Distribution Fitting and Predictions

on Heavy Tailed Data in Audit

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

an 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

an audit table

- Are the claimed value reasonable?

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

- Model the true value X based on samples of X.
- Model the claimed value $Y|X$.
  - In the modeling phase, both X and Y's are observables.
- 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 $\tilde{Y}$ (X is not observable to the auditor), he or she can use $\tilde{Y}$ along with $X|Y$ to update their opinion about true values X and then make inference on the error.

Objective in This Project

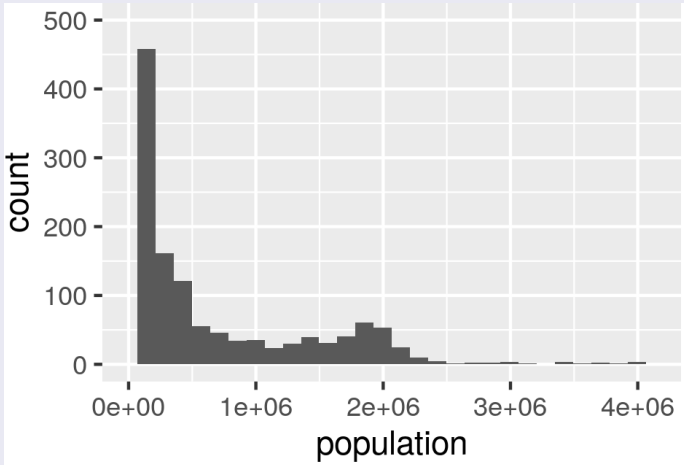
- Model the true value X by a bayesian approach.
- λ: a vector of parameters.
- $P(X|\lambda)$: a statistical model for data.
- $\sum X$: inferencial target.

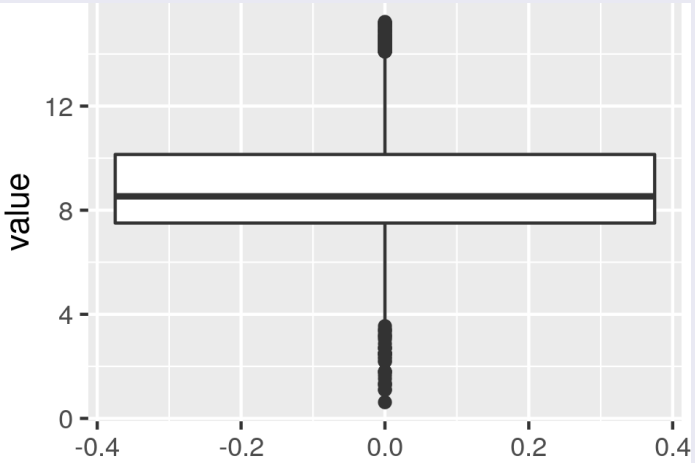
Steps

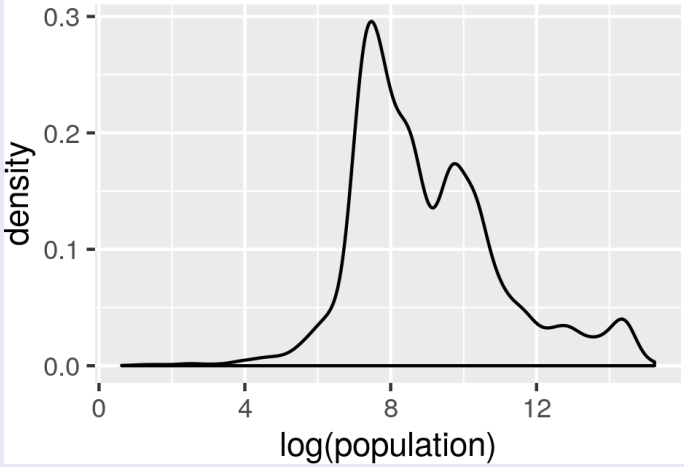
- determine priors for $P(\lambda)$ and models for $P(X|\lambda)$;
- estimate the density of X by $f(\tilde{x}|X) = \int f(\tilde{x}|\lambda)f(\lambda|X)d\lambda$;
- get posterior predictions of $\sum \tilde{X}$ from $f(\tilde{x}|X)$;
- check prediction performance.

