cos 2\pi \mu' x \\ \sin 2\pi \mu x + \sin 2\pi \mu' x \end{pmatrix}.$$

For the aforementioned non-stationary kernels, their hyper-parameters can be parameterized as positive functions described by stationary GPs. For example, we can parameterize $\theta_\ell$ as $\theta_\ell \sim GP(0, k_\ell(x, x))$. Recently, the harmonizable kernel [88] showed a novel spectral representation of the non-stationary kernel by incorporating a locally stationary kernel with an interpretation of the Wigner distribution function. In [67], another convolutional spectral kernel was proposed to give a concise representation of the input frequency spectrogram, but it shows less insight into a prespecified complex-valued radial base. To meet the development needs of a non-stationary GP, the non-separable and non-stationary kernel [89], including a varying non-separability and local structure, has a natural interpretation through the spectral representation of stochastic differential equations (SDEs).

### 4.3 Interpretable Deep Kernel

Neal[28] proved that a Bayesian neural network with infinitely many hidden neurons converges to a GP. In practice, GPs with popular kernels are mostly used as simple nonlinear interpolation models. Deep neural networks (DNNs) are demonstrated in their competent learning and representation in many application domains, including computer vision [90], speech recognition [91], language processing [92], and recommendation systems [93]. The most interesting DNN capability is feature discovery and representation. However, DNNs have a well-known interpretation imperfection.
in that the mechanism of model learning and inference is a black box, which heavily depends on hyper-parameter tuning techniques. Therefore, deep kernel GP (DKGP) [44, 94, 95] combines the nonparametric flexibility of kernel methods with the inductive biases of deep learning architectures, which presents benefits in both expressive power and interpretability. As a result, the DKGP can draw their strengths to learn a model for complicated wireless mechanisms, such as 5G and vehicle-to-everything (V2X) channel impulse responses, multipath radio signal propagation, radio feature maps (such as the signal quality, uplink/downlink traffic, wireless resource demand/supply) over time and space, and indoor pedestrian motion, etc.

For DKGP, a typical framework extracts features from DNNs and then treats the features as inputs of multiple GPs [44, 95]. The model comes from linearly mixing these GPs and jointly optimizing hyper-parameters through a marginal likelihood objective. The understanding of this kind of deep kernel is straightforward and can actually be seen as the GP using complicated feature engineering or transformation before learning. The popular structure of the deep kernel can be written as

$$k_{\text{Deep}}(x, x') \rightarrow \sum_{i=1}^{Q} k_i(g_{\text{NN}}(x, w_{\text{NN}}), g_{\text{NN}}(x', w_{\text{NN}})),$$

(8)

where $g_{\text{NN}}(x, w_{\text{NN}})$ denotes a nonlinear feature mapping given by DNN with weights $w_{\text{NN}}$. Note that the kernel $k_i$ used in Eq. (8) can be arbitrary. Similar to the DNN, the chain rule is also applicable for deep kernel learning. According to the chain rule, the derivatives of the NLML with respect to the deep kernel hyper-parameters are given as follows:

$$\frac{\partial \mathcal{L}}{\partial \Theta} = \frac{\partial \mathcal{L}}{\partial K_{\text{Deep}}} \frac{\partial K_{\text{Deep}}}{\partial \Theta},$$

(9a)

$$\frac{\partial \mathcal{L}}{\partial w_{\text{NN}}} = \frac{\partial \mathcal{L}}{\partial K_{\text{Deep}}} \frac{\partial K_{\text{Deep}}}{\partial g_{\text{NN}}(x, w_{\text{NN}})} \frac{\partial g_{\text{NN}}(x, w_{\text{NN}})}{\partial w_{\text{NN}}},$$

(9b)

where the derivative of NLML with respect to the covariance matrix is $\frac{\partial \mathcal{L}}{\partial K_{\text{Deep}}} = \frac{1}{2} \left(K_{\text{Deep}}^{-1} \mathbf{y} \mathbf{y}^\top K_{\text{Deep}}^{-1} - K_{\text{Deep}}^{-1}\right)$. Another deep kernel using the finite rank Mercer kernel function with orthogonal embeddings on the last layer has a better learning efficiency and expressiveness [96]. However, incorporating DNN into GP leads to poor interpretability due to DNN’s blackbox. To enrich the interpretability of deep kernels, the second class of deep kernels was proposed to reveal the learning dynamics of the DNN by building connections between the GP and DNN. Furthermore, considerable focus has been paid on interaction detection in DKGP to enhance its interpretability. Interestingly, a recently proposed novel optimal DKGP (see Figure 4) [45] demonstrates better model interpretability. The resulting kernel has a non-stationary dot product structure with minimized

Figure 4. The covariance structure of the optimal DKGP [45], where a multilayer fully connected feed-forward NN is applied as the universal approximator of the underlying function $f(x)$. 

(© China Communications Magazine Co., Ltd. · January 2022)
test mean squared error, shallow DNN subnetworks with feature interaction detection, much reduced hyperparameter space, and good interpretability.

4.4 Collective Intelligence Using Multi-task Kernel

In wireless communication systems, adjacent devices are not independent and must be correlated because there are shared patterns and environmental factors between them. For example, connected smartphones, robots, drones, vehicles, intelligent home systems, and NB-IoT sensors in the same wireless network may have dependent behaviors or trends impacted by the status of the wireless network. Hence, a joint learning model can make full use of data collected from adjacent devices to achieve collective intelligence. Knowledge obtained from different edge devices can be transferred to augment the overall prediction performance and system understanding. Therefore, a paradigm of multi-task learning can empower such collaboration in wireless communication.

