Bayesian Nonparametric Model Fitting and Model Comparison on Heavy Tailed Data

Abstract

Standard Bayesian analysis requires a family of parametric priors to be specified when fitting models and making inferences. This assumption leads to a substantial amount of work studying the uncertainty and sensitivity of the proposed parametric prior. In nonparametric Bayesian analysis, we are not constrained to assuming a certain family of parametric distributions. Instead, we can put a prior on different families of distributions, which is regarded as a flexible alternative to the standard Bayesian method. In this work, we estimate the density of a heavy-tailed data set from Deloitte company using a nonparametric density estimation method. It is believed that the total sum of the recorded values in an audit book is closely related to the level of the misstatement if we can model the summation of true values behind them. Obtaining a reasonable density estimation of the true values will further assist us in detecting misstatements in financial audit activities. We approach the density estimation problem by using stick-breaking priors, and develop two stick-breaking priors with different base measures. The connections and differences of those two priors are explored. We evaluate goodness of our fitting by comparing posterior predictive draws with our heavy tailed population data and compare sampling distributions of the inferential target, the sum of the true values, under various methods.

1 Introduction

A financial statement audit is defined as objective examination or evaluation of an organization's financial statements and aims to ensure fairness and accuracy of the financial records. The audit procedures are nowadays requirements for almost all companies and auditor's reports are also important references for bank and creditors before lending activities. As one part of the audit process, an auditor will examine and verify numbers recorded in company's financial statements. However, these claimed numbers are sometimes not the true values of financial transactions and therefore affect the auditor's opinion on whether material misstatements exist.

Material misstatement are those information that are sufficiently incorrect in financial statements. The occurrence of a material misstatement often has negative impact on other decisions that rely on these statements. In this project, we assume that each claimed value is greater or equal to the corresponding true value. This assumption is reasonable if we think about reporting the property's damage to the insurance company. The calimed (reported) damage, quantified in dollars, is very unlikely to be less than the true damage. And a higer claimed value may lead to a higher reimbursement. With this assumption, we can further quantify the misstatement (error) in the audit table with n records as $\sum_{i=1}^{n} C_i - \sum_{i=1}^{n} T_i$, where C represents the claimed values and T represents the true values.

One obligation for an auditor is to be able to detect material misstatement given a table filled with

the claimed values C. By modeling the true values T, we expect to get reasonably good density estimation $\hat{f}(T)$ of T. Since the summation of true values T is in part of the misstatement error, learning the distribution of T is a key step in the analysis of misstatement in audit. But estimating the density of true values T requires some techniques since true values are very "imbalanced" and heavy-tailed. The true values, in general, are financial data quantified in dollars. For example, a big company may have acquisition costs both for buying office desks and purchasing real estates or even for takeover of other companies that will highly likely to be quite large. And those types of transactions with larger values describe substantial amounts of purchasing activities in one company. One challenge in modeling imbalanced data in our case is that we have just a few data (those corresponds to the extreme large true values) through out the density estimation process. But we do need to come up with a good density estimator that takes care of the right tail since those values in the right tail affect the most in $\sum_{i=1}^{n} T_i$.

This project concentrates on modeling the true value T by estimating its density. If one learns the distribution of T well then in the future by incorporating some model $P(C|T)$, in reality the auditor can make inference on $P(T|C)$ when only the claimed values C is observed or partially observed. Finally the auditor can make inference on $p((\sum C - \sum T) > M)$, where M is a threshold for material misstatement. Given a heavy tailed dataset, the standard kernel density estimator on the data has its own limitations (Buch-Larsen et al. 2006). It will have under-smoothing effects at the tail and provide with an inaccurate estimate. Furthermore, classical kernel density estimation cannot adaptively fit the data since it uses a fixed bandwidth parameter in the estimation. We will approach the density estimation problem by using nonparametric bayesian methods to estimate the "heavy-tailed" density, in which mixture models in nonparametric bayes can be flexible enough and yield adaptive bandwidth for density estimation. One method is Dirichlet process mixture model with stick breaking priors proposed by Ishwaran and James (Ishwaran and James 2001) . We implement efficient MCMC algorithms for sampling the posteriors and then make inference on our predictive density (Ishwaran and James 2001).

In this work, we explore and evaluate how much improvements in estimating the density we obtained by using adpative bandwidth. For prediction purpose, we compare the predictions from different models and evaluate the prediction performance. After we calculate some empiricl risks and compare the sampling distribution of the point estimates, our finding is that when we include some prior knowledge for the right boundary of the data, the prediction procedure using nonparametric bayes method can be a reliable approach for predicting the sum of true values T.

In Section 2, we explore some features in our data. In section 3, we briefly review nonparametric bayesian density estimation and stick breaking priors. In section 4, we introduce DP mixture models with two priors, namely normal-gamma prior and t-betaprime prior. And we compare these two priors. Section 5 includes posterior inference we need to implement the blocked gibbs sampler. Section 6 describe the details for choosing different configurations of the hyperparameters in the priors. Section 7 shows the density estimation results and the prediction risks for all different models considered in this paper. Section 8 concludes this project.

2 Explore The Data and Kernel Density Estimator

2.1 Heavy-tailedness and multi-modality

Our dataset can be described as a collection of univariates that are nonnegative and have the characteristic of being heavy-tailed. To visualize our data, we can draw a histogram of the raw data and also a density plot on the log-transformed data. It is also not hard to notice from the plot that there's also multi-modality in our data. The multi-modality phenomenon requires our density estimation method not only to capture the overall "shape" of the distribution density but also to capture the curvature between modes. Multi-modality also suggests us to have multiple components in the density estimation when doing a mixture model.

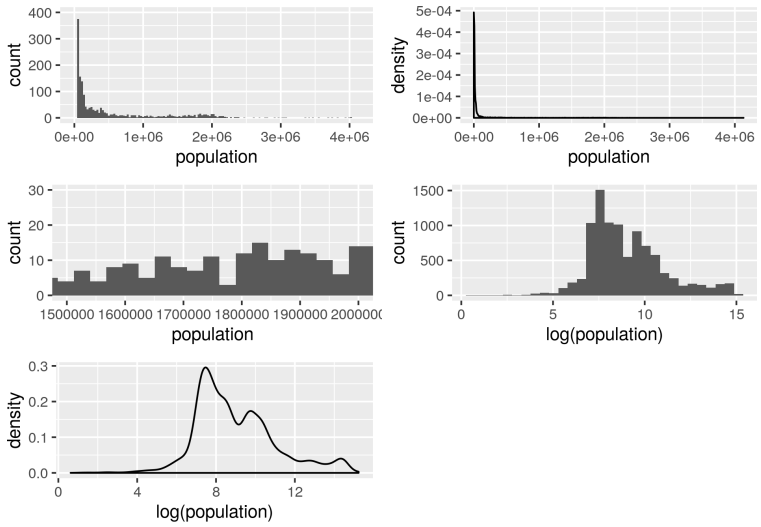


Figure 1: The regular histogram and zoomed-in histogram are intended to show that our data is heavy-tailed. The density plot is intended to show multi-modality and undersmoothed effect in the right tail.

2.2 Kernel Density Estimator

Till now, we have learned some features of our data and in terms of density estimation, perhaps the most well-known method is kernel density estimator (KDE). A gaussian kernel density estimator estimates the density function $f(x)$ by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\phi(\frac{x - x_i}{h}),$$

where h is the bandwidth and ϕ is the gaussian kernel. The guassian KDE fits a normal distribution with weights $\frac{1}{n}$ at each data point x_i and the optimal bandwidth is choosen so that the expected integrated squared error is minimized (Shalizi 2012). Essentially, finding the "optimal" bandwidth is equivalent to find a "banlanced" point in the bias and variance tradeoff. KDE will work quite well for most of the time but it may fail under some special circumstances. KDE can undersmooth the right tail if the data is heavy-tailed and very imbalanced, see an example of suicide data

(Silverman 1986). In addition to choosing a bandwidth, we can also choose the functional form of the kernel in KDE to optimize the fitting of KDE. But it is believed that the bandwidth h is more important than the choice of kernels (Scott 2015).

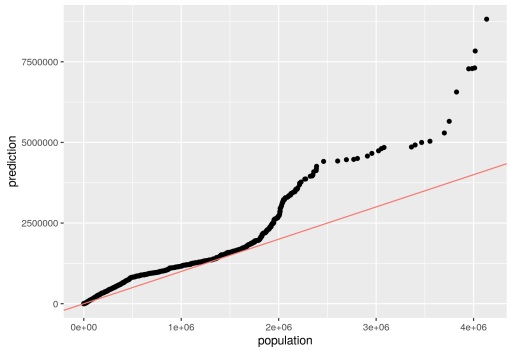


Figure 2: The QQ-plot is about predictions from KDE (using lognormal kernel) versus the population. It shows that using a single optimal bandwidth, KDE does not fit well in the right tail.

