

Identifying and Interpreting Subgroups in Health Care Utilization Data with Count Mixture Regression Models

Christoph F. Kurz¹ and Laura A. Hatfield²

¹Institute of Health Economics and Health Care Management, Helmholtz Zentrum München

²Department of Health Care Policy, Harvard Medical School

December 6, 2018

Abstract

Inpatient care is a large share of total health care spending, making analysis of inpatient utilization patterns an important part of understanding what drives health care spending growth. Common features of inpatient utilization measures such as length of stay and spending include zero inflation, over-dispersion, and skewness, all of which complicate statistical modeling. Moreover, latent subgroups of patients may have distinct patterns of utilization and relationships between that utilization and observed covariates. In this work, we apply and compare likelihood-based and parametric Bayesian mixtures of Negative Binomial and zero-inflated Negative Binomial regression models. In a simulation, we find that the Bayesian approach finds the true number of mixture components more accurately than using information criteria to select among likelihood-based finite mixture models. When we apply the models to data on hospital lengths of stay for patients with lung cancer, we find distinct subgroups of patients with different means and variances of hospital days, health and treatment covariates, and relationships between covariates and length of stay.

1 Introduction

Recent policy attention has focused on the triple aim of improving health outcomes and care quality while reducing health care spending. These efforts require a detailed understanding of the drivers of variation in spending and outcomes to target payment and delivery reform incentives. Inpatient hospital services account for the majority of total health care spending. [1] In this study, we wish to understand variation in hospital inpatient days among patients diagnosed with lung cancer. Lung cancer is the most common cancer worldwide and a major cause of cancer-related mortality. [2]

Health care utilization data such as days in the hospital or patient-level spending are non-negative and often right-skewed, heavy-tailed, and multi-modal with a point mass at zero. Models for these data must be flexible enough to accommodate these features and still produce interpretable, policy-relevant results. [3] Generalized linear models (GLMs) with exponential family distributions such as Poisson, geometric, and Negative Binomial can accommodate non-negative and right-skewed count variables. [4–6] To account for excess zeros, these count models can be augmented with zero-inflation [4, 7, 8] or hurdle components. [9] Additional flexibility such as multi-modality and over-dispersion can come from mixture models. [10]

Mixture regression models are well known (see reviews in [11,12]) and may be called switching models [13] or latent class models. [14,15] Mixtures of count distributions, such as Poisson [16] and Negative Binomial, [17,18] may also be augmented with a hurdle component for excess zeros. [19] Mixture models can also link the mixture component probabilities to covariates. [3]

Our second motivation for using mixture models is to discover latent subclasses of individuals with distinct utilization patterns, which mixture regression models also accomplish. [20] We want to identify patient subgroups with different patterns of inpatient lengths of stay. [21,22] Policymakers, payers, and clinicians seeking to improve care and reduce spending in these groups could design interventions tailored to subpopulations. Previous work has shown substantial heterogeneity in the patterns of health care utilization among patients with lung cancer. [23] In this application, both the mixture components and their parameters are of interest, as are the clusters of observations drawn from each component.¹

A key challenge in fitting mixture models is determining the number of mixture components. Too many components will over-fit the data and impair model interpretation, and too few will be insufficiently flexible. The number of components can be decided *ex ante*, by choosing a convenient and interpretable number such as two or three, or *ex post*, by calculating models with different numbers of components and comparing their fit statistics, such as Akaike Information Criterion (AIC) [24] or Bayesian Information Criterion (BIC) [25], or likelihood ratio tests. [26]

In a Bayesian approach, the number of components can be treated as a parameter and informed by both the data and prior information. [27] Algorithms to fit Bayesian mixture models are diverse. The earliest approaches used reversible jump MCMC [28,29], to accommodate the changing model size as the number of components changes across iterations, and data augmentation [30]. More recently, maximum *a posteriori* estimation [31], variational inference [32], and alternative MCMC algorithms [33,34] have been proposed.

In this paper, we define and compare two implementations of mixture models for zero-inflated count regression: maximum-likelihood-based finite mixture models (FMMs) and parametric Bayesian mixture models. Previous authors have also used likelihood-based [35] and Bayesian [36,37] models for zero-inflated health care utilization outcomes (see review in [38]). Others have fit both likelihood-based [17,39] and Bayesian mixtures models for two-part count regressions, including for health claims data. Still other authors have fit finite mixtures of count regressions. [40] Our contribution to this literature is an explicit comparison between maximum likelihood and parametric Bayesian mixture models for (zero-inflated) count data. We compare these two approaches' ability to detect the true number of mixture components and estimate component parameters, as well as the practicalities of both approaches.

The rest of the paper is organized as follows. First, we detail the count regression models and their mixture implementations in Section 2. Section 3 outlines our real data and simulations studies. We present model checks in Section 4 and results in Section 5. Finally, we conclude in Section 6 with implications of the results and suggestions for future work.

¹Throughout, we use “component” to refer to the individual distributions in the mixture and “cluster” to denote the observations drawn from each mixture component.

2 Model Definitions

2.1 Negative Binomial Regression

The Negative Binomial distribution accommodates over-dispersion in count variables with a longer, fatter tail than the Poisson distribution. [41] [42] identified numerous parameterizations of the Negative Binomial; here, we use the definition in [17]. For $n = 1, \dots, N$ observations and $d = 1, \dots, D$ covariates, the data comprise an (N, D) -dimensional covariate matrix \mathbf{X} with rows \mathbf{x}_n and an N -vector of outcomes $\mathbf{y} = (y_1, \dots, y_N)'$. For simplicity, we omit the subscript n in what follows. The density function for the $\text{NegBin}(y|\mu, \psi)$ distribution is

$$f(y | \mu, \psi) = \frac{\Gamma(y + \psi)}{\Gamma(\psi)\Gamma(y + 1)} \left(\frac{\psi}{\mu + \psi} \right)^\psi \left(\frac{\mu}{\mu + \psi} \right)^y, \quad (1)$$

where ψ is a precision parameter and we specify a regression model for the mean parameter

$$\mu = \exp(\mathbf{x}'\boldsymbol{\beta})$$

for covariate vector $\mathbf{x} = (x_1, \dots, x_D)'$ and corresponding regression coefficient $\boldsymbol{\beta} = (\beta_1, \dots, \beta_D)'$. In this specification, mean and variance are

$$\mathbb{E}(y | \mathbf{x}) = \mu, \quad \text{Var}(y | \mathbf{x}) = \mu + \psi^{-1}\mu^2,$$

which corresponds to the NB2 model definition. [43]

2.2 Zero-Inflated Negative Binomial Regression

We extend the model above with *zero-inflation*. This model combines two sources of zeros: a point mass at zero and a Negative Binomial distribution, which generates both zero and non-zero count values. We can write a ZINB model $ZINB(y | \mu, \psi, \pi)$ as

$$y \sim \begin{cases} 0, & \text{with probability } \pi \\ \text{NegBin}(\mu, \psi), & \text{with probability } (1 - \pi), \end{cases} \quad (2)$$

where $\pi \in (0, 1)$ is the probability of an observation being a structural zero, and we model this probability with a Binomial distribution. Using a canonical log link, we specify a regression model for the probability

$$\mathbb{E}(y | \mathbf{x}) = \pi + (1 - \pi) \cdot \exp(\mathbf{x}'\boldsymbol{\beta}).$$

where \mathbf{x} is a vector of covariates and $\boldsymbol{\beta}$ is a coefficient vector, as before.

2.3 Mixture Regression Models

The regression models detailed above can accommodate the zero-inflation, over-dispersion, and skewness of our count data. A mixture model implementation will allow us to discover latent subpopulations (i.e., clusters) in the data. Below we outline both likelihood-based and Bayesian mixture model implementations of (ZI)NB regressions.

2.4 Finite Mixture Models

Our first formulation is a finite mixture of $k = 1, \dots, K$ Negative Binomial distributions, each parameterized by a mean $\mu_k = \exp(\mathbf{x}\boldsymbol{\beta}_k)$ and a precision parameter ψ_k . Let the contribution of each component to the mixture be denoted c_k such that $c_k \in [0, 1]$ and $\sum_{k=1}^K c_k = 1$. The distribution of the outcome y is the weighted sum over these Negative Binomial mixture components,

$$f(y | \boldsymbol{\beta}, \boldsymbol{\psi}, \mathbf{c}, \mathbf{x}) = \sum_{k=1}^K c_k \text{NegBin}(y | \exp(\mathbf{x}'\boldsymbol{\beta}_k), \psi_k), \quad (3)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_K)'$, $\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)'$, and $\mathbf{c} = (c_1, \dots, c_K)'$. The extension to ZINB mixtures is straightforward with the addition of π_k parameters to govern the zero-inflation.

