AIMS Mathematics, 9(7): 18103–18116.
DOI:10.3934/math.2024883
Received: 22 January 2024
Revised: 18 April 2024
Accepted: 22 April 2024
Published: 28 May 2024

https://www.aimspress.com/journal/Math

Research article

Nonparametric Bayesian modeling for non-normal data through a transformation

Sangwan Kim[1], **Yongku Kim**[1,2,*] **and Jung-In Seo**[3,*]

[1] Department of Statistics, Kyungpook National University, Daegu, Korea
[2] KNU LAMP Research Center, Institute of Basic Sciences, Kyungpook National University, Daegu, Korea
[3] Department of Data Science, Andong National University, Andong, Korea

* **Correspondence:** Email: kim.1252@knu.ac.kr; leehoo1928@gmail.com.

Abstract: In many applications, modeling based on a normal kernel is preferred because not only does the normal kernel belong to the family of stable distributions, but also it is easy to satisfy the stationary condition in the stochastic process. However, the characteristic of the data, such as count or proportion, is a major obstacle to complete modeling based on a normal distribution. To solve a limited boundary or non-normal distribution problem, we provided a novel transformation method and proposed a nonparametric Bayesian approach based on a normal kernel of the transformed variable. In particular, the provided transformation transforms any probability space into a real space and is free from the constraints of the previous transformation, such as skewness, presence of power, and bounded domains. Another advantage was that it was possible to use the Dirichlet process mixture model with full conditional posterior distributions for all parameters, leading to a fast convergence rate in the Markov chain Monte Carlo. The proposed methodology was illustrated with simulated datasets and two real datasets with non-normal distribution problems. In addition, to demonstrate the superiority of the proposed methodology, the comparison with the transformed Bernstein polynomial model was made in the real data analysis.

Keywords: blocked Gibbs sampler; Dirichlet process mixture; nonparametric Bayesian; transformation; transformed Bernstein polynomial
Mathematics Subject Classification: 62G99, 82M36

1. Introduction

To implement a statistical model, it is good to derive a model based on normality. The main reason is that the normal distribution belongs to the family of stable distributions, so modeling based on a normal

kernel can easily construct a kernel density of the observed data. On the other hand, modeling assuming a non-normal distribution has limitations compared to modeling based on a normal distribution. For example, it has the disadvantage of measuring the goodness-of-fit in a generalized linear model compared to a linear model because R^2, the coefficient of determination, does not exist. Furthermore, it is difficult to satisfy a stationary condition in the stochastic process when a random variable has limited probability space such as a proportion or a countable number of possible values. For these reasons, previous studies have tried to solve the non-normal distribution problems through the various transformations of variables. Many transformations for data normality have been introduced and are continuously being proposed, such as logarithmic, reciprocal, square-root, Box-Cox [2], Modulus [9], Yeo-Johnson [15], and Dual [14]. These transformations are listed in Table 1.

Table 1. Transformation for data normality.

Transformation	y_i^P		
Id	$= x_i$		
Reciprocal	$= \frac{1}{x_i}, \quad x_i > 0$		
Box-Cox	$\begin{cases} \frac{x_i^{P-1}}{P}, & P \neq 0 \\ \log(x_i) & P = 0 \end{cases}$	$x_i > 0$	
Yeo-Johnson	$\begin{cases} \frac{(x_i+1)^{P-1}}{P}, & x_i \geq 0, P \neq 0 \\ \log(x_i+1), & x_i \geq 0, P = 0 \\ -\frac{(-x_i+1)^{2-P-1}-1}{2-P}, & x_i < 0, P \neq 2 \\ -\log(-x_i+1), & x_i < 0, P = 2 \end{cases}$		
Log	$= \log x_i, \quad x_i > 0$		
Square-Root	$= \sqrt{x_i}, \quad x_i > 0$		
Modulus	$\begin{cases} \frac{\text{sign}(x_i)\left[(\lvert x_i\rvert+1)^{P-1}\right]}{P}, & P \neq 0 \\ \text{sign}(x_i)\log(\lvert x_i\rvert+1) & P = 0 \end{cases}$	$x_i > 0$	
Dual	$\begin{cases} \frac{x_i^P - x_i^{-P}}{2P}, & P > 0 \\ \log x_i & P = 0 \end{cases}$	$x_i > 0$	

$\text{sign}(y_i) = -1$ for negative y_i and $\text{sign}(y_i) = +1$ for positive y_i.

Table 1 shows that different types of transformation notation are used according to the domain and parameter values with an unknown parameter P, which means the power of each variable. In addition, logarithmic, reciprocal, and square-root transformations can only work for positively skewed data. So, we provide a novel variable transformation that does not have the transformation's constraints such as skewness, hyperparameter, and limited domains to overcome the non-normal distribution problem. If you use this transformation, the random variable defined in any probability spaces is converted to real space $(-\infty, \infty)$. That is, the transformed variable is defined in real space, and the modeling based on a normal kernel can be easily employed.

When modeling in real space, it could be the better choice to use a nonparametric approach rather than a parametric approach to model a more flexible and smoother density function. Thereby, we propose a nonparametric Bayesian modeling through transforming non-normal distribution with the provided transformation. The nonparametric Bayesian model is implemented based on the Dirichlet process mixture model (DPM) [8]. An advantage of this approach is the mitigation of convergence issues commonly encountered in Markov chain Monte Carlo (MCMC), particularly those stemming from the assumption of non-normal distributions because it induces well-known full conditional posterior distributions for unknown parameters. Furthermore, the DPM model automatically estimates the number of latent components, which is the complexity of the observational data, and reduces the complexity of the probability distribution in the probability space of the data. Thus, it need not

determine the degree of complexity in advance.

