Received: 19 March 2020 | Revised: 6 August 2020 | Accepted: 14 August 2020

DOI: 10.1002/bimj.202000076

Biometrical Journal

RESEARCH PAPER

Isolating cost drivers in interstitial lung disease treatment using nonparametric Bayesian methods

Christoph F. Kurz[1,2] | Seth Stafford[3]

[1] Helmholtz Zentrum München, Institute of Health Economics and Health Care Management, Neuherberg, Germany

[2] Munich School of Management and Munich Center of Health Sciences, Ludwig-Maximilians-Universität München, Munich, Germany

[3] ServiceNow, Machine Learning & NLP, Santa Clara, CA, USA

Correspondence
Christoph F. Kurz, Helmholtz Zentrum München, Institute of Health Economics and Health Care Management, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.
Email: christoph.kurz@helmholtz-muenchen.de

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion; EM, expectation maximization; ILD, interstitial lung disease; KL, Kullback–Leibler; MCMC, Markov chain Monte Carlo; MLE, maximum likelihood estimation; MOVI, memoized online variational inference; VB, variational Bayes.

[Correction added on 25 September 2020, after first online publication: Projekt Deal funding statement has been added.]

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially for data confidentiality reasons.

Abstract

Mixture modeling is a popular approach to accommodate overdispersion, skewness, and multimodality features that are very common for health care utilization data. However, mixture modeling tends to rely on subjective judgment regarding the appropriate number of mixture components or some hypothesis about how to cluster the data. In this work, we adopt a nonparametric, variational Bayesian approach to allow the model to select the number of components while estimating their parameters. Our model allows for a probabilistic classification of observations into clusters and simultaneous estimation of a Gaussian regression model within each cluster. When we apply this approach to data on patients with interstitial lung disease, we find distinct subgroups of patients with differences in means and variances of health care costs, health and treatment covariates, and relationships between covariates and costs. The subgroups identified are readily interpretable, suggesting that this nonparametric variational approach to inference can discover valid insights into the factors driving treatment costs. Moreover, the learning algorithm we employed is very fast and scalable, which should make the technique accessible for a broad range of applications.

KEYWORDS

Bayesian statistics, health care costs, lung disease, mixture model, nonparametric models, variational Bayes

1 | INTRODUCTION

Efforts to improve health outcomes while simultaneously reducing health care spending require a detailed understanding of the drivers of variation in spending and outcomes. In this study, we examine variation in treatment costs among patients diagnosed with interstitial lung disease (ILD) using recently developed techniques from Bayesian statistics. Our contribution is a practical review and comparison among mixture models for continuous health care cost data that can be fitted with standard software. In simulations, we compare the models' ability to detect the true number of clusters.

1.1 Interstitial lung disease

ILD refers to a group of lung diseases that affect the interstitium. The pulmonary interstitium is the lace-like anatomic space that is bounded by the basement membranes of epithelial and endothelial cells. The pathologic features of ILD, even if originating in the interstitium, regularly include structures that are well beyond it, including the alveolar space, small airways, vessels, and even the pleura (Raghu & Brown, 2004). That makes ILD a very heterogeneous group of disorders that includes at least two dozen different types. Some forms of ILD are short-lived, while others are chronic and reversible. A U.S. study reported incidences of 31.5 per 100,000 among men and 26.1 per 100,000 among women (Coultas, Zumwalt, Black, & Sobonya, 1994). European studies have reported slightly lower ILD incidences between 4.6 and 7.6 per 100,000 inhabitants/year (Agostini et al., 2001; Schweisfurth, 1996; Thomeer et al., 2001; López-Campos & Rodríguez-Becerra, 2004). The economic burden of ILD is very high, including direct costs for hospitalization and exacerbation (Raimundo et al., 2016; Rice et al., 2017) and indirect costs like lost productivity (Kawalec & Malinowski, 2015). To better understand the heterogeneity of ILD and its associated health care expenditures, researchers have sought to identify patient subgroups with different utilization and spending patterns (Hingorani et al., 2013).

1.2 Mixture modeling

Models based on mixtures of parametric models are commonly used as a flexible way to accommodate excess zeros, overdispersion, and heavy tails common in health care utilization cost data (Mihaylova et al., 2011). These mixture models, also known as latent class models (Böhning & Seidel, 2003; Muthén & Shedden, 1999) or switching models (Frühwirth-Schnatter, 2001), are motivated by the concern that different parts of the response distribution could be differently affected by covariates (e.g., low-cost users, high-cost users). Mixture models represent a complicated density as a linear combination of simpler densities and therefore identify groups of observations with similar outcomes using unsupervised clustering. They are widely used—see reviews in McLachlan and Peel (2004) and Titterington, Smith, and Makov (1985)—and often perform better than standard generalized linear models and the hurdle model (Deb & Trivedi, 1997), especially for health care utilization (Kurz & Hatfield, 2019; Oganisian, Mitra, & Roy, 2020).

