Distribution Fitting and Predictions

on Heavy Tailed Data in Audit using Nonparametric Bayes

Andy Zhai

Financial Audit

- An independent examination of financial information of an entity[1]
- An auditor is trying to understand a company's business and provides his or her own opinion on the financial statements

Audit table

- Are the claimed value reasonable?

item	claimed value($)
a sketch pencil	1
a spiral notebook	5
an all-in-ones desktop	1, 000
⋮	⋮
a laptop	20, 000
a warehouse	400, 000

[1] https://en.wikipedia.org/wiki/Audit

Financial Audit

- An independent examination of financial information of an entity[1]
- An auditor is trying to understand a company's business and provides his or her own opinion on the financial statements

Audit table

- Are the claimed value reasonable?

item	true value($)	claimed value($)
a sketch pencil	1	1
a spiral notebook	5	5
an all-in-ones desktop	990	1, 000
⋮	⋮	⋮
a laptop	2, 000	20, 000
a warehouse	400, 000	400, 000

[1] https://en.wikipedia.org/wiki/Audit

Financial Audit

transaction id	true value $X(\$)$	claimed value $Y(\$)$
1	$x_1 = 1$	$y_1 = 1$
2	$x_2 = 5$	$y_2 = 5$
3	$x_3 = 1,000$	$y_3 = 9,000$
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

- One task: Given a table filled with claimed values, is that table successfully tell the "truth"?
- Assume each $X_i \leq Y_i$ (overstatement),

$$error = \sum_{i=1}^{n} (Y_i - X_i).$$

- If $\sum_{i=1}^{n} Y_i$ is not that different from $\sum_{i=1}^{n} X_i$, an auditor will conclude that the transaction record is "safe".

Model an Audit Process (the big picture)

Recall: X true value, Y claimed value.

- Model the true value X based on samples of X.
- Model the claimed value $Y|X$.
- Make inference on $X|Y$.

The auditor's work

- Once the auditor has the model of $P(X)$ and $P(Y|X)$, and given a set of claimed values $(y_1, y_2, ..., y_n)$, he or she can use $(y_1, y_2, ..., y_n)$ along with $P(X|Y)$ to update the opinion about true values X and then make inference on the error.

Objective in This Project

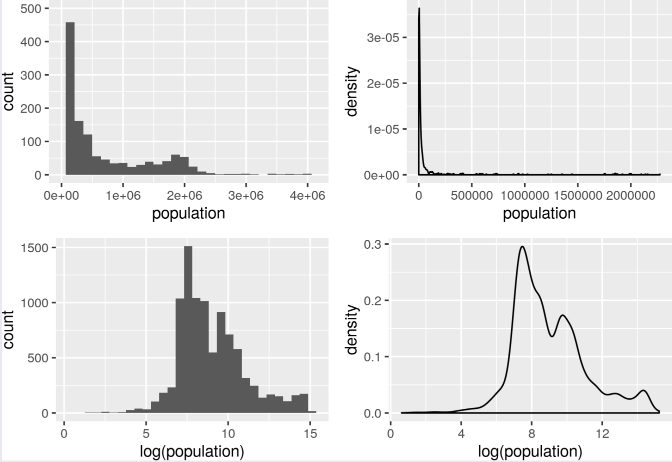
- Model the true value X by a bayesian approach.
    - λ: a vector of parameters.
    - $P(X|\lambda)$: a statistical model for true value X.
    - $\sum X$: inferencial target.

Steps

- Determine priors for $P(\lambda)$ and models for $P(X|\lambda)$;
- Derive posterior $P(\lambda|X)$;
- Estimate the density of X by $f(\tilde{X}|X) = \int f(\tilde{X}|\lambda)f(\lambda|X)d\lambda$;
- Get posterior predictions of $\sum \tilde{X}$ from $f(\tilde{X}|X)$;
- Check prediction performance.

Challenge in Modelling the Audit Data

Audit data are imbalanced and heavy-tailed.



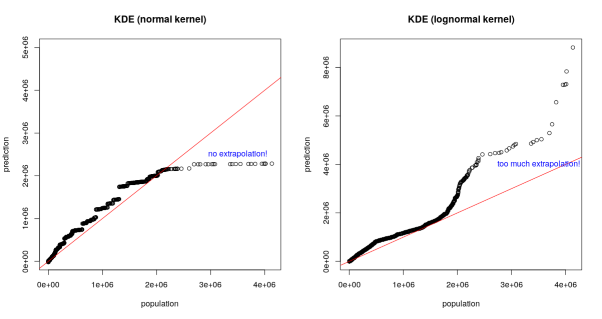
Challenge in Modelling the Audit Data

Bandwidth

- Consider classical gaussian kernel density estimation

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

- single bandwidth *h*: not sufficient for introducing local adaptiveness, especially to the right tail (multiple bandwidths).