2.3 Mixture Model

Gaussian KDE can be seen as a specific density estimate of a mixture model where for each data point, there is a normal distribution centered at that data point. A more general form of a normal mixture model is

$$f(x|\pi,\theta) = \sum_{i=1}^{m} \pi_i \phi(x|\theta_i),$$

where m is the toal number of mixture components, π_i's are the mixing probabilities and θ_i's are the parameters of gaussian kernel. To fit such mixture model, we need to decide how many components in the mixture model and their weights and the component-specific parameters. In practice, we either know how many clusters that our data intrinsically have a priori, for instance when we fit a density estimate for a dataset grouped by sex, or we need to decide how many components we want to add to the mixture model. Since there's no natural grouping structure in our data, we take a nonparametric approach to dealing with the number of components in the mixture model. Namely, we do not assume a finite number of components in advanced but rather assume infinite number of components in the mixture. The nonparametric approach will not require us to specify total number of components but let the data speak. We will also take the bayesian approach since we want our method to be included into a bigger bayesian frame. A mixture model in bayesian requires us to assign priors to (π, θ) and the nonparametric bayes density estimation will be further discussed in section 3. A single "optimal" bandwidth in KDE is not good enough to fit the right tail of our heavy-tailed dataset. Therefore, we assume that each component has its own parameters to give adaptiveness to the mixture components. This will be further discussed in section 4.

3 Nonparametric Bayesian Density Estimation

Let $X_1, ..., X_n \sim F$ where F has density f. To estimate f, we can approach this problem by using the two different priors discussed above. One example is to use Dirichelet process (DP) prior for G where the problem is described as

$$X_i|G \overset{iid}{\sim} G,$$
$$G|\alpha, H \sim DP(\alpha H).$$

Since a prior for probability measure G requires infinite dimensional parameters, this falls into the category of bayesian nonparametric problems.

However, DP prior is sometimes undesirable since it generates discrete random probability measure with probability one (Teh et al. 2006). To have an estimator that yields continuous distribution but still involves DP prior, we can use continuous kernels and fit a Dirichelet process mixture model.

In this section, we describe bayesian nonparametric mixture models with stick-breaking priors. As discussed by Ishwaran and James (Ishwaran and James 2001), $\mathcal{P}_N(\boldsymbol{\nu}, \boldsymbol{\omega})$ includes both DP and Pitman-Yor process. When $\nu_i = 1$ for every i and $\omega_i = \alpha$ for every i, the stick breaking process is consistent with the constructive definition of $DP(\alpha H)$ proposed by Sethuraman (Sethuraman 1994). To see the connection between $\mathcal{P}_N(\boldsymbol{\nu}, \boldsymbol{\omega})$ and the Pitman-Yor process (a.k.a two-parameter Poission-Dirichelet process), if we define each $\nu_k = 1 - a$ and $\omega_k = b + ka$, where $0 \leq a < 1$, $b > -a$ and $k = 1, ..., N$, we finally get a random probability measure constructed by stick breaking process with two parameters a and b. Then, under this condition, $\mathcal{P}_N(\boldsymbol{\nu}, \boldsymbol{\omega})$ is exactly the Pitman-Yor process $\mathcal{PY}(a, b)$.

3.1 Dirichelet Process Mixture Model

To fix notation, $\mathbf{X} = (X_1, ..., X_n)$ is the data in our models, which are the log-transformed version of the true values T. We use normal kernel as our smooth kernel in the DP mixture model and the mixture model is specified as follow,

$$X_i|\boldsymbol{\theta}, \boldsymbol{K} \overset{ind}{\sim} N(\theta_{k_i}), \quad where\ \theta_{K_i} = (\mu_{k_i}, \phi_{k_i}^{-1}),$$

where k_i serves as an indicator of which group, within the mixture, X_i belongs to and $\phi_{k_i}^{-1}$ is the precision.

The DP prior for such model can be specified as

$$\theta_i|G \overset{iid}{\sim} G, \quad G \sim DP(\alpha G_0),$$

where α is the concentration parameter and G_0 is the base measure in DP process. Under this model, a Polya urn Gibbs sampling method is developed by Escobar and West (Escobar and West 1994) by marginalizing out G and use the Polya urn prediction rule to update θ_i's one coordinate at a time. However, this algorithm suffers from slow mixing.

The blocked gibbs sampler was developed by Ishwaran and James that has better mixing properties than Polya urn gibbs sampler. It also avoids marginalizing our prior G when sample from the

posterior. The caveat is that we are only allowed to specify a finite N for constructing $\mathcal{P}_N(\boldsymbol{\nu}, \boldsymbol{\omega})$ measures for using blocked gibbs sampler. This truncated version of stick breaking priors forms an approximation of $\mathcal{P}_\infty(\boldsymbol{\nu}, \boldsymbol{\omega})$ and some technical details about the approximation are discussed by Ishwaran et al. Here, we differentiate two different stick-breaking priors according to the base measure in the DP process. One is the normal-gamma conjugate prior for normal models when the mean and precision are both unkown. The other is a t-betaprime prior which should be considered as a more flexible alternative to the normal-gamma prior. The differences will be demonstrated afterwards with further details.

3.2 Stick Breaking Priors

Ishwaran and James (Ishwaran and James 2001) defined the stick breaking priors to be almost surely discrete random probability measures $\mathcal{P}$ that can be represented as

$$\mathcal{P}(\cdot) = \sum_{k=1}^{N} p_K \delta_{Z_k}(\cdot), \quad 1 \leq N \leq \infty, \tag{1}$$

where $\delta_{Z_k}(\cdot)$ denotes the Dirac measure at Z_k. The p_k's are random probability weights independent from Z_k's and Z_k's are i.i.d random variables from some nonatomic distribution H. The way for constructing random probability weights, $\boldsymbol{p} = (p_1, ..., p_N)$, is defined as

$$p_1 = V_1$$

$$p_i = V_i \prod_{j=1}^{i-1} (1 - V_j), \quad i = 2, ..., N$$

$$V_i \sim Beta(\nu_i, \omega_i), \quad i = 1, ..., N.$$

If N is finite, we set $V_N = 1$ to guarantee that $\sum_{k=1}^{N} p_k = 1$. This process derives its name from an analogy to iteratively breaking the proportion V_i from the remainder of the unit-length stick, $\prod_{j=1}^{i-1}(1 - V_j)$ and set stick breaking priors apart from general random probability measures (Ishwaran et al. 2001). The flexibility of stick breaking priors is achieved by providing different $\boldsymbol{\nu} = (\nu_1, \nu_2, ...)$ and $\boldsymbol{\omega} = (\omega_1, \omega_2, ...)$. Ishwaran et al. (2001) called the stick breaking priors $\mathcal{P}$ as $\mathcal{P}_N(\boldsymbol{\nu}, \boldsymbol{\omega})$ measures and this notation connects various probability measures including the Dirichlet process (Ferguson 1973) and Pitman-Yor process (Pitman and Yor 1997).

4 Two Priors in Dirichelet Process Mixture Model

4.1 Mixture model with normal-gamma prior

$$X_i | \boldsymbol{\theta}, \boldsymbol{K} \overset{ind}{\sim} N(\theta_{k_i}), \quad where\ \theta_{K_i} = (\mu_{k_i}, \phi_{k_i}^{-1})$$

$$\mu_{k_i} | \mu_0, \kappa_0, \phi_{k_i} \sim N(\mu_0, (\kappa_0 \phi_{k_i})^{-1})$$

$$\phi_{k_i} | v_0, ss_0 \sim gamma(v_0, ss_0)$$

$$\mu_0 \sim N(a, \sigma_0^2), \quad \kappa_0 \sim gamma(g_1, g_2)$$

$$\boldsymbol{\pi} \sim \mathcal{SB}(1, \alpha), \quad \boldsymbol{K} \sim Multi(\boldsymbol{\pi})$$

$$\alpha \sim gamma(c, d)$$

4.2 Mixture model with t-betaprime prior

$$X_i|\boldsymbol{\theta}, \boldsymbol{K} \overset{ind}{\sim} N(\theta_{k_i}), \quad where\ \theta_{K_i} = (\mu_{k_i}, \phi_{k_i}^{-1})$$
$$\mu_{k_i}|\mu_0, \kappa_0, \phi_{k_i}, r_{k_i} \sim N(\mu_0, (\kappa_0 \phi_{k_i} r_{k_i})^{-1})$$
$$r_{k_i} \sim gamma(\frac{df}{2}, \frac{df}{2})$$
$$\phi_{k_i}|v_0, ss_0, h_{k_i} \sim gamma(v_0, ss_0 h_{k_i})$$
$$h_{k_i}|v_1 \sim gamma(v_1, v_1)$$
$$\mu_0 \sim N(a, \sigma_0^2), \quad \kappa_0 \sim gamma(g_1, g_2)$$
$$\boldsymbol{\pi} \sim \mathcal{SB}(1, \alpha), \quad \boldsymbol{K} \sim Multi(\boldsymbol{\pi})$$
$$\alpha \sim gamma(c, d)$$

