Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

EFFICIENT BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATION

Haidong Li

Department of Management Science
University of Chinese Academy of Sciences
3 Zhongguancun South Street
Beijing, 100190, CHINA

Long Wang

Department of Advanced Manufacturing
and Robotics
Peking University
5 Yiheyuan Road
Beijing, 100871, CHINA

Yijie Peng

Department of Management Science
and Information Systems
Peking University
5 Yiheyuan Road
Beijing, 100871, CHINA

Di Wang

Department of Industrial Engineering
and Management
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, 200240, CHINA

ABSTRACT

We consider bandwidth selection for kernel density estimation. The performance of kernel density estimator heavily relies on the quality of the bandwidth. In this paper, we propose an efficient plug-in kernel density estimator which first perturbs the bandwidth to estimate the optimal bandwidth, followed by applying a kernel density estimator with the estimated optimal bandwidth. The proposed method utilizes the zeroth-order information of kernel function and has a faster convergence rate than other plug-in methods in existing literature. Simulation results demonstrate superior finite sample performance and robustness of the proposed method.

1 INTRODUCTION

Kernel density estimation is a non-parametric method for estimating the probability density function of a random variable from samples. Owing to its flexibility, interpretability, and ease of use, kernel density estimation has become a common tool for empirical studies in various fields. In operations research, it can be used to model uncertain input parameters in simulation models (Steckley and Henderson 2003). In statistics, it is frequently employed for exploratory data analysis and estimating the density of data with unknown distributions (Silverman 1986). In econometrics, it has been used to investigate income distribution, price dynamics, and financial market volatility (Zambom and Ronaldo 2013). In environmental science, it has been utilized to analyze the spatial distribution of various phenomena, such as air pollution and forest fires (Okabe et al. 2009). In machine learning, it has been proved useful for anomaly detection, generative modeling, and fairness-aware algorithms (Cho et al. 2020).

The basic idea of kernel density estimation is to estimate the probability density function of a random variable by sampling a kernel function at each data point and then taking the sample average to obtain a smoothed estimate of the density function (Tsybakov 2009). The effective use of kernel density estimation requires the choice of a smoothing parameter, i.e., bandwidth (Wand and Jones 1994; Simonoff 2012). If

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

the bandwidth is too small, the estimate can become overly sensitive to individual data points, leading to a rough estimate with spurious features that are not representative of the underlying data-generating process. On the other hand, too large bandwidth can result in oversimplification of the estimate, leading to a loss of important features in the data that could be crucial for understanding the underlying model structure. Therefore, it is essential to select the optimal bandwidth to strike a balance between reducing noise and preserving the underlying features of the data (i.e., reducing bias).

Bandwidth selection has become one of the most widely studied topics in kernel density estimation (for details see Wand and Jones (1994), Jones et al. (1996), and Heidenreich et al. (2013)). There are two major categories of bandwidth selection methods in existing literature. Cross-validation methods, introduced by Rudemo (1982) and Bowman (1984), aim to minimize the integrated squared error, which is a stochastic process indexed by bandwidth. Plug-in methods, which trace back to Woodroofe (1970) and Nadaraya (1974), minimize the mean integrated squared error, which is a deterministic function of bandwidth. The expression of the optimal bandwidth involves some unknown parameters that need to be estimated. Plug-in methods based on the derivatives of kernel function, proposed by Park and Marron (1990) and Sheather and Jones (1991), are widely used for bandwidth selection. Recently, Tenreiro (2020) proposes a new class of Hermite series-based plug-in bandwidth selectors for kernel density estimation. Our study focuses on plug-in methods due to their faster convergence rate than cross-validation methods.

In this paper, we propose a new, zeroth-order, plug-in method to obtain kernel density estimator with a nearly optimal bandwidth. This method first estimates the unknown parameter in the expression of the optimal bandwidth. In contrast to existing plug-in methods, we utilize the zeroth-order information of the kernel function with bandwidth perturbation to derive an estimate of parameter. Then we apply the estimated parameter to obtain a nearly optimal bandwidth and conduct standard kernel density estimation. The proposed method is proved in theory to achieve a fast convergence rate that is close to the best possible rate. Finite sample performance and robustness of the proposed method are demonstrated through simulation experiments.

The rest of the paper is organized as follows. In Section 2, we introduce the setting of the kernel density estimation and bandwidth selection problems. Section 3 proposes a new plug-in kernel density estimator with asymptotic optimality. Section 4 discusses the convergence rate of bandwidth estimate. Section 5 presents simulation results. The last section concludes the paper and outlines future directions.

2 SETTING AND MOTIVATION

In this paper, we focus our discussions on the single-dimensional case. Let $X_1,\ldots,X_n$ be i.i.d. realizations of a univariate random variable with an unknown probability density function f. Suppose that the density $f(\cdot)$ is sufficiently smooth, i.e., it has bounded, integrable and continuous derivatives up to order 4. We would like to estimate the density $f(x)$ for any $x\in\mathbb{R}$.

In estimating $f(x)$, a kernel density estimator is defined as

$$\hat{f}(x) \triangleq \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{1}$$

where h is the bandwidth and kernel function $K(\cdot)$ satisfies $K(x)\geq 0$, $\int_{-\infty}^{+\infty}K(x)dx=1$, $K(x)=K(-x)$, $\int_{-\infty}^{+\infty}x^2K(x)dx<+\infty$, $\int_{-\infty}^{+\infty}K^2(x)dx<+\infty$. The positivity and normality of $K(x)$ guarantee a positive density estimate $\hat{f}(x)$. For example, for rectangle kernel function

$$K(x)=\begin{cases}1/2 & \text{if } |x|\leq 1,\\ 0 & \text{otherwise,}\end{cases}$$

kernel density estimator is a frequency histogram $f(x)=\frac{1}{2nh}\#\{i|X_i\in(x-h,x+h]\}$. More smooth and commonly used kernels are Gaussian kernel $K(x)=(2\pi)^{-1/2}e^{-x^2/2}$ and Epanechnikov kernel $K(x)=\frac{3}{4}(1-|x|^2)\mathbf{1}(|x|\leq 1)$.