The extension of GPs to multiple sources of data is known as a multi-task or multioutput Gaussian process (MTGP or MOGP). MTGP accounts for the statistical dependence across different sources of data (or tasks) [46, 47, 97]. Given $m$ tasks, the aim of the MTGP model is to jointly learn $m$ underlying functions $f_i(X) = [f_i^{(1)}(X), ..., f_i^{(m)}(X)]^\top$ and estimate their function values $y = [y^{(1)}, ..., y^{(m)}]^\top$, where $X = [x^{(1)}, ..., x^{(m)}]^\top$, $y \sim N(f_i(X), \epsilon 1)$, and $\epsilon = [\epsilon^{(1)}, ..., \epsilon^{(m)}]$ represents the noise variances of the $m$ tasks. However, similar to a single GP, underlying functions $f_i(X)$ still have a Gaussian distribution with $f_i(X) \sim GP(0, K_{MTGP}(x, x'))$. Usually, MTGP with $m$ tasks has a computational complexity of $O(m^3n^3)$ when the size of the training data in each task is $n$. If a point $x$ comes from a task $m$ and $x'$ comes from another task $m'$, then their covariance is

$$k_{MTGP}(x, x') = k_{m, m'}(x - x').$$

For MTGP, a crucial point is how to jointly encode the shared structure and difference between tasks in the kernel [98]. Kernel design should consider both the cross-covariance between tasks and auto-covariance within each task. Early MTGP approaches mainly focus on linear combinations and convolution of inde-pendent single-source GPs, which correspond to the linear model of coregionalization (LMC) framework [46, 99, 100] and convolved GP [101, 102], respectively. Many improvements and applications of MT-GPs have been introduced in previous works, such as [46, 99, 102, 103]. One method for promoting the representation ability of MTGP model is via using the SM kernels. First, the SM-LMC kernel [99] models the covariance of a single task with an SM kernel, linearly combines these single tasks with LMC and provides an interpretation of the Gaussian process regression network (GPRN) from the perspective of a neural network with

$$K_{SM-LMC}^{m,m'} = \sum_{i=1}^{Q} B_i \otimes k_{SM,i},$$

where $k_{SM,i}$ is a covariance structure shared by tasks and $B_i$ encodes the cross-covariance between tasks. Then, the cross-spectral mixture (CSM) kernel [100] additionally introduced a phase factor into $B_i$ to encode amplitude and phase for cross-covariance with

$$K_{CSM}^{m,m'} = \sum_{i=1}^{Q} B_i \otimes k_{SG,i}(\tau; \Theta_i),$$

where $k_{SG,i}(\tau; \Theta_i)$ is the phasor notation of a spectral Gaussian kernel. The multioutput spectral mixture kernel (MOSM) [103] further represents both time and phase delays in cross covariance between tasks by using complex-valued matrix decomposition.

However, MOSM has a compatibility drawback in that it cannot reduce to the SM kernel when only one task is available. Therefore, a multioutput convolution spectral mixture (MOCSM) kernel [104] was proposed to enjoy the compatibility property perfectly through cross convolution of time and phase delayed SM components. Another important extension of MTGP is multi-task generalized convolution SM (MT-GCSM) kernel [105], which models nonlinear task correlations and dependence between arbitrary components and provides a framework for heterogeneous tasks with different levels of complexity. The later convolved GP is more flexible and expressive because it allows each task to have its own kernel and complexity.
V. SCALABLE DISTRIBUTED GAUSSIAN PROCESS

The distributed Gaussian process (DGP) in wireless communication involves learning on distributed edge devices. The use of DGP can avoid frequent interactions with a central server and allow each edge device to possess a local learning model. For delay-sensitive applications such as self-driving vehicles and unmanned aircraft, a local learning model can rapidly respond to a local request in a timely manner. Particularly, the DGP can save the overall time cost of the wireless communication when the central server is not available or network congestion occurs. Therefore, DGP can be seen as a form of on-device intelligence, which addresses the major concerns of scalable computation and privacy protection in wireless communication.

In this section, we introduce the framework of DGP, which has shown significant advantages in computational efficiency [1, 35, 106–107]. There are many reasons for the selection of DGP, such as scaling ordinary GP to large datasets, applying ordinary GP to distributed edge dataset, preventing access to privacy-sensitive data and making full use of multicore high-performance computers (HPCs). In general, DGP splits big data into multiple (M) smaller pieces computed on local computing nodes to speed up the inference of the whole model [108], which refrains from centrally collecting and storing massive data. The initial aim of DGP is to make GP scalable to big data. However, with the development of multicore computing architecture and edge computing in IoT networks, DGP is gradually receiving attention from research and industrial applications because it provides a more practical machine learning framework than the existing GPs. Some representative DGP works have been published recently [1, 35, 106–109]. By using the map-reduce framework and decoupling the data conditioned on the inducing points, a distributed variational inference for GP and latent variable models (LVMs) was proposed [108]. The distributed variational inference for GP still has the limitation of scalable inference when the data size is n \geq 10^7. Another DGP is based on the mixture-of-experts (MoE) model [110]. The MoE model weights the predictions of all local expert models (node) to give the final prediction. For MoE, confusion is how to specify the number of experts and weight of each expert. Compared with MoE, product-of-GP-experts models (PoEs) [109] that multiply predictions of independent GP experts can avoid assigning weight to experts but are inevitably overconfident. The marginal likelihood p(y|X, \Theta) of PoEs is written as follows:

\[
p(y|X, \Theta) \approx \prod_{i=1}^{M} p^{(i)}(y^{(i)}|X^{(i)}, \Theta), \quad (13)
\]

where M is the number of GP experts and \( p^{(i)}(y^{(i)}|X^{(i)}, \Theta) \) is the marginal likelihood of the i-th GP expert using the i-th partition \{X^{(i)}, y^{(i)}\} of dataset \{X, y\}. Additionally, the predictive probability of PoEs is the product of all predictive probabilities of independent GP experts,

\[
p(f_s|x_s, y, X) \approx \prod_{i=1}^{M} p^{(i)}(f_s|x_s, y^{(i)}, X^{(i)}). \quad (14)
\]

Similarly, the Bayesian committee machine (BCM) [111] combines independent estimators trained on different datasets by using Bayes’ rule. BCM has a better interpretation due to considering the GP prior \( p(f_s) \). Furthermore, robust BCM [112] generalized the original BCM and PoE-GP by incorporating a GP prior and the importance of GP experts. In order to achieve a much better approximation of a full GP, other improved DGP works include: (1) asynchronously distributed variational GP [107] that uses weight-space augmentation to scale up GPs to billions of samples; (2) generalized robust BCM [113] that gains a consistent aggregated predictive distribution by randomly selecting a subset \( D^{(1)} \) as a global node for communicating with the remaining subsets; (3) nested kriging predictors that aggregates submodels based on subsets of observation points[114].

