Thailand Statistician
October 2021; 19(4): 812-824
http://statassoc.or.th
Contributed paper

Discrete Support Set Selection for Gamma Prior Density Estimation in Measurement Error Model using Empirical Bayes Deconvolution

Fevi Novkaniza [a][c], Khairil Anwar Notodiputro*[a], I Wayan Mangku [b] and Kusman Sadik [a]

[a] Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, West Java, Indonesia.

[b] Department of Mathematics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, West Java, Indonesia.

[c] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, West Java, Indonesia.

*Corresponding author; e-mail: khairilnotodiputro@gmail.com

Received: 29 November 2019
Revised: 29 April 2020
Accepted: 29 May 2020

Abstract

This paper discusses the empirical Bayes deconvolution (EBD) method in estimating gamma's prior density for count data when the true or unobserved random variable is subject to measurement error. The observed random variable W is related to the unobserved random variable X by an additive measurement error model. The count data $W_1, W_2, ..., W_n$ are assumed to follow a Poisson distribution as realizations from an unknown prior density $g(x)$. Then the EBD method is applied to estimate $g(x)$ for every discretization point in the discrete support set of X. The effect of selecting discrete support set for estimating gamma's prior density based on the EBD method is illustrated by using simulation. It is shown that by selecting discretization set for Poisson data and gamma density as a prior distribution, the larger domain, and more points in discrete support set, the smaller value of bias, and standard deviation for gamma prior density estimate. Finally, assuming that the number of high school student dropout follows Poisson distribution, the EBD method is applied to estimate the prior probability distribution for high school student dropout data in 9 cities and 18 districts in West Java province.

Keywords: Bias, conjugate, hyperparameter, loglikelihood, Poisson.

1. Introduction

In social-economic, behavioral, and environmental studies, the data frequently collected from surveys, registration systems, clinical trials, and other observational or experimental studies, which often contaminated with measurement errors. Measurement error occurs when some variables in a

statistical model of interest cannot be observed precisely, usually due to instrument or sampling error. To obtain more reliable inference, one needs to consider the measurement errors when developing statistical methods to analyze this type of data. One of the most fundamental problems in empirical studies is measurement errors. Ignoring measurement errors can produce biased estimators and lead to incorrect conclusions in data analysis (Meister 2009). Many statistical procedures have developed for statistical inference in measurement error models (Buonaccorsi 2010). The measurement error model is a combination of measurement errors and unobserved random variables where the observed value is known. Since the true random variables unobserved, it is essential to estimate the nature of the distribution of populations through the distribution of prior random variables that contain measurement errors. However, the estimation of distribution parameters from sample observation can be a problem because of the unobserved random variable measured by error. The main problem is how to estimate the density function of an unobserved random variable that contaminated with measurement error. The problem of estimating this density function is called deconvolution.

There are many parametric and nonparametric methods proposed to estimate the density function of the unobserved random variables due to measurement errors. In this paper, we develop the effect of selecting a finite discrete set for parameter space of prior distribution from unobserved random variables under the framework of measurement error models. We use the empirical Bayes deconvolution method proposed by Efron (2016) for estimating prior density distribution by combined the concept of deconvolution and empirical Bayes. The basic idea of the empirical Bayes deconvolution method is modeling prior density as a member of exponential family density. Empirical Bayes inference assumes an unknown prior density of the unobserved random variable X, $g(x)$, produces an independent observation W_i from the probability distribution of W_i given X_i. The EBD method attempts to estimate $g(x)$ using the observed sample and the distribution of W_i given W_i is specified.

Furthermore, by discretizing support set T of X_i, the parameter space of unobserved random variable X, the likelihood function is constructed to get the maximum likelihood estimation for the parameter of the prior distribution. We assume that the observed sample W_i is distributed according to Poisson distribution, and based on sample observation W_i, we estimate the density of true or unobserved random variables as prior density. We use gamma distribution as conjugate prior for Poisson distribution, and this is quite similar to the concept of the mixture distribution. If a random variable has a Poisson distribution with hyperparameter has gamma distribution, the posterior distribution will be a negative binomial distribution. Efron (2016) simplified support set T as finite discrete support, but there are no rules or guidelines for choosing support set T and the number of discretization points in T. We illustrate the performance of the discretization set for empirical Bayes deconvolution by simulation studies. Furthermore, we establish the asymptotic properties of the proposed estimators. The performance of the prior density estimation assessed from the standard deviation and estimation bias using various simulations.

