TI 2023-038/III
Tinbergen Institute Discussion Paper

Bayesian Mode Inference for Discrete Distributions in Economics and Finance

Jamie Cross[1]
Lennart Hoogerheide[2,5]
Paul Labonne[3]
Herman K. van Dijk[4,5]

1 University of Melbourne
2 Vrije Universiteit Amsterdam
3 Norwegian Business School
4 Erasmus University Rotterdam
5 Tinbergen Institute

Bayesian Mode Inference for Discrete Distributions in Economics and Finance*

Jamie Cross, Lennart Hoogerheide, Paul Labonne & Herman K. van Dijk

June 27, 2023

Abstract

Detecting heterogeneity within a population is crucial in many economic and financial applications. Econometrically, this requires a credible determination of multimodality in a given data distribution. We propose a straightforward yet effective technique for mode inference in discrete data distributions which involves fitting a mixture of novel shifted-Poisson distributions. The credibility and utility of our proposed approach is demonstrated through empirical investigations on datasets pertaining to loan default risk and inflation expectations.

JEL codes: C11, C25, C81, C82, E00, D00

Keywords: Bayesian Inference, Mixture Models, Mode Inference, Multimodality, Shifted-Poisson.

*We thank Nalan Basturk for useful discussions in the development of this research. The mode inference method can be implemented using the R package *BayesMultiMode* (Baştürk et al., 2023).

1 Introduction

Detecting heterogeneity within a population has a long tradition in economics and finance. Common examples include loan default risk in the field of credit risk analysis (e.g., Dionne et al., 1996), and the study of expectations formation among individuals or market participants (e.g., Haltiwanger and Waldman, 1985). Central to such studies is the requirement of credible mode determination. We here outline a simple method for credible mode inference on number and locations of modes and their uncertainty in such cases. The practicality of our method is demonstrated in datasets on loan default risk and inflation expectations. Details regarding computational implementation are provided in a companion R package titled *BayesMultiMode* (Baştürk et al., 2023). Our procedure may serve as a useful addition to the mode detection toolkit available to researchers, policymakers and industry practitioners within these fields.

2 A Bayesian Framework for Mode Inference

Stage 1: Estimation using a novel discrete mixture. We introduce a mixture of novel shifted-Poisson (SP) distributions specified as:

$$y_i - \kappa_k \sim \text{Poisson}(\lambda_k) \text{ if } z_{ik} = 1 \text{ for } i = 1, \ldots, n; k = 1, \ldots, K, \tag{1}$$

where $z_{ik} = 1$ if y_i belongs to cluster k, and 0 otherwise and the latent variable distribution is defined as $\Pr[z_{ik} = 1] = \pi_k$, for $i = 1, \ldots, n, k = 1, \ldots, K$, and where the shift parameter κ_k is a non-negative integer. The shift parameter κ_k is introduced to identify the amount of dispersion between the mean and variance for each component in the mixture. A single SP distribution allows for underdispersion, and mixing of multiple components accommodates overdispersion. The main advantage of the SP over a regular Poisson, is that an equidispersion restriction is not present even when the number of mixture

components is 1. The model is estimated with Bayesian methods using the following uninformative but proper priors:

$$\lambda_k \sim \text{Unif}(\lambda_{\min}, \lambda_{\max}), \tag{2}$$

$$\kappa_k \sim \text{DiscUnif}(\kappa_{\min}, \kappa_{\max}), \tag{3}$$

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \tag{4}$$

in which $[\lambda_{\min}, \lambda_{\max}] = [\kappa_{\min}, \kappa_{\max}] = [0, M]$ with $M = \max(y_i | y_i = 1, \ldots, n)$.