For any point of interest $x \in \mathbb{R}$, the mean squared error (MSE) of $\hat{f}(x)$ can be expressed as

$$\mathrm{MSE}(x) \triangleq \mathbb{E}\left[\left(\hat{f}(x)-f(x)\right)^2\right] = \left(\mathbb{E}\left[\hat{f}(x)-f(x)\right]\right)^2 + \mathrm{Var}\left[\hat{f}(x)\right].$$

The bias of kernel density estimator is given by

$$\begin{aligned}
\mathbb{E}\left[\hat{f}(x)-f(x)\right] &= \int \frac{1}{h}K(\frac{x-y}{h})f(y)dy - f(x) \\
&= \int K(z)(f(x+hz)-f(x))dz \\
&\underset{h\to 0}{=} \frac{1}{2}h^2 f''(x)\int z^2K(z)dz + o(h^2), \qquad (2)
\end{aligned}$$

and the variance of kernel density estimator is given by

$$\begin{aligned}
\mathrm{Var}\left[\hat{f}(x)\right] &= n^{-1}\int \frac{1}{h^2}K^2(\frac{x-y}{h})f(y)dy - n^{-1}\left(f(x)+\mathbb{E}\left[\hat{f}(x)-f(x)\right]\right)^2 \\
&= (nh)^{-1}\int K^2(z)f(x+hz)dz - n^{-1}\left(f(x)+\mathbb{E}\left[\hat{f}(x)-f(x)\right]\right)^2 \\
&\underset{h\to 0,\ nh\to\infty}{=} (nh)^{-1}f(x)\int K^2(z)dz + o((nh)^{-1}). \qquad (3)
\end{aligned}$$

The bandwidth $h>0$ is a tuning parameter, which controls both the bias and variance of kernel density estimator. In general, we use a common bandwidth $h>0$ for all x and aim to minimize the mean integrated squared error (MISE)

$$\begin{aligned}
\mathrm{MISE} &\triangleq \int \mathrm{MSE}(x)dx \\
&\underset{h\to 0,\ nh\to\infty}{=} \int\left[\left(\frac{1}{2}h^2f''(x)\int z^2K(z)dz + o(h^2)\right)^2 + \left((nh)^{-1}f(x)\int K^2(z)dz + o((nh)^{-1})\right)\right]dx \\
&\underset{h\to 0,\ nh\to\infty}{=} \frac{1}{4}h^4\int (f''(x))^2dx\left(\int z^2K(z)dz\right)^2 + o(h^4) + (nh)^{-1}\int K^2(z)dz + o((nh)^{-1}) \\
&\triangleq \frac{1}{4}h^4R(f'')(\sigma^2(K))^2 + (nh)^{-1}R(K) + o(h^4) + o((nh)^{-1}),
\end{aligned}$$

where $R(f) \triangleq \int f^2(x)dx$ and $\sigma^2(K) \triangleq \int x^2K(x)dx$. Therefore, the optimal bandwidth becomes

$$h_{\mathrm{opt}} = n^{-1/5}\left(\frac{R(K)}{R(f'')(\sigma^2(K))^2}\right)^{1/5}. \qquad (4)$$

When the constant $R(f'')$ is known, the optimal bandwidth h_{opt} can be determined by (4), which is the idea of plug-in bandwidth selection methods. However, $R(f'')$ is typically unknown and involves the second-order derivative of the density $f(\cdot)$, which is arguably more challenging to estimate than the density itself. In addition, it is observed in numerical experiments that choosing $R(f'')$ in an ad hoc fashion may lead to substantially different MISEs. This motivates us to investigate the issue of how to effectively estimate $R(f'')$ that ensures a kernel density estimator with small MISE, and is adaptive to different forms of $f(\cdot)$.

3 PLUG-IN KERNEL DENSITY ESTIMATOR

To tackle this issue, we propose a new plug-in kernel density estimator, which we call Kernel Density Estimator with Bandwidth Perturbation (KDE-BP). In this approach, we first perturb the bandwidth to estimate $R(f'')$, which will be discussed in detail in the rest of this section. Then we plug the estimated parameter $\hat{R}(f'')$ into the optimal bandwidth h_{opt} in (4). At last, we utilize the kernel density estimator (1) with the estimated optimal bandwidth to estimate the density $f(x)$ for any $x \in \mathbb{R}$. The entire process of implementing KDE-BP is summarized in Algorithm 1.

Algorithm 1 Kernel Density Estimator with Bandwidth Perturbation

Input: i.i.d. samples $X_i,\ i = 1,\ldots,n$ and a kernel function $K(\cdot)$;

Parameter estimation: compute $(1/h_i)K\left((x-X_i)/h_i\right),\ i = 1,\ldots,n$, where bandwidths $h_i,\ i = 1,\ldots,n$ are i.i.d. generated from a distribution $\mathscr{P}$. Then estimate $f''(x)$ and $R(f'')$ by $\hat{f}''(x)$ and $\hat{R}(f'')$ (see the rest of this section).

Density estimation: compute $(1/h_{\text{BP}})K\left((x-X_i)/h_{\text{BP}}\right),\ i = 1,\ldots,n$, where the estimated optimal bandwidth is

$$h_{\text{BP}} = n^{-1/5}\left(\frac{R(K)}{\hat{R}(f'')(\sigma^2(K))^2}\right)^{1/5}.$$

Then estimate $f(x)$ by

$$\hat{f}_{\text{BP}}(x) = \frac{1}{nh_{\text{BP}}}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_{\text{BP}}}\right).$$

Output: the estimated optimal bandwidth h_{BP} and the density estimator $\hat{f}_{\text{BP}}(x)$.