Problems to solve

- Estimate the density of the heavy-tailed financial data.
- Make the density estimator compatible under bayesian hierachy.
- One solution: mixture model with multiple bandwidths.

Priors and Models

- Prior: $P(\lambda)$, dirichlet process
- Distribution of X: $P(X|\lambda)$,lognormal mixture model
- In combined, we have a dirichlet process (DP) mixture model.

$P(X|\lambda)$: the lognormal mixture model

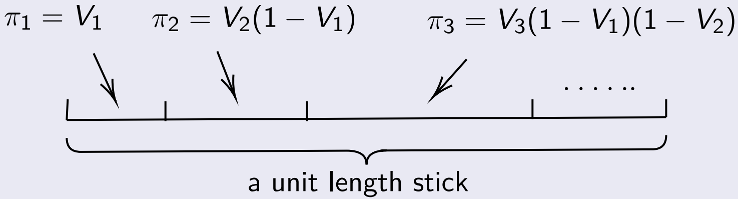
- Let X denote the sample of true values
- Let $Z = log(X)$. A normal mixture model for Z can be written as

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

- m: total number of mixture's components.
- $\theta_i = (\mu_i, \sigma_i^2 = \phi_i^{-1})$
- π_i: mixture components weight
- σ_i^2:component-specific bandwidth.
- λ :($\boldsymbol{\theta}$, $\boldsymbol{\pi}$).

Determine $P(\lambda) = P(\boldsymbol{\theta}, \boldsymbol{\pi})$

Determine $P(\boldsymbol{\pi})$: Stick-breaking prior



$\pi_1 = V_1$ $\pi_2 = V_2(1 - V_1)$ $\pi_3 = V_3(1 - V_1)(1 - V_2)$

a unit length stick

- $V_i \sim Beta(1, \alpha)$
- Stochastically decreasing sequence of probabilities π's
- After some step s, $\pi_j' s$, $j \geq s$ negligible

Why stick-breaking?

- Constructive definition of DP
- Introduce an infinite mixture
- Truncation approximation

Determine $P(\theta \mid \pi)$

- Laten variable $\boldsymbol{K}$; each $\boldsymbol{k_i} \in \{\mathbf{1}, ..., \boldsymbol{m}\}$ indexes a component in the mixture model,

$$\boldsymbol{Z_i|\theta, K \overset{ind}{\sim} N(\theta_{k_i}), \quad where\ \theta_{K_i} = (\mu_{k_i}, \sigma^2_{k_i} = \phi^{-1}_{k_i})}$$

Two Choices of $P(\theta \mid \pi)$

Normal-gamma (constrained) prior:

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

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

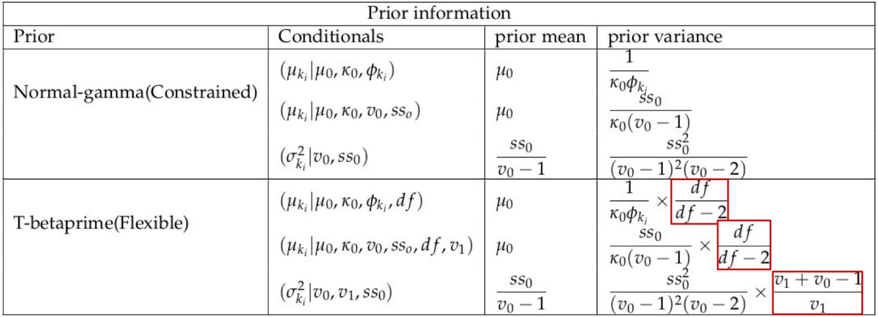
T-betaprime (flexible) prior:

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

$$\phi_{k_i}|v_0, ss_0, v_1 \sim bp(v_0, v_1, ss_0{}^{-1})$$

- T: $\mu_{k_i}|\mu_0, \kappa_0, \phi_{k_i}, r_{k_i} \sim N(\mu_0, (\kappa_0\phi_{k_i}r_{k_i})^{-1}), \quad r_{k_i} \sim gamma\left(\frac{df}{2}, \frac{df}{2}\right).$
- BP: $\phi_{k_i}|v_0, ss_0, h_{k_i} \sim gamma(v_0, ss_0 h_{k_i}), \quad h_{k_i}|v_1 \sim gamma(v_1, v_1).$