Given a number of components, K, and priors (2)–(4), it is straightforward to show that the conditional posterior distributions are given by:

$$p\left(\lambda_k | y, z, \theta_{-\lambda_k}\right) \propto \text{Gamma}_{[\lambda_{\min}, \lambda_{\max}]}\left(\frac{1}{n_k}, 1 + \sum_{i|z_{ik}=1} (y_i - \kappa_k)\right), \tag{5}$$

$$p\left(\kappa_k | y, z, \theta_{-\kappa_k}\right) \propto \frac{\lambda_k^{\sum_{i|z_{ik}=1} y_i - n_k \kappa_k}}{\prod_{i|z_{ik}=1} (y_i - \kappa_k)!}, \tag{6}$$

$$p\left(\pi | y, z, \theta_{-\pi}\right) \propto \text{Dirichlet}\left(n_1 + \alpha, \ldots, n_J + \alpha\right), \tag{7}$$

where $\text{Gamma}_{[\lambda_{\min}, \lambda_{\max}]}$ denotes the truncated Gamma density on the interval $[\lambda_{\min}, \lambda_{\max}]$, $n_k = \sum_{i=1}^{n} z_{ik}$ is the number of observations in component k and κ_k is an integer in $[\max\{\kappa_{\min}, \min_{i|z_{ik}=1}(y_i)\}, \kappa_{\max}]$. For $n_k = 0$ the conditional posteriors of λ_k and κ_k reduce to the uniform and discrete uniform priors on the intervals $[\lambda_{\min}, \lambda_{\max}]$ and $[\kappa_{\min}, \kappa_{\max}]$. Sampling from these conditional posterior densities can be done with standard MCMC.

The algorithm is completed by estimating a credible number of mixture components. We implement the sparse finite mixture SFM MCMC algorithm (Malsiner-Walli et al., 2016). The SFM approach is a simple, efficient and flexible algorithm that facilitates estimation of finite mixture models with unknown number of components. The basic idea is to deliberately overfit the mixture by specifying a larger number of components than is expected to describe the data distribution and next shrink the number of components

to a credible number with substantial posterior probability using Bayesian regularization. This is done by specifying a hyperprior of the form:

$$\alpha \sim Gamma(a_\alpha, b_\alpha), \tag{8}$$

where $E(\alpha) = a_\alpha / b_\alpha = \frac{1}{200}$ strongly favors small values. By Bayes theorem the conditional posterior distribution of α given the partition $\mathcal{P}$ of components across observations, $p(\alpha|\mathcal{P}) \propto p(\alpha)p(\mathcal{P}|\alpha)$, is given by

$$p(\alpha) \propto \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha), \tag{9}$$

$$p(\mathcal{P}|\alpha) \propto \frac{\Gamma(J\alpha)}{\Gamma(n + J\alpha)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha)}{\Gamma(\alpha)}. \tag{10}$$

Sampling from these distributions can be done with a Metropolis-Hastings step.

Stage 2: Mode inference Stage two consists of estimating the number of modes and their locations, and quantifying uncertainty around these estimates. By definition, modes must satisfy either:

1. $p_k(y_m - 1) < p_k(y_m) > p_k(y_m + 1)$, or

2. $p_k(y_m - 1) < p_k(y_m) = p_k(y_m + 1) = \ldots = p_k(y_m + l - 1) > p_k(y_m + l)$.

Case 1 is a unique mode which is clearly identified. Case 2 is a mode in which l consecutive values of the posterior predictive probability mass function are of equal value. We count this as a single mode, but keep track of each location.

3 Applications

3.1 Loan default risk

The quantification of loan default risk is common in finance. We use count data on the number of defaulted payment instalments with a total of 4329 observations in the range of 0 to 34 defaulted instalments by clients of a financial institution in Spain in 1990, see Dionne et al. (1996) and Karlis and Xekalaki (2001), Woo and Sriram (2007).