To achieve modeling here, the blocked Gibbs sampling technique is utilized to conduct a sampling from the proposed nonparametric Bayesian model. This approach is chosen for the following two significant reasons.

First, the blocked Gibbs sampler offers distinct advantages when compared to the P*ó*lya urn Gibbs sampler, particularly in situations where prediction rules are unknown. In general, a method for fitting the Dirichlet process [6] is widely recognized as the P*ó*lya Gibbs sampler. The P*ó*lya urn Gibbs sampler is a direct extension of the P*ó*lya urn sampler and was developed by Escobar [4], MacEachern [10], and Escobar and West [5]. However, the P*ó*lya urn Gibbs sampler can only be used with known prediction rules, a method is needed that can be utilized in the general case. The blocked Gibbs sampler, which is proposed by Ishwaran and James [8], provides a solution for the fitting Dirichlet process with unknown prediction rules. Thus, the blocked Gibbs sampler can be applied to more common scenarios. In the blocked Gibbs sampler, it is noteworthy that the stick-breaking prior can generate random weights $(w_1, w_2, \ldots, w_N)$ without necessitating prior knowledge of the prediction rule, making it adaptable in cases where the P*ó*lya urn characterization cannot be specified. Second, it is the usefulness of the blocked Gibbs sampler. This technique eases the expansion of the general model and alleviates concerns regarding convergence issues in MCMC simulations. Importantly, the DPM model, when used in conjunction with the blocked Gibbs sampler, benefits from having full conditional posterior distributions for all model parameters within a hierarchical framework. This property results in a notably rapid convergence rate during MCMC simulations. Consequently, for the purpose of modeling smoother and more flexible structures, the DPM model utilizing the blocked Gibbs sampler emerges as a promising choice within the realm of Bayesian nonparametric modeling.

To illustrate the merits of the proposed methodology, we perform a density function estimation on the time data. In general, a transformed Bernstein polynomial (TBP) model is widely used for modeling a random variable defined on $(0, 1)$. So, for comparison with the proposed methodology, a semi-parametric modeling based on the TBP model is developed together. Since the TBP model is modeled as a mixture of beta functions, which is called the Bernstein basis polynomial [12, 13], it has a compact support and can easily select an absolutely continuous distribution function with a continuous and smooth derivative. Mallik and Gelfand [11] provided a semi-parametric estimation method among methods to obtain a smoother function using mixing weights in the Bernstein basis function. However, there is a restriction that the degree, which is a number of the parameter, must be determined in advance. So, it is difficult to choose a model flexibly. This drawback of the TBP model motivates the need for the DPM model, which is a nonparametric Bayesian approach.

The rest of the paper is organized as follows. Section 2 provides a novel transformation in that there are no restrictions such as the shape of a density function, presence of power, and limited domain, and proposes a nonparametric Bayesian model based on a normal kernel from the transformation. The TBP model, which is mainly used as a modeling method in the bound interval $(0, 1)$, is introduced as a comparative model of the proposed methodology. In Section 3, simulations are performed to demonstrate estimating performances of the proposed methodology in a non-normal condition. Additionally, it is estimated a density function of the real datasets using the proposed methodology and the comparative methodology in Section 4. Finally, Section 5 concludes the paper.

2. Methodology

Section 2.1 provides the transformation that converts any probability space into real space. Section 2.2 proposes a nonparametric Bayesian approach through the transformation of a non-normal distribution as well as modeling based on the TBP for illustrative purposes. Figure 1 illustrates the analysis process for the methodology provided in this section.

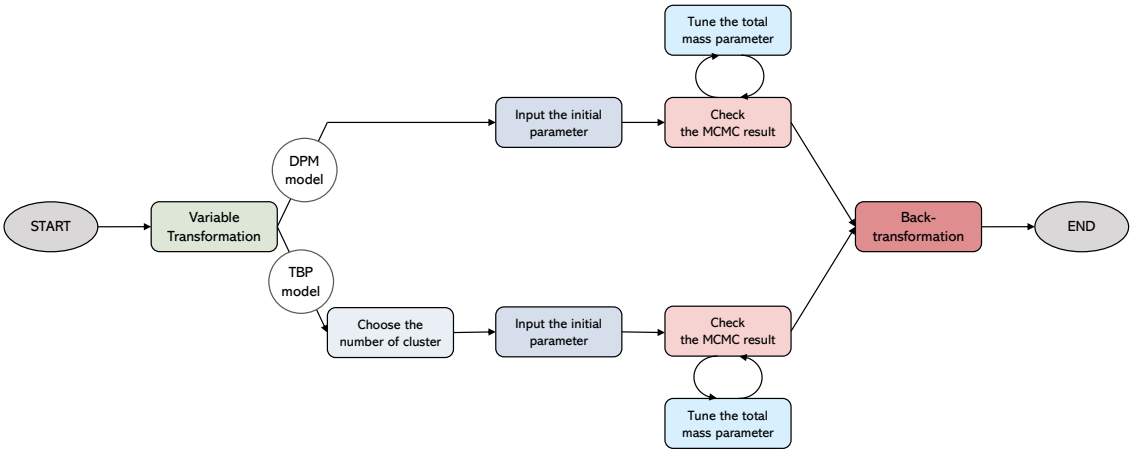

Figure 1. Flowchart for the DPM and TBP models.