4.3 Comparison of two priors

Prior information			
Prior	Conditionals	prior mean	prior variance
Normal-gamma(Constrained)	$(\mu_{k_i}\|\mu_0, \kappa_0, \phi_{k_i})$	μ_0	$\frac{1}{\kappa_0 \phi_{k_i}}$
	$(\mu_{k_i}\|\mu_0, \kappa_0, v_0, ss_o)$	μ_0	$\frac{ss_0}{\kappa_0(v_0-1)}$
	$(\sigma_{k_i}^2\|v_0, ss_0)$	$\frac{ss_0}{v_0-1}$	$\frac{ss_0^2}{(v_0-1)^2(v_0-2)}$
T-betaprime(Flexible)	$(\mu_{k_i}\|\mu_0, \kappa_0, \phi_{k_i}, df)$	μ_0	$\frac{1}{\kappa_0 \phi_{k_i}} \times \frac{df}{df-2}$
	$(\mu_{k_i}\|\mu_0, \kappa_0, v_0, ss_o, df, v_1)$	μ_0	$\frac{ss_0}{\kappa_0(v_0-1)} \times \frac{df}{df-2}$
	$(\sigma_{k_i}^2\|v_0, v_1, ss_0)$	$\frac{ss_0}{v_0-1}$	$\frac{ss_0^2}{(v_0-1)^2(v_0-2)} \times \frac{v_1+v_0-1}{v_1}$

Table 1: Prior mean and variance for the conditional distributions in both models.

From the table, we see that constrained model and flexible model, under above model specification, shares the same prior mean for μ_{k_i} and $\sigma_{k_i}^2$, but with flexible model having larger prior variance under certain conditions. On one hand, if $2 < df < \infty$, prior variance for μ_{k_i} in flexible model is larger than constrained model, but as $df \to \infty$, prior variance for mean parameter in flexible model converges to that in constrained model. On the other hand, $\frac{v_1+v_0-1}{v_1}$ is strictly greater than 1 if and only if $v_0 > 1$ and as $v_1 \to \infty$ that quantity goes to 1, which make equal prior variances for $\sigma_{k_i}^2$ in both models.

To understand the role of κ_0 in the prior, we notice that $E((\mu_{k_i} - \mu_0)^2|\mu_0, \kappa_0, \phi_{k_i}) = \frac{\sigma_{k_i}^2}{\kappa_0}$, which means that κ_0 is about the size of $\frac{\sigma_{k_i}^2}{(\mu_{k_i}-\mu_0)^2}$.

5 Posterior Inference

Before describing the blocked gibbs sampler, it is helpful to first figure out the conjugacy in DP mixture model. First, notice that the base measure G_0 in the constrained prior is conjugate to normal distribution in the mixture component. Second, generalized Dirichelet prior assigned to $\mathbf{P}$ is conjugate to the multinomial sampling model. Third, gamma prior on α is conjugate to the $Beta(1, \alpha)$ sampling model for V_i. These conjugate priors are by no means the only choice. But they are helpful for yielding easy-to-implement gibbs sampler.

For each iteration in the blocked gibbs sampler, define $\mathbf{K}^* = \{K_1^*, ..., K_m^*\}$ to be the set of current m unique values of $\mathbf{K}$ and we draw values from full conditionals of each random variable. Below are the Gibbs sampler for the constrained model.

1. Conditionals for $\boldsymbol{\theta}$: Let $\{K_1^*, ..., K_m^*\}$ denote the set of m unique values of $\mathbf{K}$. Simulate $\theta_k \overset{iid}{\sim} G_0$ for each $k \in \mathbf{K} - \{K_1^*, ..., K_m^*\}$. Also for $j = 1, ..., m$, draw $(\theta_{K_j^*} | \boldsymbol{K}, \mu_0, \kappa_0, \boldsymbol{X})$ from

$$f(\theta_{K_j^*} | \boldsymbol{K}, \mu_0, \kappa_0, \boldsymbol{X}) \propto G_0(d\theta_{K_j^*}) \times \prod_{\{i:K_i = K_j^*\}} f(X_i | \theta_{K_j^*}, \mu_0, \kappa_0) \tag{2}$$

   Without loss of generality, denote $n^{(\ell)} = |i : K_i = K_\ell^*|$ to be the size of the set. Under the ℓth block and by the conjugate result, we have again a normal-gamma posterior for the mean parameter $\mu^{(\ell)}$ and precision parameter $\phi^{(\ell)}$. To summarize,

$$\mu^{(\ell)}, \phi^{(\ell)} | \boldsymbol{K}, \mu_0, \kappa_0, \boldsymbol{X} \sim NG(\mu_0^{(\ell)}, \kappa_0^{(\ell)}, v_0^{(\ell)}, ss_0^{(\ell)}) \tag{3}$$

$$\mu_0^{(\ell)} = \frac{\kappa_0 \mu_0 + n^{(\ell)} \bar{x}^{(\ell)}}{\kappa_0 + n^{(\ell)}}, \qquad where \quad \bar{x}^{(\ell)} = \frac{\sum_{\{i:K_i = K_\ell^*\}} x_i}{n^{(\ell)}}$$

$$\kappa_0^{(\ell)} = (\kappa_0 + n^{(\ell)})^{-1}$$

$$v_0^{(\ell)} = \frac{v_0 + n^{(\ell)}}{2}, \qquad ss_0^{(\ell)} = \frac{ss_0 + \sum_{\{i:K_i = K_\ell^*\}} (x_i - \bar{x}^{(\ell)})^2}{2} + \frac{\kappa_0 n^{(\ell)} (\bar{x}^{(\ell)} - \mu_0)^2}{2(\kappa_0 + n^{(\ell)})}$$

2. Conditionals for $\mathbf{K}$: for each $i = 1, ...n$, draw K_i from

$$(K_i | \boldsymbol{\theta}, \boldsymbol{P}, \mu_0, \kappa_0, \boldsymbol{X}) \overset{ind}{\sim} \sum_{k=1}^{N} p_{k,i} \delta_k(\cdot) \tag{4}$$

$$(p_{1,i}, ..., p_{N,i}) \propto (p_1 f(X_i | \theta_1, \mu_0, \kappa_0), ..., p_N f(X_i | \theta_N, \mu_0, \kappa_0)) \tag{5}$$

3. Conditionals for $\mathbf{P}$:

$$V_k^* | \alpha, \boldsymbol{K} \overset{ind}{\sim} Beta(1 + M_k, \alpha + \sum_{l=k+1}^{N} M_l), \quad k = 1, ..., N-1 \tag{6}$$

$$p_1 = V_1^*, \quad p_k = V_k^* \prod_{i=1}^{k-1} (1 - V_i^*), \tag{7}$$

   where M_k is the number of K_i values that equal k.

4. Conditionals for α:

$$\alpha|\boldsymbol{V} \sim Gamma(c+N-1, d+\sum_{i=1}^{N-1} ln(1-V_i)) \tag{8}$$

5. Conditionals for μ_0 and κ_0:

$$f(\mu_0|\boldsymbol{\theta}, \boldsymbol{K}, \boldsymbol{X}, \kappa_0) \propto \pi(d\mu_0)p(\boldsymbol{\theta}|\mu_0, \kappa_0) \tag{9}$$

$$f(\kappa_0|\boldsymbol{\theta}, \boldsymbol{K}, \boldsymbol{X}, \mu_0) \propto \pi(d\kappa_0)p(\boldsymbol{\theta}|\mu_0, \kappa_0) \tag{10}$$

Under blocked gibbs sampler, use $\boldsymbol{\Pi} = (\boldsymbol{\theta}, \boldsymbol{K}, \boldsymbol{P}, \alpha, \mu_0, \kappa_0)$ to denote parameters in the joint posterior and the Bayesian density estimation of the DP mixture model is given by

$$f(X_{n+1}|X_{1:n}) = \int f(X_{n+1}|\boldsymbol{\Pi})dP(\boldsymbol{\Pi}|X_{1:n}) \tag{11}$$

The posterior inference under mixture model with flexible prior is similar to the procedure mentioned above but with some additional Gibbs sampling steps. For simplicity, we will only list those additional steps and the remaining steps should be the same as the previous sampling scheme.