This paper organized as follows. In Section 2, we discuss the measurement error model and the concept of empirical Bayes deconvolution. In Section 3, we establish simulation studies for the empirical Bayes deconvolution procedure and give some statistical results for density estimation of the prior distribution. We distinguish four cases according to the finite discrete set of the underlying parameter space from the prior distribution. As a real application, we applied the EBD method for high school student's dropout data in West Java during 2018 are discussed in Section 4. We offer some concluding remarks in Section 5.

2. Methods

2.1. Measurement error model

The measurement error model is a combination of measurement errors and the true random variable, which gives the observed value. Laird (1978), Fan (1991), Hall and Meister (2007) and Butucea and Comte (2009) use an additive measurement error model as follows

$$W_i = X_i + e_i, i = 1,...,n,$$

where W_i is the observed variable X_i, is an unobserved random variable, and e_i is measurement error. We assume unobserved random variable X_i and measurement errors e_i uncorrelated, and e_i is independent and identically distributed. The problem of density estimation based on contaminated data is called deconvolution.

In parametric deconvolution, the data are known to be from a specific distribution. In this case, the parameters of the distribution can be estimated by, e.g., maximum likelihood, a method of moment, and Bayesian approach like empirical Bayes. Estimation by maximum likelihood is computationally very expensive since numerical integration needs to perform for each data point for each evaluation of the likelihood function. Method of moments estimation sometimes fails to give physically meaningful estimates. The origin of this problem lies in the large sampling variations of the third moment. Since a convolution integral needed to calculate for each data point and that this must repeat for each iteration towards the maximum likelihood solution, computing cost is very high. Carroll and Hall (1988), Stefanski and Carroll (1990) introduced kernel deconvolution estimators, Carroll and Hall (2004) used the weighted kernel density for the method of density deconvolution, Delaigle and Hall (2016) proposed non and semiparametric inference on the distribution of unobserved random variable that do not assume the density of measurement error to be (fully) known. Another method for estimating the density of an unobserved random variable was proposed by Efron (2016) using the Bayesian framework, and it called empirical Bayes deconvolution (EBD) method. EBD method estimates the density of unobserved random variables X_i as prior density estimation based on the observed sample W_i. Finite support set T for X_i is choose based on the range value of W_i, and prior density $g(x)$ is assumed to belong to exponential family distribution.

2.2. Empirical Bayes deconvolution

Let unknown prior density $g(x)$ has an observed independent random sample of realizations $X_1,...,X_n$:

$$X_1,...,X_n \sim g(x).$$

Each X_i independently produces an observed random variable W_i with known probability densities for W_i given X_i, $W_i \sim p_{W_i}(w), i = 1,...,n$ and the marginal density of W_i

$$f(w) = \int p_W(w \mid x) g(x) dx .$$

According to Efron (2016), for estimating the prior density $g(x)$ using sample observation $W_1,...,W_n$, we can use EBD. The EBD method is an estimation procedure $g(x)$ based on sample observations from $f(w)$. Efron (2016) used the likelihood approach to EBD problems with prior $g(x)$, which is modeled through exponential family density in space-X, denote by T. T is assumed to be a finite discrete support set $T = (x_1,...,x_m)$ and by discretizing space-X:

$$X \in T = \left\{ x_{(1)}, x_{(2)}, ..., x_{(m)} \right\}$$

then prior distribution can be denoted as

$$g_j = \mathrm{P}(X = x_{(j)}).$$

Prior density $g(x)$ is an m-vector $g = (g_1, ..., g_m)$ which is specified the probability g_j on x_j :

$$g = g(\alpha) = \exp\left(Q\alpha - \phi(\alpha)\right)$$

with

$$\phi(\alpha) = \log \sum_{j=1}^{m} \exp\left(Q_j^T \alpha\right),$$

where α = p-dimensional vector and Q = known $m \times p$ structure matrix. The- j component of $g(\alpha)$:

$$g_j(\alpha) = \exp\left\{ Q_j^T \alpha - \phi(\alpha) \right\}, j = 1, ..., m.$$

Define $p_{ij} = p_i(W_i \mid X_i = x_j)$ and denote P_i as m-vector $P_i = \left(p_{i1}, ..., p_{im}\right)^T$, then the marginal probability for W_i :

$$f_i(\alpha) = \sum_{j=1}^{m} p_{ij} g_j(\alpha) = P_i^T g(\alpha).$$

The loglikelihood function for the parameter vector $\alpha = \left(\alpha_1, ..., \alpha_p\right)^T$ is

$$l_i(\alpha) = \log f_i(\alpha) = \log P_i^T g(\alpha),$$

with p-dimensional first derivative vector and $p \times p -$ dimensional second derivative matrix

$$\dot{l}_i(\alpha) = \left(..., \frac{\partial l_i(\alpha)}{\partial \alpha_h}, ... \right)^T, \ddot{l}_i(\alpha) = \left(..., \frac{\partial^2 l_i(\alpha)}{\partial \alpha_h \alpha_k}, ... \right)$$