VI. AN EXAMPLE OF DATA-DRIVEN WIRELESS COMMUNICATION USING GP: THE PREDICTION OF ONLINE 5G USERS

In this section, we show an exemplary case by considering a real world wireless communication dataset regarding the number of online 5G users collected from a 5G base station. This experiment can substantiate the features of data-driven wireless communication using GPs in terms of expressiveness, interpretability, and
uncertainty modeling. The dataset was collected in a southern city in China in 2021. As shown in Figure 5, there are multiple patterns with different time scales in the varying of the number of online 5G users, such as long-term (subplot (b)), short-term (subplot (c)), and mid-term (subplot (d)) trends. We split 70% of the dataset as training (in blue, up subplots) data and the rest 30% as testing (in cyan, up subplots) data. We set the SM kernel with $Q = 6$ components and initialize the hyper-parameters by fitting a Gaussian mixture model on the empirical spectral densities (in blue, bottom subplots). Subplot (a) in Figure 5 indicates that the GP model (in dashed red) can extrapolate the patterns of testing well. The predictive trend (in dash red) is very close to the ground true trend (in cyan). Moreover, the predictive 95% confidence interval (CI) (in grey shade) of the GP model shows uncertainty bounds of the prediction and completely covers the ground truth.

The interpretability of the GP model can be sniffed from the structure of spectral density (in dash red, bottom subplots). As shown in subplot (i) Figure 5, there are six significant peaks captured by the GP model, which denotes different patterns with different periods. The peaks located at low frequencies reveal long-term trends with period $\mu_i^{-1}$. On the contrary, the peaks located at high frequencies describe short-term trends. These trends with different time scales connect to the periodicity of human activities. The covariance of long-term trend in subplot (f) decays much slower than mid-term trend in subplot (g) and short-term trend in subplot (h). The amplitudes of spectral peaks determine the scales of their corresponding function values. Specifically, we exhibit three patterns (the 1st, 2nd, and 4th) learned by the GP model and their corresponding spectral densities. For instance, the 4th pattern in subplot (d) demonstrates a mid-term evolution of online 5G users.
which has a spectral density peak located at $\mu_4 = 0.04$. The sum of all these 6 spectral peaks constitutes the learned spectral structure of the GP model, namely, the Fourier transform of the covariance of the GP model. In this context, all learned patterns have clear physical interpretations. However, for neural networks we cannot get such interpretations with insights for each neuron. The expressiveness of the GP model is guaranteed due to the theories, namely, the generalization of Fourier transform and the approximation of Gaussian mixture on spectral distribution. In particular, the expressiveness of the GP model can be further improved by giving more components when learning from high complicated wireless task. By using the distributed framework shown in Section V, the scalability of the GP model is easily achieved for large wireless datasets.

VII. GP BASED WIRELESS APPLICATIONS

In this section, to show the widespread applications of GPs in wireless communication, we further review and discuss three typical applications: wireless traffic prediction, localization, and trajectory planning.

7.1 Wireless Traffic Prediction

Wireless traffic prediction has been a long-standing demand for wireless network planning and management. Highly accurate wireless traffic prediction can reduce the uncertainty of network load and reflect the traffic behavior in wireless network, which greatly matches the benefits of GP in terms of uncertainty and interpretability. There are many traffic related issues in wireless communication suitable for GP model, such as wireless traffic analysis [1], cellular traffic load prediction [115], traffic load balancing for multimedia multipath systems [116], channel prediction for communication-relay UAV [117], stochastic link modeling of static wireless sensor networks [118]. We survey some representative examples as follows:

- **Wireless traffic prediction.** In [1], a GP model with the alternating direction method of multipliers (ADMM) for distributed hyper-parameter optimization was proposed to predict 4G wireless traffic, which shows better performance than a DNN model, such as long short-term memory (LSTM).

- **Cellular traffic load prediction.** In [115], a scheme combining GP and LSTM was proposed to generate accurate cellular traffic load prediction, which is important for efficient and automatic network planning and management. Compared with benchmark schemes, the proposed scheme achieves state-of-the-art performance.

- **Traffic load balancing for multi-media multipath systems.** In [119], an adaptive load balancing algorithm using an online GP was proposed to estimate the path status and allocate traffic load to each path properly in multimedia multipath systems. The proposed GP based algorithm is helpful for offering higher reliability and stability utilizing a variety of communication media and paths.

7.2 Wireless Localization

Another representative wireless scenario using GPs is wireless localization. Wireless localization has become a cornerstone of modern life due to the increasing demand of location-based applications, e.g., shopping and industry activities. GPs have been successfully applied for indoor wireless localization due to its uncertainty modeling and nonlinear regression abilities, such as wireless tracking [35], online radio map update [120], and calibrating multichannel RSS observations for localization [116]. We review various indoor localization frameworks using GPs as follows:

- **Wireless target tracking.** In [35], a framework of distributed recursive GP was proposed to build multiple local received signal strength (RSS) maps, which has reduced computational complexity on big data generated from large-scale sensor networks. Then, a global map is constructed from the fusion of all the local RSS maps. The proposed framework shows excellent positioning accuracy in both static fingerprinting and mobile target tracking scenarios.

- **Online radio map update.** In [120], a novel scheme combining crowdsourcing and GP regression can adapt radio maps to environmental dynamics in an online fashion, which recursively fuses crowdsourced fingerprints with an existing offline radio map. The scheme has particular advantages in efficiency and scalability.
• **Calibrating multichannel RSS observations for localization.** In [116], a GP model was proposed to compensate for frequency-dependent shadowing effects and multipaths in received signal strength (RSS) observations. By applying the GP model, multichannel RSS observations can be more effectively combined for localization over a large space.