In the simplest case of mixture modeling, one begins with a natural hypothesis as to which populations might exhibit different behavior, providing both an initial choice of the number of clusters and even which data points belong to each cluster. In general, however, the problem is precisely to infer the clusters and derive hypotheses for further study *from* the model. The usual practice, in fact, amounts to either deciding the number of components ex ante, by choosing a convenient and interpretable number such as two or three, or deciding ex post, by generating models with different numbers of components and manually searching for a plausible best fit by comparing quantities such as Akaike information criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978), or likelihood ratio (McLachlan & Rathnayake, 2014). The model selection criterion applied in our study is similar, because it relies on model fit characteristics of the posterior distribution, but not in general equivalent to AIC or BIC. The advantage to our approach is that the criteria are efficiently and automatically applied during fitting, while yielding models that are very plausible according to the data.

1.3 Bayesian nonparametrics

In this paper, we will apply the nonparametric variational Bayesian (VB) approach to statistical inference. VB has been shown to be an efficient and accurate method especially suited for fitting mixture models (McGrory & Titterington, 2007). The development of advanced Markov chain Monte Carlo (MCMC) methods was a key step in making it possible to compute large hierarchical models that require integrations over hundreds of unknown parameters to perform Bayesian inference (Robert & Casella, 2005). Unfortunately, MCMC sampling can be prohibitively slow for large data sets or complicated likelihood functions (Blei & Jordan, 2006; Müller, Erkanli, & West, 1996). Recently, VB methods have been developed for fast approximation of the intractable integrals arising in Bayesian inference. VB is an alternative to MCMC sampling methods for taking a fully Bayesian approach to statistical inference over complex distributions that are difficult to directly evaluate or sample from. MCMC techniques generate a numerical approximation to the exact posterior by averaging many (often computationally expensive) samples, VB provides an exact analytical approximate posterior by sampling parameters for members of a variational family of tractable distributions. Both the sampling procedure and inference done using the resulting approximate posterior are much more efficient. VB approximation can lead to good point estimates and estimates for the marginal posterior distribution, and excellent predictive inferences. By working with a suitably flexible prior, VB mixtures can automatically select the number of mixture components. Simulation studies have shown that the estimated number of mixture components is very accurate (Tan & Nott, 2014), and that variational approximations to the posterior are competitive with MCMC for mixture models (Blei & Jordan, 2006). When it comes to model selection criteria, Beal and Ghahramani (2003) argue that VB is generally preferable to BIC and comparable to annealed importance sampling (Neal, 2001) at much lower computational cost.

Because VB inference algorithms are usually faster than MCMC, they are an excellent choice for large-scale data sets, which are becoming more and more prevalent in health economic analyses through the availability of claims data and electronic health records. Especially in clinical decision support “speed is everything” Bates et al. (2003) and MCMC methods that take several hours to run can be infeasible in practice.

Please note that in the following we refer to mixture components as *clusters* for brevity and to emphasize the unsupervised clustering done by this kind of model.

1.4 | Outline

The rest of the paper is organized as follows. First, we describe the VB approach to Bayesian inference in Section 2. Then, we detail our VB mixture regression model in Section 3. Section 4 outlines the data, and we present the results of applying VB to these data in Section 5. Finally, we conclude in Section 6 with implications of the model results.

2 | VB INFERENCE

Mixture models are a common approach to dealing with multimodal and skewed data, but we would like to take the additional step of allowing the model to find the number of clusters, which allows for the best overall fit to the data. This is the *nonparametric* approach, so-named simply because we do not specify a best guess as to the number of clusters. We begin with the prior view that our data set will look like a mixture of some number of Gaussian regressions, and then use a variational algorithm to find the best fitting number and shape of such regressions to model the data set. The specific technique we use, called memoized online variational inference (MOVI), was developed by Hughes and Sudderth (2013). It features a birth/merge technique for adaptively adjusting the number of clusters which manages to avoid local minima in some challenging cases.

Our approach is Bayesian in that we start with a diffuse prior distribution on the joint parameters which will describe our collection of approximately linear cost drivers, then use the evidence available in the data set to produce a posterior distribution from which we can sample candidate cluster assignments and regression coefficients. Maximum likelihood estimation (MLE) is also a common approach, but it is prone to singularities, which can compromise the results (Fraley & Raftery, 2007). Intuitively, pursuing a point estimate of the maximum likelihood can diverge because of unfortunate data distributions, which lead to singularities in covariance estimation. Bayesian modeling will produce estimates similar to those from MLE when applied to cleaner data sets, but is more robust when applied to the more awkward, irregular data sets which are so common in applications.