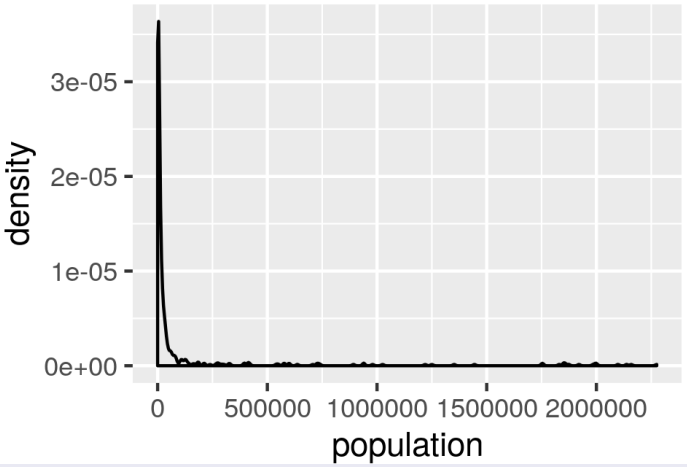
Challenge in Modelling the Audit Data

Audit data are imbalanced and heavy-tailed.







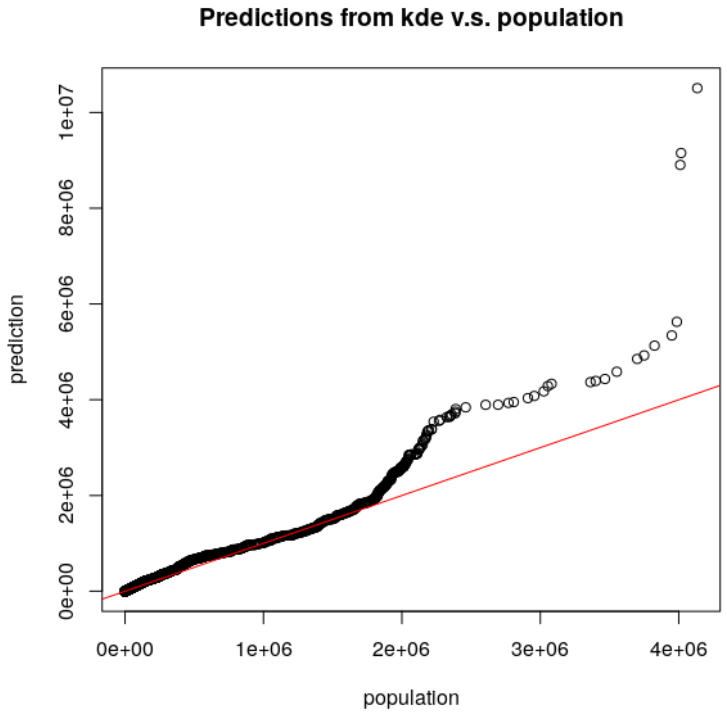


Challenge in Modelling the Audit Data

- 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).



Priors and Models

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

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

Let X denote the sample from population let $Z = log(X)$. A normal mixture model for Z can be written as

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

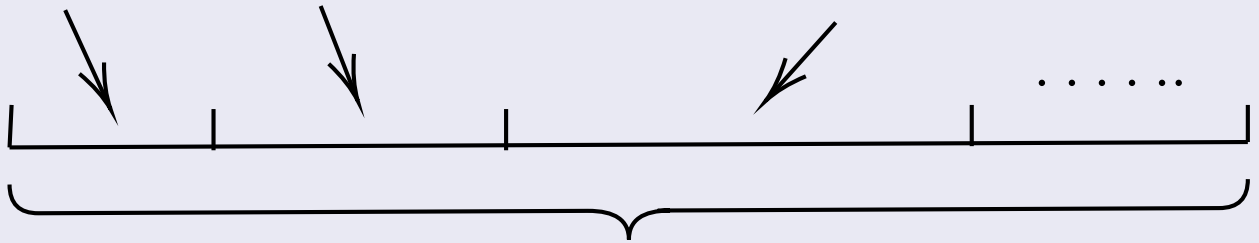
where $\theta_i = (\mu_i, \sigma_i^2 = \phi_i^{-1})$ and m is total number of mixture's components.

- σ_i^2:component-specific bandwidth.
- λ :$(\boldsymbol{\theta}, \boldsymbol{\pi})$.