From (2) and (3), we have as $h \to 0$

$$\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) = f(x) + \frac{1}{2}h^2 f''(x)\sigma^2(K) + (C(x)h^4 + o(h^4)) + \varepsilon(x,h,X),$$

where $C(x) = f''''(x)\int z^4 K(z)dz/24$ and $\varepsilon(x,h,X) \in \mathbb{R}$ is a random variable such that $\mathbb{E}[\varepsilon(x,h,X)|h] = 0$ and $\text{Var}[\varepsilon(x,h,X)|h] = (nh)^{-1}f(x)R(K) + o((nh)^{-1})$. Such expansion of kernel density estimator motivates us to use linear regression for estimating $f''(x)$. Specifically, let

$$\boldsymbol{y}(x) = \left[\frac{1}{nh_1}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right),\ldots,\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)\right]^\top,$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & \cdots & 1 \\ h_1^2\sigma^2(K)/2 & \cdots & h_n^2\sigma^2(K)/2 \end{bmatrix}^\top,$$

$$\boldsymbol{\beta}(x) = [f(x), f''(x)]^\top,$$

$$\boldsymbol{r}(x) = \left[(C(x)h_1^4 + o(h_1^4)) + \varepsilon(x,h_1,X),\ldots,(C(x)h_n^4 + o(h_n^4)) + \varepsilon(x,h_n,X)\right]^\top,$$

where $h_1,\ldots,h_n$ are the perturbed bandwidths and i.i.d. generated. Then, $\boldsymbol{y}(x) = \boldsymbol{X}\boldsymbol{\beta}(x) + \boldsymbol{r}(x)$, and we can estimate $\boldsymbol{\beta}(x)$ given by $\hat{\boldsymbol{\beta}}(x) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}(x)$. Consider the second component of $\hat{\boldsymbol{\beta}}(x)$, we have

$$\hat{f}''(x) = \frac{n\sum_{i=1}^{n} h_i^2 y_i(x) - \left(\sum_{i=1}^{n} h_i^2\right)\left(\sum_{i=1}^{n} y_i(x)\right)}{\sigma^2(K)\left(n\sum_{i=1}^{n} h_i^4 - \left(\sum_{i=1}^{n} h_i^2\right)^2\right)/2}, \tag{5}$$

where $y_i(x)$ represents the i-th component in $\boldsymbol{y}(x)$. Then we use $\hat{R}(f'') = R(\hat{f}''(x)) = \int(\hat{f}''(x))^2 dx$ to estimate $R(f'')$. The integral $\hat{R}(f'')$ can be calculated analytically since a closed-form expression for $\hat{f}''(x)$ is known.

The following theorem gives the convergence property of $\hat{R}(f'')$. It shows that the estimate $\hat{R}(f'')$ is consistent.

Theorem 1 Suppose $f(\cdot)$ has bounded, integrable and continuous derivatives up to order 4. If $h_i,\ i = 1,\ldots,n$ are i.i.d. samples of h with $n^{-2}h^{-5} \to 0$ *a.s.* and $nh^9 \to 0$ *a.s.*, we have

$$\lim_{n\to+\infty} \mathbb{E}\left[(\hat{R}(f'') - R(f''))^2\right] = 0.$$

Proof. Since random variables $h_i,\ i = 1,\ldots,n$, are i.i.d. generated, the following property is given by the Central Limit Theorem: for any $x_1, x_2 \in \mathbb{R},\ x_1 \neq x_2$,

$$\sqrt{n}\left(\begin{bmatrix} (1/n)\sum_{i=1}^n r_i(x_1) \\ (1/n)\sum_{i=1}^n h_i^2 r_i(x_1) \\ (1/n)\sum_{i=1}^n r_i(x_2) \\ (1/n)\sum_{i=1}^n h_i^2 r_i(x_2) \\ (1/n)\sum_{i=1}^n h_i^2 \\ (1/n)\sum_{i=1}^n h_i^4 \end{bmatrix} - \begin{bmatrix} C(x_1)\mathbb{E}[h^4] \\ C(x_1)\mathbb{E}[h^6] \\ C(x_2)\mathbb{E}[h^4] \\ C(x_2)\mathbb{E}[h^6] \\ \mathbb{E}[h^2] \\ \mathbb{E}[h^4] \end{bmatrix}\right) \xrightarrow[n\to+\infty]{d} N\left(\mathbf{0}, \boldsymbol{\Sigma}^2\right),$$

where

$$\boldsymbol{\Sigma}^2 = \begin{bmatrix} \mathbb{E}[(nh)^{-1}]f(x_1)R(K) & \mathbb{E}[n^{-1}h]f(x_1)R(K) & o((nh)^{-1}) & o(n^{-1}h) & C(x_1)\mathrm{Cov}(h^4,h^2) & C(x_1)\mathrm{Cov}(h^4,h^4) \\ \mathbb{E}[n^{-1}h]f(x_1)R(K) & \mathbb{E}[n^{-1}h^3]f(x_1)R(K) & o(n^{-1}h) & o(n^{-1}h^3) & C(x_1)\mathrm{Cov}(h^6,h^2) & C(x_1)\mathrm{Cov}(h^6,h^4) \\ o((nh)^{-1}) & o(n^{-1}h) & \mathbb{E}[(nh)^{-1}]f(x_2)R(K) & \mathbb{E}[n^{-1}h]f(x_2)R(K) & C(x_2)\mathrm{Cov}(h^4,h^2) & C(x_2)\mathrm{Cov}(h^4,h^4) \\ o(n^{-1}h) & o(n^{-1}h^3) & \mathbb{E}[n^{-1}h]f(x_2)R(K) & \mathbb{E}[n^{-1}h^3]f(x_2)R(K) & C(x_2)\mathrm{Cov}(h^6,h^2) & C(x_2)\mathrm{Cov}(h^6,h^4) \\ C(x_1)\mathrm{Cov}(h^2,h^4) & C(x_1)\mathrm{Cov}(h^2,h^6) & C(x_2)\mathrm{Cov}(h^2,h^4) & C(x_2)\mathrm{Cov}(h^2,h^6) & \mathrm{Cov}(h^2,h^2) & \mathrm{Cov}(h^2,h^4) \\ C(x_1)\mathrm{Cov}(h^4,h^4) & C(x_1)\mathrm{Cov}(h^4,h^6) & C(x_2)\mathrm{Cov}(h^4,h^4) & C(x_2)\mathrm{Cov}(h^4,h^6) & \mathrm{Cov}(h^4,h^2) & \mathrm{Cov}(h^4,h^4) \end{bmatrix}$$

and we use $n^{-2}h^{-1} \to 0$ *a.s.* and $nh^9 \to 0$ *a.s.* to argue the negligibility of higher-order terms.