We add subscripts $n = 1, \dots, N$ to identify individual observations of the outcome y_n and the covariates \mathbf{x}_n . Then we augment these observed data with cluster membership indicators z_{nk} , which equal 1 if observation n is drawn from component k and 0 otherwise. An observation can come from only one component, so $\sum_k z_{nk} = 1$. The complete data likelihood is therefore

$$f(\mathbf{y} | \boldsymbol{\beta}, \boldsymbol{\psi}, \mathbf{z}, \mathbf{x}) = \prod_{n=1}^N \prod_{k=1}^K \text{NegBin}(y_n | \exp(\mathbf{x}'_n \boldsymbol{\beta}_k), \psi_k)^{z_{nk}}, \quad (4)$$

where $\mathbf{y} = (y_1, \dots, y_N)'$ and $\mathbf{z} = (\mathbf{z}'_1, \dots, \mathbf{z}'_N)'$ is the collection of N -vectors $\mathbf{z}_k = (z_{1k}, \dots, z_{Nk})'$.

Of course, the membership indicators for each observation are missing, because we do not know from which component any observation is drawn. Thus we use an expectation-maximization (EM) algorithm to fit the model. [44–46] The EM algorithm iterates between two steps to overcome the missing data problem. In each E step, we take expectation of the complete data log likelihood with respect to the conditional distribution of the missing component assignment indicators given the current parameter values and the observed data. By averaging over these for each component, we obtain prior probabilities of each component. Then in the M step, we maximize the expected log likelihood in the parameters.

In addition to mixing probabilities and parameter estimates for each component, the EM algorithm also produces posterior probabilities of each observation belonging to each component, i.e.,

$$Pr(z_{nj} = 1 | \boldsymbol{\beta}, \boldsymbol{\psi}, \mathbf{c}, \mathbf{x}) = \frac{c_j \text{NegBin}(y_n | \exp(\mathbf{x}'_n \boldsymbol{\beta}_j), \psi_j)}{\sum_{k \neq j} c_k \text{NegBin}(y_n | \exp(\mathbf{x}'_n \boldsymbol{\beta}_k), \psi_k)}. \quad (5)$$

We can use the posterior probabilities to “hard classify” each observation into a cluster using the component for which it has the highest posterior probability of membership. Alternatively, we can weight observations by their posterior probabilities of being in each cluster. Using either method, the goal is to produce summaries of the clusters of observations.

The extension of the model above to mixtures of ZINB regressions is straightforward by replacing the Negative Binomial model in Eqs. (3) and (4) with the corresponding ZINB model,

$$f(y | \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\pi}, \mathbf{c}, \mathbf{x}) = \sum_{k=1}^K c_k \text{ZINB}(y | \exp(\mathbf{x}'\boldsymbol{\beta}_k), \psi_k, \pi_k), \quad (6)$$

and

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\pi}, \mathbf{z}, \mathbf{x}) = \prod_{n=1}^N \prod_{k=1}^K \text{ZINB}(y_n \mid \exp(\mathbf{x}'_n \boldsymbol{\beta}_k), \psi_k, \pi_k)^{z_{nk}} . \quad (7)$$

To fit the models in Eqs. (3) through (7), we modify the `FLXMRnegbin` function from the `countreg` package [47,48], which is a driver for the general EM mixture model fitting algorithms of the `flexmix` package. [49–51] This package also implements stochastic EM (SEM), which uses a sample from the conditional distribution of the membership indicators rather than their expectation. To choose the number of mixture components K , we fit the model for various values of K and choose the one with the best fit statistics. We use AIC and BIC to compare models with different numbers of components.

2.5 Bayesian Mixture Models

We next turn to Bayesian mixture models, for which we specify a prior distribution on component membership indicators in Eq. (4). A popular choice is a multinomial distribution on the (binary, sum-to-one) indicators and a (conjugate) Dirichlet prior on the mixture probabilities.

$$p(\mathbf{z} \mid \mathbf{c}) \propto \prod_{n=1}^N \prod_{k=1}^K c_k^{z_{nk}} \quad (8)$$

$$p(\mathbf{c}) = \text{Dir}(\mathbf{c} \mid \alpha_0) = \frac{\Gamma(K\alpha_0)}{\Gamma(\alpha_0)^K} \prod_{k=1}^K c_k^{\alpha_0 - 1} , \quad (9)$$

where $\Gamma(\cdot)$ is the Gamma function and α_0 is a hyperparameter. We fix the maximum number of components at $K = 20$. In practice, the number of components with non-trivial posterior mixing proportions is less than K , resulting in mixture model that has only as many components as the data require. See Section 6 for more discussion of this choice.

To complete the specification, we need priors for the regression coefficients and precision parameters of the component regression models. We chose Normal and Log Normal distributions,

$$\begin{aligned} \beta_{kd} &\sim \mathcal{N}(m_0, s_0) \\ \psi_k &\sim \mathcal{LN}(a_0, b_0) \end{aligned}$$

for $k = 1, \dots, K$ and $d = 1, \dots, D$ and hyperparameters $\alpha_0 = 0.1$, $m_0 = 0.0$, $s_0 = 10.0$, $a_0 = 0.0$, and $b_0 = 2.0$, that is, weakly informative priors. [52]

As before, this mixture model produces posterior estimates of the prevalence of each component and the parameters that govern it, plus posterior probabilities of membership in each mixture component for each observation.

The extension to mixtures of ZINB regressions follows as before, putting a multinomial-Dirichlet prior on the component indicators in Eq. (7). We refer to the models as MD-NB for the Multinomial-Dirichlet Negative Binomial and MD-ZINB for the zero-inflated version.

We implement this model in STAN [53] with 2000 iterations and a warm-up of 1000. Because this No-U-Turn sampler [54] does not support discrete latent variables, we marginalize over the component assignment variables.

3 Data

3.1 Monte Carlo Simulation

In a small simulation study, we compared the performance of the 2 modeling techniques described above: 1) finite mixture models with fixed numbers of components fit via EM and 2) Bayesian mixture models. For this, we generated data from Negative Binomial mixtures with 2, 3, 4, and 5 components and varying amounts of overlap. We first generated three covariates $\mathbf{x} = (x_1, x_2, x_3)'$ from Normal distributions $N(0, 0.5)$. The corresponding intercept and regression coefficients $\beta_k = (\beta_{0k}, \beta_{1k}, \beta_{2k})'$ and dispersion parameters ψ_k were chosen to produce components with high, medium, and low overlap. Then we generated the outcome as $y \sim \text{NegBin}(\exp(\mathbf{x}\beta_k), \psi_k)$. These choices (summarized in Appendix Table 9) produce a variety of shapes of the Negative Binomial mixture components with different degrees of overlap. Figure 1 shows densities of the true mixture components in each scenario.

3.2 AOK Data Set

We analyzed health care billing claims provided by the AOK Research Institute. AOK covers around 30% of the German resident population. The data set contains patient-level information on inpatient and outpatient diagnoses and procedures from 2009 to 2012, as well as service utilization. We used a study population previously derived from this data set consisting of patients with incident lung cancer in 2009. More information on the data set and the selection criteria can be found in Schwarzkopf et al. [55]

The outcome of interest was the total number of inpatient hospital days for each patient in the year after diagnosis. Inpatient hospital days are defined as the number of days from formal admission to hospital until discharge (i.e., hospital outpatient procedures do not count as hospital days), summed over all hospitalizations in a year. Admission and discharge on the same day count as one hospital day, so only individuals who were never admitted have zero hospital days. We included only individuals who survived for the full year, resulting in 7118 individual observations. The mean number of hospital days is 44, with a maximum of 296 and 59 zero observations.

We included the following covariates in the model: age, sex, treatment type during the course of the disease (chemotherapy, radiation therapy, or surgery), number of other tumor sites at diagnosis, number of metastases at diagnosis, Charlson comorbidity index, and district type of residence (major city, urban district, rural district, or thinly populated rural district). The Charlson comorbidity index was calculated using ICD-10 codes as in [56] with the slight modification of excluding the diagnosis of lung cancer out of the group “solid tumor without metastases”.

Figure 2 summarizes the sample in a tableplot. [57] The skewness of the hospital days outcome is apparent in the leftmost panel, with wide variation at the highest quantiles of the distribution. Patients were mostly older than 60 years, male, and urban, and these demographic features did not appear to be strongly related to length of stay. Number of metastases, Charlson scores, chemotherapy, and surgery were all positively correlated with length of stay.

4 Model Evaluation

4.1 Graphical Posterior Predictive Checks

To check the fit of the model, we use three different graphical checks. The first two are based on the posterior predictive, [58] which is the distribution of the outcome conditional on the posterior of the parameters, that is, after updating our beliefs about the parameters using the observed data. From the posterior predictive, conditional on the observed covariates \mathbf{X} , we repeatedly simulate N observations (as in the observed data), y_m^{rep} for $m = 1, \dots, 1000$. Then we compare the distribution of each y_m^{rep} to the distribution of the observed outcome y . Similar shapes indicate good model fit.