Finally, our approach is variational in nature because the family of possible distributions formed from convex combinations of an arbitrary number of component distributions is itself very large and the natural measure of goodness of fit—generally the Kullback–Leibler (KL) divergence—defines a nonconvex loss. It is not feasible to directly compute a unique best fitting mixture, so we must find iterative approximations that reduce the KL divergence from the true posterior distribution. We will provide a brief exposition of the variational approach before focusing on the specifics of our model, and recommend Blei, Kucukelbir, and McAuliffe (2017) and Ormerod and Wand (2010) as excellent reviews of the topic for readers seeking more insight.

Our goal is to assign our data to clusters that are each well approximated by Gaussian linear regressions and estimate the best fit for each of those regressions. In effect, we are choosing a variational family $\mathcal{D}$ of analytically tractable distributions parameterized by a choice of subsets and corresponding regression coefficients, and we want to learn which member $q(\mathbf{z})$ of this family best approximates the true posterior. If the distributions involved were simpler, we might do this by directly maximizing the posterior distribution of the parameters given the evidence—traditional maximum a posteriori probability estimation. But given the complexity, we first approximate this true posterior with the closest fitting member of our variational family.

Writing our parameters for now as some latent variables $\mathbf{z} = z_{1:m}$ to be specified later, and the observations as $\mathbf{x} = x_{1:n}$, we want to approximate the posterior $p(\mathbf{z}|\mathbf{x})$ with a variational distribution $q(\mathbf{z})$:

$$p(\mathbf{z}|\mathbf{x}) \approx q(\mathbf{z}),$$

and then choose $\mathbf{z}$ to maximize the selected $q(\mathbf{z})$. We find $q(\mathbf{z})$ by minimizing the KL-divergence between $q(\mathbf{z})$ and the true posterior $p(\mathbf{z}|\mathbf{x})$, resulting in the following optimization problem:

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z})\in\mathcal{D}}{\arg\min}\, \mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})).$$

Recalling the definition of KL divergence:

$$\begin{aligned}\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x})) &= \mathbb{E}_q\left[\ln \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right]\\ &= \mathbb{E}_q[\ln q(\mathbf{z})] - \mathbb{E}_q[\ln p(\mathbf{z}|\mathbf{x})]\\ &= \mathbb{E}_q[\ln q(\mathbf{z})] - \mathbb{E}_q[\ln p(\mathbf{z},\mathbf{x})] + \ln p(\mathbf{x}),\end{aligned}$$

where $\|$ marks the statistical distance that emphasizes the order of the arguments. Note that the variational strategy is to minimize $\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x}))$ rather than the reverse $\mathrm{KL}(p(\mathbf{z}|\mathbf{x})\|q(\mathbf{z}))$. These quantities are not always the same and $\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x}))$ has the advantage of using expectation with respect to the tractable $q(\mathbf{z})$ distribution. Note also that $\ln p(\mathbf{x})$ is constant as we vary $q(\mathbf{z}) \in \mathcal{D}$, so we can focus instead on maximizing the equivalent objective (with reversed sign):

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\ln p(\mathbf{z},\mathbf{x})] - \mathbb{E}_q[\ln q(\mathbf{z})].$$

Noting that $\mathrm{KL}(\cdot\|\cdot) \geq 0$ and rearranging terms above, we find that:

$$\begin{aligned}\ln p(\mathbf{x}) &\geq \mathbb{E}_q[\ln p(\mathbf{z},\mathbf{x})] - \mathbb{E}_q[\ln q(\mathbf{z})]\\ &= \mathrm{ELBO}(q),\end{aligned}$$

which accounts for the name ELBO: short for *evidence lower bound*.

3 | MODEL DEFINITION

We now focus on the specific model for our application case, adopting the notation of Hughes and Sudderth (2013) and Hughes (2016) as far as possible to make it easier to relate our model to their theoretical work. To identify patient subpopulations with distinctive health care costs and effects of covariates on costs, we model the cost data as a mixture of Bayesian linear regressions. The model is hierarchical with two stages: first, an *allocation* step which assigns each data point to a cluster, and second, an *observation* process that estimates the costs y_n by regression on the known covariates $\mathbf{x}_n$ within each cluster. Formally, for $n = 1, \ldots, N$ observations, the data set consists of pairs $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$, where $\mathbf{x}_n$ is a covariate vector of length D and outcome y_n is scalar. For each pair there is a latent variable z_n indicating the cluster assignment. Each cluster $k \in \{1, \ldots, K\}$ has linear regression coefficients $\boldsymbol{w}_k$ (including intercept w_0) and precision scalar δ_k so the full mixture model has the form:

$$y_n \sim \sum_{k=1}^{K} r_{nk}\mathcal{N}\left(w_{0k} + \sum_{d=1}^{D} w_{dk} \cdot x_{nd}, \delta_k^{-1} \right),$$

r_{nk} is the probability that covariate vector x_n is assigned to cluster k, or $P(z_n = k)$. These are soft cluster assignments that evolve as the learning algorithm experiments with possible clusters. Then after training, we substitute hard cluster assignments in which each z_n is set to the cluster which has the max responsibility r_{nk}. Typically, the learning process will move r_{nk} values close to either 0 or 1, so these hard assignments reflect a decision to accept the current soft assignments as having "converged." In general, there cannot be more clusters than observations.

For the allocation step, we need a diffuse prior for the latent z_n variables. Because the posterior distribution for z_n should be a categorical (aka multinomial) on K clusters, the conjugate prior is the Dirichlet distribution on the K-dimensional simplex. This poses the awkward question of which dimension to work in—how many clusters to use—which we had hoped to avoid. The Dirichlet process (Blei & Jordan, 2006) is an elegant generalization of this, which folds all finite dimensional cases into an abstract infinite dimensional container, while leaving each finite dimensional truncation (first m dimensions) identical to the corresponding finite-dimensional categorical/Dirichlet conjugate pair.

More formally, the factorized variational family from which we will seek a good approximation to the true posterior $p(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta} \mid \mathbf{x})$ is then:

$$q(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta}) = \prod_{n=1}^{N} \mathrm{Cat}(z_n \mid \hat{r}_{n1}, \dots, \hat{r}_{nK}) \cdot \prod_{k=1}^{K} \mathcal{W}_1(\boldsymbol{\delta}_k \mid \hat{\nu}_k, \hat{\tau}_k)\mathcal{N}_{D+1}(\mathbf{w}_k \mid \hat{\mathbf{w}}_k, \delta_k^{-1}\hat{P}^{-1}),$$

where the *responsibility* $\hat{r}_{nk}$ is the probability that $z_n = k$, the count $\hat{\nu}_k$ and location $\hat{\tau}_k$ are parameters of the Wishart prior on precision δ_k, $\hat{P}$ is the precision matrix for the ($D + 1$-dimensional) Normal prior on $\mathbf{w}_k$, and the hat notation indicates a variational parameter undergoing updates. The MOVI algorithm (Hughes & Sudderth, 2013; Hughes, 2018) we employ alternates between computing expectations of the local parameters $\hat{r}_{nk}$ and updating global parameters $\hat{\nu}_k$, $\hat{\tau}_k$, $\hat{\mathbf{w}}_k$, $\hat{P}_k$ to maximize the evidence lower bound (ELBO) given the updated $\hat{r}_{nk}$ values. The model is initialized with $K = 10$, global parameters: $\hat{\nu}_k = 0.01$, $\hat{\tau}_k = 0.01$, $\hat{\mathbf{w}}_k = 0$, $\hat{P}_k = diag(1.0)$, and $\hat{r}_{nk}$ are inferred from initial samples of the global parameters using Mahalanobis distance between each $\mathbf{x}_n$ and cluster distribution k.

Note that the variational family as presented is specific to a choice of K. In their paper, Hughes and Sudderth (2013) are careful to use a nested truncation of the Dirichlet Process to ensure that as K is adjusted in the search for the best overall fit, the amount of probability mass assigned to most existing clusters will not need to be recomputed. The family of $q(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta})$ for a given K corresponds exactly to the subset of q's in dimension $K + 1$ in which the last component has mass zero.

In the first iteration of the variational algorithm, each data point is assigned to one of the $K = 10$ initial clusters at random according to a truncated stick breaking process. In subsequent iterations, the points are reallocated only when the clusters are identified as needing improvement (birth/merge process described below). This procedure works because of two things: first, that we have an update procedure (learning algorithm) that refines the cluster assignment and attaches meaningful estimates of mean and precision to more sensible clusters of data points, and second, that a procedure for proposing new clusters and rejecting others is able to avoid settling on poorly chosen clusters. The learning algorithm (MOVI) is a variant of expectation maximization (EM) (Dempster, Laird, & Rubin, 1977), which reliably (provably) finds local optima, and the birth/merge procedure helps identify opportunities to find better (ideally global) optima.

