ISSN: 2307-7743
www.scienceasia.asia

Asian J. Math. Appl. (2023) 2023:1

A Short Note on Comparing Bayesian Density Estimation with Univariate Kernel Density Estimation

Rui Gong[1], John Chan[2], Jinwei Liu[3]

[1]Department of Informatics and Mathematics, Mercer University, GA, USA
[2]School of Mathematical and Statistical Sciences, Clemson University, SC, USA
[3]Department of Computer and Information Sciences, Florida A&M University, FL, USA

Correspondence should be addressed to Rui Gong: gong_r@mercer.edu

Abstract
Besides the frequentist methods, Bayesian approaches are applied to solve the non-parametric problems. This work focuses on comparing the univariate Kernel density estimation(KDE) with Bayesian density estimation using Dirichlet process mixtures(DPM) of Gaussians, which are non-parametric ways to estimate the probability density function. The results of simulation show Bayesian density estimations are better than Kernel density estimations in most cases.

Keywords: Bayesian Density Estimation, Univariate Kernel Density Estimation, Dirichlet Process Mixtures.

Introduction

The kernel estimators are widely used in many nonparametric models. David and Stephan described the application of diverse nonparametric kernel methods to reveal hidden structure in data [1]. Those kernel approaches were utilized in the nonparametric regression for exploring data [2] and the graphical analysis of data [3, 4]. Among nonparametric kernel estimation in many forms, Wolfgang et al. showed the univariate Kernel density estimation(KDE) was one of the commonest nonparametric and semiparametric modeling techniques to undercover both the statistical characteristics and the probability density function of a random variable [5], though computational complexity of calculations, especially for data-based bandwidth selection and adaptation of bandwidth coefficient, makes this way inefficient and possibly inaccurate [6]. There have been "first generation" methods and "second generation" methods in the past two decades to select bandwidth for KDE [7], but the optimal bandwidth selection method may vary for the different data set [8, 9]. Because of the difficulty in making one general optimal bandwidth selection method, many other statistical ideas were developed to estimate density.

Nonparametric Bayesian methods are increasingly popular for density estimation due to the practicality and validity. Escobar and West used Dirichlet processes mixture(DPM) models in Bayesian method for density estimation [10]. Steven and Peter expanded the scope of DPM models to nonconjugate situations [11]. Furthermore, they proposed the computational strategies for normal-normal DPM models. Diverse Packages in R were created to implement Bayesian nonparametric models and contain functions for model comparison [12–14], but there is one gap to compare the KDE with Bayesian density estimation. In this research work,

Received: 2022-11-24

we focus on comparing the accuracy in density estimation between the univariate KDE and Bayeisan density estimation using DPM of Gaussians.

Section 2 provides the details of KDE and Bayeisan DPM of Gaussians. In Section 3, two density estimation methods are applied to the data simulated from Cauchy distribution, F distribution and Mixed Normal distribution with the different sample sizes and the results are discusses. In Section 4, there is the conclusion.

Methodology

Univariate Kernel Density Estimation

Kernels are used in KDE to estimate random variables density function $f(x)$ [15, 16], The general kernel estimator for the real density function $f(x)$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-X_i}{h}). \tag{1}$$

where $K(x)$ denotes one kernel function, h is the bandwidth and X_i is the ith sample.

In order to overcome the convergence issue, all kernel must satisfy three conditions. The first condition is the integral of one kernel on the domain is equal to one and all codomains are nonnegative $K(x) \geq 0$ & $\int_R K(x)dx = 1$; the second condition is the integral of the multiplication of the kernel and each element in domain is zero $\int_R xK(x)dx = 0$; the third one is the integral of the multiplication of the square of each element in domain and kernel is finite $\int_R x^2K(x)dx < \infty$.