for the maximum likelihood calculation. For W_i with n observation, the total loglikelihood $l(\alpha) = \sum_{i=1}^{n} \dot{l}_i(\alpha)$ has the first and the second derivative which is

$$\dot{l}(\alpha) = \sum_{i=1}^{n} \dot{l}_i(\alpha) = Q^T \sum_{i=1}^{n} B_i(\alpha) = Q^T B_+ \alpha,$$

where

$$B_i(\alpha) = \left\{ b_{i1}(\alpha), ..., b_{im}(\alpha) \right\}^T, \ b_{ij}(\alpha) = g_j(\alpha) \left\{ \frac{p_{ij}}{f_i(\alpha)} - 1 \right\}$$

and

$$-\ddot{l}_i(\alpha) = Q^T \left[B_i(\alpha) B_i(\alpha)^T + B_i(\alpha) g(\alpha)^T + g(\alpha) B_i(\alpha)^T - diag\left\{ B_i(\alpha) \right\} \right] Q.$$

Efron (2016) proposed the maximum likelihood estimate α which is satisfied

$$Q^T B_+ \alpha = 0.$$

3. Simulation

We created some Poisson simulation scenarios for the EBD method. Suppose $W_i \sim Poisson\left(X_i\right)$ and $X_i \sim Gamma\left(\alpha, \beta\right), i = 1, ..., 100.$ We generated 100 observations by first generating X_i and then

creating a dataset (X_i, W_i). We did 1,000 simulations, each with 100 observations independently, and we got the 100×100 data matrix. By taking the various setting of discrete support set $T = [0,...,5]$, $T = [0,...,10], T = [0,...,20], T = [0,...,30]$ and we estimated gamma prior density using EBD method and compute bias for 1,000 simulations. In Figures 1-4, we created a plot of gamma density and its deconvolution for some discretization set of T as follows:

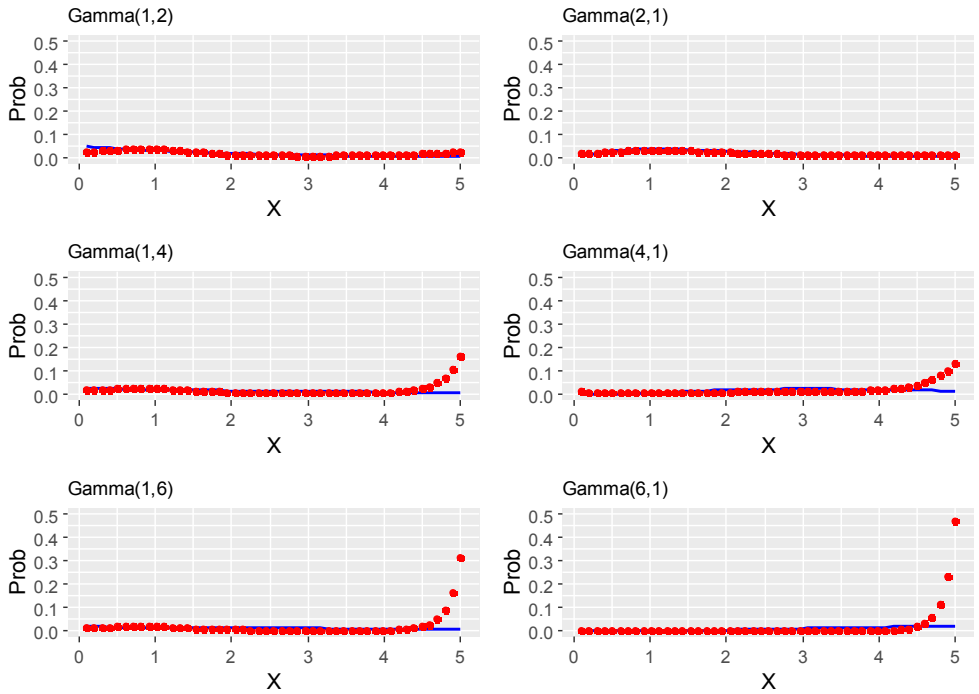


Figure 1 The plot of gamma prior density (blue line) and its estimation (red dotted line) using EBD for support set $T = [0,...,5]$

In this simulation studies, we also computed bias and standard deviation of gamma's prior density estimation for various discretization set of T which are shown in in Tables 1 and 2.