We also compute a test quantity on each replicated data set, $q_m = q(y_m^{rep})$, and plot the distribution of this test quantity compared with the observed test quantity $q = q(y)$. Specifically, we use the mean as our test quantity of interest, $q_m = \bar{y}_m^{rep}$. A distribution of simulated means centered around the observed mean supports good model fit.

Rootograms [59] show goodness-of-fit for count data by comparing observed frequencies from the model to expected frequencies. We can use these to check for over-dispersion, skewness, and excess zeros. We employ hanging rootograms in which the bars of observed frequencies are “hanging” from the curve representing the expected frequencies. Rootograms use “soft” component assignment weights, i.e., probabilities of belonging to a certain component. Ideal fit is indicated when the bars are close to the horizontal zero line.

5 Results

5.1 Component Identification in Simulated Data

Table 5.1 presents the results of the Monte Carlo simulation using the AIC, BIC, and MD-NB to estimate the number of components in the mixture.

AIC and BIC selection finds the true number of components in two of the simulations, for low and medium overlap with 4 components. On the other hand, MD-NB is very accurate in detecting the true number of components in all scenarios with 5 components. It also identifies the true 2 components for medium overlap and the true 4 components for high overlap.

All methods tend to overestimate the number of components in most scenarios but MD-NB is still closer to the truth. When the true number of components are estimated correctly, the regression parameter estimates and mixture weights are also close to the true values (see Table 9 and Table 9 in appendix).

5.2 Mixture Regression Analysis of AOK Data

For the AOK data set, the MD-NB finds 3 components as having the highest expected posterior mixture weights. In the following, we only show results based on this final model with 3 components. Figure 7 displays histogram plots of five replicated outcomes $y_1^{rep}, \dots, y_5^{rep}$ using the posterior predictive distribution. All replicates exhibit similar shape to the true distribution y . The means of the replicated data are centered around the mean of the observed data (Figure 8).

Truth	high overlap			medium overlap			low overlap		
	AIC	BIC	MD-NB	AIC	BIC	MD-NB	AIC	BIC	MD-NB
2	1	1	4	3	3	2	6	3	5
3	5	5	4	6	4	5	5	5	5
4	6	6	4	4	4	6	4	4	5
5	6	6	5	4	4	5	6	6	5

Table 1: True and estimated number of components based on AIC and BIC model selection criteria and on MD-NB for the simulated data sets. Bold indicates cases where the selected number of components matches the truth.

In Figure 5, we show rootograms for the MD-NB (top row) and MD-ZINB (bottom row). We see that the MD-NB under-fits count 0 and over-fits the subsequent counts 1 and 2 in the first component, which is typical for data with excess zeros. The rootograms for the MD-ZINB fit the data better in the first component. Rootograms nicely exhibit the different means and variances for each component.

Component 1 contains only 6% (430/7118) of all observations and corresponds to individuals who spend, on average, fewer days in hospital but with very high variance. The mode of hospital days for component 1 is 4 for the MD-NB and 7 for the MD-ZINB (zero mode omitted). Component 2 is the largest, with 59% (4172/7118) of individuals; they stay longer in hospital (the mode is at 24 days) with less variance. Component 3 comprises 35% (2517/7118) of the population, and these patients have the most hospital days (mode at 39 days) and again high variance.

Figure 3 shows β_k parameter estimates for each component of the MD-NB as incidence rate ratios (IRRs) alongside the Bayesian high probability density intervals. These coefficients can be interpreted as the multiplicative increase in the expected number of hospital days for every one unit increase in the predictor. For example, in components 1 and 2, treatment is associated with more hospital days, across all modalities. The IRR for the combination of all three treatments is the largest, 7.9. That is, compared to patients who receive no treatment, those who receive chemotherapy, radiation, and surgery have 7.9 times as many expected hospital days. In general, all treatment combinations have higher IRRs in component 1 than in components 2 and 3. The only exception is radiotherapy, which has an IRR of 4.1 in component 2, higher than 0.6 in component 3, and 3.8 in component 1. The IRR for the number of metastases is uneven over the components: 1.24 and 1.22 in 1 (MD-NB and MD-ZINB), 1.51 in 2, and 0.95 in 3. On the other hand, the IRR for the number of multiple tumors is increasing from 0.69 to 0.91 to 0.94 across components 1, 2, and 3. In component 3, radiation is associated with fewer hospital days and surgery is null, whereas chemotherapy and combinations are associated with more hospital days. Demographic factors and baseline health were less strongly associated with hospital days. Age and sex appear to have no relationship to hospital days (the IRRs are around 1.0 in all components). Regional factors are only important for individuals in component 3 and only for urban districts, where the IRR is 0.64.

We display the regression coefficient estimates for component 1 of the MD-ZINB in Figure 4, separately for the negative binomial part of the model (left) and the binomial part of the model (right). Note that the binomial coefficients are relative risks, while the negative binomial coef-

ficients are IRRs. In the negative binomial component, the pattern of coefficients is nearly the same as the MD-NB fit, but the coefficients are slightly different. For example, the IRR for receiving all three treatments (chemotherapy, radiotherapy, and surgery) is 7.6 in the NB components of the MD-ZINB compared to 7.9 in the MD-NB. We do not show regression coefficient estimates for components 2 and 3 of the MD-ZINB because they are essentially the same as the MD-NB coefficients.

When we use hard assignments to classify individuals into components according to the highest posterior probability, we see that treatment patterns are very different across components (see Figure 6). Chemotherapy plus radiation is the most common treatment in all components, but individuals in component 1 are far more likely to receive this combination (42% compared to 22% in component 2 and 24% in component 3). Surgery alone is the second most common treatment in components 2 (22%) and 3 (17%), and chemotherapy alone is the second most common treatment in component 1 (19%), but it is infrequent in 2 (13%) and 3 (11%). Compared to people in component 1, those in components 2 and 3 are more likely to receive chemotherapy combined with surgery or surgery and radiation. Figure 9 shows greater comorbidity burden and slightly older age in component 3. The Charlson comorbidity burden increased across the components: median 1 (IRQ 0–3) in component 1; median 2 (IQR 0–3) in component 2; and median 2 (IQR 1–4) in component 3. Metastases were more common in patients in component 1 (median 1, IQR 0–1) than in components 2 and 3 (median 0, IQR 0–1).

6 Discussion

This paper compares parametric Bayesian models and finite mixture models for count data on health care utilization. The advantage of Bayesian mixture modeling is to allow the number of mixture components to be estimated from the data. In a simulation study, we show that selecting the number of mixture components in a finite mixture model using model fit statistics such as AIC and BIC is not a very accurate method for finding the true number of mixture components. Instead, the posterior mixture component probabilities from the Bayesian model are closer to the truth, though slightly overestimate the number of components, as seen by other authors. [60]

However, our simulation only covered a small fraction of possible scenarios. Furthermore, the data-generating process closely resembles the model specification, which is rarely the case in the real world. While the MD-NB is clearly more accurate in finding the true number of components than AIC or BIC model selection, it still misses the exact components in many of the simulations. In this simulation, “truth” is defined by the data-generating process, but it has been argued that the idea of “truth” in component analysis depends on the context and the application. [61]

On the AOK data set, graphical model checks show that the model fits the data well, particularly the zero-inflated model. This is not clear from the histograms, as there is no apparent spike at zero, but becomes obvious in the rootograms. For the AOK data set, the MD-NB and MD-ZINB models find three components of individuals with strikingly different distributions of hospital days and treatment patterns.

In the treatment of lung cancer, surgery offers the best prospect of cure. If diagnosed at

an early stage, it is possible to remove the tumor as a whole, such that no further treatment is necessary. If the tumor is already bigger, surgical resection with chemotherapy and radiation is the treatment of choice. In metastatic lung cancer, palliative chemotherapy, possibly accompanied by radiation therapy for individual metastases, may alleviate symptoms and prolong survival. [62]

Component 1 has the fewest hospital days on average. In this component, we find many patients with chemotherapy only, and chemotherapy in combination radiation therapy. This, and the lack of surgery, likely indicate that these patients were already in an advanced (metastatic) stage at diagnosis. For these patients, it is likely that therapy had a palliative intent with a focus on improving quality of life. In contrast, patients in components 2 and 3 were more likely to have surgery only, surgery and chemotherapy, and the combination of all three treatments. This indicates diagnosis at an earlier stage and more aggressive treatment.