Based on the kernel estimation function (1) the derivative function of kernel density estimator is derived directly for gradient estimation and it is shown as follows:

$$\hat{f}_h^{(r)}(x) = \frac{\mathrm{d}^r}{\mathrm{d}x^r}\frac{1}{nh}\sum_{i=1}^{n}K(\frac{x-X_i}{h}) = \frac{1}{nh^{r+1}}\sum_{i=1}^{n}K^{(r)}(\frac{x-X_i}{h}). \tag{2}$$

As seen in the equations (1) and (2), when working with a kernel estimator of the density function or its rth derivative function two variables are determined first: the kernel function K and the smoothing parameter or bandwidth h. In practice, the choice of K is a problem of less importance because K is not sensitive to the shape of estimator; the choice of an efficient computation h method, is a crucial problem for an observed data sample because of the effect of the bandwidth on the shape of the corresponding density estimator. When the chose bandwidth is relatively smaller than the optional one, we will obtain an under-smoothed estimator with high variability. On the contrary, when the value is significantly bigger than the optimal one, the resulting estimator will be over-smoothed and further apart from the real density function. There are diverse methods to find the optimal h for different kernels [17–19]. In this paper, we use the Gaussian distribution as the kernel and take one number which can minimize the Mean Integrated Squared Error (MISE) to be as the optimal h for every distribution because most literature suggested this method when the difference between the real density and the density estimator is known.

Bayesian Density Estimation Using DPM of Normals

A mixture model method for estimating a density is as follows:

$$f(x) = \sum_{j=1}^{n} w_j f(x;\theta_j). \tag{3}$$

In general, (3) can be extended to the infinite mixture as follows:

$$f(x) = \sum_{j=1}^{\infty} w_j f(x;\theta_j). \tag{4}$$

The priors should be determined for the parameters $\theta_1, \cdots, \theta_n$ and $w_1, \cdots, w_n$ respectively in (3) & (4) when Bayesian method is utilized. In our work, $\theta_1, \cdots, \theta_n$ are drawn from one normal-inverted Wishart distribution G_0 and $w_1, \cdots, w_n$ are drawn from the stick breaking prior. DPM is also denoted as the random distribution $DP(\alpha, G_0)$; the probability function $F = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$, where δ_{θ_j} is the point mass distributions. So, when n sample points are taken the model becomes as follows:

$$G \sim DP(\alpha, G_0)$$
$$\theta_1, \cdots, \theta_n | G \sim G$$
$$X_i | \theta_i \sim f(x|\theta_i), \qquad i = 1, \cdots, n$$

where α is the hyperparameter and X_i is the ith observed data point, which is the ith sample point from the real distribution in the next section. The normal distribution N is added to DMP [10] and the updated model is as follows:

$$G \mid \alpha, G_0 \sim DP(\alpha, G_0)$$
$$(\mu_i, \sum_i) \mid G \sim G$$
$$y_i \mid \mu_i, \sum_i \sim N(\mu_i, \sum_i), \qquad i = 1, \cdots, p$$

where μ_i and $\sum_i$ are the mean and variance of normal distribution, y_i is the estimated point by Bayesian method, the conjugate normal-inverted-Wishart G_0 is the baseline distribution and p is the dimension of G_0 as follows:

$$G_0 = N(\boldsymbol{\mu} \mid \boldsymbol{m_1}, (1/k_0) \sum) IW(\sum \mid v_1, \boldsymbol{\psi_1}),$$

where $\boldsymbol{\mu}$ is one $p \times 1$ vector, $\sum$ is one $p \times p$ matrix, m_1 is one hyperparemeter, v_1 is one real number which must be larger than $p-1$ based on the assumption of Inverse-Wishart distribution (IW), $\boldsymbol{\psi_1}$ is the $p \times p$ sample matrix taken from (IW) and $\boldsymbol{m_1} = m_1 \times [1, \cdots, 1]^T$ with $p \times 1$ dimension.

If $\boldsymbol{\mu} \sim N(\boldsymbol{m_1}, (1/k_0) \sum)$, then their posterior distribution is proportional as follows:

$$\begin{aligned} &\pi(\boldsymbol{\mu}, \boldsymbol{m_1}, (1/k_0) \sum) \\ &\propto \det((1/k_0) \sum)^{-p/2} \exp\{-(1/k_0) \sum_i^p (\mu_i - m_1)^T \sum^{-1} (\mu_i - m_1)/2\} \\ &\propto \det((1/k_0) \sum)^{-n/2} \exp\{-(1/k_0) tr(\boldsymbol{\psi_{m_1}^{-1}} \sum^{-1}/2)\}, \end{aligned} \tag{5}$$

with $\boldsymbol{\psi_{m_1}} = (\mu_i - m_1)(\mu_i - m_1)^T$ and tr represents trace.