Determine $P(\lambda) = P(\theta, \pi)$

Determine $P(\pi)$

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



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

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

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

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

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:

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

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

$\Updownarrow$

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

$r_{k_i} \sim gamma(\frac{df}{2}, \frac{df}{2})$

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

$h_{k_i}|v_1 \sim gamma(v_1, v_1)$

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

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

- shares the same prior mean for μ_{k_i} and $\sigma^2_{k_i}$
- flexible prior having larger prior variance in general
- flexible prior having 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:

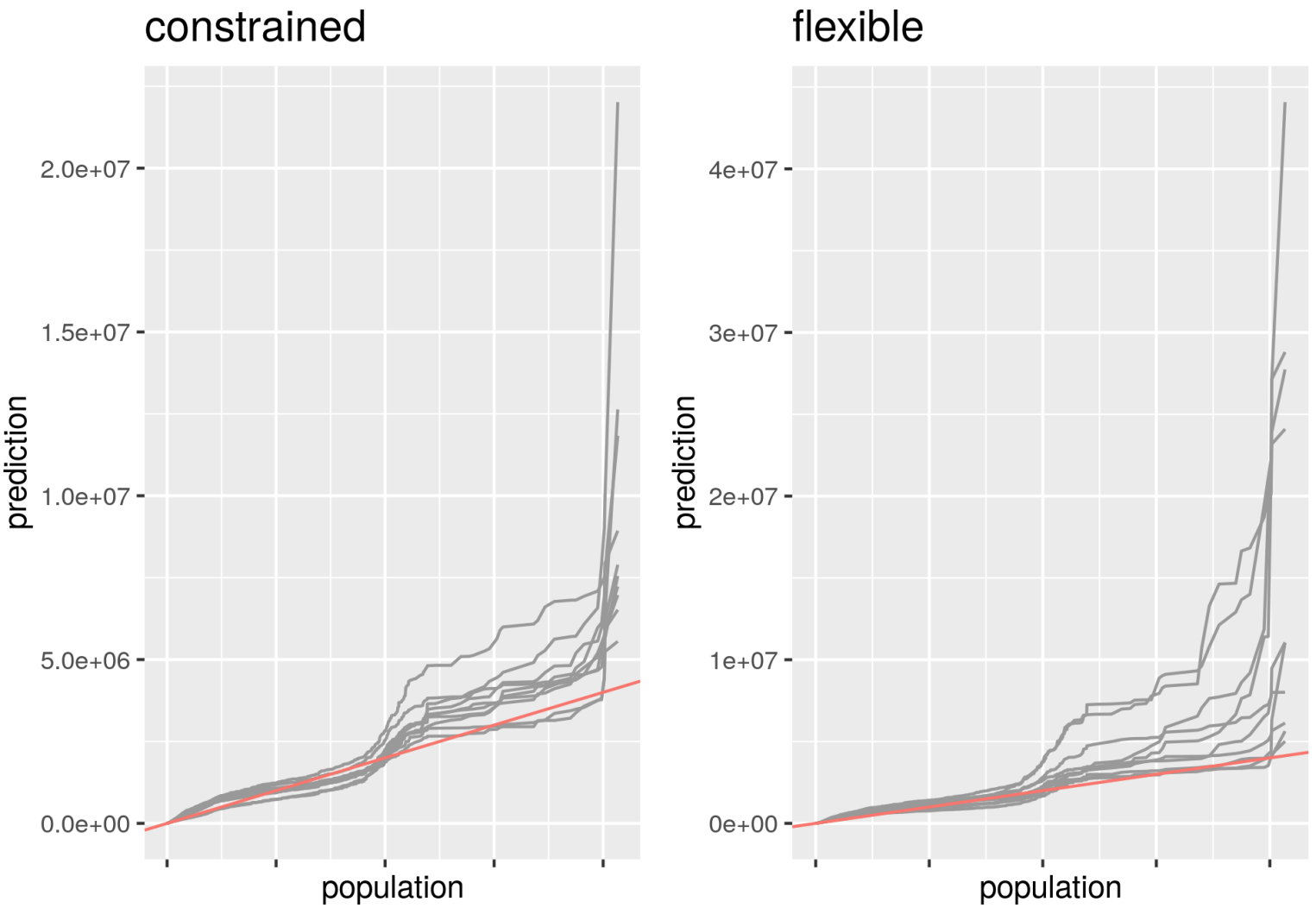
$$\mu_{k_i}|\mu_0, \kappa_0, \phi_{k_i}, df \sim T(\mu_0, (\kappa_0\phi_{k_i})^{-1}, df)$$