1. Conditionals for $h_{k_j^*}$:

$$(h_{k_j^*}|v_0, ss_0, v_1, \phi_{k_j^*}) \sim gamma(v_0+v_1, ss_0\phi_{k_j^*}+v_1) \tag{12}$$

2. Conditionals for $\mu_{k_j^*}$:

$$(\mu_{k_j^*}|\boldsymbol{K}, \boldsymbol{X}, \phi_{k_j^*}, \kappa_0, r_{k_j^*}) \sim N\left(\frac{\phi_{k_j^*}n^{(j)}\bar{x}^{(j)}+\kappa_0\phi_{k_j^*}r_{k_j^*}\mu_0}{\phi_{k_j^*}n^{(j)}+\kappa_0\phi_{k_j^*}r_{k_j^*}}, (\phi_{k_j^*}n^{(j)}+\kappa_0\phi_{k_j^*}r_{k_j^*})^{-1}\right) \tag{13}$$

3. Conditionals for $\phi_{k_j^*}$:

$$(\phi_{k_j^*}|\mu_{k_j^*}, h_{k_j^*}, v_0, \mu_0, \kappa_0) \sim gamma\left(\frac{n^{(j)}+2v_0+1}{2}, r\right) \tag{14}$$

where $r = \frac{1}{2}(\sum_{\{i:k_i=k_j^*\}}(x_i-\mu_{k_j^*})^2 + (\mu_{k_j^*}-\mu_0)^2\kappa_0 r_{k_j^*} + 2ss_0h_{k_j^*})$.

4. Conditionals for $r_{k_j^*}$:

$$(\kappa_{k_j^*}|\phi_{k_j^*}, df, \mu_0, \mu_{k_j^*}, \kappa_0) \sim Gamma\left(\frac{df+1}{2}, \frac{df+\kappa_0\phi_{k_j^*}(\mu_{k_j^*}-\mu_0)^2}{2}\right) \tag{15}$$

5. Conditionals for κ_0:

$$(\kappa_0|g_1, g_2, \mu_0, \boldsymbol{\phi}, \boldsymbol{r}) \sim Gamma\left(\frac{N+2g_1}{2}, \frac{2g_2+\sum_{i=1}^{N}\phi_i r_i(\mu_i-\mu_0)^2}{2}\right) \tag{16}$$

6. Conditionals for μ_0:

$$(\mu_0|\boldsymbol{\mu}, \boldsymbol{r}, \boldsymbol{\phi}, \kappa_0) \sim N\left(\frac{\kappa_0 \sum_{i=1}^{N} \phi_j r_j \mu_j + a{\sigma_0}^{-2}}{\kappa_0 \sum_{i=1}^{N} \phi_j r_j + {\sigma_0}^{-2}}, (\kappa_0 \sum_{i=1}^{N} \phi_j r_j + {\sigma_0}^{-2})^{-1}\right) \tag{17}$$

6 Data-based Simulation Design

The "acquisition value" dataset contains $n = 9483$ data points from which we separate the data into two parts. We use the training data (of sample size 500) as input to fit the mixture model and leave the holdout dataset for testing. Method of obtaining training samples are just simple random sampling. The experiments are done under multiple simple random samples (500 different samples), through which we get the sampling distribution of the summation of T. Heavy tailed characteristic can be observed both from the qqplot and density plot so that our data is extremely imbalanced with few data at the right tail.

Although the $DP(\alpha G_0)$ and stick-breaking process $\mathcal{P}_N(1, \alpha)$ are equivalent only if $N \to \infty$, for computational consideration, we need some finite N to do the fittings. Any finite N will induce a finite stick-breaking procedure which can be considered as an approximation to the infinite stick-breaking process. In our experiment, we fix $N = 50$. It is also worth mentioning that the stick-breaking process might be vulnerable to numerical underflow issue when generating beta random variables. To be more concrete, the underflow issue, related to our problem, can occur when generate some $beta(a, b)$ random variables with very small b values. Therefore we adopt a scheme where we first generate random log-gamma variables and then generate the target beta random variables. The algoritm developed by Marsaglia and Tsang (Marsaglia and Tsang 2000) is included in appendix A.

The data-based simulation experiment aims to choosing different but reasonable values for hyperparameters in both priors and to see its impact on our inferential target. We will choose hyperparameter values that guarantee the existence of the first and second moments in table 1. Some prior information tables are included in appendix B.

6.1 Design for Constrained Model

Constrained prior are parameterized by 8 hyperparameters in total. To be more specific, (v_0, ss_0) mainly control the prior for $\sigma^2_{k_i}$; (a, A) control grand mean μ_0; (g_1, g_2) controls κ_0; (c, d) control the concentration parameter α.

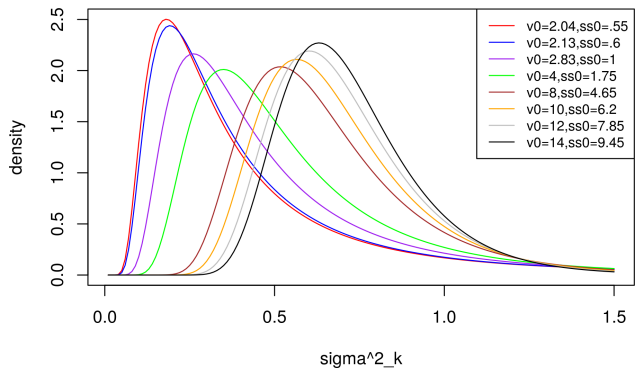
Parameter	(v_0, ss_0)	(g_1, g_2)	(a, A)	(c, d)
Value	(2.13, 0.6) (4, 1.75) (10, 6.2)	(1, 4) (0.5, 2)	(0, 100)	(3, 0.35) (11, 1)

Table 2: Design table for constrained model.

6.1.1 v_0 and ss_0

- To guarantee existence of prior variance of $(\sigma^2_{k_i}|v_0, ss_0)$, we need a $v_0 > 2$.

- We saw that almost all estimated $\sigma^2_{k_i}$'s are less than 1 from the finite mixture model fitting, and we will incorporate that knowledge into the prior $\sigma^2_{k_i} \sim invgamma(v_0, ss_0)$ as a form of tail event probability. To be more specific, we do not want to specify a prior for $\sigma^2_{k_i}$ that has $p(\sigma^2_{k_i} > 1) > 0.1$. The treshold probability 0.1 is subjective and varys from person to person.
- Multiple density curves for $(\sigma^2_{k_i}|v_0, ss_0)$



Above plot are the prior density plots for different paris of (v_0, ss_0)'s that yield roughly the same tail event probability $p(\sigma^2_{k_i} > 1) \approx 0.1$. v_0 is the main parameter controlling the tail shape for $(\sigma^2_{k_i}|v_0, ss_0)$ while ss_0 mainly takes charge the behavior around the origin. From the density plot, with an increasing sequence both for v_0 and ss_0, we observe that the prior becomes more light in tail and at the origin. Personally, I will take my three choices to be $(v_0 = 2.13, ss_0 = 0.6)$, $(v_0 = 4, ss_0 = 1.75)$, $(v_0 = 10, ss_0 = 6.2)$. Here are some reasons. First, I want the prior density does not decay too fast around the origin. Small in-group variances might correspond to the rare(large) values at the tail, which are hard to be grouped into a component largely formed by "small" values since the rare ones are far away from the centroid. Second, from the prior density plot, we may say that the choice of $(v_0 = 2.13, ss_0 = 0.6)$ and $(v_0 = 10, ss_0 = 6.2)$ yield curvatures that are at two ends of the "spectrum" of the prior density plot, while $(v_0 = 4, ss_0 = 1.75)$ serves as a middle case.

6.1.2 g_1 and g_2

- $\kappa_0 \sim gamma(g_1, g_2)$; I will mainly tune the shape parameter g_1 since for κ_0, we are more interested in those κ_0's around the origin, which is mainly controlled by g_1. Small values for κ_0 allow more adaptation for fitting those large μ_k's at the right tail since those μ_k's are more or less far away from the grand mean μ_0.
- Since κ_0 is about the size of $\dfrac{\sigma^2_{k_i}}{(\mu_{k_i} - \mu_0)^2}$, which can be considered as the square of the (translated) coefficient of variation, we obtain the plug-in estimates for that (translated) coefficient of variation from the finite mixture model fittings. It turns out that the estimated mean for κ_0 is 0.25, I will choose two sets of (g_1, g_2) that match that quantity.
- $g_1 = 1, g_2 = 4$ introduces mode at 0 and matches the estimated mean of κ_0 from the finite mixture model.