If $\sum \sim IW(v_1, \boldsymbol{\psi_1})$, then its posterior distribution is proportional as follows:

$$\pi(\sum) \propto \det(\sum)^{-(v_1+p+1)/2} \exp\{-tr(\boldsymbol{\psi_1^{-1}} \sum^{-1})/2\}. \tag{6}$$

Thus,

$$\begin{aligned} &\pi(\mu, \sum \mid X, m_1, (1/k_0), \psi_1) \\ &\propto \pi(\sum) \pi(\mu \mid m_1, (1/k_0), \psi_1) \pi(X \mid \mu, \sum) \\ &\propto \det(\sum)^{-(v_1+p+1)/2} \exp\{-tr(\boldsymbol{\psi_1^{-1}} \sum^{-1})/2\} \\ &\times \det((1/k_0) \sum)^{-p/2} \exp\{-\sum_i^p (\mu_i - m_1)^T \sum^{-1} (\mu_i - m_1)/2\} \\ &\times \det(\sum)^{-n/2} \exp\{-\sum_i^p (X_i - \mu_i)^T \sum^{-1} (X_i - \mu_i)/2\} \\ &\propto \det(\sum)^{-(n+p+v_1+p+1)/2} \exp\{-tr((\boldsymbol{\psi_{m_1}^{-1}} + \boldsymbol{\psi_\mu^{-1}} + \boldsymbol{\psi_1^{-1}}) \sum^{-1}/2)\}, \end{aligned} \tag{7}$$

with $\boldsymbol{\psi_\mu} = (X_i - \mu_i)(X_i - \mu_i)^T$.

To complete the model specification, independent hyperpriors are assumed as follows:

$$\alpha \mid a_0, b_0 \sim Gamma(a_0, b_0),$$

$$m_1 \mid m2, s_2 \sim N(m_2, s_2),$$

$$k_0 \mid \tau_1, \tau_2 \sim Gamma(\tau_1/2, \tau_2/2),$$

$$\boldsymbol{\psi_1} \mid v_2, \boldsymbol{\psi_2} \sim IW(v_2, \boldsymbol{\psi_2}),$$

where $a_0, b_0, m_2, s2, \tau_1, \tau_2, v_2$ are real numbers and $\boldsymbol{\psi_2}$ is one $p \times p$ matrix.

Simulation

In this section, we make the simulation of three distributions: one is Cauchy distribution with parameters $x_0 = -2$ and $\gamma = 1$, one is F distribution with parameters $d_1 = 5$ and $d_2 = 2$ and the last one is the combination of two normal distributions with μ_1=1.5, σ_1=0.8 and μ_2=4.0, σ_2=0.6 respectively. We draw the graphs in R to compare the results of density estimations between KDE and Bayesian approach using DPM of Gaussians. In addition, we change the sample size n to check how it influences the accuracy of estimations.

When Bayesian approach is applied, formula (5), (6) and (7) are used to take sample points for estimation. For convenience, we follow [12] and take $a_0 = 2, b_0 = 1, m_2 = 0, s_2 = 10000, v_1 = 5, v_2 = 4, \tau_1 = 4, \tau_2 = 2$ and $\boldsymbol{\psi_2} = 2$ with one dimension. Note that the choices of those numbers do not change the results of density estimation significantly when the convergence is guaranteed.

For the Cauchy distribution, Figure 1, 2 and 3 show how density estimations change when n changes from 1000 to 10000. By comparing two different estimating methods we can see that Bayesian density estimation is always better than KDE for different n. When n is 1000, KDE is most different from the true distribution. Though KDE becomes closer to the true distribution as n becomes larger, Bayesian density estimation is much better.