The treatments received by people in components 2 and 3 are quite similar, with only the proportion of surgery being slightly higher in 2 than in 3. However, the coefficients governing the relationships between treatment type and hospital days in these components are quite different. While radiation therapy is associated with significantly more hospital days in component 2, it has the opposite association in component 3. Moreover, the strong and positive association of surgery with hospital days in component 2 fades in component 3, where surgery has no relationship to hospital days. Together, these results suggest that people in components 2 and 3 get similar treatment combinations, but for different reasons.

There are several limitations to this study. First, mixture models present computational challenges. For example, care must be taken when fitting Bayesian mixture models to avoid the so-called “label-switching problem” caused by the model being invariant under permutations of the indices of the components (i.e., the indices of the model components may be permuted across chains). [33,63] It is crucial to run multiple Markov chains, inspect the resulting posterior samples, and apply posterior checks, as we have done here. We also enforced an ordering constraint on the component means. Other authors have proposed more sophisticated methods, using loss functions, [64,65] exploratory analysis of unconstrained posterior samples from a permutation sampler, [13] or highest posterior density. [66] The number of selected components might vary, as mixture models with different values of K can provide good representations of the same data. [67]

Our choice of $K = 20$ balances computational burden with the the goal of our analysis, which is to describe the parameters of the mixture components and the corresponding clusters of observations. More than 20 components/clusters would be unwieldy in our applied setting. [68] The appropriate limit on the number of mixture components would be different in a mixture model intended for flexible density estimation, for which the approximation to the infinite mixture occurs at values of K nearer to 70 or 100. [69–71] In our Bayesian model, components that contribute very little will have mixing proportions that go to zero, thus being effectively removed from the model. This allows us to make a single training run in which we start with a relatively large initial value of K , and allow surplus components to be pruned out of the model. [32]

In our applications, fitting this model to the AOK data took from approximately 4 hours up to 12 hours, depending on the number of observations, on a current quadcore CPU with 32GB RAM. The finite mixture model with a pre-specified number of components took only 3 minutes to compute. Further research should investigate how variational Bayesian methods

could improve speed and how this affects the accuracy of the estimates. Variational Bayesian mixture models have been found to accurately detect the number of mixture components [67, 72] and suffer less from identifiability issues. [73]

We limited our consideration to mixtures of parametric (zero-inflated) Negative Binomial regression models. Previous authors have recommended more flexible kernels for count data. [74, 75] The ZINB distribution is more flexible than the Poisson, but still parametric. [37] suggested a semi-parametric alternative to two-part models for zero-inflated data.

We are also limited by the parametric form of the mixing distribution. [76] proposed a semi-parametric approach to modeling flexibly the relationship of covariates to mixing proportions. There exists an extensive literature on *nonparametric* Bayesian mixture modeling approaches, (see reviews in [77–80] and references therein). The Dirichlet prior is the most widely used prior for mixture components, [81, 82] though [74] found the Pitman–Yor process can be more robust. [83] notes that the Dirichlet distribution is limited the negative correlation structures and proposes to use the Beta-Liouville distribution. Relevant to our application, [84] find that Bayesian nonparametric mixture models of medical claims outperform analogous mixture models, though their purpose was prediction rather than interpretation and they modeled semi-continuous, rather than zero-inflated count, outcomes. [85] noted the connection between mixtures of Poisson and negative binomial and Dirichlet processes.

Some authors have argued that interpreting the parameters of the mixture coefficients should be avoided. [71, 86] In addition, interpretation is necessarily more difficult in complex models such as these. In the case of the MD-ZINB model, with three mixture components in each of two sub-models (i.e., the Binomial part and the Negative Binomial part), the number of regression coefficients is six times the number of covariates (assuming each covariate is in each sub-model). However, inference on multiple parameters simultaneously is relatively straightforward in Bayesian models, which is another advantage of this approach.

7 Conclusion

This work presents Bayesian and likelihood-based clustering mixture models for count data with many zeros (here, hospital days) that can be used to find subgroups of patients. In contrast to clustering methods based on finite mixture models, the Bayesian approach avoids under- and over-fitting while still being fully interpretable. We apply this method to study hospital days for patients with lung cancer and demonstrate that it can find subgroups with specific properties that correspond well to the different number of hospital days in each component. Clustering models are useful and practical methods for understanding heterogeneity in inpatient hospital services, an important component of total health care spending.

8 Acknowledgments

The authors thank Bret Zeldow for reviewing the manuscript, Michael Betancourt and Bob Carpenter for help with the Stan code, and Larissa Schwarzkopf for interpreting the clusters.

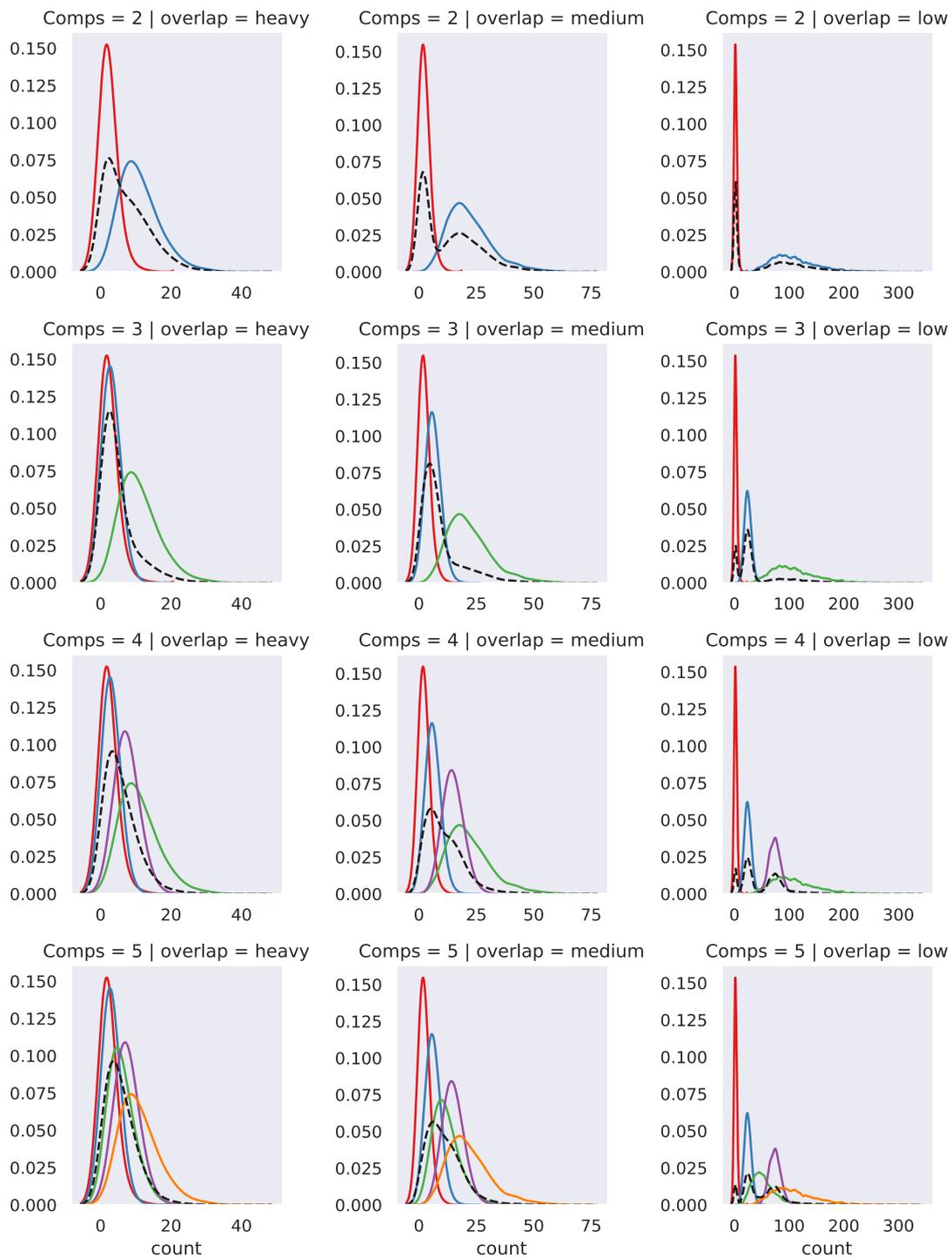


Figure 1: Density histograms for the simulated data sets with 2, 3, 4, and 5 mixture components with varying amounts of overlap. The black dashed line marks the combined density.