Note that

$$\hat{\boldsymbol{\beta}}(x) - \boldsymbol{\beta}(x) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}(x) - \boldsymbol{\beta}(x) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\beta}(x) + \boldsymbol{r}(x)) - \boldsymbol{\beta}(x) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{r}(x),$$

then we have

$$\hat{f}''(x) - f''(x) = \frac{\sum_{i=1}^n h_i^2 r_i(x)/n - \left(\sum_{i=1}^n h_i^2/n\right)\left(\sum_{i=1}^n r_i(x)/n\right)}{\sigma^2(K)\left(\sum_{i=1}^n h_i^4/n - \left(\sum_{i=1}^n h_i^2/n\right)^2\right)/2}.$$

With the multivariate delta method, we obtain

$$\sqrt{n}\left(\begin{bmatrix} \hat{f}''(x_1) - f''(x_1) \\ \hat{f}''(x_2) - f''(x_2) \end{bmatrix} - \begin{bmatrix} \frac{C(x_1)(\mathbb{E}[h^6] - \mathbb{E}[h^2]\mathbb{E}[h^4])}{\sigma^2(K)(\mathbb{E}[h^4] - (\mathbb{E}[h^2])^2)/2} \\ \frac{C(x_2)(\mathbb{E}[h^6] - \mathbb{E}[h^2]\mathbb{E}[h^4])}{\sigma^2(K)(\mathbb{E}[h^4] - (\mathbb{E}[h^2])^2)/2} \end{bmatrix}\right) \xrightarrow[n\to+\infty]{d}$$

$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{((\mathbb{E}[h^2])^2\mathbb{E}[h^{-1}] - 2\mathbb{E}[h]\mathbb{E}[h^2] + \mathbb{E}[h^3])f(x_1)R(K)}{n(\sigma^2(K)/2)^2(\mathbb{E}[h^4] - (\mathbb{E}[h^2])^2)^2} & o(n^{-1}h^{-5}) \\ o(n^{-1}h^{-5}) & \frac{((\mathbb{E}[h^2])^2\mathbb{E}[h^{-1}] - 2\mathbb{E}[h]\mathbb{E}[h^2] + \mathbb{E}[h^3])f(x_2)R(K)}{n(\sigma^2(K)/2)^2(\mathbb{E}[h^4] - (\mathbb{E}[h^2])^2)^2} \end{bmatrix}\right),$$

where we use $n^{-2}h^{-5} \to 0$ *a.s.* and $nh^9 \to 0$ *a.s.* to argue the negligibility of higher-order terms.

Note that we have

$$\hat{R}(f'') - R(f'') = \int(\hat{f}''(x))^2 - (f''(x))^2 dx = \int(\hat{f}''(x) - f''(x))^2 + 2f''(x)(\hat{f}''(x) - f''(x))dx.$$

With the multivariate delta method, we obtain

$$\sqrt{n}\left([\hat{R}(f'')-R(f'')]-\left[\frac{(\mathbb{E}[h^6]-\mathbb{E}[h^2]\mathbb{E}[h^4])^2\int(C(x))^2dx}{(\sigma^2(K)(\mathbb{E}[h^4]-(\mathbb{E}[h^2])^2)/2)^2}+\frac{(\mathbb{E}[h^6]-\mathbb{E}[h^2]\mathbb{E}[h^4])\int 2f''(x)C(x)dx}{\sigma^2(K)(\mathbb{E}[h^4]-(\mathbb{E}[h^2])^2)/2}\right]\right)$$
$$\xrightarrow[n\to+\infty]{d} N\left(0,\int\int 4f''(x_1)f''(x_2)\mathrm{Cov}\left(\hat{f}''(x_1)-f''(x_1),\hat{f}''(x_2)-f''(x_2)\right)dx_1dx_2\right).$$

Specifically, the bias of $\hat{R}(f'')$ is given by

$$\begin{aligned}
&\mathbb{E}\left[\hat{R}(f'')-R(f'')\right]\\
=\;& O\left(\frac{(\mathbb{E}[h^6]-\mathbb{E}[h^2]\mathbb{E}[h^4])^2\int(C(x))^2dx}{(\sigma^2(K)(\mathbb{E}[h^4]-(\mathbb{E}[h^2])^2)/2)^2}+\frac{(\mathbb{E}[h^6]-\mathbb{E}[h^2]\mathbb{E}[h^4])\int 2f''(x)C(x)dx}{\sigma^2(K)(\mathbb{E}[h^4]-(\mathbb{E}[h^2])^2)/2}\right)+O(n^{-1/2-\xi})\\
=\;& O(h^2)+O(n^{-1/2-\xi}),
\end{aligned}$$

where $\xi>0$ depicts the order of gap between convergence in distribution and moment convergence, then $\lim_{n\to+\infty}\mathbb{E}\left[\hat{R}(f'')-R(f'')\right]=0$ due to $h\to 0$ *a.s.*; the variance of $\hat{R}(f'')$ is given by

$$\begin{aligned}
\mathrm{Var}\left[\hat{R}(f'')-R(f'')\right] &= O\left(\frac{1}{n}\int\int 4f''(x_1)f''(x_2)\mathrm{Cov}\left(\hat{f}''(x_1)-f''(x_1),\hat{f}''(x_2)-f''(x_2)\right)dx_1dx_2\right)\\
&= O(n^{-2}h^{-5}),
\end{aligned}$$

then $\lim_{n\to+\infty}\mathrm{Var}\left[\hat{R}(f'')-R(f'')\right]=0$ due to $n^{-2}h^{-5}\to 0$ *a.s.*. Therefore, $\lim_{n\to+\infty}\mathbb{E}\left[(\hat{R}(f'')-R(f''))^2\right]=0$. □

We note that the proposed method utilizes the zeroth-order information of kernel function as shown in (1) and (5). In contrast to derivatives of kernel function used in Park and Marron (1990) and Sheather and Jones (1991) or Hermite series used in Tenreiro (2020), kernel function itself is much easier to compute. Therefore, our proposed method can ease the inherent computational burden of plug-in methods.

4 CONVERGENCE RATE OF BANDWIDTH ESTIMATE

We plug $\hat{R}(f'')$ into (4) to obtain a nearly optimal bandwidth and then estimate the density $f(\cdot)$ by (1). The following theorem gives the relative convergence rate between h_{BP} and h_{opt}.