The MOVI algorithm is based on earlier work by Ueda and Ghahramani (2002) where they described a VB split and merge procedure to optimize an objective function that simultaneously estimates the parameters and number of cluster in a Gaussian mixture while avoiding local optima. This was improved upon by Wu, McGrory, and Pettitt (2012), which added a delete move to trim out unnecessary clusters. These deletion moves were previously developed by Attias (1999) and Corduneanu and Bishop (2001).

In MOVI the birth/merge process runs as follows: for birth, each cycle includes Collection, Creation, and Adoption steps. *Collection* subsamples data $\mathbf{x}'$ largely explained by a specific target cluster (meaning responsibility $\hat{r}_{nk}$ above a threshold), *Creation* allocates these data points to K' (eg. 10) new clusters, and *Adoption* updates local and global parameters for all $K + K'$ clusters. Merges are handled by selecting a target cluster k_a at random, then a second cluster k_b most similar to

TABLE 1 Data-generating parameters for the simulation study

Cluster	Mean	Cov. matrix
1	$\begin{pmatrix} 0 \\ -0.1 \end{pmatrix}$	$\begin{pmatrix} 0.7 & 0 \\ 0 & 0.3 \end{pmatrix}$
2	$\begin{pmatrix} 2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.2 \end{pmatrix}$
3	$\begin{pmatrix} -1.6 \\ 0.6 \end{pmatrix}$	$\begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$
4	$\begin{pmatrix} -2.8 \\ -0.5 \end{pmatrix}$	$\begin{pmatrix} 0.9 & 0.5 \\ 0.5 & 0.1 \end{pmatrix}$
5	$\begin{pmatrix} 0.5 \\ 1.9 \end{pmatrix}$	$\begin{pmatrix} 1.5 & 0 \\ 0 & 0.5 \end{pmatrix}$

k_a. A merger of k_a and k_b into k_m is tested by summing their responsibilities: $\hat{r}_{nk_m} = \hat{r}_{nk_a} + \hat{r}_{nk_b}$, recomputing the ELBO, and accepting the merge if it results in an improvement. Consult Hughes and Sudderth (2013) for details.

4 | DATA

4.1 | Monte Carlo simulation

To compare the ability of the VB approach to find the true number of clusters, we performed a small simulation study. We generated data from Gaussian mixtures with two, three, four, and five components. Data were drawn from bivariate Gaussian distributions with different mean and covariance matrices (summarized in Table 1) to produce a variety of shapes of Gaussian distributions. We limited the simulation study to two dimensions for simplicity and for visualization of the results. We compared the VB model to a finite mixture model that computes models for 2, 3, 4, and 5 clusters and selects the ideal number of clusters according to the lowest AIC value.

4.2 | AOK data set

We analyzed a data set provided by the AOK Research Institute consisting of health care billing claims. AOK is a German health insurer that covers around 30% of the German resident population. The data set contains patient-level information on inpatient and outpatient diagnoses and procedures from 2010 to 2013, as well as health care expenditures in different categories. The study population in this data set consists of patients with ILD in this observation period.

The outcome of interest was the total health care expenditures for each patient in the year after diagnosis. Expenditures include inpatient, outpatient, and medication expenditures. We included only individuals who survived for the full year, resulting in $N = 8972$ individual observations. Individual medical expenditures range from 0 to 96874, with a mean of 8409. All monetary values are in 2013-€ .

We included the following covariates in the model: age, sex, ILD type (nine different subgroups of ILD that were present in this data set), indicator variables for lung cancer, gastroesophageal reflux disease (GERD), and living in a nursing home, Charlson comorbidity index, and Elixhauser comorbidity index. The Charlson comorbidity index was calculated using ICD-10 codes as in Sundararajan et al. (2004) with the slight modification of excluding the diagnosis of lung cancer out of the group “solid tumor without metastases.”

The nine different subgroups of ILD in this data set were: idopathic interstitial pneumonia (IIP), emphysema sclerosis (Emphys.), sarcoidosis (Sarc.), drug-related ILD (Drug.), pneumoconiosis (Pneu.), radiotherapy associated ILD (Radio.), eosinophilic pneumonia (Eosino.), hypersensitivity pneumonitis (Hyper.), and rheumatic and connective tissue associated ILD (Rheum.).

Figure 1 summarizes the sample in a tableplot (Tennekes, de Jonge, & Daas, 2013). This plot shows that some ILD types, the Charlson comorbidity index, the Elixhauser index, and age are positively correlated with increasing costs. We provide `python` code of our analysis online (Kurz, 2018).