For varying gamma parameter and support discrete set in Figures 1-4, when scale parameter becomes smaller, the plot of gamma density estimation using EBD method look-alike pdf plot of gamma distribution. It was not surprising because the scale parameter is controlling the spread out of the distribution. The more significant value of scale parameter, the more considerable value of gamma variance, and it will be more difficult for estimating the density function by the EBD method. However, from Figures 1-4, we also see that in the initial discretization point of T still underestimate or overestimate, even though we choose the broader support set. From Tables 1 and 2, bias and standard deviation for gamma density estimation become smaller when we enlarge the support discrete set T.

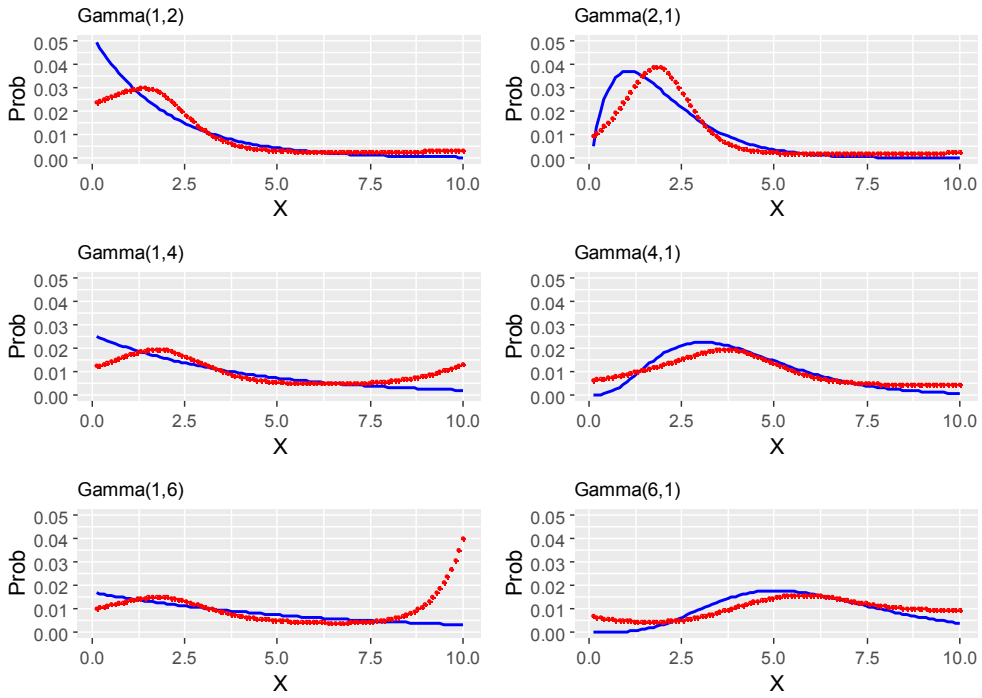


Figure 2 The plot of gamma prior density (blue line) and its estimation (red dotted line) using EBD for support set $T = [0,...,10]$

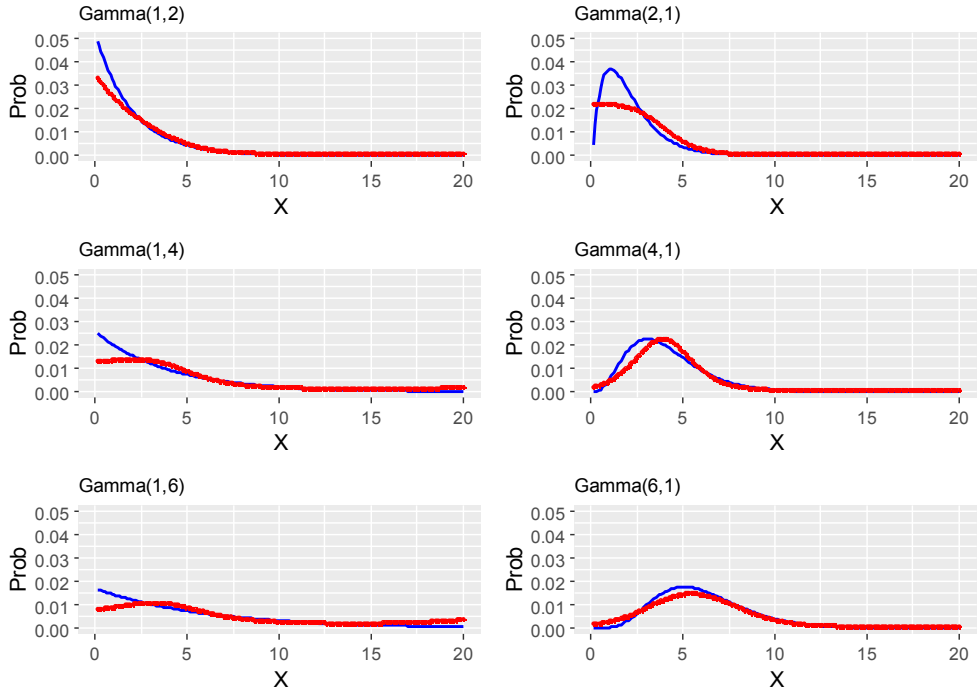


Figure 3 The plot of gamma prior density (blue line) and its estimation (red dotted line) using EBD for support set $T = [0,...,20]$