Theorem 2 Suppose $f(\cdot)$ has bounded, integrable and continuous derivatives up to order 4. If h_i, $i=1,\ldots,n$ are i.i.d. samples of h with $h=O(n^{-2/9})$, we have

$$\frac{h_{\mathrm{BP}}-h_{\mathrm{opt}}}{h_{\mathrm{opt}}}=O(n^{-4/9}).$$

Proof. Note that

$$\frac{h_{\mathrm{BP}}-h_{\mathrm{opt}}}{h_{\mathrm{opt}}}=\frac{(\hat{R}(f''))^{-1/5}-(R(f''))^{-1/5}}{(R(f''))^{-1/5}}.$$

With the multivariate delta method, we obtain

$$\sqrt{n}\left((\hat{R}(f''))^{-1/5}-\left[R(f'')+\frac{(\mathbb{E}[h^6]-\mathbb{E}[h^2]\mathbb{E}[h^4])^2\int(C(x))^2dx}{(\sigma^2(K)(\mathbb{E}[h^4]-(\mathbb{E}[h^2])^2)/2)^2}+\frac{(\mathbb{E}[h^6]-\mathbb{E}[h^2]\mathbb{E}[h^4])\int 2f''(x)C(x)dx}{\sigma^2(K)(\mathbb{E}[h^4]-(\mathbb{E}[h^2])^2)/2}\right]^{-1/5}\right)$$
$$\xrightarrow[n\to+\infty]{d} N\left(0,\frac{1}{25}(R(f''))^{-12/5}\int\int 4f''(x_1)f''(x_2)\mathrm{Cov}\left(\hat{f}''(x_1)-f''(x_1),\hat{f}''(x_2)-f''(x_2)\right)dx_1dx_2\right).$$

Specifically, the bias of $(\hat{R}(f''))^{-1/5}$ is given by

$$\begin{aligned}
& \mathbb{E}\left[(\hat{R}(f''))^{-1/5} - (R(f''))^{-1/5}\right] \\
= \; & O\left(\frac{(\mathbb{E}[h^6] - \mathbb{E}[h^2]\mathbb{E}[h^4])^2 \int (C(x))^2 dx}{(\sigma^2(K)(\mathbb{E}[h^4] - (\mathbb{E}[h^2])^2)/2)^2} + \frac{(\mathbb{E}[h^6] - \mathbb{E}[h^2]\mathbb{E}[h^4]) \int 2f''(x)C(x)dx}{\sigma^2(K)(\mathbb{E}[h^4] - (\mathbb{E}[h^2])^2)/2}\right) + O(n^{-1/2-\xi}) \\
= \; & O(h^2) + O(n^{-1/2-\xi}),
\end{aligned}$$

where $\xi > 0$ depicts the order of gap between convergence in distribution and moment convergence; the variance of $(\hat{R}(f''))^{-1/5}$ is given by

$$\begin{aligned}
\mathrm{Var}\left[\hat{R}(f'') - R(f'')\right] &= O\left(\frac{1}{n}\int\int 4f''(x_1)f''(x_2)\mathrm{Cov}\left(\hat{f}''(x_1) - f''(x_1), \hat{f}''(x_2) - f''(x_2)\right) dx_1 dx_2\right) \\
&= O(n^{-2}h^{-5}).
\end{aligned}$$

In order to obtain a fast rate of convergence of h_{BP}, we should balance the squared bias term and the variance term to the same order in terms of n. Therefore, h is set to be order $n^{-2/9}$, otherwise by perturbing the order of h either of the two terms would increase. When $h = O(n^{-2/9})$, we have $(h_{\mathrm{BP}} - h_{\mathrm{opt}})/h_{\mathrm{opt}} = O(n^{-4/9})$. □

Table 1 displays the relative rates of convergence of bandwidth selection in various existing plug-in methods. Park and Marron (1990) demonstrate that their bandwidth has a convergence rate of $n^{-4/13}$. Sheather and Jones (1991) prove that the relative rate of convergence of their bandwidth is of order $O(n^{-5/14})$, which is slightly better than that of Park & Marron's plug-in. Consider plug-in for (4), the order $n^{-2/5}$ is achieved for the rate of convergence by Tenreiro (2020). Theorem 2 establishes that the bandwidth h_{BP} given by KDE-BP has a faster convergence rate than other plug-in methods in existing literature. Note that Hall and Marron (1991) show the best possible rate of convergence is $n^{-1/2}$, and thus we call h_{BP} nearly optimal.

Table 1: Comparison on the relative rates of convergence among plug-in methods.

	Bandwidth Perturbation	Park & Marron	Sheather & Jones	Hermite Series
$(\hat{h} - h_{\mathrm{opt}})/h_{\mathrm{opt}}$	$O(n^{-4/9})$	$O(n^{-4/13})$	$O(n^{-5/14})$	$O(n^{-2/5})$

5 SIMULATION RESULTS

In this section, we conduct simulation experiments to test the finite sample performance of the proposed KDE-BP method. Consider the fact that any density can be approximated arbitrarily closely by a normal mixture. Hence, we employ four normal mixture densities in (Marron and Wand 1992) as our experimental examples:

1. Skewed Unimodal Density: $\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2}, (\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{12}, (\frac{5}{9})^2)$.
2. Bimodal Density: $\frac{1}{2}N(-1, (\frac{2}{3})^2) + \frac{1}{2}N(1, (\frac{2}{3})^2)$.
3. Asymmetric Bimodal Density: $\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2}, (\frac{1}{3})^2)$.
4. Asymmetric Claw Density: $\frac{1}{2}N(0,1) + \sum_{\ell=-2}^{\ell=2}(2^{1-\ell}/31)N(\ell + \frac{1}{2}, (2^{-\ell}/10)^2)$.

The above density functions are visualized in Figure 1.