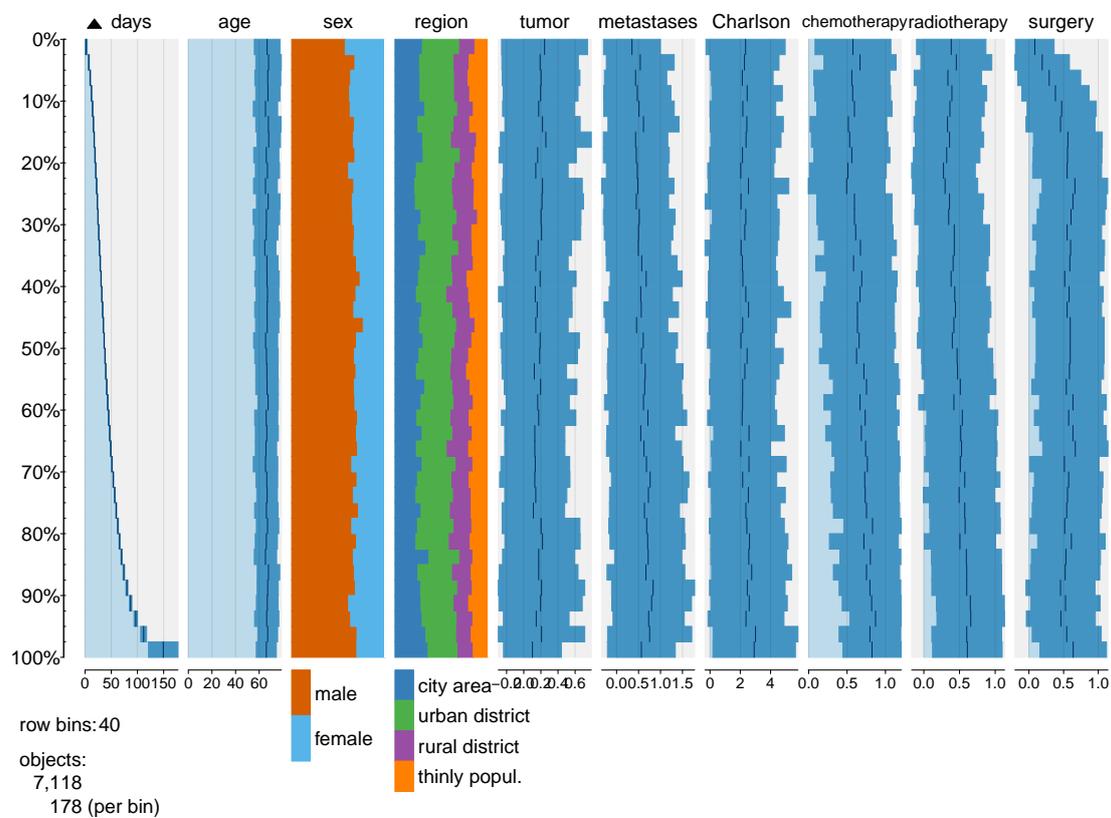


Figure 2: Tableplot of the AOK data set. The whole data set is sorted by hospital days (left side) in increasing order. All other variables are grouped into row bins, where numeric variables are displayed as bar charts and categorical variables as stacked bar charts.

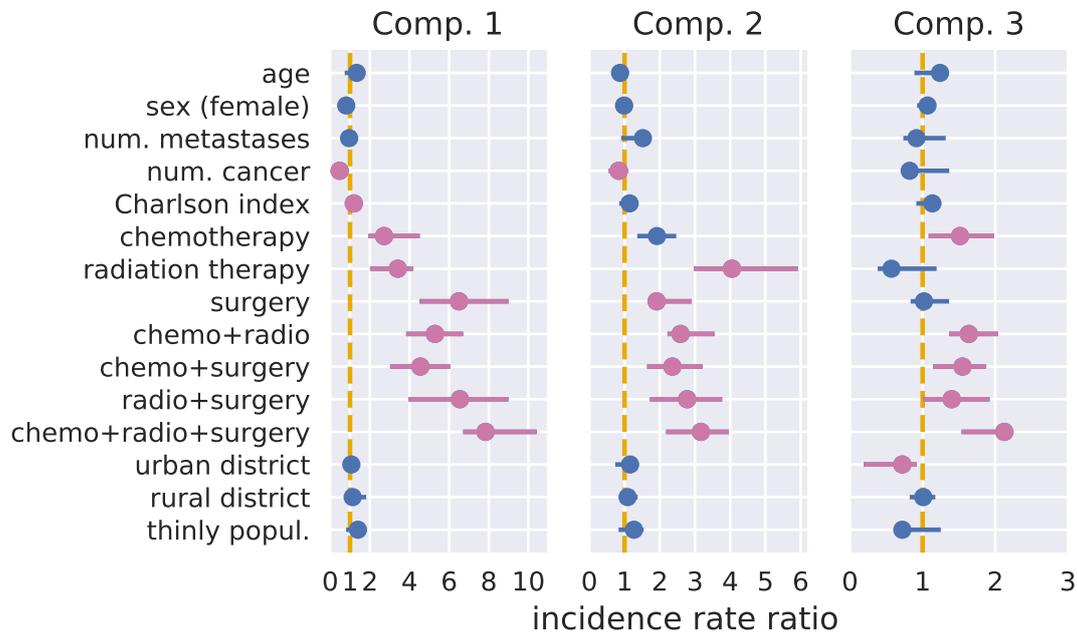


Figure 3: NP-NB estimation results for all three components on the AOK data set. Parameter estimates are presented as incidence rate ratios and 95% highest probability density intervals. Intervals that exclude the 1 are highlighted in purple. Intercept is not shown.

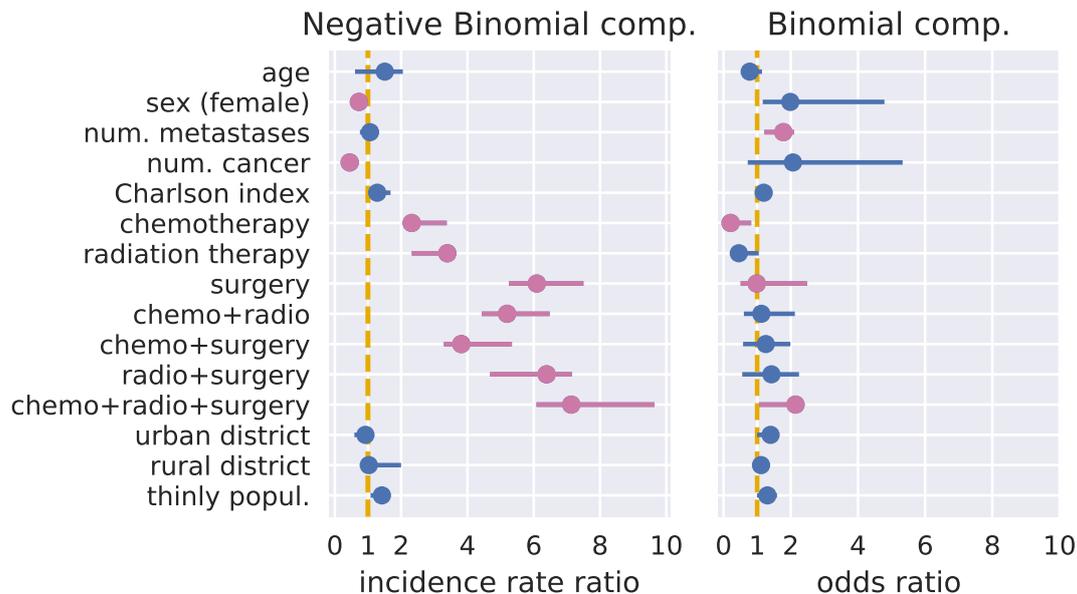


Figure 4: NP-ZINB estimation results for the first components only on the AOK data set. Parameter estimates are presented as incidence rate ratios for the count part and odds ratios for the zero part. 95% highest probability density intervals are shown for both parts. Intervals that exclude the 1 are highlighted in purple. Intercept is not shown.