### 7.3 Wireless Trajectory Planning

UAV has been an emerging technology strongly correlated with wireless communication. Due to the complex communication environment and irregular time delay in wireless network, UAV trajectory planning is rapidly gaining attention in wireless communication. To meet the requirement of expressiveness appearing in wireless network, popular RL and DNN methods [6] have been recently considered as panaceas for wireless trajectory planning. In addition, the GP-based approaches with uncertainty representation and sample efficiency are also impressive for wireless trajectory planning because of their unique features, namely, model expressiveness, uncertainty representation, and sample efficiency. We review recent advances of wireless trajectory planning using GPs as follows:

- **GP-based runtime planning, learning, and recovery for safe UAV operations.** In [121], a recursive dynamic GP regression-based framework with fast online planning, learning, and recovery approach was proposed for safe UAV operations under unknown runtime disturbances. The framework can estimate the behavior of the UAV system and provide safe plans at runtime under unseen disturbances.

- **Networked operation of a UAV using GPs.** In [122], a novel GP model predictive control (MPC) scheme for path planning was introduced to deal with the time-varying network delay, non-linearity, and time-sensitive characteristics of multirotor-type UAVs. The scheme increased the accuracy of path planning and state estimation for multirotor-type UAVs.

- **Obstacle-aware informative path planning using GPs for UAV-based target search.** In [123], an algorithm leverages a layered planning strategy using a GP-based model of target occupancy to generate informative paths in continuous 3D space. The algorithm can achieve a balance between information gain, field coverage, sensor performance, and collision avoidance for efficient target detection.

### VIII. GP FOR FUTURE WIRELESS COMMUNICATION

From the motivations of using GP for wireless communication, we note that there are many emerging difficulties. We outline a few challenging open issues of GP models for future data-driven wireless communication.

- **Ultra large-scale distributed GPs for dense and decentralized wireless communication systems.** In future data-driven wireless communication, the widely existing sensors gather considerable data at all times, which leads to large considerable data transmission and storage. An effective and pragmatic solution reducing the cost of data transmission and storage is to perform ultra large-scale distributed machine learning. Even though the scalability of GP is available currently. However, ultra large-scale distributed GP is still an open research issue.

- **GPs for multimodal data in wireless communication systems.** Currently, the GP model can only learn from structured data generated from wireless communication systems. There are also multimodal data collected from different types of sensors, such as numerical raw data from smartphones, ultrasound data from UAV ultrasonic sensors, images and videos from surveillance cameras, and natural language from speech sensors. Particularly, the signaling and data transmitted via interfaces of both LTE/5G wireless and core networks, are always nonstructured. Therefore, learning from nonstructured and multimodal data in wireless communication systems is another challenge for GP model.

- **Highly interpretable GPs with deep structure in wireless communication systems.** The deep kernel of a GP has difficulties in that it increases the flexibility of the GP model as well as the difficulty of model interpretation. From both the theorems of stationary and non-stationary kernels, the mathematical definition of deep kernels in the frequency domain remains unclear. Similar to DNN,
sacrificing interpretability in data-driven wireless communication is usually the compromise option between learning and understanding the network, which is less tolerable for high complexity state prediction. Hence, pursuing a highly interpretable GP with deep structures will be a critical open issue in future data-driven wireless communication.

• Privacy-enhancing GPs to address privacy issues and legal concerns. Since machine learning models depend on the sharing of the required training and prediction data, the concerns of potential privacy disclosure must be addressed. One type of privacy-enhancing GP solution is the fully or partially homomorphic encryption (HE) [124], and the other type of privacy-enhancing GP solution adopts differential privacy (DP) technique [125]. However, the HE techniques may introduce heavy computational overheads, while DP techniques may degrade the performance of the training process. Besides, an emerging type of privacy-enhancing technique is through the exploitation of federated learning [126, 127], which collaboratively learn a shared model while keeping all the training data on local devices. However, design of efficient privacy-preserving federated learning-based GPs for wireless communications remains an open challenge.

IX. CONCLUSION

In this paper, we comprehensively review data-driven wireless communication using GPs in terms of motivation, definition and construction of a GP model, GP expressiveness using different kernels, and distributed GP scalability. A GP with a Bayesian nature can model a large class of wireless communication systems through the designation of its covariance function. By using a distributed approach, GP models are capable of performing scalable inference on big data in a wireless network.

Data-driven wireless communication systems using GPs can achieve desired properties, expressiveness, scalability, interpretability, and uncertainty modeling. These characteristics become crucial for models in wireless communication due to the collected rich data and the modeling complexity in wireless networks. In particular, interpretability and uncertainty modeling are inherent advantages of GPs due to their mathematical definition. From existing applications of the GP models in wireless communication, we present that the GP models can cover the aforementioned properties of data-driven wireless communication very well, which has been successfully proven to be valuable.

ACKNOWLEDGMENT

The work was supported in part by the National Key R&D Program of China with grant No. 2018YFB1800800, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, by Natural Science Foundation of China (NSFC) with grants No. 92067202 and No. 62106212, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, and by China Postdoctoral Science Foundation with grant No. 2020M671899.