The proposed KDE-BP method is compared with three other plug-in methods: Park & Marron's plug-in, Sheather & Jones' plug-in, and Hermite series-based plug-in. In particular,

- Park & Marron's plug-in (PM): consider $\hat{f}''(x) = (1/(ng^3))\sum_{i=1}^n K''((x - X_i)/g)$ and $\hat{R}(f'') = R(\hat{f}''(x)) - (1/(ng^5))R(K'')$. The bandwidth $(g_{\mathrm{PM}}, h_{\mathrm{PM}})$ is yielded by numerically solving $g =$

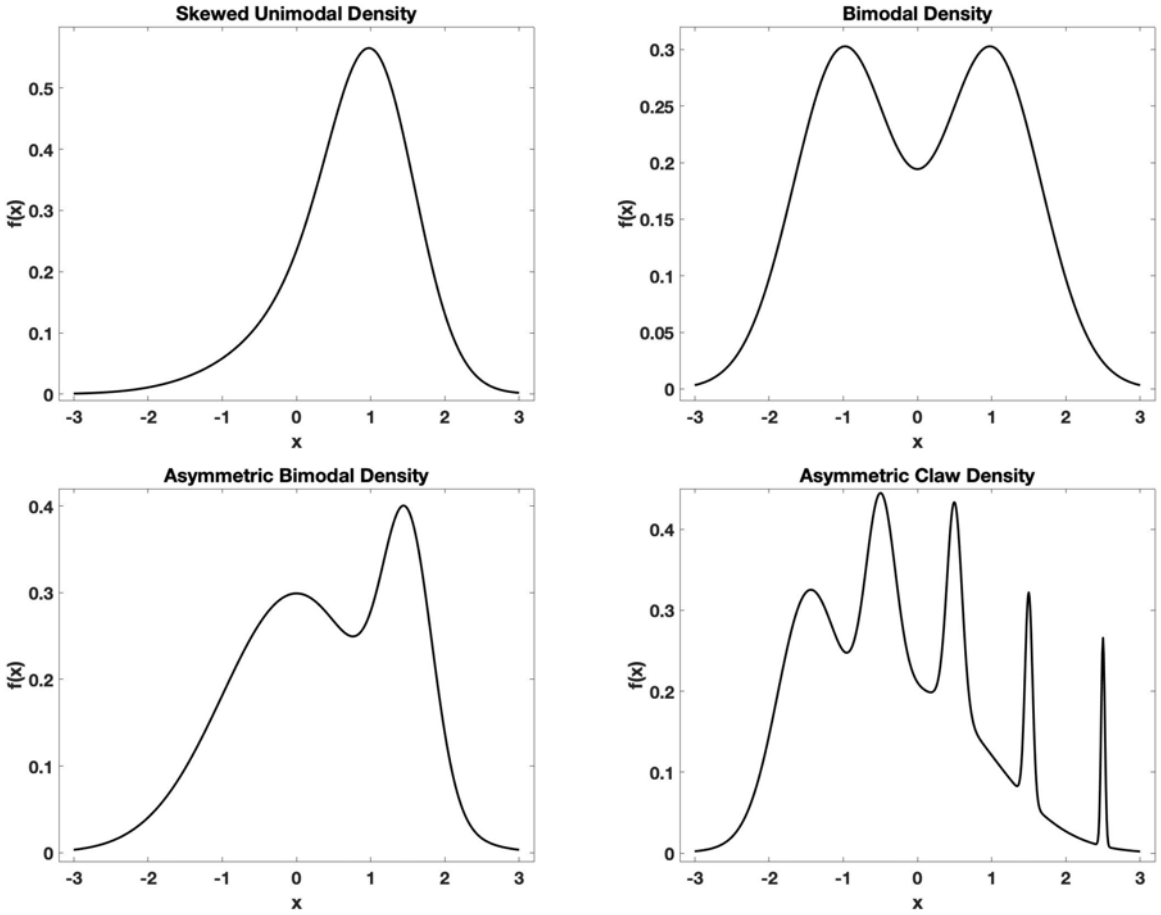

Figure 1: Normal mixture densities for simulation experiments.

$C_1(K)C_2(f)h^{10/13}$ and $h = n^{-1/5}(R(K)/((\sigma^2(K))^2\hat{R}(f'')))^{1/5}$, where $C_1(K)$ is known and $C_2(f)$ is estimated using Silverman's rule-of-thumb. Then it estimates $f(x)$ by $\hat{f}_{\mathrm{PM}}(x) = (1/(nh_{\mathrm{PM}}))\sum_{i=1}^{n} K((x - X_i)/h_{\mathrm{PM}})$.

- Sheather & Jones' plug-in (SJ): consider $\hat{f}''(x) = (1/(ng^3))\sum_{i=1}^{n} L''((x - X_i)/g)$ and $\hat{R}(f'') = R(\hat{f}''(x))$. The bandwidth $(g_{\mathrm{SJ}}, h_{\mathrm{SJ}})$ is yielded by numerically solving $g = C_3(K,L)C_4(f)h^{5/7}$ and $h = n^{-1/5}(R(K)/((\sigma^2(K))^2\hat{R}(f'')))^{1/5}$, where $C_3(K,L)$ is known and $C_4(f)$ is estimated using Silverman's rule-of-thumb. Then it estimates $f(x)$ by $\hat{f}_{\mathrm{SJ}}(x) = (1/(nh_{\mathrm{SJ}}))\sum_{i=1}^{n} K((x - X_i)/h_{\mathrm{SJ}})$.
- Hermite series-based plug-in (HS): consider $\hat{R}(f'') = \sum_{k=0}^{m} \hat{a}_k^2$, where $m = m(n)$ is a sequence of integers converging to infinity with n, $\hat{a}_k = (1/n)\sum_{i=1}^{n} h_k''(X_i)$ is an estimate of the k-th Hermite coefficient of $f''(x)$, $h_k(x) = (2^k k!\pi^{1/2})^{-1/2}(-1)^k e^{x^2}(d^k/dx^k)e^{-3x^2/2}$ is the Hermite orthonormal basis of L_2. Then it uses the bandwidth $h_{\mathrm{HS}} = n^{-1/5}(R(K)/((\sigma^2(K))^2\hat{R}(f'')))^{1/5}$ and estimates $f(x)$ by $\hat{f}_{\mathrm{HS}}(x) = (1/(nh_{\mathrm{HS}}))\sum_{i=1}^{n} K((x - X_i)/h_{\mathrm{HS}})$.