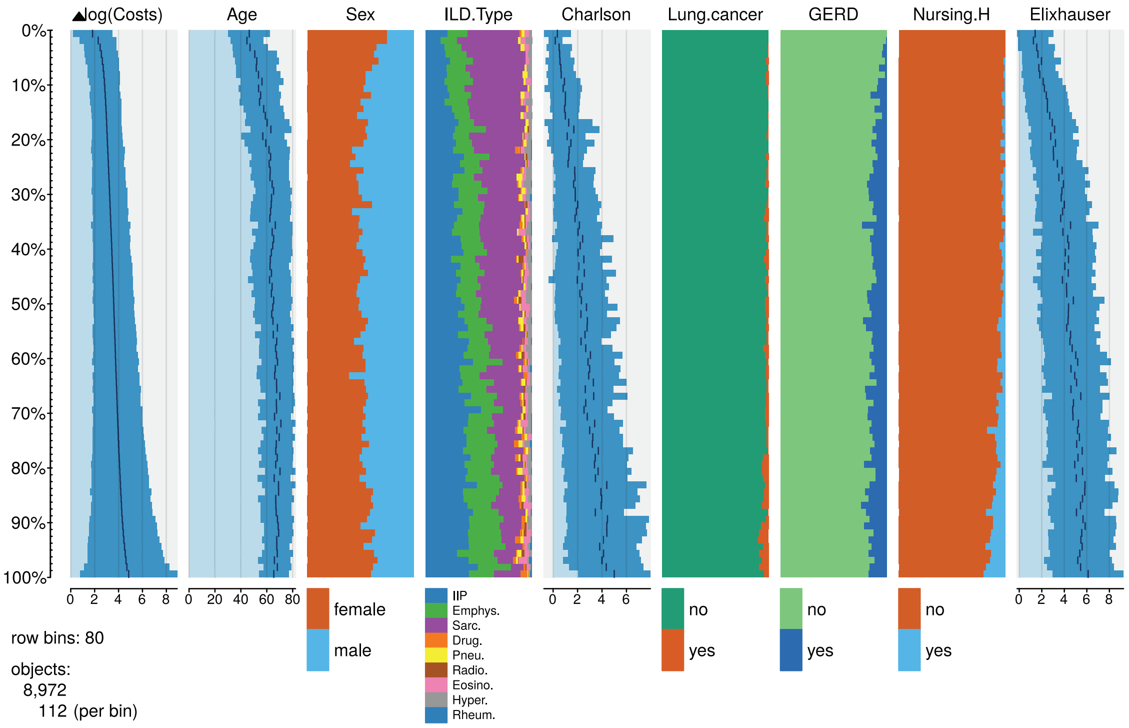

FIGURE 1 Tableplot of the AOK data set. Each column represents a variable. The whole data set is sorted by log costs (left side) in increasing order. All other variables are grouped into row bins of size 112, where numeric variables are displayed as bar charts and categorical variables as stacked bar charts. For the numeric variables, the black line in each bin marks the mean, and the dark blue bars the standard deviation

5 | RESULTS

5.1 | Cluster identification in simulated data

Figure 2 presents the number of clusters in the mixture estimated by the VB and finite mixture models in our simulation study. The VB model identified the true number of clusters in all of the simulations (2, 3, 4, and 5 components), while the finite mixture model is only accurate in the four-cluster scenario.

5.2 | AOK data set

For the AOK data set, the VB model finds five clusters (mixture components) as having the highest expected posterior mixture weights. In the following, we only show results based on this final model with five clusters.

Cluster 1 contains 24% (2139/8972) of all observations and corresponds to individuals who exhibit, on average, the least amount of costs. The mean costs in this cluster are 873 with a standard deviation of 514. Cluster 2 comprises 18% (2139/8972) of the population with mean costs of 2241 and standard deviation of 1017. Cluster 3 is the largest, with 27% (2397/8972) of individuals. Their mean costs in this cluster are 5012 with a standard deviation of 2201. The last two clusters capture the cases with the highest costs: 22% (1963/8972) of individuals are in cluster 4 and their mean costs are 12585 (standard deviation 4957). Cluster 5 has the highest spenders but is also the smallest cluster with only 10% (885/8972) of the population. Mean costs in cluster 5 are 37,724 with a standard deviation of 16,195. Please note that the percentages correspond to the cluster weights in the VB model specification.

In Figure 3, we show density histograms (including the kernel density estimate) for the distribution of costs in all clusters. This plot nicely exhibits the different means and variances for each cluster.