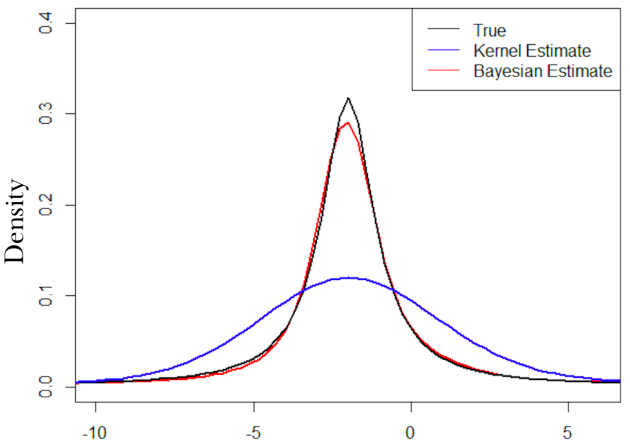

Figure 1: Kernel and Bayesian estimations for Cauchy distribution with $n = 1000$

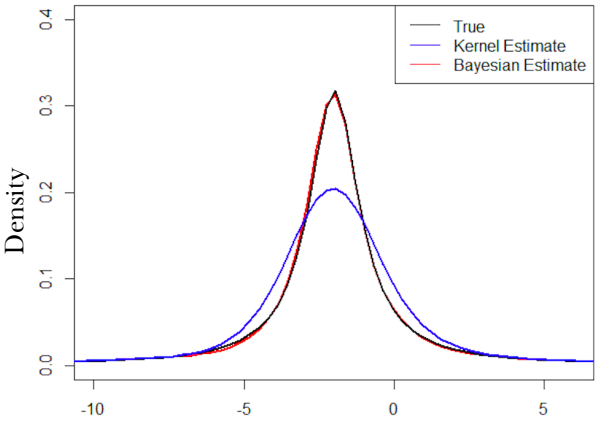

Figure 2: Kernel and Bayesian estimations for Cauchy distribution with $n = 5000$

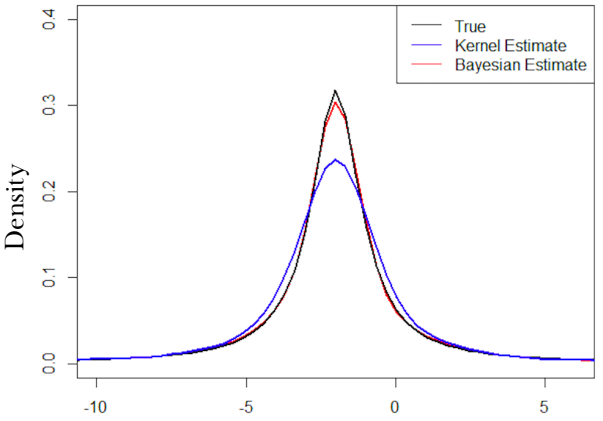

Figure 3: Kernel and Bayesian estimations for Cauchy distribution with $n = 10000$

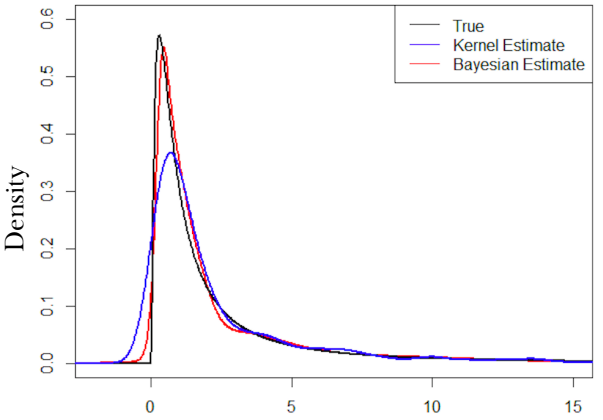

Figure 4: Kernel and Bayesian estimations for F distribution with $n = 1000$

For the F distribution, we have the results in Figure 4, 5 and 6 when sample sizes are 1000, 5000 and 10000 respectively. When n becomes larger Bayesian density estimation becomes better while KDE becomes worse, so for F distribution DPM makes the error smaller. In addition, when n become larger we obtain an under smoothed estimator by KDE and Bayesian estimation overcomes such issue.