In all simulation experiments, the performance of each tested plug-in method is measured by the MISE. Empirical MISE is estimated by 1,000 independent experimental replications. The sample size n is set as $n = 25 \times 2^k$, $k = 0, \ldots, 7$, and Gaussian kernel is used. The MISE is reported as a function of $k = \log_2(n/25)$ in each experiment.

The behaviour of our proposed KDE-BP method and three other plug-in methods is presented in Figures 2, 3, 4 and 5. In each figure, we can see that our proposed KDE-BP method is the most efficient

plug-in method among the four, whereas Park & Marron's plug-in is the worst. Hermite series-based plug-in performs slightly better than Sheather & Jones' plug-in for skewed unimodal density and bimodal density, while both methods perform similarly for asymmetric bimodal density and asymmetric claw density. In other words, KDE-BP is more robust against different shapes of density. Moreover, as the sample size increases, the MISE of KDE-BP decreases and gets close to zero. Such observation is in accord with the consistency of KDE-BP in Theorem 1.

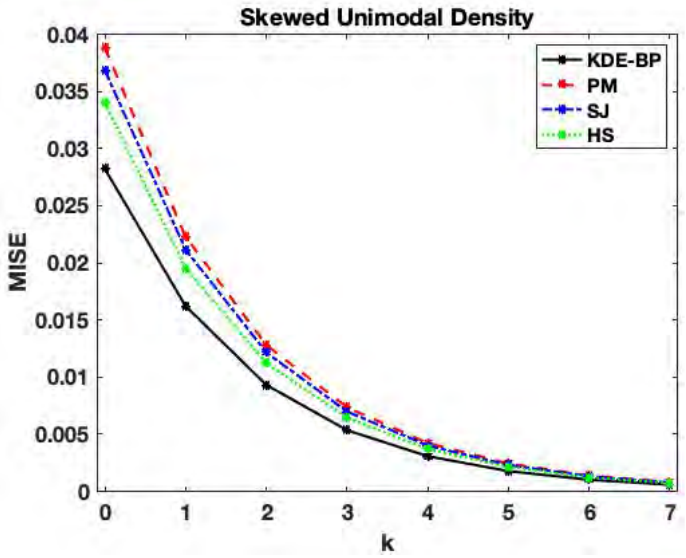

Figure 2: Empirical MISE of all tested plug-in methods for skewed unimodal density.

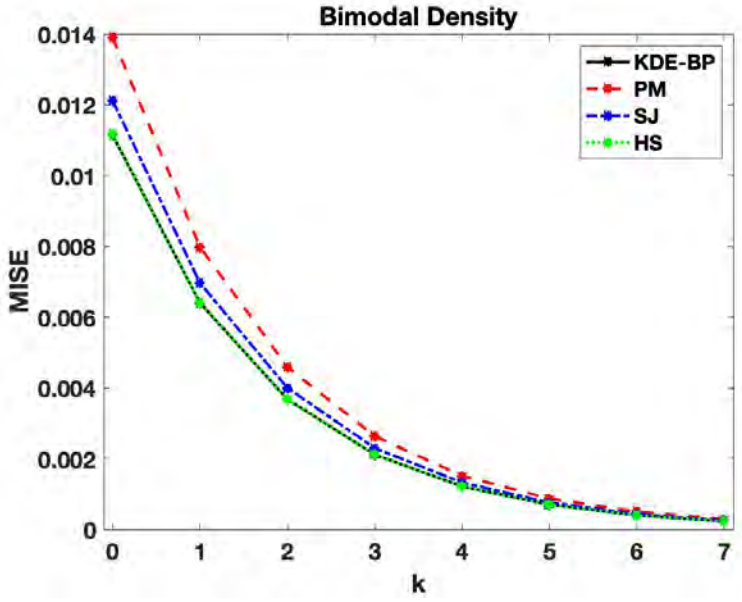

Figure 3: Empirical MISE of all tested plug-in methods for bimodal density.

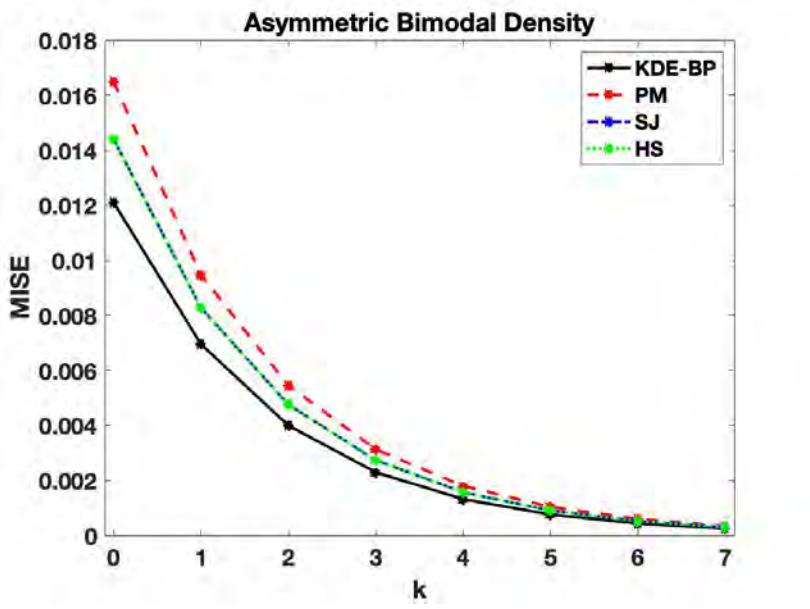

Figure 4: Empirical MISE of all tested plug-in methods for asymmetric bimodal density.