A closer look at two priors of $P(\theta \mid \pi)$



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

- Share the same prior mean for μ_{k_i} and $\sigma^2_{k_i}$
- Flexible prior has larger prior variance in general
- Flexible prior has heavier tail

Data-based simulation study

Experiment

- Given 500 data from the population, we want to use those data to approximate nonparametric bayesian density estimates under different priors and predict the sum of the "holdout" population, then check the prediction performance.

Implementation

- Blocked gibbs sampling method for stick-breaking prior (Ishwaran and James 2001)[a] with total number of iterations for each MCMC chain to be 100,000.

[a] http://people.ee.duke.edu/~lcarin/Yuting3.3.06.pdf

Comparisions

- Density estimation: QQ plots; KS goodness of fit test
- Prediction: root mean square logarithmic error (RMSLE)

Data-based simulation study: choosing prior's hyperparameter

Normal-gamma (constrained) prior:

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

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

T-betaprime (flexible) prior:

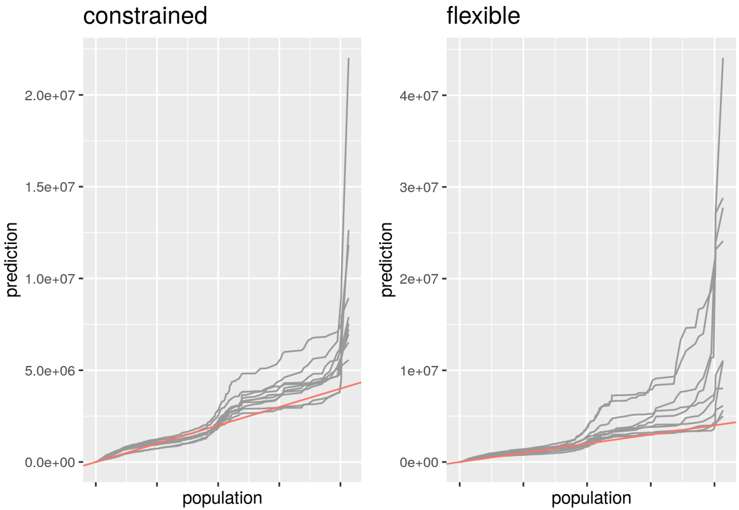
$$\mu_{k_i}|\mu_0, \kappa_0, \phi_{k_i}, df \sim T(\mu_0, (\kappa_0\phi_{k_i})^{-1}, df)$$

$$\phi_{k_i}|v_0, ss_0, v_1 \sim bp(v_0, v_1, {ss_0}^{-1})$$

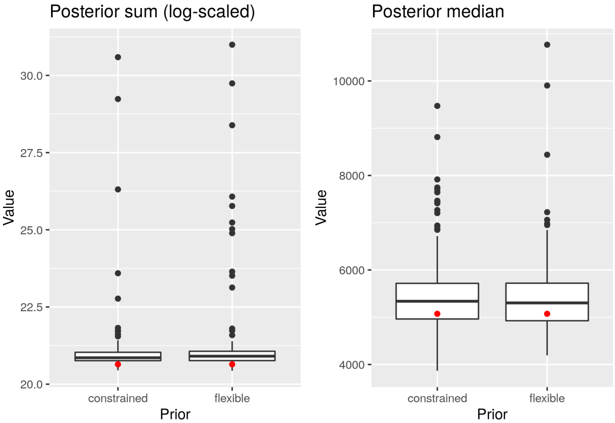
$$\mu_0 \sim N(a, \sigma_0^2), \kappa_0 \sim gamma(g_1, g_2)$$
$$\pi \sim SB(1, \alpha), K \sim Multi(\pi)$$
$$\alpha \sim gamma(c, d)$$

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

Posterior predictions extrapolate too much!