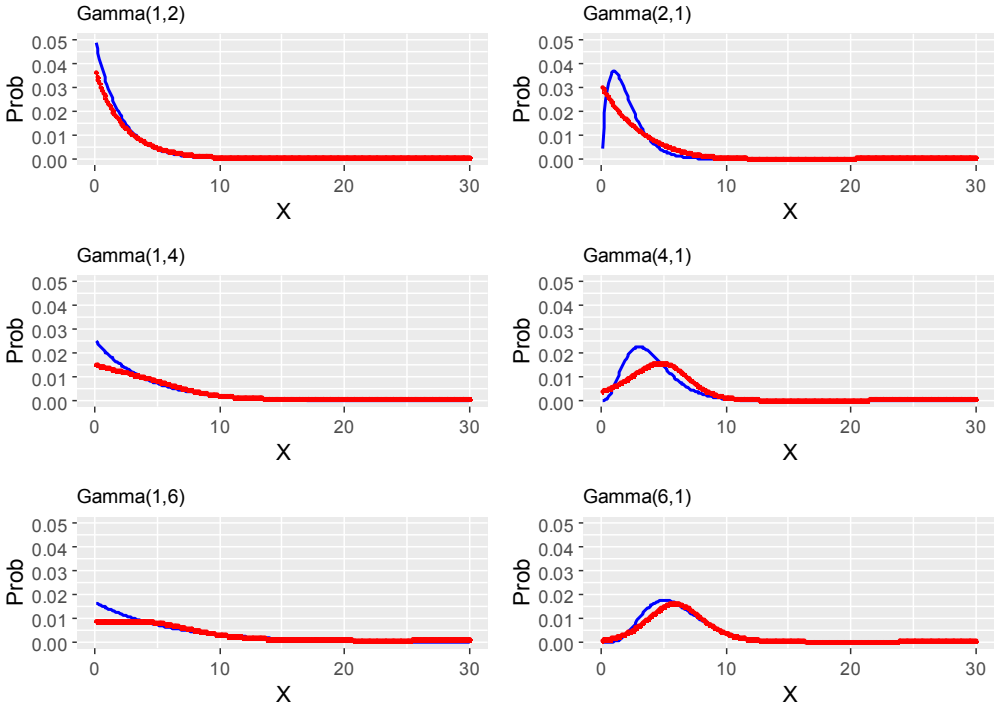


Figure 4 The plot of gamma prior density (blue line) and its estimation (red dotted line) using EBD for support set $T=[0,...,30]$

Table 1 Bias for gamma prior density estimation

Support set	Gamma parameter (α,β)	Bias	Gamma parameter (α,β)	Bias
$T=[0,...,5]$	(1,2)	−0.00122	(2,1)	−0.00046
	(1,4)	−0.00553	(4,1)	−0.00519
	(1,6)	−0.00866	(6,1)	−0.01250
$T=[0,...,10]$	(1,2)	0.00017	(2,1)	0.00008
	(1,4)	−0.00071	(4,1)	−0.00007
	(1,6)	−0.00181	(6,1)	−0.00061
$T=[0,...,20]$	(1,2)	0.00017	(2,1)	0.00011
	(1,4)	0.00003	(4,1)	0.00001
	(1,6)	−0.00014	(6,1)	0.00001
$T=[0,...,30]$	(1,2)	0.00013	(2,1)	0.00010
	(1,4)	0.00005	(4,1)	0.00001
	(1,6)	0.00001	(6,1)	0.00000

Table 2 Standard deviation of gamma prior density estimation

Support set	Gamma parameter (α,β)	Standard deviation	Gamma parameter (α,β)	Standard deviation
$T=[0,...,5]$	(1,2)	0.00421	(2,1)	0.00407
	(1,4)	0.00451	(4,1)	0.00581
	(1,6)	0.00329	(6,1)	0.00201
$T=[0,...,10]$	(1,2)	0.00207	(2,1)	0.00177
	(1,4)	0.00190	(4,1)	0.00215
	(1,6)	0.00222	(6,1)	0.00221
$T=[0,...,20]$	(1,2)	0.00072	(2,1)	0.00067
	(1,4)	0.00090	(4,1)	0.00065
	(1,6)	0.00086	(6,1)	0.00085
$T=[0,...,30]$	(1,2)	0.00037	(2,1)	0.00034
	(1,4)	0.00050	(4,1)	0.00038
	(1,6)	0.00055	(6,1)	0.00037