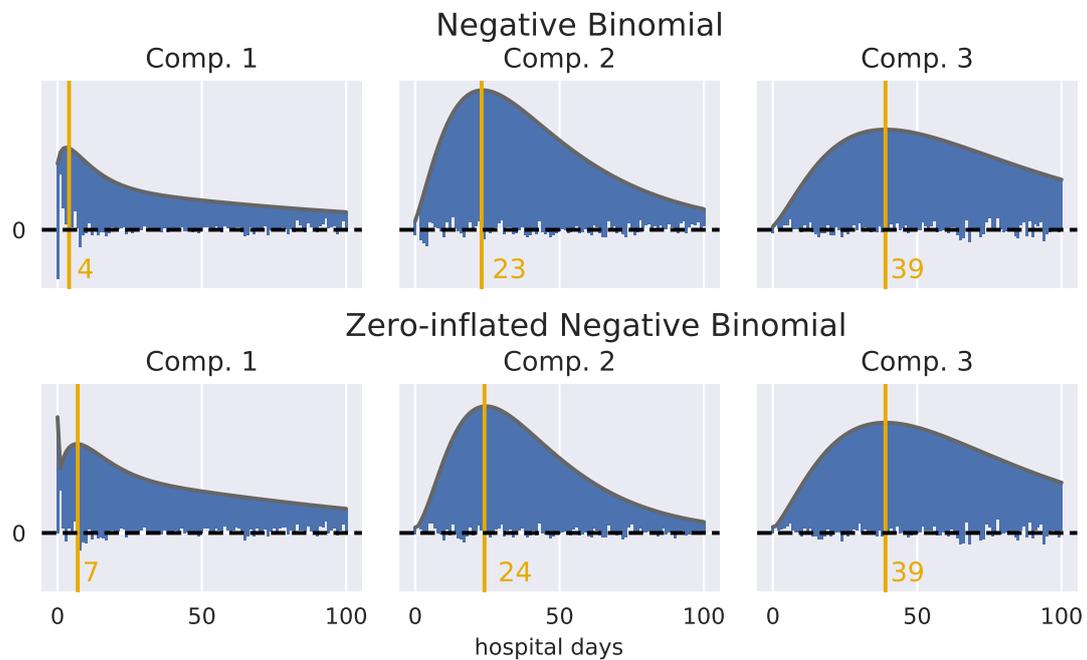


Figure 5: Hanging rootograms showing goodness-of-fit by comparing observed frequencies from the model to expected frequencies for NP-NB (top row) and NP-ZINB (bottom row) for all three components based on component assignment weights from the posterior. Ideal fit is indicated when the bars do not overlap or underlie the horizontal zero line. The vertical line marks the mode.

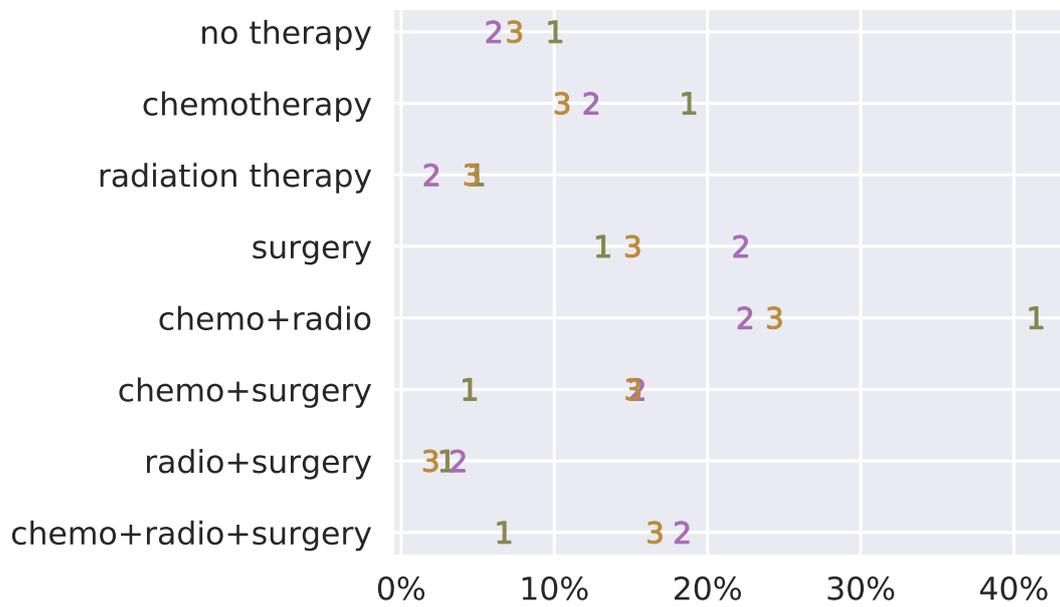


Figure 6: Plot for treatment types as percentages after using hard component assignments based on the MD-ZINB for the AOK data set. Numbers mark the three components.

9 Appendix

Component	N	high overlap				medium overlap				low overlap			
		β_0	β_1	β_2	ψ	β_0	β_1	β_2	ψ	β_0	β_1	β_2	ψ
1	3000	0.0	0.5	-0.5	1.00	0.0	0.5	-0.5	1.00	0.0	0.5	-0.5	1.00
2	5000	1.0	0.0	0.0	13.25	1.8	0.0	0.0	25.75	3.2	0.0	0.0	50.75
3	3000	1.7	-0.5	0.5	25.50	2.4	-0.5	0.5	50.50	3.9	-0.5	0.5	100.50
4	4000	2.0	0.0	0.0	37.75	2.7	0.0	0.0	75.25	4.3	0.0	0.0	150.25
5	2000	2.3	0.5	-0.5	50.00	3.0	0.5	-0.5	100.00	4.6	0.5	-0.5	200.00

Table 2: Data-generating parameter for the simulation study. N indicates the number of draws in each component.

2 Comp.	Comp. 1				Comp. 2				
	med. overlap	β_0	β_1	β_2	c_1	β_0	β_1	β_2	c_2
estimate	0.0	0.4	-0.5	0.60	3.0	0.5	-0.5	0.40	
truth	0.0	0.5	-0.5	0.60	3.0	0.5	-0.5	0.40	

4 Comp.	Comp. 1				Comp. 2				Comp. 3				Comp. 4			
	high overlap	β_0	β_1	β_2	c_1	β_0	β_1	β_2	c_2	β_0	β_1	β_2	c_3	β_0	β_1	β_2
estimate	-0.6	0.7	-1.7	0.10	1.2	0.1	-0.1	0.59	2.1	-0.0	0.0	0.15	2.3	0.5	-0.5	0.16
truth	0.0	0.5	-0.5	0.21	1.0	0.0	0.0	0.36	2.0	0.0	0.0	0.29	2.3	0.5	-0.5	0.14

5 Comp.	Comp. 1				Comp. 2				Comp. 3				Comp. 4				Comp. 5			
	high overlap	β_0	β_1	β_2	c_1	β_0	β_1	β_2	c_2	β_0	β_1	β_2	c_3	β_0	β_1	β_2	c_4	β_0	β_1	β_2
estimate	2.1	0.5	-0.5	0.10	1.2	0.1	0.0	0.35	1.6	-0.4	0.4	0.27	2.0	0.1	-0.1	0.17	2.3	0.5	-0.5	0.12
truth	0.0	0.5	-0.5	0.18	1.0	0.0	0.0	0.29	1.7	-0.5	0.5	0.18	2.0	0.0	0.0	0.24	2.3	0.5	-0.5	0.12

5 Comp.	Comp. 1				Comp. 2				Comp. 3				Comp. 4				Comp. 5			
	med. overlap	β_0	β_1	β_2	c_1	β_0	β_1	β_2	c_2	β_0	β_1	β_2	c_3	β_0	β_1	β_2	c_4	β_0	β_1	β_2
estimate	1.0	0.3	-0.5	0.25	1.9	-0.1	0.0	0.23	2.4	-0.5	0.5	0.19	2.7	0.0	0.0	0.21	3.0	0.5	-0.5	0.12
truth	0.0	0.5	-0.5	0.18	1.8	0.0	0.0	0.29	2.4	-0.5	0.5	0.18	2.7	0.0	0.0	0.24	3.0	0.5	-0.5	0.12

5 Comp.	Comp. 1				Comp. 2				Comp. 3				Comp. 4				Comp. 5			
	low overlap	β_0	β_1	β_2	c_1	β_0	β_1	β_2	c_2	β_0	β_1	β_2	c_3	β_0	β_1	β_2	c_4	β_0	β_1	β_2
estimate	0.1	0.5	-0.6	0.22	3.2	0.0	0.0	0.22	3.9	-0.5	0.5	0.22	4.3	0.0	0.0	0.22	4.6	0.5	-0.5	0.12
truth	0.0	0.5	-0.5	0.18	3.2	0.0	0.0	0.29	3.9	-0.5	0.5	0.18	4.3	0.0	0.0	0.24	4.6	0.5	-0.5	0.12

Table 3: Comparison of true and estimated parameters for cases where the number of mixture components was accurately estimated by MD-NB. β 's indicate intercept and regression coefficients, c 's are the mixture weights. All estimates are based on posterior medians. Because the estimated mixture components are not always in the same order, estimated components were hand-matched to their corresponding best-fitting component.