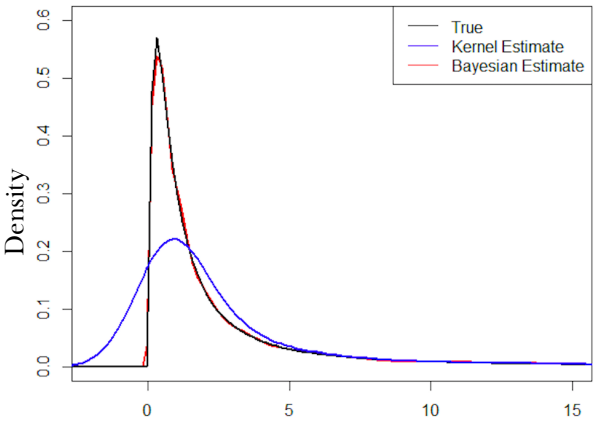

Figure 5: Kernel and Bayesian estimations for F distribution with $n = 5000$

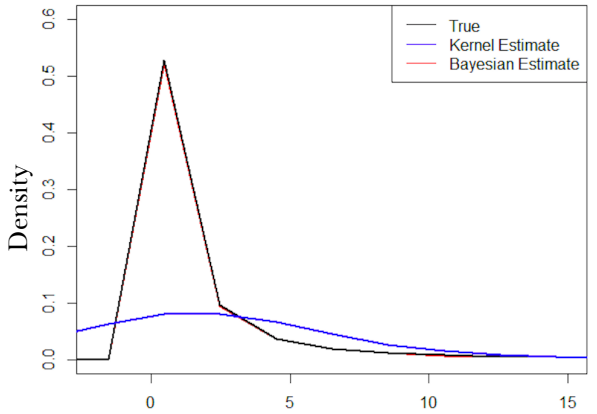

Figure 6: Kernel and Bayesian estimations for F distribution with $n = 10000$

For the Mixed Normal distribution, Figure 7, 8 and 9 show how two density estimations change as n changes. When n is not large enough ($n = 1000$), both estimations do not match the true distribution well: MISE between KDE and the true density is 0.185, and MISE between Bayesian density estimation and the true one is 0.178. However, when n becomes larger Bayesian density estimation is improved while KDE does not change much.

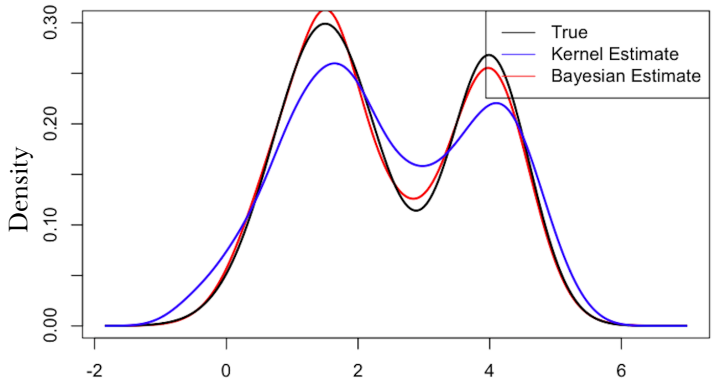

Figure 7: Kernel and Bayesian estimations for Mixed Normal distribution with $n = 1000$

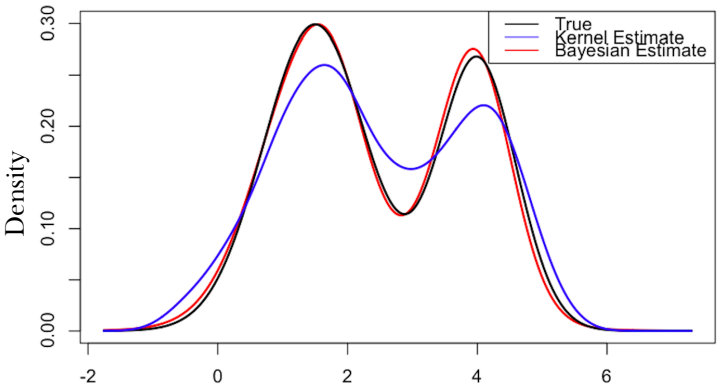

Figure 8: Kernel and Bayesian estimations for Mixed Normal distribution with $n = 5000$