The empirical distribution of the data along with output from our mode inference procedure is provided in Figure 1. The left panel shows that these data are a typical example of zero-inflated count data, but also possess a fat tail. A standard Poisson mixture may therefore fail to approximate this data distribution. In contrast, the fitted mixture suggests that our SP mixture is well suited. The center panel of the figure shows that our mode inference procedure provides strong evidence for the existence of two modes. In contrast, application of three well-known frequentist tests to the data fails to reject the null of unimodality (see row 1 of Table 1), while the Bayesian mode inference procedure indicates a very low probability of one mode. A major benefit of our approach over frequentist tests is the credible determination of number and locations of modes. The right panel of Figure 1 shows strong evidence for the existence of such modes at zero and four (or five), respectively, emphasizing that the mode in the empirical distribution at four is probably not just a random result, but that there is truly a second mode in the underlying distribution at four (or five). There is a small posterior probability of a third mode between 14 and 19, but it is more likely that the large observations stem from a fat tail than from an actual mode at such a high value.

In further research with microdata on explanatory variables, differences between individual characteristics can be used to detect which types of clients fall into these categories and probabilities of default can be inferred. Thus, our methods may be used as a tool for risk assessment and sharper institution's policies surrounding the granting of loans.

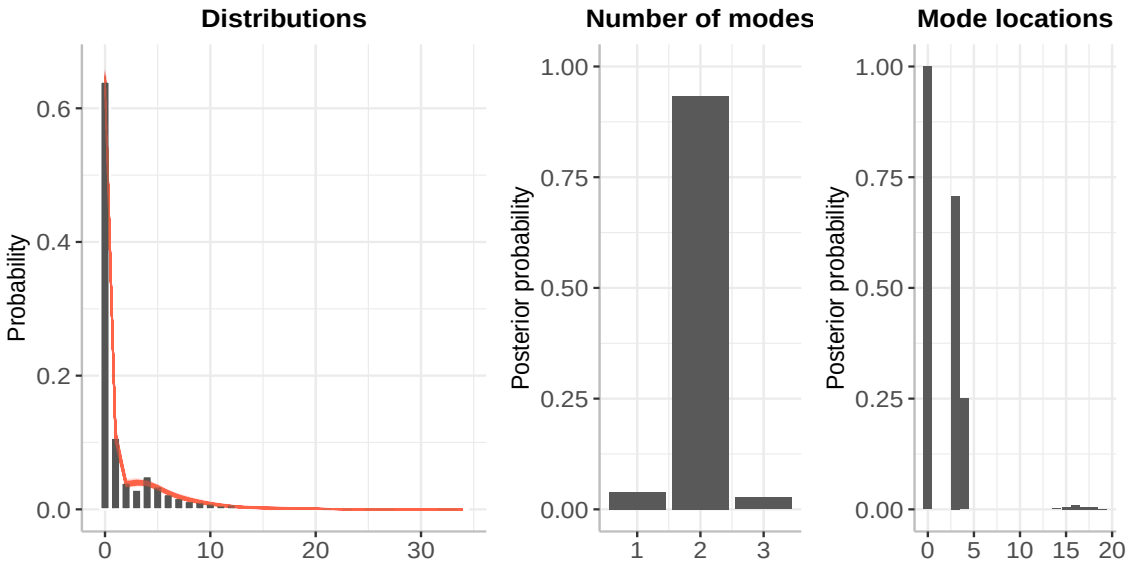

Figure 1: Empirical distribution of defaulted payment installments and estimated probability mass function (left), number of modes (center) and mode locations (right) using 1000 stored iterations of the MCMC algorithm.

Table 1: P-values from three frequentist tests with null hypothesis of unimodality alongside the posterior probability of unimodality from our Bayesian mode inference (BMI)

	SI	HY	HH	BMI
Default data	0.50	0.10	0.33	0.04
Michigan survey Feb 2020	0.17	0.01	0.03	0.03
Michigan survey Feb 2023	0.28	0.01	0.00	0.00

Note: SI = Silverman (1981), HY = Hall and York (2001), HH = Hartigan and Hartigan (1985)