References

[1] Y. Xu, F. Yin, et al., “Distributed Gaussian process: New paradigm and application to wireless traffic prediction,” in 2019 IEEE International Conference on Communications, ICC 2019, Shanghai, China, May 20-24, 2019. IEEE, 2019, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICC.2019.8761264
[2] Y. Xu, F. Yin, et al., “Scalable learning paradigms for data-driven wireless communication,” IEEE Commun. Mag., vol. 58, no. 10, 2020, pp. 81–87. [Online]. Available: https://doi.org/10.1109/MCOM.2020.9014763
[3] Y. Liu, S. Bi, et al., “When machine learning meets big data: A wireless communication perspective,” IEEE Vehicular Technology Magazine, vol. 15, no. 1, 2019, pp. 63–72.
[4] M. Z. Chowdhury, M. Shahjalal, et al., “6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions,” IEEE Open Journal of the Communications Society, vol. 1, 2020, pp. 957–975.
[5] Q. Yan, W. Chen, et al., “Big data driven wireless communications: A human-in-the-loop pushing technique for 5G systems,” IEEE Wireless Communications, vol. 25, no. 1, 2018, pp. 64–69.
[6] J. Hu, H. Zhang, et al., “Cooperative Internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning,” IEEE Transactions on Communications, vol. 68, no. 11, 2020, pp. 6807–6821.
[7] H. Zhou, W. Xu, et al., “Evolutionary V2X technologies toward the Internet of vehicles: Challenges and opportunities,” Proceedings of the IEEE, vol. 108, no. 2, 2020, pp. 308–332.
[8] L. Jin, G. Zhang, et al., “Quantitative survivability analysis using probability model checking: A study of cluster-based vehicle networks with dual cluster heads,” China Communications, vol. 17, no. 2, 2020, pp. 206–219.
[9] J. Xie, S. Xiao, et al., “A throughput-aware joint vehicle route and access network selection approach based on SMDP,” China Communications, vol. 17, no. 5, 2020, pp. 243–265.

[10] F. Samie, L. Bauer, et al., “From cloud down to things: An overview of machine learning in Internet of Things,” IEEE Internet of Things Journal, vol. 6, no. 3, 2019, pp. 4921–4934.

[11] G. Casolla, S. Cuomo, et al., “Exploring unsupervised learning techniques for the Internet of Things,” IEEE Transactions on Industrial Informatics, vol. 16, no. 4, 2019, pp. 2621–2628.

[12] L. Wang, S. Liu, et al., “Dynamic uplink transmission scheduling for satellite Internet of Things applications,” China Communications, vol. 17, no. 10, 2020, pp. 241–248.

[13] W. Feng, J. Tang, et al., “NOMA-based UAV-aided networks for emergency communications,” China Communications, vol. 17, no. 11, 2020, pp. 54–66.

[14] W. Chen, N. Deligiannis, et al., “On the energy self-sustainability of IoT via distributed compressed sensing,” China Communications, vol. 17, no. 12, 2020, pp. 37–51.

[15] X. Qiao, P. Ren, et al., “Web AR: A promising future for mobile augmented reality—state of the art, challenges, and insights,” Proceedings of the IEEE, vol. 107, no. 4, 2019, pp. 651–666.

[16] X. Qiao, P. Ren, et al., “Mobile web augmented reality in 5G and beyond: Challenges, opportunities, and future directions,” China Communications, vol. 16, no. 9, 2019, pp. 141–154.

[17] J. Zhang, Y. Zeng, et al., “Receding horizon optimization for energy-efficient UAV communication,” IEEE Wireless Communications Letters, vol. 9, no. 4, 2019, pp. 490–494.

[18] M. Hua, Y. Wang, et al., “Energy-efficient cooperative secure transmission in multi-UAV-enabled wireless networks,” IEEE Transactions on Vehicular Technology, vol. 68, no. 8, 2019, pp. 7761–7775.

[19] K. A. Bousu, W. Zhou, et al., “Optimal power allocation with limited feedback of channel state information in multi-user MIMO systems,” China Communications, vol. 17, no. 2, 2020, pp. 163–175.

[20] Y. Chen, J. Yang, et al., “Energy efficiency optimization for heterogeneous cellular networks modeled by Matérn hard-core point process,” China Communications, vol. 17, no. 8, 2020, pp. 70–80.

[21] R. Chen, H. Zhou, et al., “Spectral and energy efficiency of line-of-sight OAM-MIMO communication systems,” China Communications, vol. 17, no. 9, 2020, pp. 119–127.

[22] M. Lin and Y. Zhao, “Artificial intelligence-empowered resource management for future wireless communications: A survey,” China Communications, vol. 17, no. 3, 2020, pp. 58–77.

[23] J. Zhu, M. Zhao, et al., “Exploring the road to 6G: ABC-foundation for intelligent mobile networks,” China Communications, vol. 17, no. 6, 2020, pp. 51–67.

[24] W. Xia, X. Zhang, et al., “The interplay between artificial intelligence and fog radio access networks,” China Communications, vol. 17, no. 8, 2020, pp. 1–13.

[25] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[26] D. MacKay, “Introduction to Gaussian processes,” 1998.

[27] C. K. Williams and C. E. Rasmussen, Gaussian processes for machine learning. MIT press Cambridge, MA, 2006.

[28] R. M. Neal, Bayesian learning for neural networks. Springer Science & Business Media, 2012, vol. 118.

[29] A. Ghazal, Y. Yuan, et al., “A non-stationary IMT-advanced MIMO channel model for high-mobility wireless communication systems,” IEEE Transactions on Wireless Communications, vol. 16, no. 4, 2016, pp. 2057–2068.

[30] R. He, B. Ai, et al., “Mobility model-based non-stationary mobile-to-mobile channel modeling,” IEEE Transactions on Wireless Communications, vol. 17, no. 7, 2018, pp. 4388–4400.

[31] Y. LeCun, Y. Bengio, et al., “Deep learning,” nature, vol. 521, no. 7553, 2015, p. 436.

[32] A. Zappone, M. Di Renzo, et al., “Wireless networks design in the era of deep learning: Model-based, AI-based, or both?” IEEE Transactions on Communications, vol. 67, no. 10, 2019, pp. 7331–7376.

[33] Z. Ning, K. Zhang, et al., “Joint computing and caching in 5G-envisioned Internet of vehicles: A deep reinforcement learning-based traffic control system,” IEEE Transactions on Intelligent Transportation Systems, 2020.

[34] Y. Xu, F. Yin, et al., “Wireless traffic prediction with scalable Gaussian process: Framework, algorithms, and verification,” IEEE Journal on Selected Areas in Communications, vol. 37, no. 6, 2019, pp. 1291–1306.

[35] F. Yin and F. Gunnarsson, “Distributed recursive Gaussian processes for RSS map applied to target tracking,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 3, 2017, pp. 492–503.

[36] Y. Xu, F. Yin, et al., “Scalable Gaussian process using inexact ADMM for big data,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 7495–7499.