- $g_1 = 1/2, g_2 = 2$ serves as another choice that introduce more "sharp" asymptote around 0 but still have the prior mean of κ_0 to be 0.25, which matches the estimated mean. The purpose of reducing g_1 is to see whether more asymptote around the origin in the prior have "significantly big" impact on the fitting.

6.1.3 a and σ_0^2

- $\mu_0 \sim N(a = 0, \sigma_0^2 = 100)$; a somehow diffuse prior; since μ_0 lies in the second hierarchy in both models' prior (relatively far away from μ_k's and σ_k's), I think different priors for μ_0 will not have hugely different impact on the fittings. Also since we work on log-scaled data, a variance of 100 in the normal prior might be large enough for the grand mean μ_0.

6.1.4 c and d

- Conditioned on $N = 50$ (total number of mixture) and $n = 500$ (sample size) in the databased experiment, I investigate the role of concetration parameter α with prior $\alpha \sim gamma(c, d)$.
- Since we choose $N = 50$ to do the databased experiment, we may not want those α's that can induce too big N in the mixture model, since that situation may potentially hurt the fitting due to the argument that $N = 50$ is too small.
- One way to get how large total number of mixture will be under an α is to calculate the prior probability $P(N_{induced} > N = 50)$ in our case. In principle, we do not want this probability to be too big provided the user-specified N.
- It might also be useful to think about the DP mixture as the Chinese resturant process (Teh 2010) with number of customers $n = 500$ and concentration parameter α. To choose hyperparameter values for c and d in $\alpha \sim gamma(c, d)$, we first look at the distrubution of total number of nonempty tables obtained by Monte Carlo simulation.

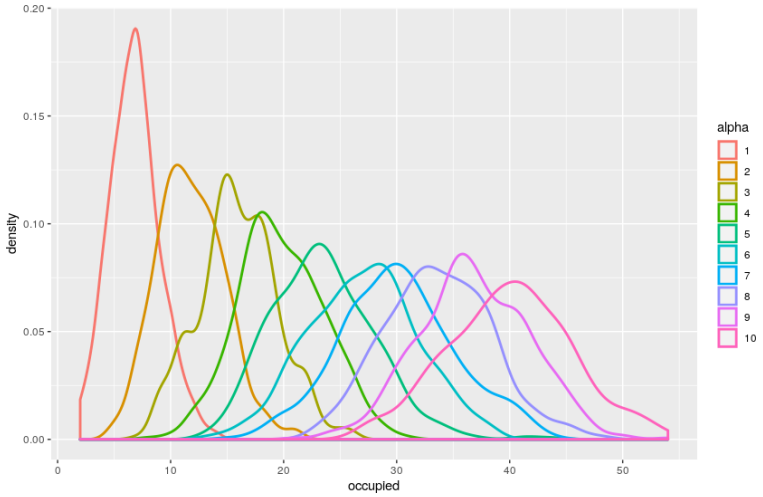
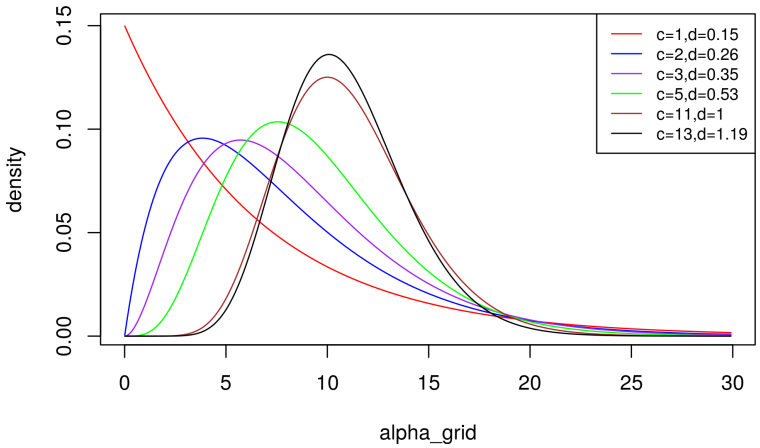


Figure 3: Density plots for number of occupied tables in the restraunt process with colors representing different values in alpha.

From above facet plot, we may conclude that α's that are greater than 15 are too big for $N = 50$ since the prior mode is already greater than 50. To choose values for c and d, we search for choices that yield roughly the same probability of $\alpha > 15$. In other words, we also don't want the α in prior to be bigger than 15 with "large" probability to have too many occupied components. The probability of the tail event, $p(\alpha > 15)$, is set to be 0.1. We choose pairs of (c, d)'s according to that tail event probability.

- Multiple density curves for α



At first, we may have a coarse classification of those priors to be whether a parior has mode 0. Obviously, the choice $(c = 1, d = 0.15)$ gives a prior whose mode is 0 while the other choices do not. I will try to aviod putting a prior for α with mode 0 since any small enough α may make the MCMC chain stuck. From the remaining density curves, I will choose two configurations $(c = 3, d = 0.35)$ and $(c = 11, d = 1)$.

6.2 Design for Flexible Model

In the prior of flexible model, we get two additional hyperparameters df and v_1. The common parameters between constrained and flexible models are keep the same for the purpose of comparability.

Parameter	(v_0, ss_0)	(g_1, g_2)	(a, σ_0^2)	(c, d)	df	v_1
Value	(2.13, 0.6)	(1, 4)	(0, 100)	(3, 0.35)	3	1.5
	(4, 1.75)	(0.5, 2)		(11, 1)	12	3
	(10, 6.2)				22	

Table 3: Design table for flexible model.

- v_1 can be regarded as the amplification factor for prior variance of $(\sigma_{k_i}^2 | v_0, ss_0)$. Since we've already choose values for v_0, v_1 can be calculated correspondingly provided how much the amplification we want for the prior variance. It maybe reasonable that we want that prior variance in flexible model to be mainly 2 times the prior variance in constrained model. Following this, we choose $v_1 = 1.5, 3$.

- $df = 3$ is the smallest interger value for making the flexible model indeed have larger prior variance (boundary case). Remaing choices are considered as ones that still make differences when comparing t-distribution and normal distribution.

7 Density Estimation and Prediction

7.1 Density Estimation

After we set up different prior configurations, we can compare the posterior predictives to the holdout population data. Goodness of fit can be checked by the qqplot. We can also obtain a sampling distribution for the population sum from bayesian mixture models and see whether the sampling distribution cover the true value of population sum. Total number of iterations for one MCMC chain is set to be 100,000.

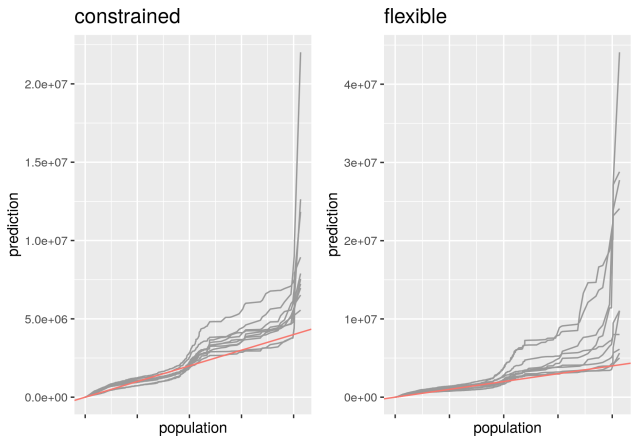


Figure 4: Two qqplots of holdout population data against posterior predictions; The red line is the diagonal reference line and grey lines are samples from the posterior predictive distribution. Although the qqplot is one drawn under one of the simulation configurations, the issue here is obvious. Both mixture models under constrained prior and flexible prior have the ability to do the data extrapolation, which may generate perdictions that are extremely large.

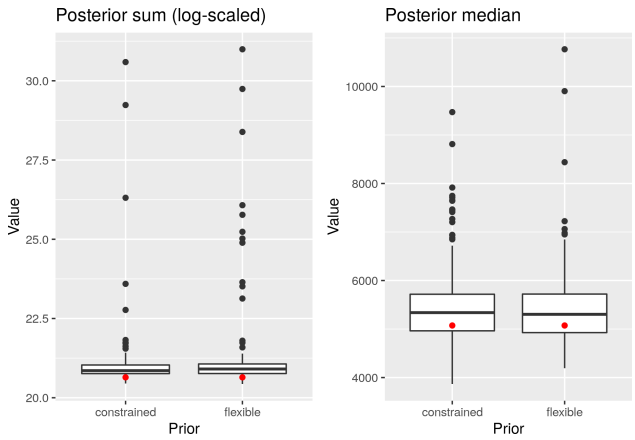


Figure 5: Box plots for the log of predictive sum and the predictive median. Red dots are the true value of the corresponding statistic from the holdout dataset. The predicted sum are heavyly affected by the extremely large predictions. But when we look at the boxplot for the predictive median, which is more robust to outliers compared to mean, the true value is actually clost to the median of those predictive medians.