2.1. Transformation

The transformation is mainly used for the normality of a random variable with a non-normal distribution problem. Which transformation to use depends on the data structure, the presence of power, and the domain. Removing these constraints improves the utility of the transformation because it can be applied to any probability space. A variable transformation is now provided for this purpose.

Let $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ be a set of n random variables, which is defined with any probability space $\mathcal{S}$, then we consider the following transformation:

$$Y = \log\left(\frac{X - X_{min}}{X_{max} - X}\right), \quad X \in \mathcal{S} \tag{2.1}$$

where X_{max} is the maximum value of $\boldsymbol{X}$ and X_{min} is the minimum value of $\boldsymbol{X}$. A transformed set $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ is defined with the real space.

Furthermore, unlike the transformations listed in Table 1, this transformation does not have any restrictions on the domain of the variables, nor does it involve any powers such as P as shown in Table 1. This allows for a much shorter computation time compared to other methods during the MCMC process. These advantages make it possible to model variables in any probability space by transforming them into real space. In the next section, we provide a nonparametric Bayesian approach to derive full conditional posterior functions for all parameters based on the transformation (2.1).

2.2. The nonparametric Bayesian modeling

As mentioned earlier, we provide a nonparametric Bayesian model based on the transformation (2.1) in this section.

2.2.1. Dirichlet process mixture model

Suppose a sequence of random variable $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ has a non-normal distribution with any support. In this case, to estimate its kernel density, we can assume a DPM model to explain the observed data as follows:

$$\begin{aligned} X_i|\theta_i &\overset{\text{ind}}{\sim} \mathfrak{F}(\cdot \mid \theta_i) \\ \theta_i \mid G &\overset{\text{iid}}{\sim} G \\ G &\sim DP(\lambda G_0), \end{aligned}$$

where $\mathfrak{F}$ is any probability distribution with parameter θ_i, λ is the total mass parameter, G_0 is the centering measure, and *iid* denotes an abbreviation for independent and identical distribution. Since $\boldsymbol{X}$ has a non-normal distribution, it is a very hard task deriving a well-known full conditional posterior distribution. However, the provided transformation can easily develop a nonparametric Bayesian model that derives well-known full conditional posterior distributions for all parameters. By the transformation, it can be supposed that a transformed variable Y has a normal kernel with mean μ and variance σ^2 because of $Y \in \mathcal{R}$. From the assumption, the density function of Y is written as

$$\begin{aligned} f_0^{DPM}(y) &= \sum_{k=1}^{\infty} w_k N\left(y \mid \mu_k, \sigma_k^2\right) \\ &= \sum_{k=1}^{\infty} w_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right), \end{aligned}$$

where w_k is a weight, then the kth kernel function in terms of X is derived as

$$g(x; \mu_k, \sigma_k^2) = J(x) \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(\log\left(\frac{x-x_{min}}{x_{max}-x}\right) - \mu_k\right)^2}{2\sigma_k^2}\right),$$

where $J(x) = \left|\frac{x_{max}-x_{min}}{(x-x_{min})(x_{max}-x)}\right|$ is the Jacobian of the transformation. In addition, the density function with the original scale is given by

$$\begin{aligned} f_0^{DPM}(x) &= \sum_{k=1}^{\infty} w_k g(x; \mu_k, \sigma_k^2) \\ &= \sum_{k=1}^{\infty} w_k J(x) \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\left(\log\left(\frac{x-x_{min}}{x_{max}-x}\right) - \mu_k\right)^2}{2\sigma_k^2}\right). \end{aligned} \tag{2.2}$$

Note that the model (2.2) has an advantage of the posterior MCMC sampling for all parameters since the full conditional posterior distributions have well-known distributions through the blocked Gibbs sampling technique that provides a tractable and efficient Gibbs sampler.

Let $\theta = 1/\sigma^2$, then the model (2.2) is given as the following nonparametric Bayesian hierarchical model

$$X_i \mid \boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{Z}, \mathbf{w} \overset{\text{ind}}{\sim} f_0^{DPM}(x), \quad i = 1, 2, \ldots, n$$

$$
\begin{aligned}
Z_i \mid \mathbf{w} &\overset{\text{iid}}{\sim} \sum_{k=1}^{K} w_k \delta_{Z_k}(\cdot), \\
w_k &= \pi_k \prod_{l<k} (1-\pi_l), \ \ \pi_k \overset{\text{iid}}{\sim} Beta(1,\lambda), \ \ k = 1,2,\ldots,K \\
\mu_j &\overset{\text{iid}}{\sim} N(\nu, \sigma_\mu^2), \\
\theta_j &\overset{\text{iid}}{\sim} Gamma(\nu_1, \nu_2), \ \ j = 1,2,\ldots,K \\
\lambda &\sim Gamma(\eta_1, \eta_2).
\end{aligned} \tag{2.3}
$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K), \boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)$, K is an unknown value that is automatically derived in MCMC, and $\mathbf{w}$ is defined by the stick-breaking. Sampling for each parameter in the hierarchical model (2.3) is implemented using the blocked Gibbs algorithm, which is detailed in the Appendix.

To demonstrate the validity and superiority of the DPM model based on the provided transformation, we develop the model for a specific time period. For comparison with the proposed DPM model, a semi-parametric modeling of the TBP model is derived together because the TBP model is commonly used for modeling a random survival function on $(0, 1)$. Additionally, the TBP model follows in the next section.

2.2.2. Transformed Bernstein polynomial model

This model estimates a function using a linear combination of the Bernstein basis polynomial function that it is a beta density with the parameters a_k and b_k for $k = 1, \ldots, K$. In this model, the K, which is called "degree", determines the number of basis functions and is a known value, unlike the proposed DPM model.