[37] F. Pérez-Cruz, S. Van Vaerenbergh, et al., “Gaussian processes for nonlinear signal processing: An overview of recent advances,” IEEE Signal Processing Magazine, vol. 30, no. 4, 2013, pp. 40–50.

[38] G. Rios and F. Tobar, “Compositionally-warped Gaussian processes,” Neural Networks, vol. 118, 2019, pp. 235–246.

[39] J. Chen, H. Avron, et al., “Hierarchically compositional kernels for scalable nonparametric learning,” The Journal of Machine Learning Research, vol. 18, no. 1, 2017, pp. 2214–2255.

[40] A. Wilson and R. Adams, “Gaussian process kernels for pattern discovery and extrapolation,” in Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 1067–1075.

[41] K. Chen, T. van Laarhoven, et al., “Incorporating dependencies in spectral kernels for Gaussian processes,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2019, pp. 565–581.

[42] S. Remes, M. Heinonen, et al., “Non-stationary spectral kernels,” in Advances in Neural Information Processing Systems, 2017, pp. 4642–4651.

[43] F. Yin, L. Pan, et al., “Linear multiple low-rank kernel based stationary Gaussian processes regression for time series,” IEEE Trans. Signal Process., vol. 68, 2020, pp. 5260–5275.
[44] A. G. Wilson, Z. Hu, et al., “Deep kernel learning,” in Artificial intelligence and statistics, 2016, pp. 370–378.
[45] Y. Dai, T. Zhang, et al., “An interpretable and sample efficient deep kernel for Gaussian process,” in Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), ser. Proceedings of Machine Learning Research, J. Peters and D. Sontag, Eds., vol. 124. Virtual: PMLR, 03–06 Aug 2020, pp. 759–768. [Online]. Available: http://proceedings.mlr.press/v124/dai20a.html
[46] E. V. Bonilla, K. M. Chai, et al., “Multi-task Gaussian process prediction,” in Advances in neural information processing systems, 2008, pp. 153–160.
[47] W. Ruan, A. B. Milstein, et al., “Multiple output Gaussian process regression algorithm for multi-frequency scattered data interpolation,” in Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS), Jul. 2017, pp. 3992–3995.
[48] C. E. Rasmussen and H. Nickisch, “Gaussian processes for machine learning (GPML) toolbox,” Journal of Machine Learning Research, vol. 11, no. Nov, 2010, pp. 3011–3015.
[49] C. K. Williams and M. Seeger, “Using the Nyström method to speed up kernel machines,” in Advances in Neural Information Processing Systems, 2001, pp. 682–688.
[50] C. K. Williams and C. E. Rasmussen, Gaussian processes for machine learning. MIT press Cambridge, MA, 2006.
[51] A. Wilson and H. Nickisch, “Kernel interpolation for scalable structured Gaussian processes (KISS-GP),” in International Conference on Machine Learning, 2015, pp. 1775–1784.
[52] Y. Saatçi, “Scalable inference for structured Gaussian process models,” Ph.D. dissertation, Citeseer, 2012.
[53] V. Tresp, “A Bayesian committee machine,” Neural Computation, vol. 12, no. 11, 2000, pp. 2719–2741.
[54] J. Quiñonero-Candela and C. E. Rasmussen, “A unifying view of sparse approximate Gaussian process regression,” Journal of Machine Learning Research, vol. 6, no. Dec, 2005, pp. 1939–1959.
[55] J. Hensman, A. Matthews, et al., “Scalable variational Gaussian process classification,” Journal of Machine Learning Research, 2015.
[56] J. Hensman, N. Durrande, et al., “Variational Fourier features for Gaussian processes,” Journal of Machine Learning Research, vol. 18, no. 151, 2017, pp. 1–151.
[57] C. K. Ling, K. H. Low, et al., “Gaussian process planning with Lipschitz continuous reward functions: Towards unifying Bayesian optimization, active learning, and beyond,” in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17, 2016, Phoenix, Arizona, USA, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 1860–1866.
[58] X. Cheng, L. Fang, et al., “Mobile big data: The fuel for data-driven wireless,” IEEE Internet Things J., vol. 4, no. 5, 2017, pp. 1489–1516. [Online]. Available: https://doi.org/10.1109/JIOT.2017.2714189
[59] N. Zhang, P. Yang, et al., “Synergy of big data and 5G wireless networks: opportunities, approaches, and challenges,” IEEE Wireless Communications, vol. 25, no. 1, 2018, pp. 12–18.
[60] Z. Liu, J. Liu, et al., “On covert communication with interference uncertainty,” in 2018 IEEE International Conference on Communications (ICC). IEEE, 2018, pp. 1–6.
[61] B. Su, Q. Ni, et al., “Robust transmit beamforming for swipt-enabled cooperative NOMA with channel uncertainties,” IEEE Transactions on Communications, vol. 67, no. 6, 2019, pp. 4381–4392.
[62] P. Yang, Y. Xiao, et al., “5G wireless communications: Vision and potential techniques,” IEEE Network, vol. 33, no. 4, 2019, pp. 70–75.
[63] Z. Chang, L. Lei, et al., “Learn to cache: Machine learning for network edge caching in the big data era,” IEEE Wireless Communications, vol. 25, no. 3, 2018, pp. 28–35.
[64] H. Zou, Y. Zhou, et al., “Adaptive localization in dynamic indoor environments by transfer kernel learning,” in 2017 IEEE wireless communications and networking conference (WCNC). IEEE, 2017, pp. 1–6.
[65] D. Duvenaud, J. R. Lloyd, et al., “Structure discovery in nonparametric regression through compositional kernel search,” arXiv preprint arXiv:1302.4922, 2013.
[66] S. Remes, M. Heinonen, et al., “Non-stationary spectral kernels,” in Advances in Neural Information Processing Systems, 2017, pp. 4645–4654.
[67] Z. Shen, M. Heinonen, et al., “Learning spectrograms with convolutional spectral kernels,” arXiv preprint arXiv:1905.09917, 2019.
[68] X. Wang, X. Wang, et al., “Deepmap: Deep Gaussian process for indoor radio map construction and location estimation,” in 2018 IEEE Global Communications Conference (GLOBECOM). IEEE, 2018, pp. 1–7.
[69] A. Jung, G. Tauböck, et al., “Compressive spectral estimation for nonstationary random processes,” IEEE Transactions on Information Theory, vol. 59, no. 5, 2013, pp. 3117–3138.
[70] F. Teng, W. Tao, et al., “Localization reliability improvement using deep Gaussian process regression model,” Sensors, vol. 18, no. 12, 2018, p. 4164.
[71] S. Salari, I.-M. Kim, et al., “Distributed cooperative localization for mobile wireless sensor networks,” IEEE Wireless Communications Letters, vol. 7, no. 1, 2017, pp. 18–21.
[72] C. M. Bishop, Pattern recognition and machine learning, 5th Edition, ser. Information science and statistics. Springer, 2007. [Online]. Available: https://www.worldcat.org/oclc/71008143
[73] M. Stein, “Interpolation of spatial data: some theory for Kriging,” 1999.
[74] S. Bochner, Lectures on Fourier Integrals. Princeton University Press, 2016, vol. 42.
[75] D. K. Duvenaud, H. Nickisch, et al., “Additive Gaussian processes,” in Advances in neural information processing systems, 2011, pp. 226–234.
[76] P. A. Jang, A. Loeb, et al., “Scalable Lévy process priors for spectral kernel learning,” in Advances in Neural Information Processing Systems, 2017, pp. 3943–3952.
[97] K. Chen, T. van Laarhoven, et al., “Multitask gaussian process with hierarchical latent interactions,” 2021.