Figure 4 shows the standardized parameter estimates from the regression model for each component together with Bayesian 95% credible intervals (CIs). Standardized coefficients ignore the independent variable's scale of units, making

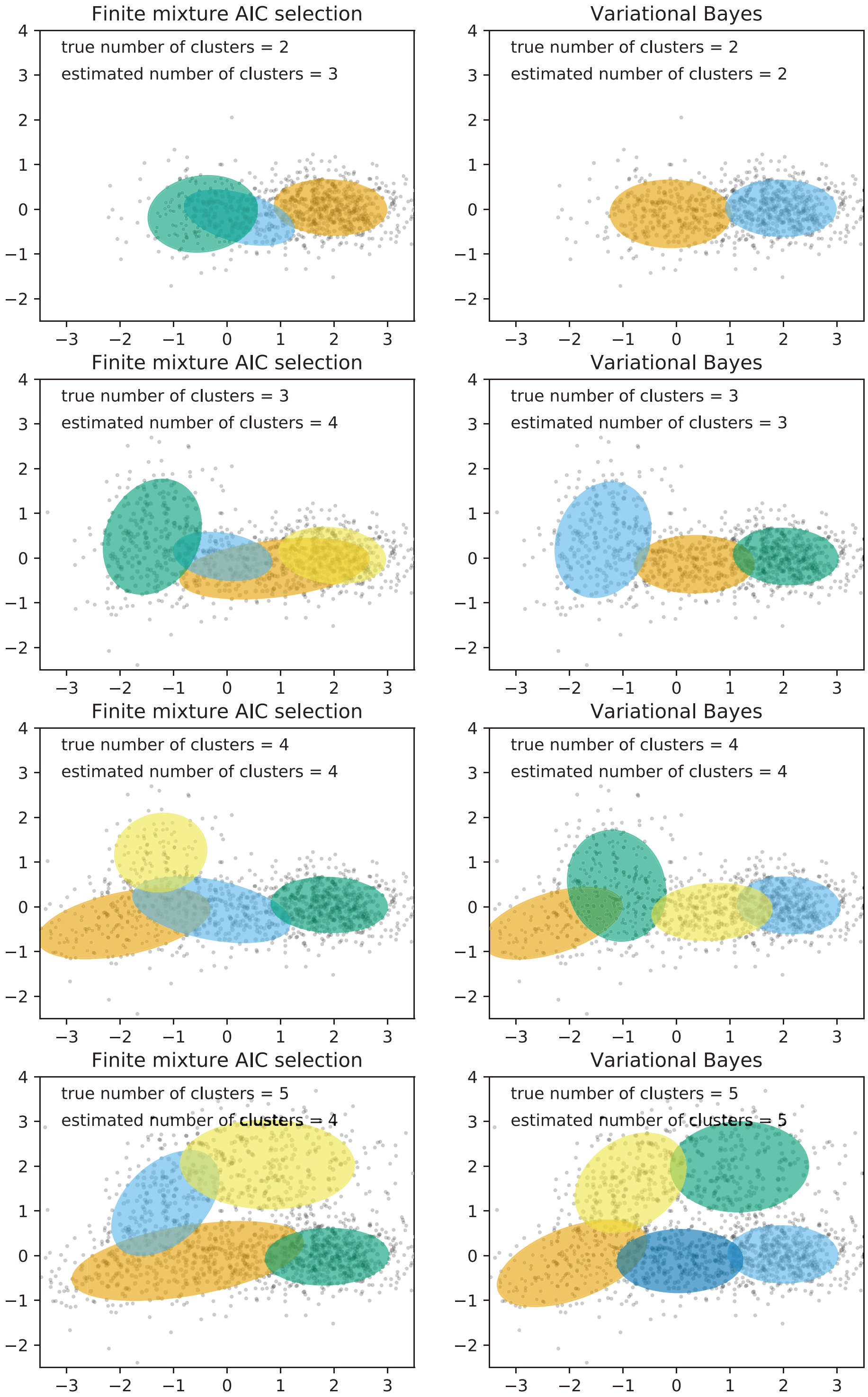

FIGURE 2 True and estimated number of clusters based on the VB model and the finite mixture model (with AIC model selection). Ellipsoids represent the 95% confidence region for each cluster as estimated by the model, gray dots are the simulated data points

it easier to compare the values. They can be interpreted as an increase in percentage of the mean in this covariate if costs increase by one standard deviation. For example, in cluster 1, the cluster with the lowest average costs, radiotherapy associated ILD has the highest impact on costs. This means, a one standard deviation increase in radiotherapy implies an increase in health care costs equal to 33% of mean health care costs. Other than the indicator for living in a nursing home, most covariates in cluster 1 only have a small (but very often significant, in terms of CIs not overlapping zero) impact on the total costs.