Since both of our nonparametric mixture models aim to model the whole population, we may add another assumption about the population (true values of the statements) and the observations (claimed values of the statements). If we assume that the claimed value is always greater than its corresponding true value, then the upper bound of all the observables can be our prior information about the upper bound for the predictions, therefore allowing us to do the truncated posterior predictions. The purpose for doing truncated predictions is also to downplay the impact of predicted outliers.

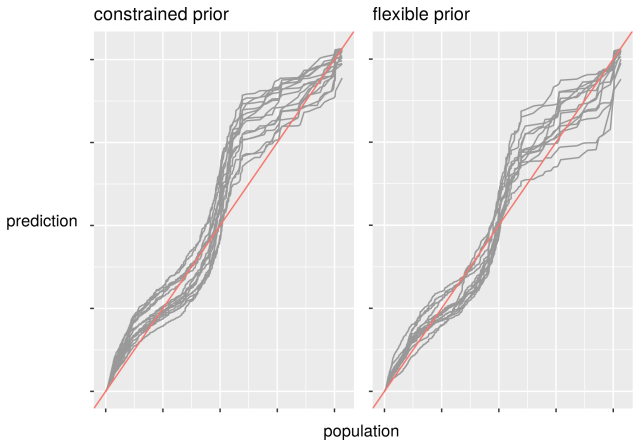


Figure 6: Two sample QQ plot comparison for fitting from different priors; X-axis represents hold-out population data; Y-axis represents predictions; Hyperparameter choice for the plot: $(v_0 = 10, ss_0 = 6.2, g_1 = 0.5, g_2 = 2, c = 3, d = 0.35, df = 3, v_1 = 1.5)$.

To evaluate the goodness of fit for each different prior configurations, we use the test statistic

related to the Kolmogorov–Smirnov test. We make use of our predictions and the holdout data to do a two sample K-S test. Two smaple K-S test measures the largest deviation among two empirical cumulative distribution function. Formally, the statistic D is defined as

$$D = \sup_x |\hat{F}_1(x) - \hat{F}_2(X)|,$$

where $\hat{F}_1$ and $\hat{F}_2$ are two empirical distribution.

Denote $d^{(j)}$ as the two sample K-S test statistic estimated at iteration j ($j = 1, ..., M$). Then a way to compare the goodness of fit between models is to use the arithematic mean $\bar{d}$ among iterations.

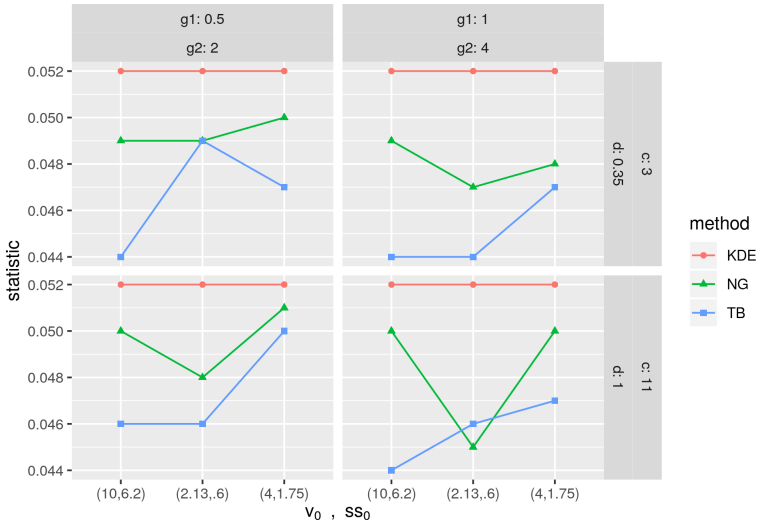


Figure 7: KS test statistic plot; A smaller test statistic value means better fit. NG stands for normal-gamma prior and TB stands for t-betaprime prior, df = 3, v1 = 1.5 for TB.

After we do the truncated prediction, the posterior predictions of the population are more prone to lie on the referenced diagonal line, although biases existed there. The choosen configuration for t-betaprime with $(df = 3, v_1 = 1.5)$ is the most "flexible" one in terms of prior information from table 1, which have the smallest degrees of freedom and largest prior variance inflation factor. If we increase df and v_1, we expect that the behavior of flexible priors will be more similar to that of constrained prior.

To compare methods, we also include the classic gaussian kernel density estimation, which is a method widely used in density estimation problem. Note that this frequntist method's estimation is fixed given the random sample, so that the K-S test statistics are all the same under the "kde" column. We can see from the K-S test statistic plot, nonparametric bayesian mixture models give us a better density estimation, although the actual difference in $\bar{d}$ is relatively small. That difference in terms of goodness of fit can be explained by the adaptivities in the component-specific variances introduced by the mixture model setup, while a kernel density estimator usually use a single bandwidth.

7.2 Prediction Performance

In this subsection, we evaluate the ability to predicting the population sum under different methodology. Besides the bayesian approach described before, there are also many different methods for predicting/estimating the total sum of a data set given that we can only observe part of the population data. Using the sample mean to estimate the population mean is valid by central limit theorem. Also if we know the total number of records in a financial book, estimating the sum is equivalent in estimating the mean. We can also fit a finite mixture model using the EM algorithm and have predictions of population sum from that.

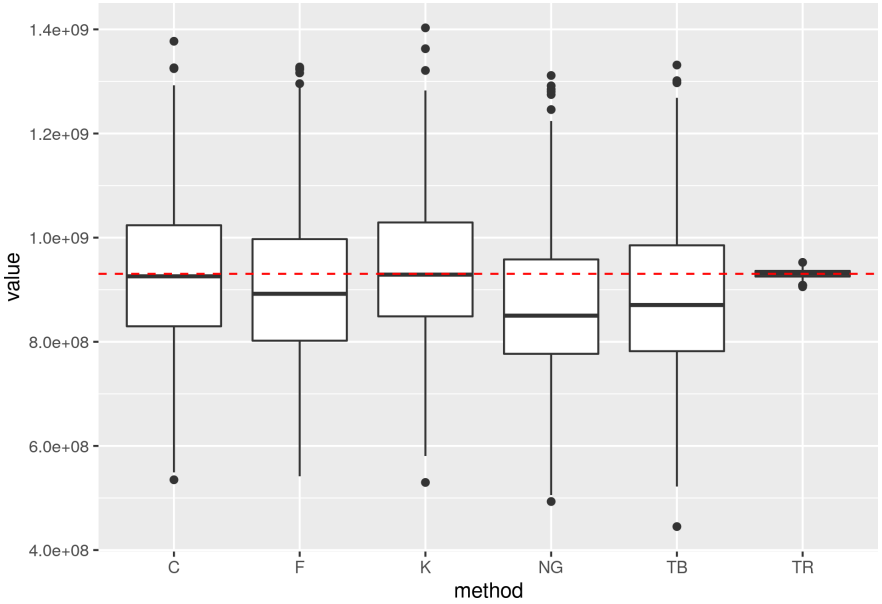


Figure 8: Boxplot for predictions of population sum. Lables on X-axis are as follows: T-betaprime(TB); Normal-gamma(NG); KDE(K); Central limit theorem(C); Finite mixture model(F); True value(TR). Hyperparameter choice for the plot: $(v_0 = 10, ss_0 = 6.2, g_1 = 0.5, g_2 = 2, c = 3, d = 0.35, df = 3, v_1 = 1.5)$. Red dashed line is the true value of the population sum.

Here besides the bayesian mixture model, every other methods are frequntist method and give us point estimates of the population sum. The sampling distribution of the point estimate is given by observing different random samples from the population. To evaluate the prediction procedure given by a nonparametric bayesian mixture models, we also use a point estimate of the predictive population sum given by the output from MCMC and use that to form another sampling distribution. The gibbs sampler for each configuration of prior is ran on 500 different simple random samples each with 100,000 iterations, through which we obtain 500 posterior means of the predictive population sums.