4. Application

In Indonesia, education has become the primary policy to foster growth developed, and Indonesia's education system has gradually improved, and enrolment rates have significantly increased over the last 50 years. However, Indonesia still faces some barriers for a student to get a better education, such as dropping out of school before graduating. Not only in Indonesia, but school dropout is also considered a global problem (Ajaja 2012, Sang et al. 2013), especially at the high school level. Some studies on dropout have been conducted in Indonesia to explain factors contributing to dropping out of high school students (Setyadharma et al. 2018). Setyadharma et al. (2018) had collected primary data from 439 former high school students, and 878 parents/guardians participated in Central Java province, Indonesia. In their study, they found that female student and having more family members indicates increasing the probability of dropout. However, a lower level of dropout happened when household heads have a university degree, when student academic activities supported by mothers (but not by fathers), and when poor students receive government cash transfers.

West Java is a province of Indonesia on the western part of the island of Java. West Java is the most populous province of Indonesia, with a population of 48,683,861 as of 2018. However, the level of dropout in West Java in 2018 is 37,971 students from primary school until high school. Children who dropped out of primary school reached 5,627 students, junior high schools reached 9,621 students, high schools reached 5,403, and the worst were vocational students who totaled 17,320 students dropping out of school. Based on data from Regional Education Balance 2018 for 9 cities and 18 districts in West Java, we focus on the number of high school student dropout and the statistics descriptive of the data are shown in Tables 3 and 4.

Table 3 High school students dropout data in West Java province 2018

City/District	Number of high school students DO	Proportion of high school students DO
Kota Banjar	15	0.0043
Kota Cimahi	19	0.0021
Kab Pangandaran	24	0.0057
Kota Sukabumi	25	0.0031
Kab Sumedang	34	0.0021
Kota Tasikmalaya	39	0.0026
Kota Depok	52	0.0021
Kota Cirebon	54	0.0044
Kab Ciamis	60	0.0040
Kab Majalengka	62	0.0037
Kota Bekasi	124	0.0030
Kab Kuningan	134	0.0074
Kota Bogor	148	0.0075
Kab Indramayu	174	0.0095
Kab Subang	177	0.0077
Kab Cirebon	200	0.0092
Kab Bekasi	236	0.0048
Kab Bandung Barat	237	0.0090
Kab Purwakarta	245	0.0161
Kab Tasikmalaya	265	0.0127
Kota Bandung	326	0.0054
Kab Bandung	327	0.0062
Kab Karawang	346	0.0111
Kab Sukabumi	411	0.0118
Kab Cianjur	471	0.0145
Kab Bogor	573	0.0083
Kab Garut	625	0.0140

Table 4 Statistics descriptive for high school students dropout data in West Java province 2018

Summary statistics	
Mean	200.11
St. Dev	172.97
Variance	29,920.03
n	27
Minimum	15
1st Quartile	52
Median	174
3rd Quartile	326
Maximum	625

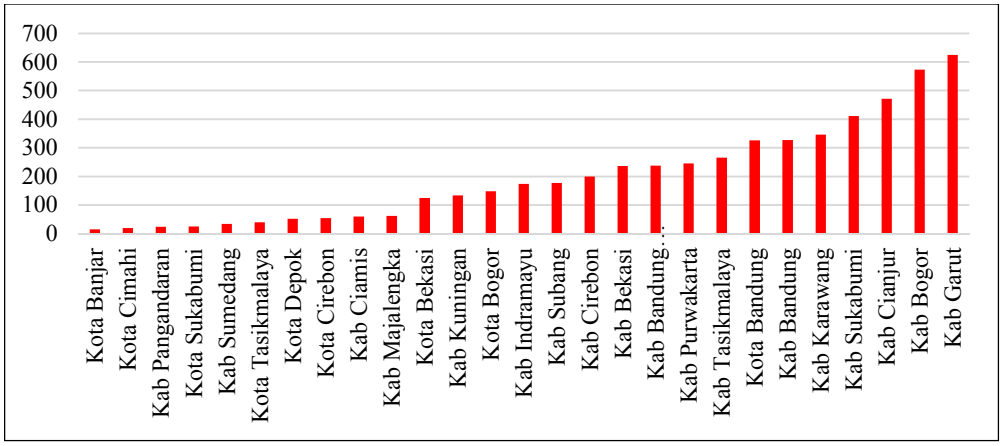


Figure 5 Frequency of dropout high school students in 27 City/District in West Java 2018