[98] R. Yu, G. Li, et al., “Tensor regression meets Gaussian processes,” in International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain, ser. Proceedings of Machine Learning Research, A. J. Storkey and F. Pérez-Cruz, Eds., vol. 84. PMLR, 2018, pp. 482–490. [Online]. Available: http://proceedings.mlr.press/v84/yu18a.html

[99] A. G. Wilson, D. A. Knowles, et al., “Gaussian process regression networks,” arXiv preprint arXiv:1110.4411, 2011.

[100] K. R. Ulrich, D. E. Carlson, et al., “GP kernels for cross-spectrum analysis,” in Advances in neural information processing systems, 2015, pp. 1999–2007.

[101] M. A. Álvarez, L. Rosasco, et al., “Kernels for vector-valued functions: A review,” Foundations and Trends® in Machine Learning, vol. 4, no. 3, 2012, pp. 195–266.

[102] M. A. Álvarez and N. D. Lawrence, “Computationally efficient convolved multiple output Gaussian processes,” Journal of Machine Learning Research, vol. 12, no. May, 2011, pp. 1459–1500.

[103] G. Parra and F. Tobar, “Spectral mixture kernels for multi-output Gaussian processes,” in Advances in Neural Information Processing Systems, 2017, pp. 6684–6693.

[104] K. Chen, T. van Laarhoven, et al., “Multitask convolution spectral mixture for Gaussian processes,” IEEE Transactions on Neural Networks and Learning Systems, 2019.

[105] K. Chen, T. van Laarhoven, et al., “Generalized convolution spectral mixture for multitask Gaussian processes,” IEEE Transactions on Neural Networks and Learning Systems, 2020.

[106] M. TavassoliPour, S. A. Motahari, et al., “Learning of Gaussian processes in distributed and communication limited systems,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, 2020, pp. 1928–1941. [Online]. Available: https://doi.org/10.1109/TPAMI.2019.2906207

[107] H. Peng, S. Zhe, et al., “Asynchronous distributed variational Gaussian process for regression,” in Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 2788–2797. [Online]. Available: http://proceedings.mlr.press/v70/peng17a.html

[108] Y. Gal, M. van der Wilk, et al., “Distributed variational inference in sparse Gaussian process regression and latent variable models,” in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, Z. Ghahramani, M. Welling, et al., Eds., 2014, pp. 3257–3265.
Biographies

Kai Chen received his Ph.D. degree from the University of Chinese Academy of Sciences in computer science (2019). Since 1 January 2020, he is a postdoctoral researcher at the Chinese University of Hong Kong, Shenzhen, China. His current research interests include data driven wireless communication, stochastic process, multitask and transfer learning, and automatic machine learning.

Qinglei Kong received her Ph.D. degree from the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore, in 2018; the M.Eng. degree in electronic and information engineering from Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, in 2015; and the B.Eng. degree in communication engineering from Harbin Institute of Technology, Harbin, China, in 2012. She used to work in Cyber Security Cluster, Institute for Infocom Research, Singapore and Tencent Security, Shenzhen, as a research scientist. Now she is working in The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China, as a postdoc researcher. Her research interests include applied cryptography, blockchain, VANET, and game theory.