From the boxplot, we see that the point estimates for the sum from the nonparametric bayes method have relatively similar sampling distribution compared with sampling distribution of the sample mean. If we use the root mean square logarithmic error (RMSLE) to quantify the

prediction error, where RMSLE R for n predictions is defined as

$$R = \sqrt{\frac{\sum_{i=1}^{n}(log(pred^{(i)}) - log(true^{(i)}))^2}{n}}.$$

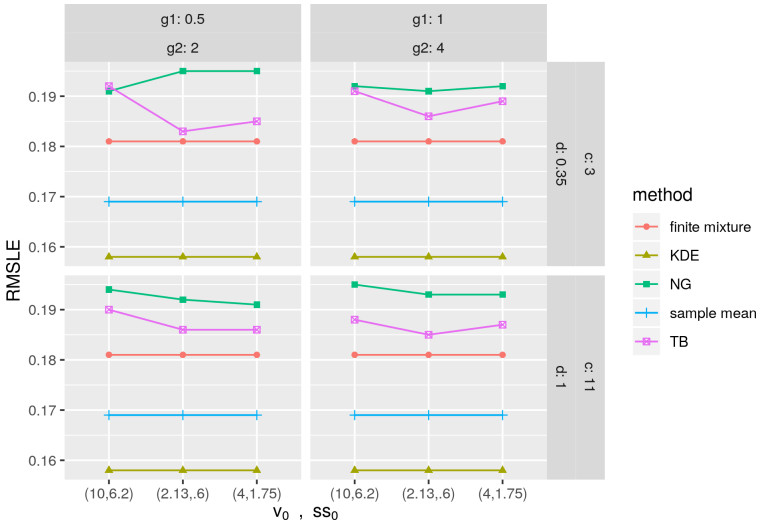


Figure 9: RMSLE plot; NG stands for normal-gamma prior and TB stands for t-betaprime prior, df = 3, v1 = 1.5 for TB.

From the RMSLE table, we see that mixture model with a flexible prior generally has lower prediction error.

8 Discussion

In this work, we learn the distribution of commonly seen heavy-tailed data in audit using bayesian mixture models with stick breaking priors. Two different base distributions in the nonparametric prior are also discussed. Comparison between density estimates from nonparametric models and gaussian kernel density estimates tells us that by using models which allow adptiveness in the variance of each components in the mixture models, we end up with slightly better or similarly good density estimation for the imbalanced data. When we use nonparametric models to do predictions, our finding is that if we do not constrain the ability of extrapolation of the models, our estimates will be highly influenced by the outliers and thus impairs the prediction performance. However, if we incorporate the prior knowledge as the upper bound for predictions, the truncated predictions from all nonparametric methods in this work become more reasonable. Learning the density of the true value T is only the very first step towards modeling the misstatements in audit. One also need, at least, a model about error generation from the true values to the claimed values. To better understand how different density estimation of true values T will influence the inference of material misstatement, we need to further put the distribution of T, $P(T)$, into the "big picture". From there, we can explore how uncertainty from modeling true values T can be

propagated to predicting the underlying (hidden) true values given the claimed values C. Then, a more comprehensive view of our bayesian mixture models with stick breaking priors can be obtained.

Appendix A: Sampling Method for Log of Beta Random Variables.

- Generate two log-gamma random variables A and B according to Marsaglia and Tsang's method(Marsaglia and Tsang 2000), where $A \sim loggamma(\alpha, 1)$ and $B \sim loggamma(\beta, 1)$.
- Denote $C = max(A, B)$, and calculate $A_2 = A - C$ and $B_2 = B - C$
- To generate $beta(\alpha, \beta)$ random variables V, by definition, we generate its logrithm by calculating

$$\begin{aligned} log(\frac{e^A}{e^A + e^B}) &= A - log(e^A + e^B) \\ &= A_2 + C - log(e^{A_2+C} + e^{B_2+C}) \\ &= A_2 + C - log(e^C(e^{A_2} + e^{B_2})) \\ &= A_2 - log(e^{A_2} + e^{B_2}) \\ &= A_2 - log(1 + e^C), \end{aligned}$$

whose value is taken to be the value of $log(V)$. And by the same proof, we can also generate random variable $log(1 - V)$ as $B_2 - log(1 + e^C)$.By this strategy, we generate log of $beta$ random variables which we can use them directly in the Gibbs sampling algorithm.

Appendix B: Additional Prior Information

Calculated prior information for $(\sigma_{k_i}^2 \mid v_0, v_1, ss_0)$			
Configuration	Prior	mean	sd
$(v_0 = 2.13, ss_0 = 0.6)$	constrained	0.53	1.47
$(v_0 = 4, ss_0 = 1.75)$	constrained	0.58	0.41
$(v_0 = 10, ss_0 = 6.2)$	constrained	0.69	0.24
$(v_0 = 2.13, ss_0 = 0.6, v_1 = 1.5)$	flexible	0.53	1.47×1.3
$(v_0 = 2.13, ss_0 = 0.6, v_1 = 3)$	flexible	0.53	1.47×1.2
$(v_0 = 4, ss_0 = 1.75, v_1 = 1.5)$	flexible	0.58	0.41×1.7
$(v_0 = 4, ss_0 = 1.75, v_1 = 3)$	flexible	0.58	0.41×1.4
$(v_0 = 10, ss_0 = 6.2, v_1 = 1.5)$	flexible	0.69	0.24×2.6
$(v_0 = 10, ss_0 = 6.2, v_1 = 3)$	flexible	0.69	0.24×2

Table 4: Calculated prior mean and sd for component-specific variance σ_k^2.

Prior information for α				
Configuration	prior mean	prior sd	$P(\alpha > 15)$	$E(N_{occ}\|\alpha = prior \quad mean)$
$(c = 3, d = 0.35)$	8.6	4.9	0.1	36
$(c = 11, d = 1)$	11.0	3.3	0.1	43

Table 5: Prior information related to α.

Appendix C: Hyperparameter Table

Table 6: Simulation table for constrained model.

configuration	v0	ss0	a	sigma_0	g1	g2	c	d
1	2.13	0.60	0	10	1.0	4	3	0.35
2	2.13	0.60	0	10	1.0	4	11	1.00
3	2.13	0.60	0	10	0.5	2	3	0.35
4	2.13	0.60	0	10	0.5	2	11	1.00
5	4.00	1.75	0	10	1.0	4	3	0.35
6	4.00	1.75	0	10	1.0	4	11	1.00
7	4.00	1.75	0	10	0.5	2	3	0.35
8	4.00	1.75	0	10	0.5	2	11	1.00
9	10.00	6.20	0	10	1.0	4	3	0.35
10	10.00	6.20	0	10	1.0	4	11	1.00
11	10.00	6.20	0	10	0.5	2	3	0.35
12	10.00	6.20	0	10	0.5	2	11	1.00

Table 7: Simulation table for flexible model.