The Bernstein basis polynomial function is written as

$$
\beta(u; a_k, b_k) = \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} u^{a_k - 1}(1-u)^{b_k - 1}, \qquad u \in (0,1)
$$

with $\Gamma(\cdot)$ as the gamma function.

In addition, a Bernstein polynomial with the degree K can be expressed by

$$
f_{\mathbf{w}}(x) = \sum_{k=1}^{K} w_k^{TBP} \beta(x; a_k, b_k)
$$

where $w_k^{TBP} = (w_1^{TBP}, w_2^{TBP}, \ldots w_K^{TBP})'$ is a vector of weights summing to unity such that $0 < w_k^{TBP} < 1$ for $k = 1, \ldots, K$. Denote $B(x)$ is a cumulative distribution function of $\beta(x)$, then the baseline density function and survival function based on the TBP model are given by

$$
f_0^{TBP}(t) = -\frac{\partial}{\partial t} S_\phi(t) \sum_{k=1}^{K} w_k^{TBP} \beta(S_\phi(t); a_k, b_k)
$$

and

$$
S_0^{TBP}(t) = \sum_{k=1}^{K} w_k^{TBP} B(S_\phi(t); a_k, b_k) \tag{2.4}
$$

with $E\{S_0(t)\} = S_\phi(t)$ for all $t > 0$, respectively. In applications for data analysis, the following settings are used.

The weight w_k^{TBP} is defined as

$$w_k^{TBP} = \frac{e^{\nu_k}}{1 + \sum_{k=1}^{K-1} e^{\nu_k}}, \quad k = 1, 2, \ldots, K$$

from the multivariate logistic-normal distribution [1], then under $\mathbf{w} \sim Dirichlet(\lambda \mathbf{1}_K)$, where $\mathbf{1}_K$ is a vector of size K with all components equal to one, the corresponding prior on $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{K-1})'$ with $\nu_K \stackrel{\text{def}}{=} 0$ is given by

$$p(\nu) = \frac{\Gamma(\lambda K)}{[\Gamma(\lambda)]^K} \prod_{k=1}^{K} w_k^{\lambda}$$

In addition, $S_\phi(t)$ is a survival density and assumes the Weibull family. Finally, to produce K evenly spaced in the Bernstein basis function in (2.4), the parameters a_k and b_k are fixed as k and $K - 1$, respectively. Note that the parameter λ controls how close the weight w_k^{TBP} are to $1/K$. For example, as the parameter λ increases, $S_0^{TBP}(t)$ is concentrated about $S_\phi(t)$.

Coincidentally, the degree K in the TBP model must be determined in advance. This constraint prevents the model from becoming more flexible and smooth. Thus, this drawback about the degree K leads to the need for the DPM model, which determines the complex degree of data automatically.

3. Simulation

This section conducts a simulation to demonstrate our model's estimating performances of a density function under non-normal conditions. To illustrate this, a performance verification method based on a goodness-of-fit test is considered. To perform a goodness-of-fit test, it needs to compare the true parameter with its estimate. However, since the proposed DPM model focuses on estimating a distribution rather than estimating unknown parameters, a general goodness-of-fit test cannot be applied. So, we implement a goodness-of-fit test using sample quantiles, which is an idea borrowed from Dunn and Smyth [3]. Concerning this, the ith sample quantile is defined by

$$r_i = \Phi^{-1}\{E(F_0(x_i))\}, \tag{3.1}$$

where Φ and $F_0(x_i)$ are the cumulative distribution functions of the standard normal distribution and a random variable X_i, respectively. If X_i is a continuous random variable, then $E(F_0(x_i))$ is a uniformly distribution on $(0, 1)$. This implies that the distribution of the sample quantile r_i converges to the standard normal distribution. Therefore, the Q-Q plot using the sample quantiles can be employed for a goodness-of-fit test.

To conduct the simulation, simulated datasets are generated from the mixture of two beta distributions, which are Bernstein basis functions. In conclusion, simulated datasets are generated from a density function that satisfies non-normal conditions and is closer to the comparative model than the proposed methodology. Specifically, a density function used for the simulation is defined as

$$f^{sim}(t) = \frac{1}{2}\beta(S_\phi(t); 40, 1) + \frac{1}{2}\beta(S_\phi(t); 50, 50),$$

where $S_\phi(\cdot)$ is a Weibull distribution. In this simulation, we are supposed to consider various situations ranging from small to large. Thus, the sizes of the datasets used are 30, 50, and 70.

After generating simulated datasets, we estimate density functions using the proposed DPM and TPB models. This process is repeated 1,000 times. Consequently, the simulation datasets are generated from 1,000 different density functions.

Figure 2 shows the posterior mean and the highest posterior density (HPD) credible interval (CrI) of each quantile from the cumulative distribution functions estimated by the proposed DPM and the comparative TBP models. If we estimate the density function well, the black points, which are the posterior mean, lie close to the black solid line. Additionally, the gray plane means the 95% HPD CrI.