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

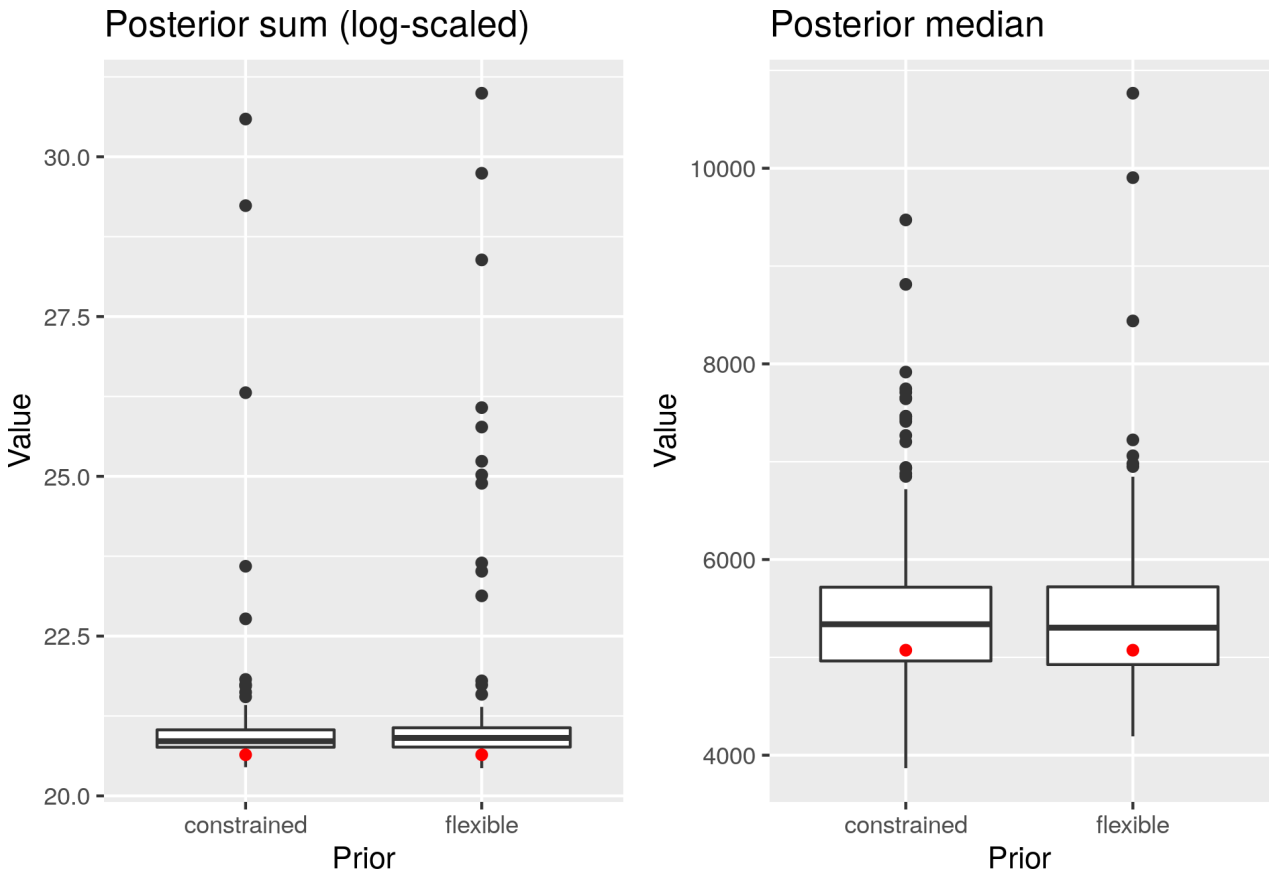
$$\mu_0 \sim N(a, \sigma_0^2), \kappa_0 \sim gamma(g_1, g_2)$$
$$\pi \sim SB(1, \alpha), K \sim Multi(\pi)$$
$$\alpha \sim gamma(c, d)$$

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!