configuration	v0	ss0	a	sigma_0	g1	g2	c	d	df	v1
1	2.13	0.60	0	10	1.0	4	3	0.35	3	1.5
2	2.13	0.60	0	10	1.0	4	3	0.35	12	1.5
3	2.13	0.60	0	10	1.0	4	3	0.35	22	1.5
4	2.13	0.60	0	10	1.0	4	3	0.35	3	3.0
5	2.13	0.60	0	10	1.0	4	3	0.35	12	3.0
6	2.13	0.60	0	10	1.0	4	3	0.35	22	3.0
7	2.13	0.60	0	10	1.0	4	11	1.00	3	1.5
8	2.13	0.60	0	10	1.0	4	11	1.00	12	1.5
9	2.13	0.60	0	10	1.0	4	11	1.00	22	1.5
10	2.13	0.60	0	10	1.0	4	11	1.00	3	3.0
11	2.13	0.60	0	10	1.0	4	11	1.00	12	3.0
12	2.13	0.60	0	10	1.0	4	11	1.00	22	3.0
13	2.13	0.60	0	10	0.5	2	3	0.35	3	1.5
14	2.13	0.60	0	10	0.5	2	3	0.35	12	1.5
15	2.13	0.60	0	10	0.5	2	3	0.35	22	1.5
16	2.13	0.60	0	10	0.5	2	3	0.35	3	3.0
17	2.13	0.60	0	10	0.5	2	3	0.35	12	3.0
18	2.13	0.60	0	10	0.5	2	3	0.35	22	3.0
19	2.13	0.60	0	10	0.5	2	11	1.00	3	1.5
20	2.13	0.60	0	10	0.5	2	11	1.00	12	1.5
21	2.13	0.60	0	10	0.5	2	11	1.00	22	1.5
22	2.13	0.60	0	10	0.5	2	11	1.00	3	3.0
23	2.13	0.60	0	10	0.5	2	11	1.00	12	3.0
24	2.13	0.60	0	10	0.5	2	11	1.00	22	3.0
25	4.00	1.75	0	10	1.0	4	3	0.35	3	1.5
26	4.00	1.75	0	10	1.0	4	3	0.35	12	1.5
27	4.00	1.75	0	10	1.0	4	3	0.35	22	1.5
28	4.00	1.75	0	10	1.0	4	3	0.35	3	3.0
29	4.00	1.75	0	10	1.0	4	3	0.35	12	3.0
30	4.00	1.75	0	10	1.0	4	3	0.35	22	3.0
31	4.00	1.75	0	10	1.0	4	11	1.00	3	1.5
32	4.00	1.75	0	10	1.0	4	11	1.00	12	1.5
33	4.00	1.75	0	10	1.0	4	11	1.00	22	1.5
34	4.00	1.75	0	10	1.0	4	11	1.00	3	3.0
35	4.00	1.75	0	10	1.0	4	11	1.00	12	3.0
36	4.00	1.75	0	10	1.0	4	11	1.00	22	3.0
37	4.00	1.75	0	10	0.5	2	3	0.35	3	1.5
38	4.00	1.75	0	10	0.5	2	3	0.35	12	1.5
39	4.00	1.75	0	10	0.5	2	3	0.35	22	1.5
40	4.00	1.75	0	10	0.5	2	3	0.35	3	3.0
41	4.00	1.75	0	10	0.5	2	3	0.35	12	3.0
42	4.00	1.75	0	10	0.5	2	3	0.35	22	3.0
43	4.00	1.75	0	10	0.5	2	11	1.00	3	1.5
44	4.00	1.75	0	10	0.5	2	11	1.00	12	1.5
45	4.00	1.75	0	10	0.5	2	11	1.00	22	1.5
46	4.00	1.75	0	10	0.5	2	11	1.00	3	3.0
47	4.00	1.75	0	10	0.5	2	11	1.00	12	3.0
48	4.00	1.75	0	10	0.5	2	11	1.00	22	3.0
49	10.00	6.20	0	10	1.0	4	3	0.35	3	1.5
50	10.00	6.20	0	10	1.0	4	3	0.35	12	1.5
51	10.00	6.20	0	10	1.0	4	3	0.35	22	1.5
52	10.00	6.20	0	10	1.0	4	3	0.35	3	3.0
53	10.00	6.20	0	10	1.0	4	3	0.35	12	3.0
54	10.00	6.20	0	10	1.0	4	3	0.35	22	3.0
55	10.00	6.20	0	10	1.0	4	11	1.00	3	1.5
56	10.00	6.20	0	10	1.0	4	11	1.00	12	1.5
57	10.00	6.20	0	10	1.0	4	11	1.00	22	1.5
58	10.00	6.20	0	10	1.0	4	11	1.00	3	3.0
59	10.00	6.20	0	10	1.0	4	11	1.00	12	3.0
60	10.00	6.20	0	10	1.0	4	11	1.00	22	3.0
61	10.00	6.20	0	10	0.5	2	3	0.35	3	1.5
62	10.00	6.20	0	10	0.5	2	3	0.35	12	1.5
63	10.00	6.20	0	10	0.5	2	3	0.35	22	1.5
64	10.00	6.20	0	10	0.5	2	3	0.35	3	3.0
65	10.00	6.20	0	10	0.5	2	3	0.35	12	3.0
66	10.00	6.20	0	10	0.5	2	3	0.35	22	3.0
67	10.00	6.20	0	10	0.5	2	11	1.00	3	1.5
68	10.00	6.20	0	10	0.5	2	11	1.00	12	1.5
69	10.00	6.20	0	10	0.5	2	11	1.00	22	1.5
70	10.00	6.20	0	10	0.5	2	11	1.00	3	3.0
71	10.00	6.20	0	10	0.5	2	11	1.00	12	3.0
72	10.00	6.20	0	10	0.5	2	11	1.00	22	3.0

Appendix D: Comparison of Fittings from Constrained Prior and Flexible Prior with Largest df and v_1

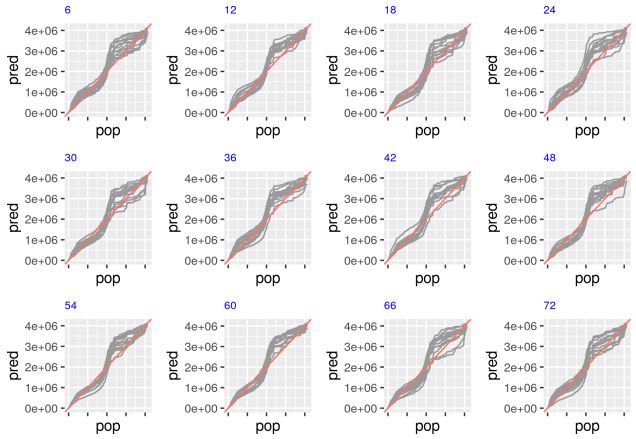


Figure 10: Two sample QQ plot for different configurations with flexible model (fitted based on random sample 1); X-axis represents hold-out population data; Y-axis represents predictions; Subtitles represent different configurations, which can be referenced in appendix C. These configurations are with the largest degrees of freedom and largest v1.

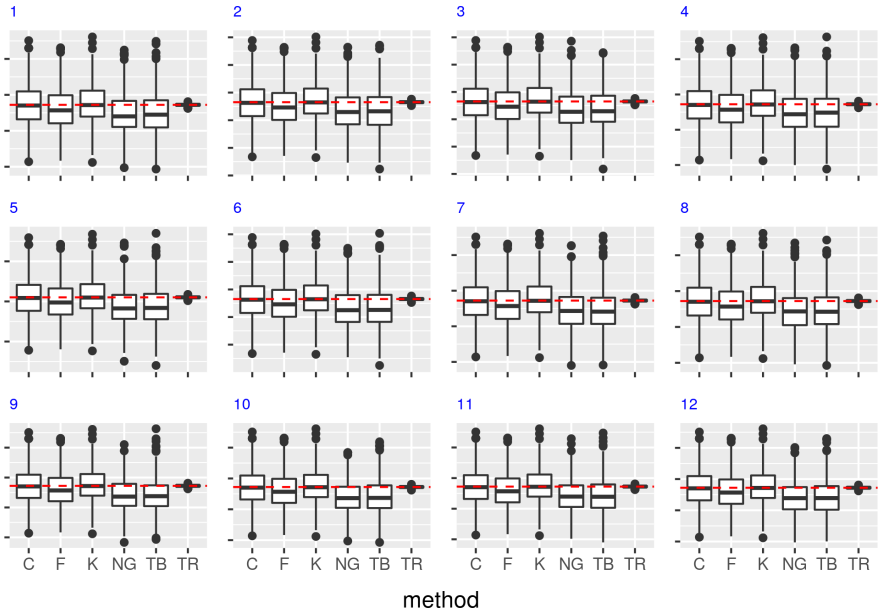


Figure 11: Boxplot for predictions of population sum. Lables on X-axis are as follows: T-betaprime(TB); Normal-gamma(NG); KDE(K); Central limit theorem(C); Finite mixture model(F); True value(TR). Hyperparameter choice for the plot: $(v_0 = 10, ss_0 = 6.2, g_1 = 0.5, g_2 = 2, c = 3, d = 0.35, df = 22, v_1 = 3)$. Red dashed line is the true value of the population sum.

References

Buch-Larsen, Tine, Jens Nielsen, Montserrat Guillen, and Catalina Bolancé. 2006. "Kernel Density Estimation for Heavy-Tailed Distributions Using the Champernowne Transformation." *Statistics* 39 (January): 503–18. https://doi.org/10.2139/ssrn.704903.

Escobar, Michael D., and Mike West. 1994. "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association* 90: 577–88.

Ferguson, Thomas S. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *Ann. Statist.* 1 (2). The Institute of Mathematical Statistics: 209–30. https://doi.org/10.1214/aos/1176342360.

Ishwaran, Hemant, and Lancelot F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association* 96: 161–73.

Marsaglia, George, and Wai Wan Tsang. 2000. "A Simple Method for Generating Gamma Variables." *ACM Trans. Math. Softw.* 26 (3). New York, NY, USA: Association for Computing Machinery: 363–72. https://doi.org/10.1145/358407.358414.

Pitman, Jim, and Marc Yor. 1997. "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator." *Ann. Probab.* 25 (2). The Institute of Mathematical Statistics: 855–900. https://doi.org/10.1214/aop/1024404422.

Scott, D.W. 2015. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley. https://books.google.com/books?id=XZo3BwAAQBAJ.

Sethuraman, Jayaram. 1994. "A Constructive Definition of Dirichlet Priors." *Statistica Sinica* 4: 639–50.

Shalizi, C. 2012. "Advanced Data Analysis from an Elementary Point of View." In.

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

Teh, Yee Whye. 2010. "Dirichlet Process." In *Encyclopedia of Machine Learning*, edited by Claude Sammut and Geoffrey I. Webb, 280–87. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8_219.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101 (476): 1566–81. http://www.gatsby.ucl.ac.uk/~ywteh/research/npbayes/jasa2006.pdf.