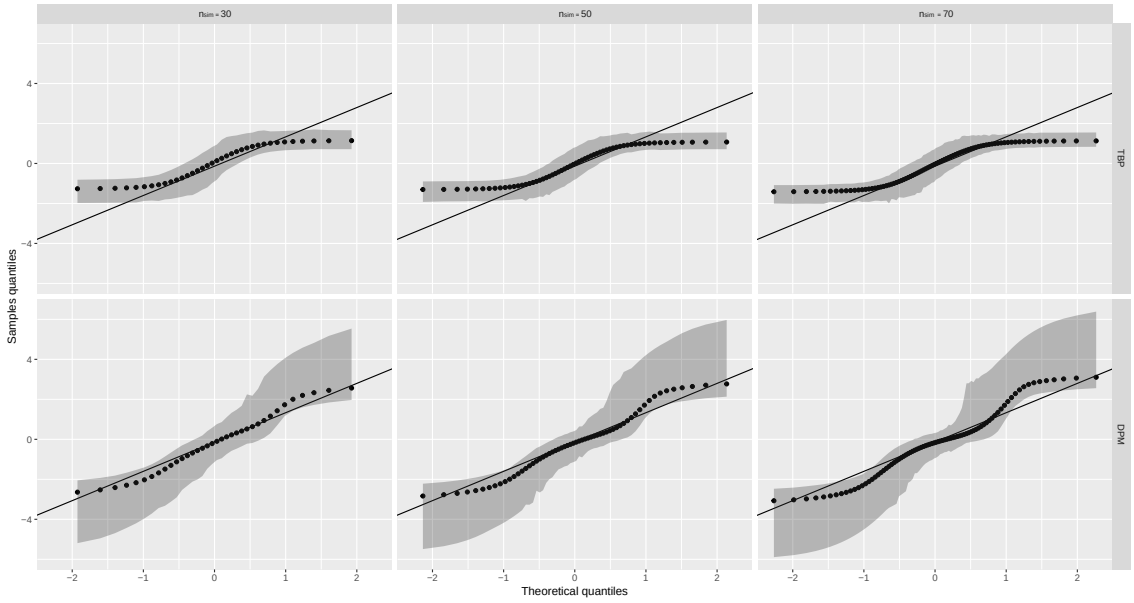

Figure 2. Q-Q plots for the simulated datasets modeling based on the TPB (top) and proposed DPM (bottom) models.

In Figure 2, the top panels represent modeling results on the TPB model for the different sample sizes n_{sim}, which is the number of generated random variables, and the bottom panels represent modeling results on the proposed DPM model for the different sample sizes n_{sim}. The 95% HPD CrI of the proposed DPM model often encompasses the black solid line within all intervals. In contrast, the TBP model is estimated in a form that is close to the logit link function, and it is evident that the estimation performance deteriorates at the edges. Therefore, the proposed DPM model appears to have better estimation performance than the TPB model in the simulation study.

4. Application to real data

To show our proposed methodology works well in any probability space and to demonstrate its superiority, it is performed to modeling for two real datasets, the survival of acute myeloid leukemia (LeukSurv) data and primary biliary cholangitis (PBC) data.

First, we construct a visualization that draws the histogram of the observed data and the plot of the posterior density function modeled by the DPM and TBP models developed in Section 2. Additionally,

a goodness-of-fit test mentioned in the previous section is performed to confirm how well an estimated function describes the observed data.

4.1. LeukSurv

This section provides modeling for a dataset on the LeukSurv, first analyzed by Henderson et al. [7]. This dataset contains 1,043 patients in northwest England with 24 administrative districts. In the example, the survival time (in days) of 28 uncensored patients in the first district is considered for modeling.

To implement the MCMC algorithm, the TBP model has 150,000 iterations and 50,000 burn-in, while the proposed DPM model based on the provided transformation has 15,000 iterations and 5,000 burn-in. The results of the modeling are reported in Figure 2. The real data plots as histograms and the last 300 posterior draws of each model are a gray dashed line. Additionally, a black solid line implies the total posterior mean and a gray dotted line is the 95% HPD CrI.

Figure 3 shows that the proposed DPM model estimates a density function close to real data with multi-modal, despite having a much shorter convergence period than the TBP model. In contrast, the TBP model failed to estimate the modal between 100 and 150. Additionally, the proposed DPM model estimated a heavy-tailed density function compared to the TBP model, as seen in the results above 400.

For more in-depth reviews and comparisons, the Q-Q plots are presented in Figure 3. As mentioned earlier, sample quantile (r_i) is computed by (3.1).

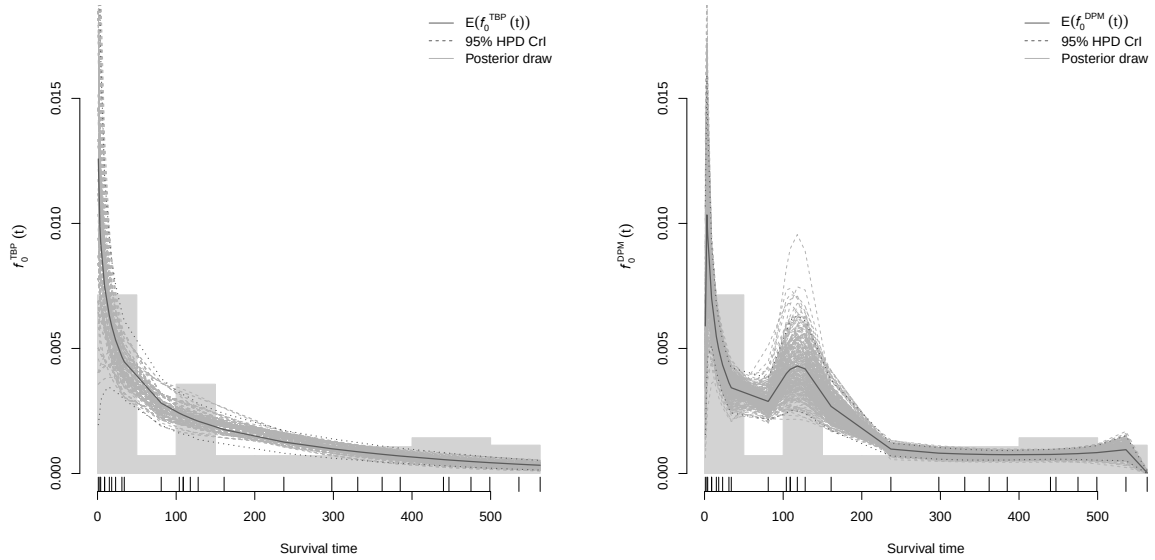

Figure 3. Posterior inference results for the LeukSurv dataset modeling based on the TBP (left) and proposed DPM (right) models.