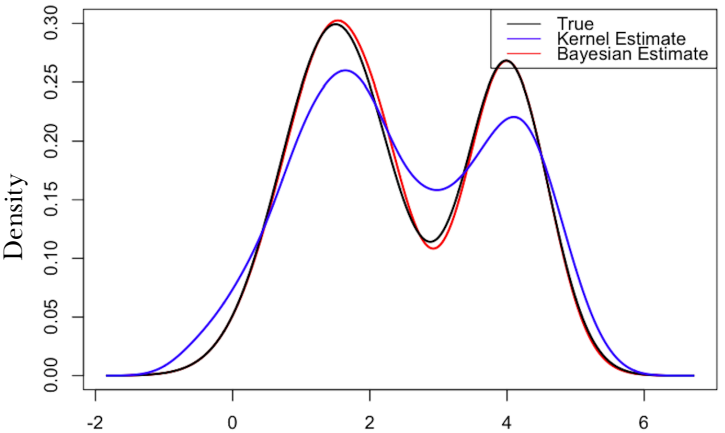

Figure 9: Kernel and Bayesian estimations for Mixed Normal distribution with $n = 10000$

Conclusion

We apply two widely used non-parametric methods for density estimation in this work, one mothod is the univariate KDE and the other one is Bayesian density estimation using DPM of Gaussians. Gaussian distribution is taken as the kernel and MISE is implemented to choose the optimal h for the univariate KDE. Compared with the univariate KDE, Bayesian density estimation is better most times for Cauchy distribution, F distribution and Mixed normal distribution in the simulation. Furthermore, when the number n becomes larger, Bayesian density estimation is closer to the true density function in the simulation, while Kernel density estimation becomes better, worse and keeps almost the same for Cauchy distribution, F distribution and Mixed normal distribution respectively.

References

[1] W. S. David, R. S. Stephan, Multidimensional density estimation, Handbook of Statistics 24 (2005), 229–261.

[2] W. B. Adrian, A. Adelchi, Applied smoothing techniques for data analysis : the kernel approach with S-plus illustrations, OUP Oxford, 1997.

[3] W. Venables, D. B. Ripley, Modern applied statistics with S, Springer, 2002.

[4] A. C. Guidoum, Kernel estimator and bandwidth selection for density and its derivatives: The kedd package (2020).

[5] H. Wolfgang, M. Marlene, S. Stefan, W. Axel, Nonparametric and Semiparametric Models, Springer Series in Statistics, Springer, 2004.

[6] L. Szymon, Parallel computing of kernel density estimates with mpi, International Conference on Computational Science (2007), 726–733.

[7] M. C. Jone, J. S. Marron, S. J. Scheather, A brief survey of bandwidth selection for density estimation, Journal of the American Statistical Association, 91 (433) (1996), 401–407.

[8] C. Su, Optimal bandwidth selection for kernel density functionals estimation, Journal of Probability and Statistics, 2015 (2015), 242683.

[9] E. H. Khadijetou, L. Djamal, Optimal bandwidth selection in kernel density estimation for continuous time dependent processes, Statistics & Probability Letters 138 (2018), 9–19.

[10] D. E. Michael, W. Mike, Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association 90 (1995), 577–588.

[11] N. M. Steven, M. Peter, Estimating mixture of dirichlet process models, Journal of Computational and Graphical Statistics 7 (2) (1998), 223–238.

[12] J. Alejandro, E. H. Timothy, A. Q. Fernando, M. Peter, L. R. Gary, Dppackage: Bayesian semi- and nonparametric modeling in r, Journal of statistical software 40 (5) (2011), 1–30.

[13] C. Riccardo, C. Antonio, N. Bernardo, Bnpmix: An r package for bayesian nonparametric modeling via pitman-yor mixtures, Journal of statistical software 100 (2021), 1–30.

[14] C. Jacinto, G. Salvador, d. M. R. María, H. Francisco, rnpbst: An r package covering non-parametric and bayesian statistical tests, Hybrid Artificial Intelligent Systems 10334 (2017), 281–292.

[15] M. P. Wand, M. C. Jones, Kernel Smoothing, Chapman & Hall, London, 1995.

[16] W. Bernard, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London, 1986.

[17] A. Björn, A. v. D. Alina, Improving the bandwidth selection in kernel equating, Journal of Educational Measurement 51 (3) (2014), 223–238.

[18] H. Jenny, W. Marie, Optimal bandwidth selection in observed-score kernelequating, Journal of Educational Measurement 51 (2) (2014), 201–211.

[19] C. Shean-Tsong, Bandwidth selection for kernel density estimation, The Annals of Statistics 19 (4) (1991), 1883–1905.