4 Comp.	Comp. 1				Comp. 2				Comp. 3				Comp. 4			
med. overlap	β_0	β_1	β_2	c_1	β_0	β_1	β_2	c_2	β_0	β_1	β_2	c_3	β_0	β_1	β_2	c_4
estimate	2.3	0.2	-0.2	0.26	2.3	0.2	-0.2	0.38	2.3	0.1	-0.2	0.28	2.3	0.2	-0.2	0.08
truth	0.0	0.5	-0.5	0.21	1.0	0.0	0.0	0.36	2.0	0.0	0.0	0.29	2.3	0.5	-0.5	0.14
4 Comp.	Comp. 1				Comp. 2				Comp. 3				Comp. 4			
low overlap	β_0	β_1	β_2	c_1	β_0	β_1	β_2	c_2	β_0	β_1	β_2	c_3	β_0	β_1	β_2	c_4
estimate	-0.3	0.5	-0.6	0.20	3.9	0.2	-0.2	0.77	4.4	0.2	-0.2	0.02	4.5	0.5	-0.5	0.01
truth	0.0	0.5	-0.5	0.21	1.0	0.0	0.0	0.36	2.0	0.0	0.0	0.29	2.3	0.5	-0.5	0.14

Table 4: Comparison of true and estimated parameters for cases where the number of mixture components was accurately estimated by AIC and BIC. β 's indicate intercept and regression coefficients, c 's are the mixture weights.

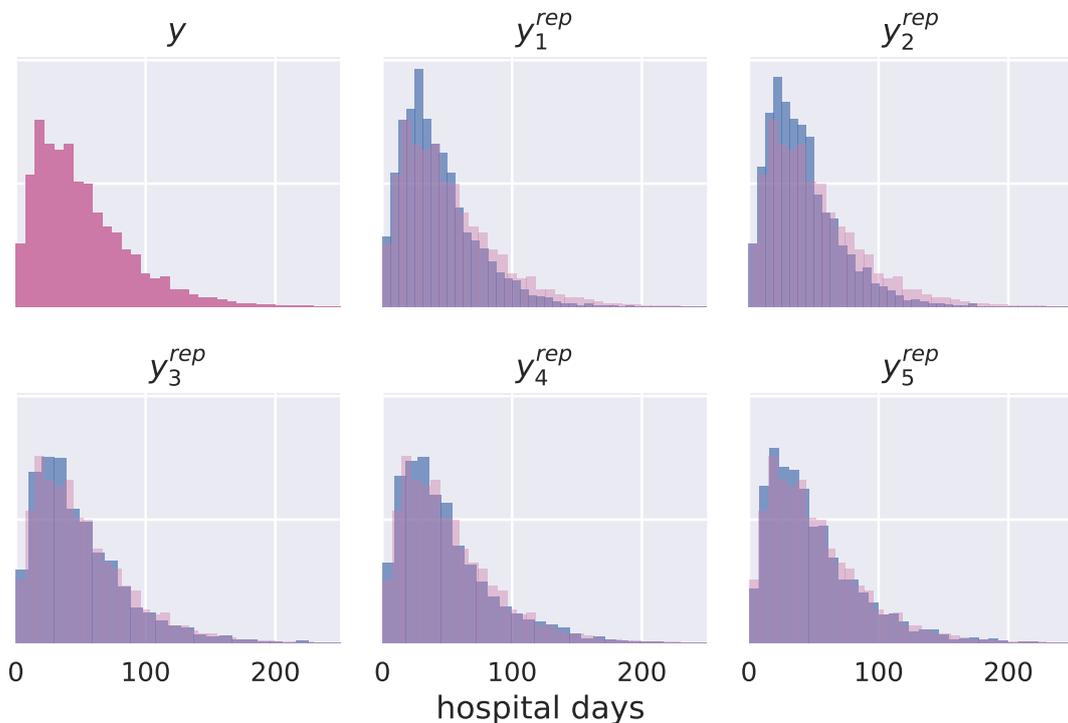


Figure 7: Density histograms for replicated outcome y^{rep} from simulating data from the posterior predictive distribution using the observed predictors. True outcome y from the AOK data set for comparison.

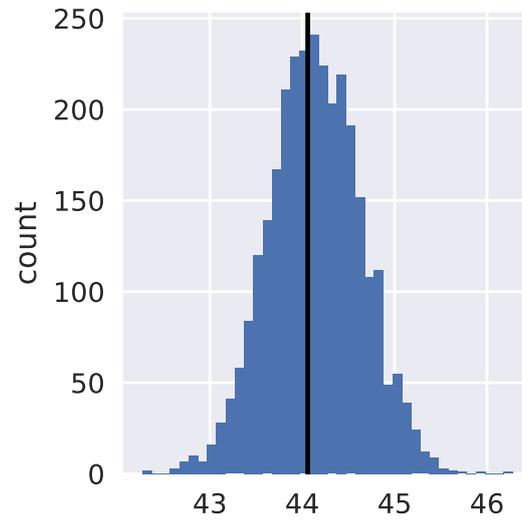


Figure 8: Histogram of means based on 1000 replicated data sets from the posterior predictive distribution. The black line marks the observed mean of hospital days from the AOK data set.

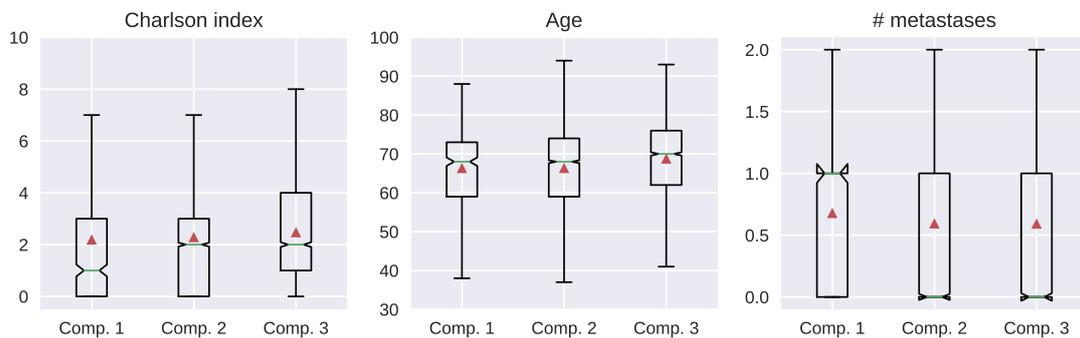


Figure 9: Boxplots for Charlson comorbidity index, age, and the number of metastases after using hard component assignments based on the MD-ZINB for the AOK data set. The red triangle marks the mean.

References

- [1] D. Kashihara and K. Carper. National Health Care Expenses in the U.S. Civilian Non-institutionalized Population, 2007. Technical Report Statistical Brief #272, Agency for Healthcare Research and Quality, Rockville, MD, December 2009.
- [2] Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108, 2015.
- [3] Borislava Mihaylova, Andrew Briggs, Anthony O’Hagan, and Simon G Thompson. Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20(8):897–916, 2011.
- [4] William Greene. Accounting for excess zeros and sample selection in Poisson and Negative Binomial regression models. Technical report, New York University, Leonard N. Stern School of Business, Department of Economics, 1994.
- [5] William Greene and others. Functional form and heterogeneity in models for count data. *Foundations and Trends in Econometrics*, 1(2):113–218, 2007.
- [6] Joseph M Hilbe. *Modeling Count Data*. Cambridge University Press, 2014.
- [7] Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.
- [8] Kelvin K. W. Yau, Kui Wang, and Andy H. Lee. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. *Biometrical Journal*, 45(4):437–452, June 2003.
- [9] Naihua Duan, Willard G Manning, Carl N Morris, and Joseph P Newhouse. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, 1(2):115–126, 1983.
- [10] Peiming Wang, Martin L. Puterman, Iain Cockburn, and Nhu Le. Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics*, 52(2):381, June 1996.
- [11] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. John Wiley & Sons, 2004.
- [12] D. M. Titterington, Adrian F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.
- [13] Sylvia Frühwirth-Schnatter. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209, 2001.
- [14] Dankmar Böhning and Wilfried Seidel. Editorial: Recent developments in mixture models. *Computational Statistics & Data Analysis*, 41(3-4):349–357, January 2003.
- [15] Bengt Muthén and Kerby Shedden. Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm. *Biometrics*, 55(2):463–469, June 1999.