Figure 4 shows that the gray dots are close to the black solid line in the center portion of both Q-Q plots. In contrast, comparing the edge portion of the plot, the gray dot in the proposed DPM model's Q-Q plot is closer to the black solid line than in the TPB model. Apparently, the proposed DPM model estimates a density function on the LeukSurv dataset rather than the TBP model.

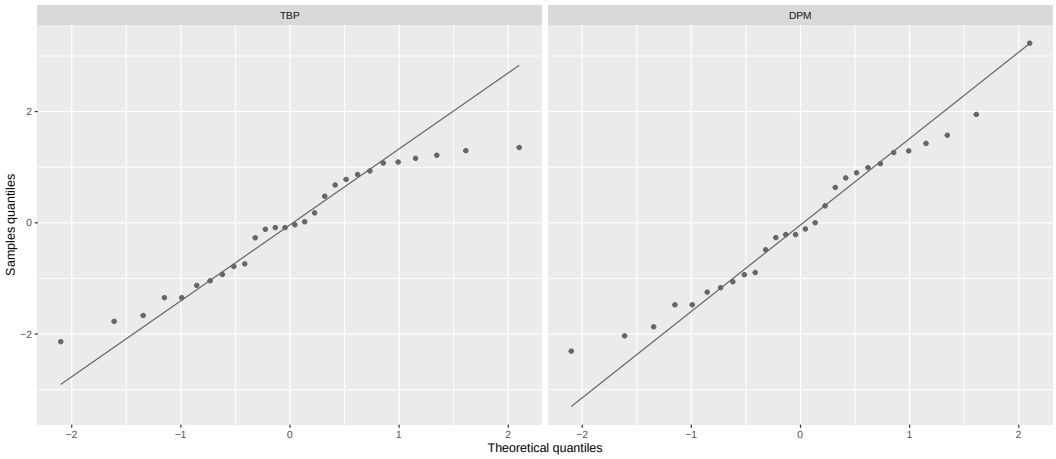

Figure 4. Q-Q plots for the LeukSurv dataset modeling based on the TPB (left) and proposed DPM (right) models.

4.2. Primary biliary cirrhosis

As another example, we utilize the PBC dataset obtained from Mayo Clinic trials conducted during the period spanning from 1974 to 1984. The PBC dataset pertains to a condition known as PBC, an autoimmune disease that results in the progressive destruction of the small bile ducts in the liver. This chronic ailment typically exhibits a slow and inexorable progression, ultimately culminating in the development of cirrhosis and liver decompensation. Within the PBC dataset, one can find a key variable denoted as 'time,' which quantifies the number of days from the time of patient registration to either the occurrence of mortality, transplantation, or the conclusion of the study in July 1986. This dataset encompasses a total of 418 patients. Of the total 418 patients, the number of patients whose variable *status* is "dead" is 161, and modeling is performed using the variable *time* representing the survival time. The MCMC setting is the same as the previous section, and the results are as follows.

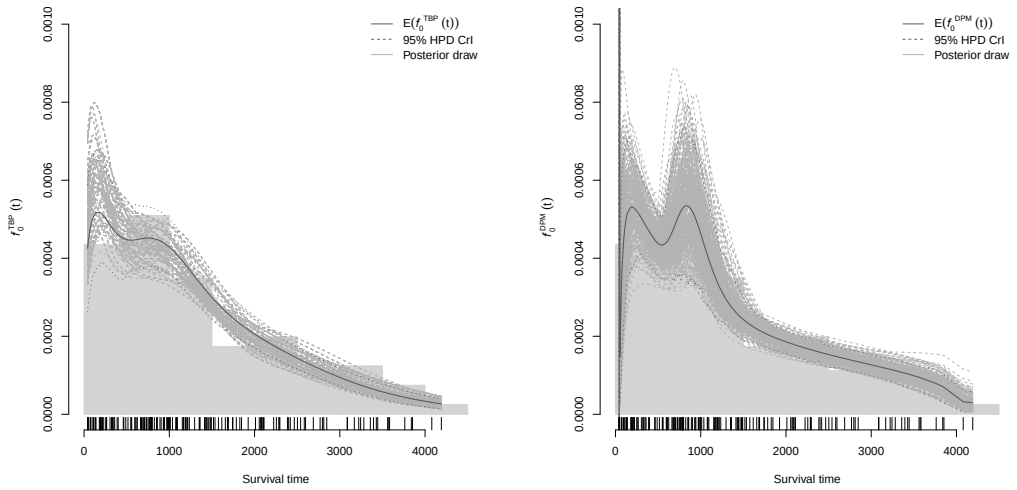

Figure 5. Posterior inference results for the PBC dataset modeling based on the TBP (left) and proposed DPM (right) models.