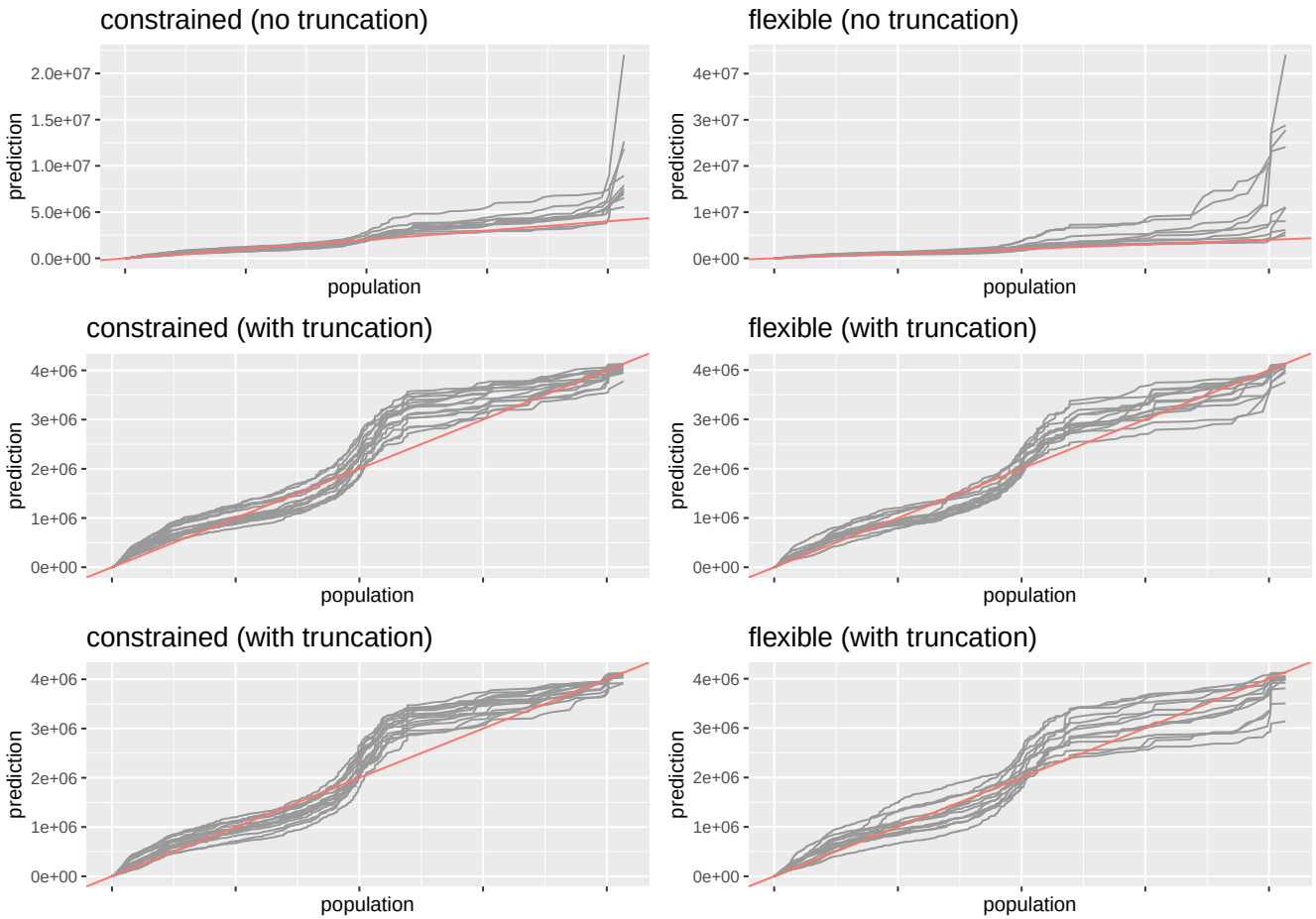


Posterior predictions extrapolate too much!