[109] J. W. Ng and M. P. Deisenroth, “Hierarchical mixture-of-experts model for large-scale Gaussian process regression,” arXiv preprint arXiv:1412.3078, 2014.
[110] T. V. Nguyen and E. V. Bonilla, “Fast allocation of Gaussian process experts,” in Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014. ser. JMLR Workshop and Conference Proceedings, vol. 32. JMLR.org, 2014, pp. 145–153. [Online]. Available: http://proceedings.mlr.press/v32/nguyena14.html
[111] V. Tresp, “A Bayesian committee machine,” Neural Comput., vol. 12, no. 11, 2000, pp. 2719–2741. [Online]. Available: https://doi.org/10.1162/089976600300014908
[112] M. P. Deisenroth and J. W. Ng, “Distributed Gaussian processes,” in Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, ser. JMLR Workshop and Conference Proceedings, P. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 1481–1490. [Online]. Available: http://proceedings.mlr.press/v37/deisenroth15.html
[113] H. Liu, J. Cai, et al., “Generalized robust Bayesian committee machine for large-scale Gaussian process regression,” in Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 3137–3146.
[114] D. Rullière, N. Durrande, et al., “Nested Kriging predictions for datasets with a large number of observations,” Stat. Comput., vol. 28, no. 4, 2018, pp. 849–867. [Online]. Available: https://doi.org/10.1007/s11222-017-9766-2
[115] W. Wang, C. Zhou, et al., “Cellular traffic load prediction with LSTM and Gaussian process regression,” in ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020, pp. 1–6.
[116] Z. He, Y. Li, et al., “Calibrating multi-channel RSS observations for localization using Gaussian process,” IEEE Wireless Communications Letters, vol. 8, no. 4, 2019, pp. 1116–1119.
[117] P. Ladosz, H. Oh, et al., “Gaussian process based channel prediction for communication-relay UAV in urban environments,” IEEE Transactions on Aerospace and Electronic Systems, vol. 56, no. 1, 2019, pp. 313–325.
[118] A. Shahahngi, Y. Yang, et al., “On the stochastic link modeling of static wireless sensor networks in ocean environments,” in IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE, 2019, pp. 1144–1152.
[119] J. Kim, J. Park, et al., “Gaussian process regression-based traffic load balancing for multimedia multipath systems,” IEEE Transactions on Network and Service Management, vol. 17, no. 2, 2019, pp. 1211–1223.
[120] Z. Xu, B. Huang, et al., “Online radio map update based on a marginalized particle Gaussian process,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 4624–4628.
[121] E. Yel and N. Bezzo, “GP-based runtime planning, learning, and recovery for safe UAV operations under unforeseen disturbances,” in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 2173–2180.
[122] D. Jang, J. Yoo, et al., “Networked operation of a UAV using Gaussian process-based delay compensation and model predictive control,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 9216–9222.
[123] A. A. Meera, M. Popović, et al., “Obstacle-aware adaptive informative path planning for UAV-based target search,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 718–724.
[124] P. Fisener and E. Pyzer-Knapp, “Privacy-preserving Gaussian process regression - A modular approach to the application of homomorphic encryption,” in 2020 Advancement of Artificial Intelligence (AAAI), 2020, pp. 3866–3873.
[125] M. T. Smith, M. Álvarez, et al., “Differentially private regression with Gaussian processes,” in 2018 International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 84. PMLR, 2018, pp. 1195–1203.
[126] F. Yin, Z. Lin, et al., “Fedloc: Federated learning framework for data-driven cooperative localization and location data processing,” IEEE Open Journal of Signal Processing, vol. 1, 2020, pp. 187–215.
[127] Q. Kong, F. Yin, et al., “Privacy-preserving aggregation for federated learning-based navigation in vehicular fog,” IEEE Transactions on Industrial Informatics, vol. 17, no. 12, 2021, pp. 8453–8463.
Yijue Dai received the M.S. degree from Xiangtan University in 2017. She is currently a Ph.D. candidate in the Chinese University of Hong Kong, Shenzhen, China. Her current research interests mainly focus on Gaussian process, Kernel learning with respect to data efficiency, interpretability, scalability and optimality.

Yue Xu received the Ph.D. degree from Beijing University of Post and Telecommunication (BUPT), Beijing, China, in 2020. He is currently a research scientist at Alibaba Group. He has been a Visiting Researcher with the Department of Electrical and Computer Engineering, University of California at Davis, Davis, CA, USA, the Chinese University of Hong Kong, Shenzhen, China, and Shenzhen Research Institute of Big Data, Shenzhen. His research interests include data-driven wireless network management, machine learning, large-scale data analytics, and system control.

Feng Yin is an assistant professor at The Chinese University of Hong Kong, Shenzhen, China. He received the B.Sc. degree from Shanghai Jiao Tong University, China, and the Ph.D. degree from Technische Universität Darmstadt, Germany, in 2014. His research interests include statistical signal processing, Bayesian deep learning, and sensory data fusion. He was a recipient of the Chinese Government Award for Outstanding Self-Financed Students Abroad in 2013 and the European Union Marie Curie Scholarship in 2014. He is currently serving as the Associate Editor of the Elsevier Signal Processing journal. He is a senior member of IEEE.

Lexi Xu received M.S. and Ph.D. degrees from Beijing University of Posts and Telecommunications, Beijing, China, and Queen Mary University of London, London, United Kingdom, in 2009 and 2013, respectively. From 2013 to 2020, he was a senior engineer at Network Technology Research Institute, China United Network Communications Corporation (China Unicom). Since August 2020, he is now a senior engineer at Research Institute, China Unicom. He is also a China Unicom delegate in ITU, ETSI, 3GPP, CCSA. His research interests include big data, self-organizing networks, radio resource management in communication systems.

Shuguang Cui received his Ph.D. in Electrical Engineering from Stanford University, California, USA, in 2005. Afterwards, he has been working as assistant, associate, full, Chair Professor in Electrical and Computer Engineering at the Univ. of Arizona, Texas A&M University, UC Davis, and CUHK at Shenzhen respectively. He has also been the Vice Director at Shenzhen Research Institute of Big Data. His current research interests focus on data driven large-scale system control and resource management, large data set analysis, IoT system design, energy harvesting based communication system design, and cognitive network optimization. He was selected as the Thomson Reuters Highly Cited Researcher and listed in the World’s Most Influential Scientific Minds by ScienceWatch in 2014. He was the recipient of the IEEE Signal Processing Society 2012 Best Paper Award. He has served as the general co-chair and TPC co-chairs for many IEEE conferences. He has also been serving as the area editor for IEEE Signal Processing Magazine, and associate editors for IEEE Transactions on Big Data, IEEE Transactions on Signal Processing, IEEE JSAC Series on Green Communications and Networking, and IEEE Transactions on Wireless Communications. He has been the elected member for IEEE Signal Processing Society SPCOM Technical Committee (2009-2014) and the elected Chair for IEEE ComSoc Wireless Technical Committee (2017-2018). He is a member of the Steering Committee for IEEE Transactions on Big Data and the Chair of the Steering Committee for IEEE Transactions on Cognitive Communications and Networking. He was also a member of the IEEE ComSoc Emerging Technology Committee. He was elected as an IEEE Fellow in 2013 and an IEEE ComSoc Distinguished Lecturer in 2014.