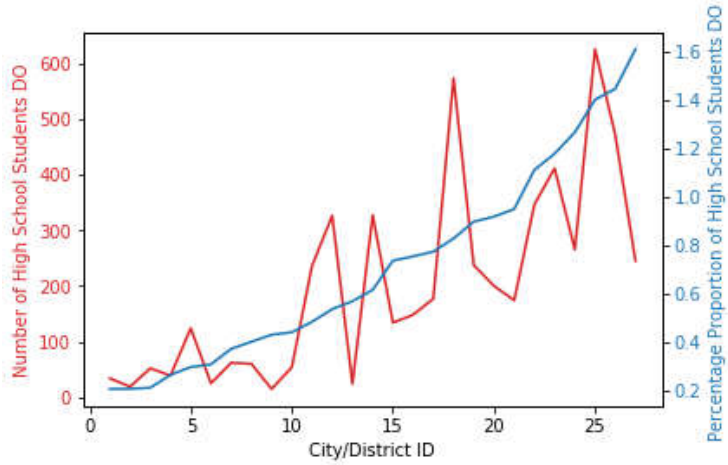
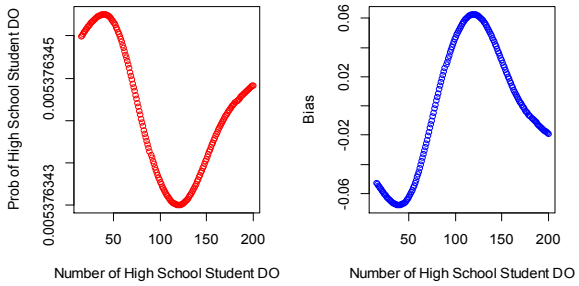
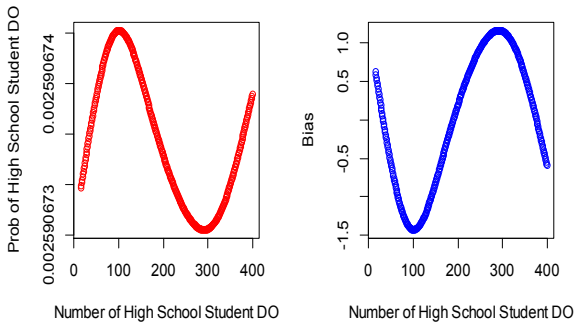


Figure 6 Number of high school students dropout and percentage proportion of high school students DO in West Java 2018

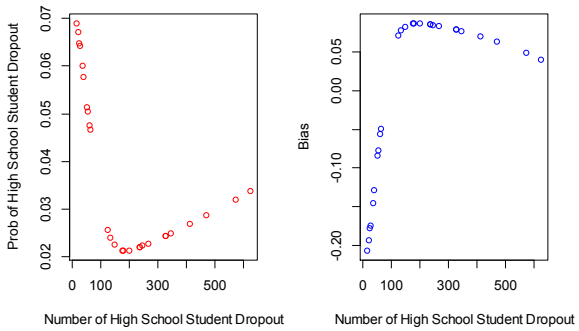
From Figures 5 and 6, Kab Garut has the highest number of high school student dropouts, i.e., 625 students from a total of 44,609 high school students, but the highest proportion of the number of high school students dropout is 0.0161 in Kab Purwakarta. Based on the percentage proportion of high school students drop out in Figure 8, we can see that the number of high school dropout students (W_i) has Poisson characteristics. Because of the data taken from a survey, we believe that the true number of high school students dropout data (X_i) measured with error and W_i as the realization of the exact number of high school students dropout. We assumed the number of high school students dropout has a Poisson distribution with the Poisson parameter has prior distribution. We estimated the probability of the true number of high school students drop out $g(x_i) = \mathrm{P}(X = x_{(i)})$ using the EBD method for the various number of high school students drop out in a support set. We choose three finite discrete support sets based on range value of W_i by EBD method in DeconvolveR package as in Figure 7.



(a) $T = [15,16,17,...,200]$



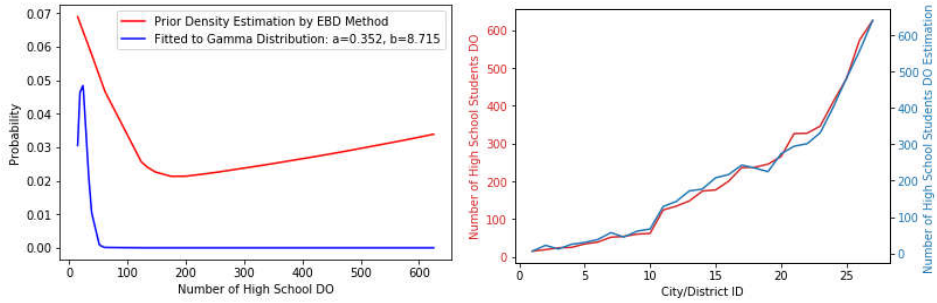
(b) $T = [15,16,17,...,400]$



(c) $T = [15,19,24,...,625]$

Figure 7 (a)-(c) Probability estimation of high school students dropout in West Java and its bias using EBD method for 3 types of discrete support set