In Figure 5, the proposed DPM model estimates a density function closer to real data with multi-modal than the TBP model despite the much shorter convergence period. The TBP model estimates a smooth density function but lacks a multi-modal density function. On the other hand, the proposed DPM model accurately estimates the density function between 500 and 1500, and better captures heavy tails of the real data.

The goodness-of-fit test is performed through the Q-Q plot (Figure 5) for more detailed reviews and comparisons.

Figure 6 shows that the gray dot in the proposed DPM model's Q-Q plot is closer to the black solid line than the TPB model overall. Even on this dataset, as in the example of the LeukSurv, the DPM model estimates a density function rather than the TBP model.

As confirmed in the previous two examples, both the TBP model and the proposed DPM model estimate smooth functions based on nonlinearity. However, the modeling based on the proposed DPM model is more flexible than the TBP model because it can estimate the degree freely. Furthermore, the proposed DPM model has an excellent ability to estimate a density function with multi-modal.

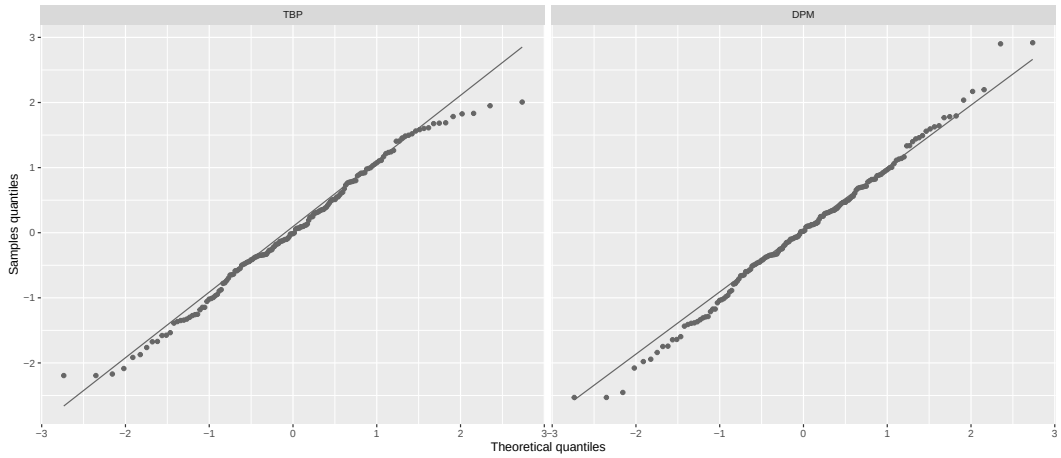

Figure 6. Q-Q plots for the PBC dataset modeling based on the TBP (left) and proposed DPM (right) models.

5. Conclusions

In this study, we proposed a nonparametric Bayesian modeling based on the provided transformation to take advantage of modeling based on a normal distribution. The provided transformation can convert any probability space into real space, unlike the methods in Table 1. Moreover, the provided transformation has no constraints such as skewed direction, presence of power, and limited domain.

Additionally, the proposed methodology derives well-known full conditional posterior distributions for unknown parameters, eliminating concerns about convergence problems in MCMC that can arise from assuming a non-normal distribution kernel. Moreover, the methodology estimates a more flexible density function by automatically determining the number of latent components.

To demonstrate the validity and superiority of the proposed methodology, we used simulated datasets, as well as the LeukSurv and PBC datasets, for survival time modeling. We employed the TBP

model, which is primarily used for modeling closed intervals of (0, 1), as a benchmark for performance comparison. In summary, the proposed DPM model estimated a more flexible function and modeled a density function with multi-modal compared to the TBP model. It is possible to derive a heavy-tailed distribution. Additionally, the DPM model automatically estimates the degree of complexity in data using MCMC, while the TBP model requires determining the degree in advance. This implies that estimating the multi-modal density function from the TBP model is limited. Furthermore, the proposed DPM model offers a faster convergence rate compared to the TBP model when using MCMC with blocked Gibbs sampling due to the full conditional posterior distribution of all parameters.

Therefore, despite the DPM modeling needing to not decide the number of latent components in advance, the DPM modeling based on the provided transformation shows better performance and estimation ability than modeling based on the TBP model. Ultimately, this proposed methodology's major advantage is automatically derived a kernel density by the nonparametric Bayesian approach, that is, a more flexible and smoother function than the estimation function by parametric methodology in any probability space. For the demonstration of the generality of the proposed methodology, we attempt to apply a multivariate extension, space extension, and the autocorrelation variable in further studies.

Author contributions

S.K.: Conceptualization, Methodology, Writing–original draft; Y.K.: Data curation, Formal Analysis, Writing–original draft, Writing–review & editing; J.-I.S.: Conceptualization, Methodology, Writing–original draft, Writing–review & editing. All three authors read and approved the final version of the article.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by Learning & Academic research institution for Master's·PhD students, and Postdocs (LAMP) Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301914).