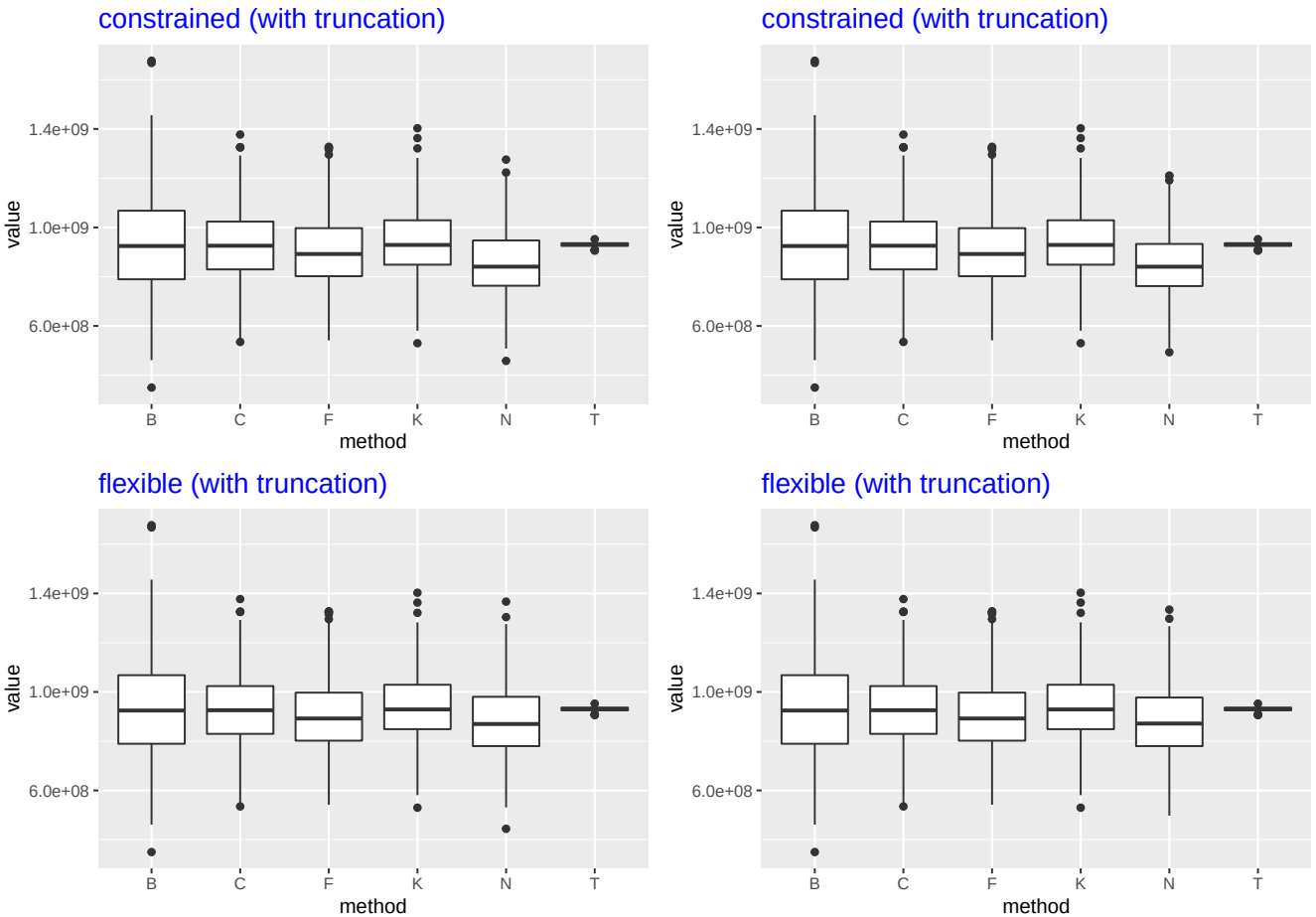
- By the assumption of overstatement, truncated posterior predictions are made from the bayesian density estimates.

Posterior predictions



Posterior prediction variability

- targeting on point estimats of population sum, $\sum_i X_i$
- Predicted sum's are obtained through doing the experiment on 500 different simple random samples.



Empirical prediction risk

- root mean square logarithmic error (RMSLE):

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

RMSLE table for bayesian mixture model.

constrained	flexible
0.191	0.186
0.193	0.185
0.195	0.183
0.192	0.186
0.192	0.189
0.193	0.187
0.195	0.185
0.191	0.186
0.192	0.191
0.195	0.188
0.191	0.192
0.194	0.190

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.