Based on prior probability estimation for every city and district in support set $T = [15,19,24,...,625]$, we fitted that prior probability estimation into pdf of gamma distribution because gamma distribution is a conjugate prior for Poisson distribution. We also estimated the parameter of gamma distribution, and we got shape parameter estimation = 0.352 and scale parameter estimation = 8.715. We compared the plot of prior density estimation $\hat{g}(x)$ from the EBD method and fitted pdf plot of gamma distribution based on W_i in Figure 8(a). In Figure 8(b), we estimated the number of high school student dropouts for 27 cities/districts in West Java 2018 with prior parameter distribution of Poisson was fitted gamma distribution, and it was very close to W_i implying the suitability of the obtained prior.



(a) Prior density estimation vs. fitted gamma pdf (b) Estimation of number of high school DO

Figure 8 (a)-(b) Number of high school students dropout in West Java and its estimation using gamma prior probability estimation based on EBD method

5. Conclusions

In this paper, we presented the effect of selecting the discrete support set for estimating prior gamma density function based on the empirical Bayes deconvolution method (Efron 2016). From some simulation scenario, if we choose a broader discrete set for the domain of gamma density, the estimation of gamma's prior density will look like the specific gamma density function, but still, underestimate or overestimate in initial discretization points. However, after generated 1,000 datasets and did the same scenario, we computed standard deviation and bias for gamma's prior density estimation in a short and broader discretization set. We conclude that bias will have a smaller value for a larger discretization set of domain gamma density. As a real application, we implemented the EBD method for estimating the prior probability for the number of high school students drop out in West Java 2018. Then we also compute an estimation of the number of high school student dropouts in West Java. However, in high school student's dropout data, we do not consider or include any covariate that can use as possible predictive or explanatory variable of the number of high school students dropout. In the next research, we will investigate the underestimate or overestimate problem for prior density estimation in some initial point in support set based on empirical Bayes deconvolution, not only for conjugate prior for Poisson distribution.

Acknowledgments

The first authors thank LPDP and KemendikbudDikti for Beasiswa Unggulan Dosen Indonesia (BUDI), which financially supported the study.

References

Ajaja OP. School dropout pattern among senior secondary schools in Delta State, Nigeria. Int Educ Stud. 2012; 5(2): 145-153.

Buonaccorsi JP. Measurement error: models, methods, and applications. Portland: Chapman & Hall/CRC; 2010.

Butucea C, Comte F. Adaptive estimation of linear functionals in the convolution model and applications. Bernoulli. 2009; 15: 69-98.

Carroll RJ, Hall P. Optimal rates of convergence for deconvolving a density, J Am Stat Assoc. 1988; 83(404): 1184-1186.

Carroll RJ, Hall P. Low order approximations in deconvolution and regression with errors in variables, J R Stat Soc Series B Stat Methodol. 2004; 66(1): 31-46.

Delaigle A, Hall P. Methodology for non-parametric deconvolution when the error distribution is unknown. J R Stat Soc Series B Stat Methodol. 2016; 78(1): 231-252.

Efron B. Empirical Bayes Deconvolution estimates. Biometrika. 2016; 103(1): 1-20.

Fan J. On the optimal rates of convergence for nonparametric deconvolution problems. Ann Stat. 1991; 19(3): 1257-1272.

Fuller WA. Measurement error models. New York: John Wiley & Sons; 1987.

Hall P, Meister A. A ridge-parameter approach to deconvolution. Ann Stat. 2007; 35(4): 1535–1558.

Laird N. Nonparametric maximum likelihood estimation of a mixed distribution. J Am Stat Assoc 1978; 73(364): 805-811.

Meister A. Deconvolution problems in nonparametric statistics. New York: Springer; 2009.

Sang AKA, Koros PKA, Bosire JN. An analysis of dropout levels of public secondary schools in Kericho district in relation to selected school characteristics. Int Educ Stud. 2013; 6(7): 247-259.

Setyadharma A, Engelbrecht HJ, Balli HO. Analysis of upper secondary school dropout in Central Java Province, Indonesia: Preliminary Results and Insights, School of Economics and Finance, Massey University; 2018.

Stefanski LA, Carroll RJ. Deconvolving kernel density estimators. Statistics. 1990; 21(2): 169-184.