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. J. Aitchison, S. M. Shen, Logistic-normal distributions: some properties and uses, *Biometrika*, **67** (1980), 261–272. https://doi.org/10.1093/biomet/67.2.261
2. G. E. P. Box, D. R. Cox, An analysis of transformations, *J. Royal Stat. Soc. B*, **26** (1964), 211–252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

3. P. K. Dunn, G. K. Smyth, Randomized quantile residuals, *J. Comput. Graph. Stat.*, **5** (1996), 236–244. https://doi.org/10.1080/10618600.1996.10474708

4. M. D. Escobar, Estimating normal means with a Dirichlet process prior, *J. Amer. Stat. Assoc.*, **89** (1994), 268–277. https://doi.org/10.2307/2291223

5. M. D. Escobar, M. West, Bayesian density estimation and inference using mixtures, *J. Amer. Stat. Assoc.*, **90** (1995), 577–588. https://doi.org/10.2307/2291069

6. T. S. Ferguson, Prior distributions on spaces of probability measures, *Ann. Statist.*, **2** (1974), 615–629. https://doi.org/10.1214/aos/1176342752

7. R. Henderson, S. Shimakura, D. Gorst, Modeling spatial variation in leukemia survival data, *J. Amer. Stat. Assoc.*, **97** (2002), 965–972. https://doi.org/10.1198/016214502388618753

8. H. Ishwaran, L. F. James, Gibbs sampling methods for stick-breaking priors, *J. Amer. Stat. Assoc.*, **96** (2001), 161–173. https://doi.org/10.1198/016214501750332758

9. J. A. John, N. R. Draper, An alternative family of transformations, *J. Royal Stat. Soc. C Appl. Statist.*, **29** (1980), 190–197. https://doi.org/10.2307/2986305

10. S. N. Maceachern, Estimating normal means with a conjugate style Dirichlet process prior, *Commun. Stat.-Simulat. Comput.*, **23** (1994), 727–741. https://doi.org/10.1080/03610919408813196

11. B. K. Mallick, A. E. Gelfand, Generalized linear models with unknown link functions, *Biometrika*, **81** (1994), 237–245. https://doi.org/10.1093/biomet/81.2.237

12. S. Petrone, Bayesian density estimation using Bernstein polynomials, *The Canadian Journal of Statistics*, **27** (1999), 105–126. https://doi.org/10.2307/3315494

13. S. Petrone, Random Bernstein polynomials, *Scand. J. Stat.*, **26** (1999), 373–393. https://doi.org/10.1111/1467-9469.00155

14. Z. Yang, A modified family of power transformations, *Econ. Lett.*, **92** (2006), 14–19. https://doi.org/10.1016/j.econlet.2006.01.011

15. I. K. Yeo, R. A. Johnson, A new family of power transformations to improve normality or symmetry, *Biometrika*, **87** (2000), 954–959. https://doi.org/10.1093/biomet/87.4.954

Appendix

Blocked Gibbs sampling

Let $\left\{Z_1^*, \ldots, Z_m^*\right\}$ denote the current m unique values of $\mathbf{Z}$, then the Gibbs sampling procedure for each parameter in the hierarchical model (2.3) can be summarized as follows:

(a) Conditional draw for μ : For each $j \in \left\{Z_1^*, \ldots, Z_m^*\right\}$,

$$\mu_j \mid \mathbf{Z}, \boldsymbol{x} \overset{\text{ind}}{\sim} N\left(\mu_j^*, \sigma_j^{2*}\right),$$

where

$$\mu_j^* = \frac{v\sigma^2 + \sigma_\mu^2 \sum_{\{i:Z_i^*=j\}} y_i}{\sigma^2 + n_j\sigma_\mu^2},$$
$$\sigma_j^{2^*} = \frac{\sigma_\mu^2\sigma^2}{\sigma^2 + n_j\sigma_\mu^2},$$

and n_j is the number of times Z_j^* occurs in $\mathbf{Z}$. In addition, for each $j \in \mathbf{Z} - \{Z_1^*, \ldots, Z_m^*\}$, independently simulate $\mu_j \sim N(v, \sigma_\mu^2)$.

(b) Conditional draw for θ : For each $j \in \left\{Z_1^*, \ldots, Z_m^*\right\}$,

$$\theta_j \mid \boldsymbol{\mu}, \mathbf{Z}, \boldsymbol{x} \overset{\text{ind}}{\sim} Gamma\left(\nu_1 + n_j/2,\ \nu_2 + \sum_{\{i:Z_i^*=j\}} (y_i - \mu_j)^2/2\right).$$

In addition, for each $j \in \mathbf{Z} - \{Z_1^*, \ldots, Z_m^*\}$, independently simulate $\theta_j \sim Gamma(\nu_1, \nu_2)$.

(c) Conditional draw for $\mathbf{Z}$:

$$Z_i \mid \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{x} \overset{\text{iid}}{\sim} \sum_{k=1}^{K} w_{k,i}\delta_{Z_i}(\cdot), \quad i = 1, \ldots, n.$$

where

$$(w_{1,i}, \ldots, w_{K,i}) \propto \left(w_1 g(x_1 \mid \mu_1, \sigma_1^2), \ldots, w_K g(x_K \mid \mu_K, \sigma_K^2)\right).$$

(d) Conditional draw for $\mathbf{w}$:

$w_1 = \pi_1^*$ and $w_k = \pi_k^* \prod_{l<k}(1 - \pi_l^*),\ k = 2, \ldots, K-1,$

$$\pi_k^* \overset{\text{ind}}{\sim} Beta\left(1 + \#\{Z_i = k\}, \lambda + \#\{Z_i > k\}\right)$$

where $\#\{Z_i = k\}$ is the number of Z_i equal to k and $\#\{Z_i > k\}$ is the number of Z_i greater than k for $k = 1, \ldots, K-1$, then w_K is calculated as $w_K = 1 - \Sigma_{k=1}^{K-1} w_k$.

(e) Conditional draw for λ :

$$\lambda \mid \mathbf{w} \sim Gamma\left(K + \eta_1 - 1, \eta_2 - \sum_{k=1}^{K-1} \log(1 - \pi_k^*)\right)$$

for $k = 1, \ldots, K-1$.

AIMS Press

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)