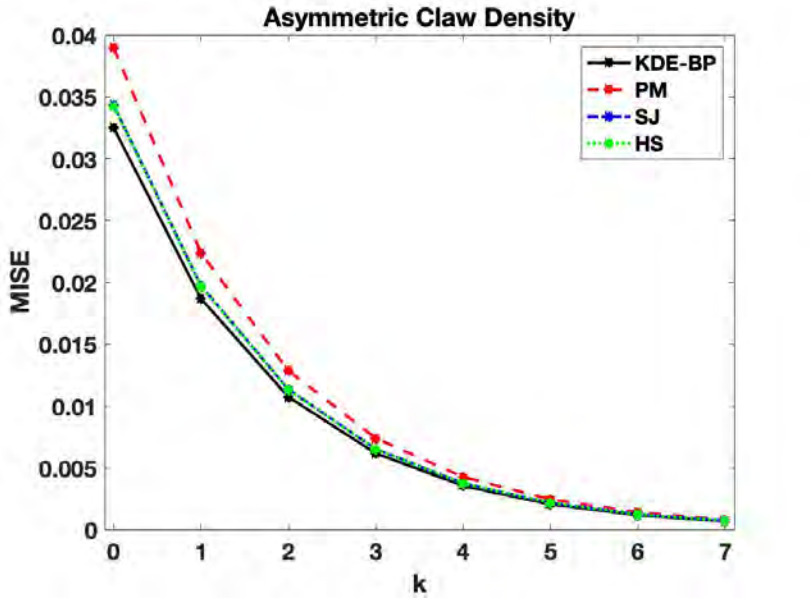

Figure 5: Empirical MISE of all tested plug-in methods for asymmetric claw density.

6 CONCLUSION

This paper studies bandwidth selection for kernel density estimation. We propose an efficient plug-in kernel density estimator named KDE-BP, which utilizes the zeroth-order information of kernel function and has a nearly optimal convergence rate. Simulation experiments demonstrate that KDE-BP has a good finite sample performance and is robust to different density functions. In future work, we will explore the estimator with the best possible convergence rate, investigate multi-dimensional generalizations, consider higher-order kernels, and conduct more extensive simulations.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation of China (NSFC) under Grants 72201006, 72022001, 92146003, 71901003.

REFERENCES

Bowman, A. W. 1984. "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates". *Biometrika* 71(2):353–360.

Cho, J., G. Hwang, and C. Suh. 2020. "A Fair Classifier Using Kernel Density Estimation". In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, 15088–15099. Cambridge, Massachusetts: The Massachusetts Institute of Technology Press.

Hall, P., and J. S. Marron. 1991. "Lower Bounds for Bandwidth Selection in Density Estimation". *Probability Theory and Related Fields* 90(2):149–173.

Heidenreich, N.-B., A. Schindler, and S. Sperlich. 2013. "Bandwidth Selection for Kernel Density Estimation: a Review of Fully Automatic Selectors". *AStA Advances in Statistical Analysis* 97:403–433.

Jones, M. C., J. S. Marron, and S. J. Sheather. 1996. "A Brief Survey of Bandwidth Selection for Density Estimation". *Journal of the American Statistical Association* 91(433):401–407.

Marron, J. S., and M. P. Wand. 1992. "Exact Mean Integrated Squared Error". *The Annals of Statistics* 20(2):712–736.

Nadaraya, E. 1974. "On the Integral Mean Square Error of Some Nonparametric Estimates for the Density Function". *Theory of Probability & Its Applications* 19(1):133–141.

Okabe, A., T. Satoh, and K. Sugihara. 2009. "A Kernel Density Estimation Method for Networks, Its Computational Method and a GIS-Based Tool". *International Journal of Geographical Information Science* 23(1):7–32.

Park, B. U., and J. S. Marron. 1990. "Comparison of Data-Driven Bandwidth Selectors". *Journal of the American Statistical Association* 85(409):66–72.

Rudemo, M. 1982. "Empirical Choice of Histograms and Kernel Density Estimators". *Scandinavian Journal of Statistics* 9:65–78.

Sheather, S. J., and M. C. Jones. 1991. "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation". *Journal of the Royal Statistical Society: Series B (Methodological)* 53(3):683–690.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Simonoff, J. S. 2012. *Smoothing Methods in Statistics*. New York: Springer.

Steckley, S. G., and S. G. Henderson. 2003. "A Kernel Approach to Estimating the Density of a Conditional Expectation". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 383–391. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Tenreiro, C. 2020. "Bandwidth Selection for Kernel Density Estimation: a Hermite Series-Based Direct Plug-In Approach". *Journal of Statistical Computation and Simulation* 90(18):3433–3453.

Tsybakov, A. B. 2009. *Introduction to Nonparametric Estimation*. London: Springer.

Wand, M. P., and M. C. Jones. 1994. *Kernel Smoothing*. New York: Chapman & Hall.

Woodroofe, M. 1970. "On Choosing a Delta-Sequence". *The Annals of Mathematical Statistics* 41(5):1665–1671.

Zambom, A. Z., and D. Ronaldo. 2013. "A Review of Kernel Density Estimation with Applications to Econometrics". *International Econometric Review* 5(1):20–42.

AUTHOR BIOGRAPHIES

HAIDONG LI is an Assistant Professor in the Department of Management Science at University of Chinese Academy of Sciences, Beijing, China. He received his B.S. Degree from the Department of Engineering Mechanics at Peking University, and his Ph.D. Degree from the Department of Industrial Engineering and Management at Peking University. His research interests include simulation optimization, network analysis, and stochastic gradient estimation. His email address is haidong.li@pku.edu.cn.

LONG WANG is a Professor in the Department of Advanced Manufacturing and Robotics at Peking University, Beijing, China. He received his Bachelor, Master, and Doctor's degrees in Dynamics and Control from Tsinghua University and Peking University in 1986, 1989, and 1992, respectively. His research interests are in the fields of networked systems, hybrid systems, swarm dynamics, cognitive science, collective intelligence, and bio-mimetic robotics. His email address is longwang@pku.edu.cn.

YIJIE PENG is an Associate Professor in the Department of Management Science and Information Systems in Guanghua School of Management at Peking University, Beijing, China. He received the B.S. degree in mathematics from Wuhan University, Wuhan, China, in 2007, and the Ph.D. degree in management science from Fudan University, Shanghai, China, in 2014, respectively. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data

analytics, and healthcare. His email address is pengyijie@pku.edu.cn.

DI WANG is an Assistant Professor with the Department of Industrial Engineering and Management, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China. She received the B.S. degree in industrial engineering from Nankai University, Tianjin, China, in 2015, and the Ph.D. degree in management science and engineering from Peking University, Beijing, China, in 2020. Her research interests include statistical modeling and artificial intelligence of process modeling, monitoring, and prognostics. Her email address is d.wang@sjtu.edu.cn.