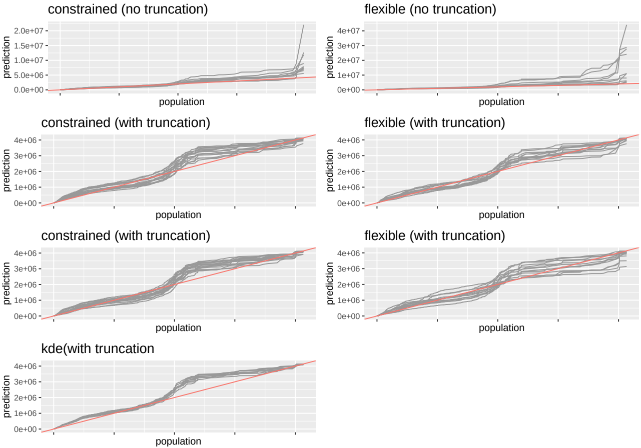


Posterior predictions extrapolate too much!



- By the assumption $X_i \leq Y_i$, we truncate posterior predictions.

Posterior predictions with truncation



constrained (no truncation)

flexible (no truncation)

constrained (with truncation)

flexible (with truncation)

constrained (with truncation)

flexible (with truncation)

kde(with truncation

- Constrain the extrapolation
- Similar to KDE with lognormal kernel

Posterior prediction variability

- Considering point estimats of predicted sum, $\sum_i \tilde{X}_i$
- The experiments are done with 500 different simple random samples.

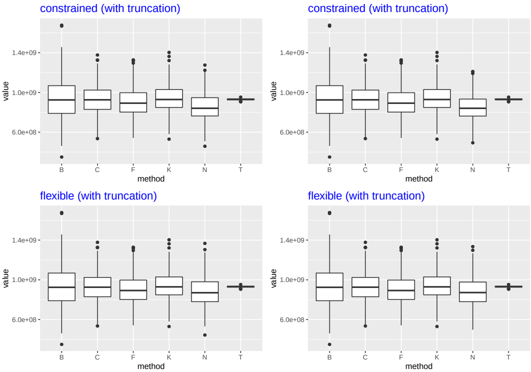
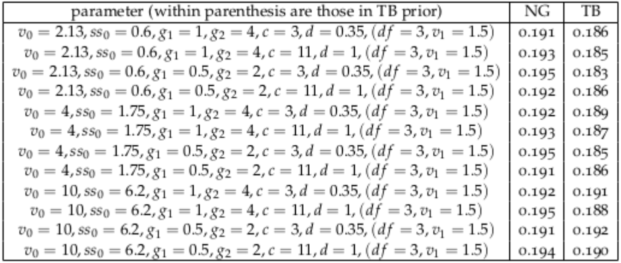


Figure 1: Selected boxplot for predicted sum. Lables on X-axis are: Bootstrapping(B); Nonparametric(N); KDE(K); Central limit theorem(C); Finite mixture model(F); True value(T).

Empirical prediction risk

- root mean square logarithmic error (RMSLE):