- [16] Winfried Pohlmeier and Volker Ulrich. An econometric model of the two-part decisionmaking process in the demand for health care. *Journal of Human Resources*, pages 339–361, 1995.
- [17] Partha Deb and Pravin K. Trivedi. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12(3):313–336, may 1997.
- [18] Byung-Jung Park and Dominique Lord. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention*, 41(4):683–691, July 2009.
- [19] Teresa Bago d’Uva. Latent class models for utilisation of health care. *Health Economics*, 15(4):329–343, 2006.
- [20] Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, June 2002.
- [21] Aroon D Hingorani, D. A. van der Windt, Richard D Riley, Keith Abrams, Karel GM Moons, Ewout W Steyerberg, Sara Schroter, Willi Sauerbrei, Douglas G Altman, and Harry Hemingway. PROGnosis RESearch Strategy (PROGRESS) 4: Stratified medicine research. *BMJ*, 346:e5793, 2013.
- [22] S. Nicole Hastings, Heather E. Whitson, Richard Sloane, Lawrence R. Landerman, Carolyn Horney, and Kimberly S. Johnson. Using the past to predict the future: Latent class analysis of patterns of health service use of older adults in the emergency department. *Journal of the American Geriatrics Society*, 62(4):711–715, 2014.
- [23] Megan S Schuler, Nina R Joyce, Haiden A Huskamp, Elizabet B Lamont, and Laura A Hatfield. Medicare beneficiaries with advanced cancer experience diverse patterns of care from diagnosis to death. *Health Affairs*, 36(7):1193–1200, 2017.
- [24] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.
- [25] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [26] Geoffrey J McLachlan and Suren Rathnayake. On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355, 2014.
- [27] John W Lau and Peter J Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558, 2007.
- [28] Sylvia Richardson and Peter J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, 1997.

- [29] Daiane Aparecida Zuanetti and Luis Aparecido Milan. A generalized mixture model applied to diabetes incidence data. *Biometrical Journal*, 59(4):826–842, 2017.
- [30] Jean Diebolt and Christian P. Robert. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 56(2):363–375, 1994.
- [31] Zoran Zivkovic and Ferdinand van der Heijden. Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):651–656, 2004.
- [32] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.
- [33] Matthew Stephens. Bayesian Analysis of Mixture Models with an Unknown Number of Components- An Alternative to Reversible Jump Methods. *The Annals of Statistics*, 28(1):40–74, 2000.
- [34] Qiwei Li, Michele Guindani, Brian J Reich, Howard D Bondell, and Marina Vannucci. A bayesian mixture model for clustering and selection of feature occurrence rates under mean constraints. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):393–409, 2017.
- [35] R. Neal Axon, Mulugeta Gebregziabher, Charles J. Everett, Paul Heidenreich, and Kelly J. Hunt. Dual health care system use is associated with higher rates of hospitalization and hospital readmission among veterans with heart failure. *American Heart Journal*, 174:157 – 163, 2016.
- [36] Brian H Neelon, A James O’Malley, and Sharon-Lise T Normand. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling: An International Journal*, 10(4):421–439, 2010.
- [37] Emanuela Dreassi and Emilia Rocco. A Bayesian semiparametric model for non negative semicontinuous data. *Communications in Statistics - Theory and Methods*, 46(10):5133–5146, 2017.
- [38] Brian Neelon, A James O’Malley, and Valerie A Smith. Modeling zero-modified count and semicontinuous data in health services research part 2: Case studies. *Statistics in Medicine*, 35(27):5094–5112, 2016.
- [39] A Matthew Prina, Martijn Huisman, Bu B Yeap, Graeme J Hankey, Leon Flicker, Carol Brayne, and Osvaldo P Almeida. Association between depression and hospital outcomes among older men. *Canadian Medical Association Journal*, 185(2):117–123, 2013.
- [40] Yajie Zou, Yunlong Zhang, and Dominique Lord. Application of finite mixture of Negative Binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention*, 50:1042–1051, January 2013.

- [41] Mei-Chen Hu, Martina Pavlicova, and Edward V Nunes. Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse*, 37(5):367–375, 2011.
- [42] J.M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, 2 edition, 2011.
- [43] A. Colin Cameron and Pravin K. Trivedi. Econometric models based on count data. Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1(1):29–53, 1986.
- [44] Michel Wedel and Wayne S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(1):21–55, March 1995.
- [45] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 39(1):1–38, 1977.
- [46] R. C. Jansen. Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics*, 49(1):227–231, 1993.
- [47] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8):1–25, 2008.
- [48] Achim Zeileis and Christian Kleiber. *Countreg: Count Data Regression*. 2018.
- [49] Friedrich Leisch. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18, 2004.
- [50] Bettina Grün and Friedrich Leisch. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51(11):5247–5252, 2007.
- [51] Bettina Grün and Friedrich Leisch. FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35, 2008.
- [52] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.
- [53] Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37, 2016.
- [54] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.
- [55] Larissa Schwarzkopf, Margarethe Wacker, Rolf Holle, Reiner Leidl, Christian Günster, Jürgen-Bernhard Adler, and Rudolf Maria Huber. Cost-components of lung cancer care within the first three years after initial diagnosis in context of different treatment regimens. *Lung Cancer*, 90(2):274–280, 2015.

- [56] Vijaya Sundararajan, Toni Henderson, Catherine Perry, Amanda Muggivan, Hude Quan, and William A Ghali. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *Journal of Clinical Epidemiology*, 57(12):1288–1294, 2004.
- [57] Martijn Tennekes, Edwin de Jonge, Piet JH Daas, and others. Visualizing and inspecting large datasets with tableplots. *Journal of Data Science*, 11(1):43–58, 2013.
- [58] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC, Boca Raton, FL, 2014.
- [59] Christian Kleiber and Achim Zeileis. Visualizing count data regressions using rootograms. *The American Statistician*, 70(3):296–303, 2016.
- [60] Akio Onogi, Masanobu Nurimoto, and Mitsuo Morita. Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods. *BMC Bioinformatics*, 12(1):263, 2011.
- [61] Christian Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015.
- [62] C Gridelli, F Perrone, F Nelli, S Ramponi, and F De Marinis. Quality of life in lung cancer patients. *Annals of Oncology*, 12(suppl 3):S21–S25, 2001.
- [63] Ramsés H Mena and Stephen G Walker. On the Bayesian mixture model and identifiability. *Journal of Computational and Graphical Statistics*, 24(4):1155–1169, 2015.
- [64] Merrilee Hurn, Ana Justel, and Christian P. Robert. Estimating Mixtures of Regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79, 2003.
- [65] Riccardo Rastelli and Nial Friel. Optimal Bayesian estimators for latent variable cluster models. *Statistics and Computing*, pages 1–18, October 2017.
- [66] Weixin Yao and Bruce G. Lindsay. Bayesian mixture labeling by highest posterior density. *Journal of the American Statistical Association*, 104(486):758–767, 2009.
- [67] Burton Wu, Clare A McGrory, and Anthony N Pettitt. A new variational Bayesian algorithm with application to human mobility pattern modeling. *Statistics and Computing*, 22(1):185–203, 2012.
- [68] David I Ohlssen, Linda D Sharples, and David J Spiegelhalter. Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons. *Statistics in Medicine*, 26(9):2088–2112, 2007.
- [69] Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.
- [70] Hemant Ishwaran and Lancelot F. James. Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532, 2002.
- [71] Abel Rodriguez and David B Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1), 2011.

- [72] Siew Li Tan and David J Nott. Variational approximation for mixtures of linear mixed models. *Journal of Computational and Graphical Statistics*, 23(2):564–585, 2014.
- [73] Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.
- [74] Antonio Canale and Igor Prünster. Robustifying Bayesian nonparametric mixtures for count data. *Biometrics*, 73(1):174–184, 2017.
- [75] Antonio Canale and David B. Dunson. Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539, December 2011.
- [76] Mian Huang and Weixin Yao. Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724, 2012.
- [77] Peter Orbanz and Yee Whye Teh. Bayesian Nonparametric Models. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*. Springer, Boston, MA, 2 edition, 2017.
- [78] N.L. Hjort, C. Holmes, P. Müller, and S.G. Walker, editors. *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.
- [79] J.K. Ghosh and R.V. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics. Springer New York, New York, NY, 2006.
- [80] Peter Müller and Riten Mitra. Bayesian nonparametric inference—why and how. *Bayesian Analysis*, 8(2), 2013.
- [81] Alan E Gelfand and Athanasios Kottas. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11(2):289–305, June 2002.
- [82] C. B. Storlie, S. M. Myers, S. K. Katusic, A. L. Weaver, R. G. Voigt, P. E. Croarkin, R. E. Stoeckel, and J. D. Port. Clustering and variable selection in the presence of mixed variable types and missing data. *Statistics in Medicine*, 37(19):2884–2899, 2018.
- [83] N. Bouguila. Count Data Modeling and Classification Using Finite Mixtures of Distributions. *IEEE Transactions on Neural Networks*, 22(2):186–198, February 2011.
- [84] Gilbert W Fellingham, Athanasios Kottas, and Brian M Hartman. Bayesian nonparametric predictive modeling of group health claims. *Insurance: Mathematics and Economics*, 60:1–10, 2015.
- [85] M. Zhou and L. Carin. Negative Binomial Process Count and Mixture Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, February 2015.

- [86] Thomas S Ferguson. Bayesian density estimation by mixtures of normal distributions. In M. H. Rizvi, J. S. Rustagi, and D. Siegmund, editors, *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, pages 287–302. Academic Press, 1983.