3.2 Inflation expectations

Heterogeneity within the joint distribution of private agents' inflation expectations has been linked with learning (Pfajfar and Santoro, 2010) and economic literacy (Burke and Manz, 2014). In these studies multimodality is typically "eyeballed" from the data but our framework can be used to formally detect and explore this phenomenon. To that end, we use discrete data from survey responses to the question: "By what percentage do you expect prices to go up, on average, during the next 12 months?", from the Survey

Research Center (SRC) at the University of Michigan. The first row in Figure 2 contains response data in February 2020 and 2023 – 558 and 533 observations, respectively – along with fitted mixtures from our mode inference procedure. It is notable that a substantial subgroup of respondents select round numbers 5%, 10%, 15% or even 50%. For example, 27 out of the 558 respondents expect 10% inflation in 2020. Binder (2017) suggests that such behavior is indicative of high uncertainty.

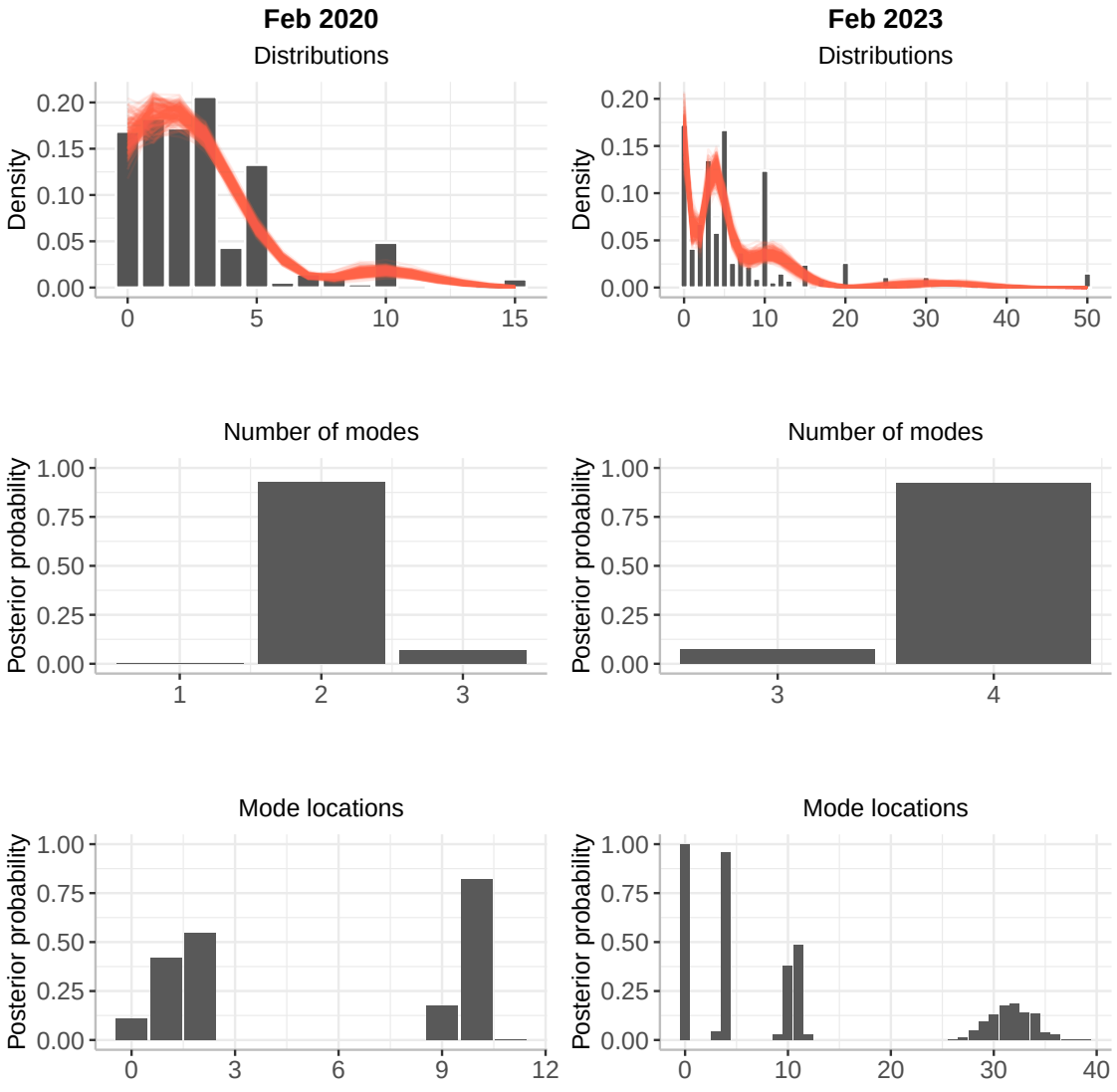

Figure 2: Empirical distribution of inflation expectations data and estimated probability mass function (top row), mode locations (center row) and number of modes (bottom row) using 1000 stored iterations of the MCMC algorithm.

On the detection of heterogeneity via multimodality, the frequentist tests in rows 2 and 3 of Table 1 provide conflicting results. The SI test does not reject unimodality in 2020 and 2023. The HI test gives borderline results at the 1 percent significance level for 2020 and 2023 while the HH test does not reject at the 1 percent significance level in 2020.

Our procedure provides credible information regarding number of modes and their locations. Results in rows 2 and 3 of Figure 2, provide strong evidence for multimodality, with increased heterogeneity of responses in 2023 exhibited by the increase from two modes to four. In 2020 there are likely two modes; one around 1%-2% and another one centered around higher inflation at 9%-10%. The latter figure seems excessive in normal times and the existence of this subgroup may stem from economic illiteracy (Burke and Manz, 2014). In 2023, the lower mode is split into two modes at 0% and 3-4%, suggesting a de-anchoring