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

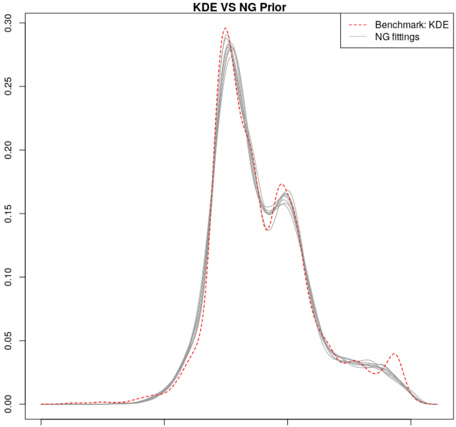


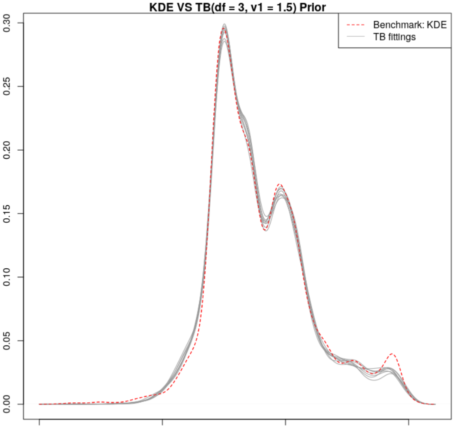
parameter (within parenthesis are those in TB prior)	NG	TB
$v_0 = 2.13, ss_0 = 0.6, g_1 = 1, g_2 = 4, c = 3, d = 0.35, (df = 3, v_1 = 1.5)$	0.191	0.186
$v_0 = 2.13, ss_0 = 0.6, g_1 = 1, g_2 = 4, c = 11, d = 1, (df = 3, v_1 = 1.5)$	0.193	0.185
$v_0 = 2.13, ss_0 = 0.6, g_1 = 0.5, g_2 = 2, c = 3, d = 0.35, (df = 3, v_1 = 1.5)$	0.195	0.183
$v_0 = 2.13, ss_0 = 0.6, g_1 = 0.5, g_2 = 2, c = 11, d = 1, (df = 3, v_1 = 1.5)$	0.192	0.186
$v_0 = 4, ss_0 = 1.75, g_1 = 1, g_2 = 4, c = 3, d = 0.35, (df = 3, v_1 = 1.5)$	0.192	0.189
$v_0 = 4, ss_0 = 1.75, g_1 = 1, g_2 = 4, c = 11, d = 1, (df = 3, v_1 = 1.5)$	0.193	0.187
$v_0 = 4, ss_0 = 1.75, g_1 = 0.5, g_2 = 2, c = 3, d = 0.35, (df = 3, v_1 = 1.5)$	0.195	0.185
$v_0 = 4, ss_0 = 1.75, g_1 = 0.5, g_2 = 2, c = 11, d = 1, (df = 3, v_1 = 1.5)$	0.191	0.186
$v_0 = 10, ss_0 = 6.2, g_1 = 1, g_2 = 4, c = 3, d = 0.35, (df = 3, v_1 = 1.5)$	0.192	0.191
$v_0 = 10, ss_0 = 6.2, g_1 = 1, g_2 = 4, c = 11, d = 1, (df = 3, v_1 = 1.5)$	0.195	0.188
$v_0 = 10, ss_0 = 6.2, g_1 = 0.5, g_2 = 2, c = 3, d = 0.35, (df = 3, v_1 = 1.5)$	0.191	0.192
$v_0 = 10, ss_0 = 6.2, g_1 = 0.5, g_2 = 2, c = 11, d = 1, (df = 3, v_1 = 1.5)$	0.194	0.190

Table 5: RMSLE table for bayesian mixture model.

- T-betaprime prior yields smaller prediction error.

Density Plot Comparison





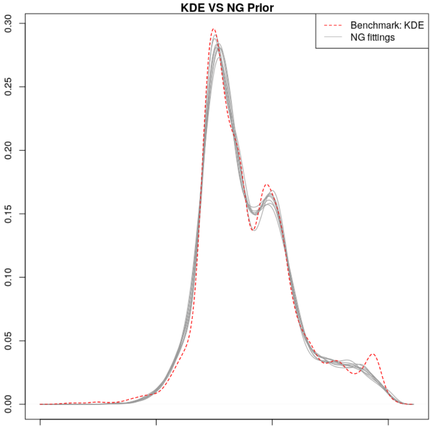
- Both nonparametric fittings smooth the tail
- TB fittings "closer" to the benchmark

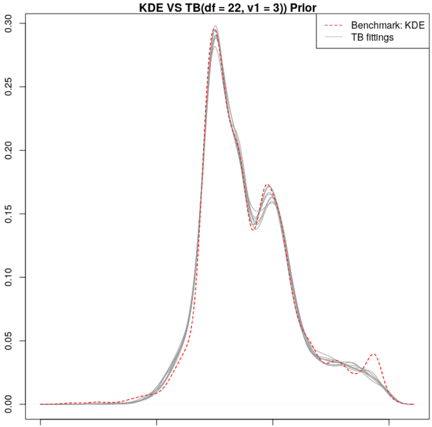
Conclusion

- Learn the distribution of commonly seen heavy-tailed data in audit using bayesian mixture models with stick breaking priors.
- Normal-gamma prior is compared with t-betaprime prior and t-betaprime prior fits the data better and leads to smaller prediction error.
- By truncation, we can fix the extrapolation problem. In terms of point estimates of $\sum \tilde{X}$, our method mimics the behavior of using lognormal KDE, finite mixture of lognormals, and sample sum. But our method maintains the merit of being integrated into a bigger bayesian hierarchy.

Codes available at
https://github.com/HongxuanZhai/DataAnalysisProject.git

Backup


KDE VS NG Prior
Benchmark: KDE
NG fittings
0.00
0.05
0.10
0.15
0.20
0.25
0.30


KDE VS TB(df = 22, v1 = 3)) Prior
Benchmark: KDE
TB fittings
0.00
0.05
0.10
0.15
0.20
0.25
0.30