of inflation expectations away from the central bank's target of 2%. There is also a new mode around very high inflation expectations of 34-35% albeit with large dispersion. Overall, our analysis suggests that inflation expectations were well-anchored around 1 to 2% in 2020, but exhibit important changes with widespread disagreement in 2023, both within the subgroup with 'credible' expectations and the subgroup with 'incredible' expectations. Our results provide a useful starting point for a more detailed study into the causes of heterogeneity in inflation expectation formation during this period.

References

Baştürk N, Cross J, de Knijff P, Hoogerheide L, Labonne P, van Dijk H. 2023. *BayesMultiMode: Bayesian Mode Inference*. R package version 3.5.0.

Binder CC. 2017. Measuring uncertainty based on rounding: New method and application to inflation expectations. *Journal of Monetary Economics* **90**: 1–12.

Burke MA, Manz M. 2014. Economic literacy and inflation expectations: Evidence from a laboratory experiment. *Journal of Money, Credit and Banking* **46**: 1421–1456.

Dionne G, Artís M, Guillén M. 1996. Count data models for a credit scoring system. *Journal of Empirical Finance* **3**: 303–325.

Hall P, York M. 2001. On the calibration of Silverman's test for multimodality. *Statistica Sinica* **11**: 515–536.

Haltiwanger J, Waldman M. 1985. Rational expectations and the limits of rationality: An analysis of heterogeneity. *The American Economic Review* **75**: 326–340.

Hartigan JA, Hartigan PM. 1985. The DIP test of unimodality. *The Annals of Statistics* **13**: 70–84.

Karlis D, Xekalaki E. 2001. Robust inference for finite Poisson mixtures. *Journal of Statistical Planning and Inference* **93**: 93–115.

Malsiner-Walli G, Frühwirth-Schnatter S, Grün B. 2016. Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing* **26**: 303–324.

Pfajfar D, Santoro E. 2010. Heterogeneity, learning and information stickiness in inflation expectations. *Journal of Economic Behavior & Organization* **75**: 426–444.

Silverman BW. 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)* **41**: 97–99.

Woo MJ, Sriram T. 2007. Robust estimation of mixture complexity for count data. *Computational Statistics & Data Analysis* **51**: 4379 – 4392.