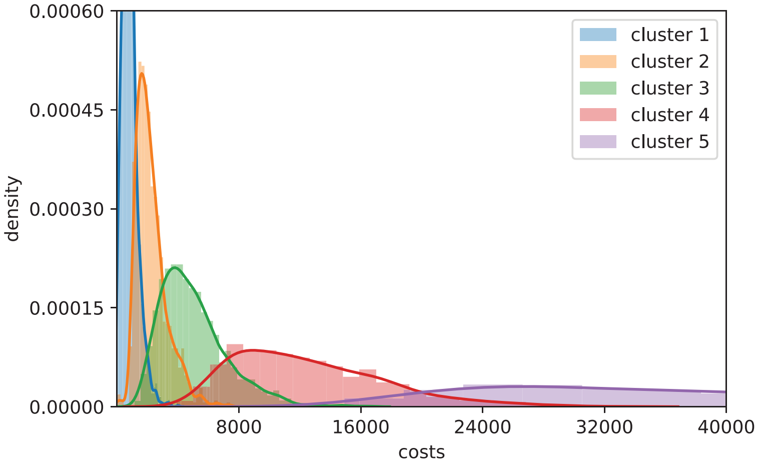

FIGURE 3 Density histograms and kernel density estimates for the costs in each clusters. Values on the x-axis (costs in €) are truncated for better visibility

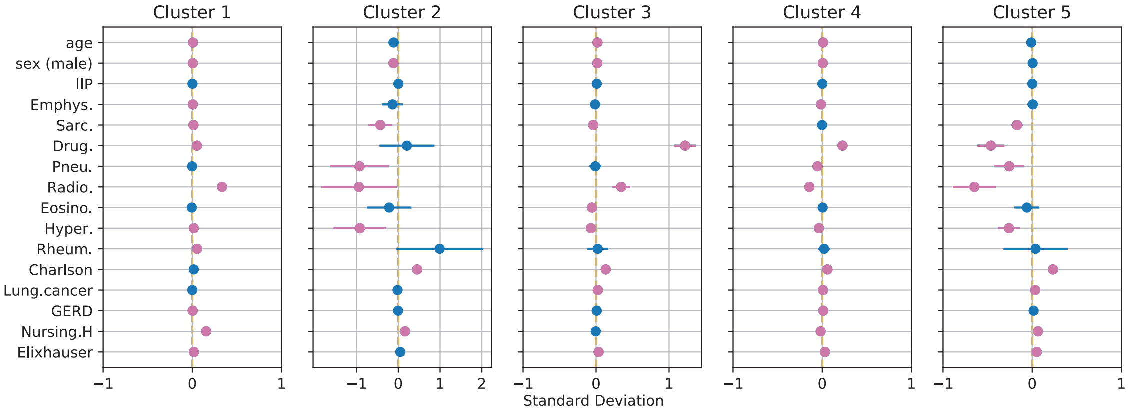

FIGURE 4 Estimation results for the five components on the AOK data set. Estimates are standardized linear regression estimates with 95% Bayesian credible intervals. Intervals that exclude the zero are shown in purple. Intercept is not shown

On the other hand, cluster 2 shows strong heterogeneity for effects of ILD subtypes on costs. Here, the subtypes sarcoidosis, pneumoconiosis, radiotherapy-associated ILD, and hypersensitivity pneumonitis show a strong negative effect of up to 90% on costs, while rheumatic ILD can increase costs up to 95%, although the uncertainty interval for this estimate is very wide. This cluster also exhibits a strong positive effect of the number of comorbidities through the Charlson index. Cluster 3 is again a very homogeneous cluster with two notable outliers: An increase of one standard deviation of drug-related ILD increases costs over 100%. Also radiation therapy–induced ILD accounts for a significant increase in costs in this cluster. Cluster 4 is similar to cluster 3 but the effect of drug-related ILD is smaller, while radiation ILD even decreases average costs slightly. The cluster with the highest average costs, cluster 5, has cost-saving effects for drug-related ILD, pneumoconiosis, radiation ILD, and hypersensitivity pneumonitis. The main reason for increased average costs in this cluster seems to be the number of comorbidities.

Figure 5 show ILD subtypes in the different clusters after classifying individuals based on their highest posterior probability. We see different ILD subtype patterns in each cluster. Cluster 1 is very heterogeneous, with a higher proportion of eosinophilic pneumonia and hypersensitivity pneumonitis but with less rheumatic ILD. Cluster 2 has two apparent spikes with high numbers of rheumatic ILD and drug-related ILD. Cluster 3 is overall very homogeneous with an almost equal representation of all subtypes. Cluster 4 has less individuals with sarcoidosis but more with drug-related ILD. In the high-cost cluster 5, sarcoidosis, pneumoconiosis, and hypersensitivity pneumonitis are the most prevalent subtypes.

6 | DISCUSSION

This paper uses a VB model for fitting an infinite mixture regression model for continuous cost data. The advantage of this nonparametric model and the efficient implementation we used is that it automatically performs a search for the most

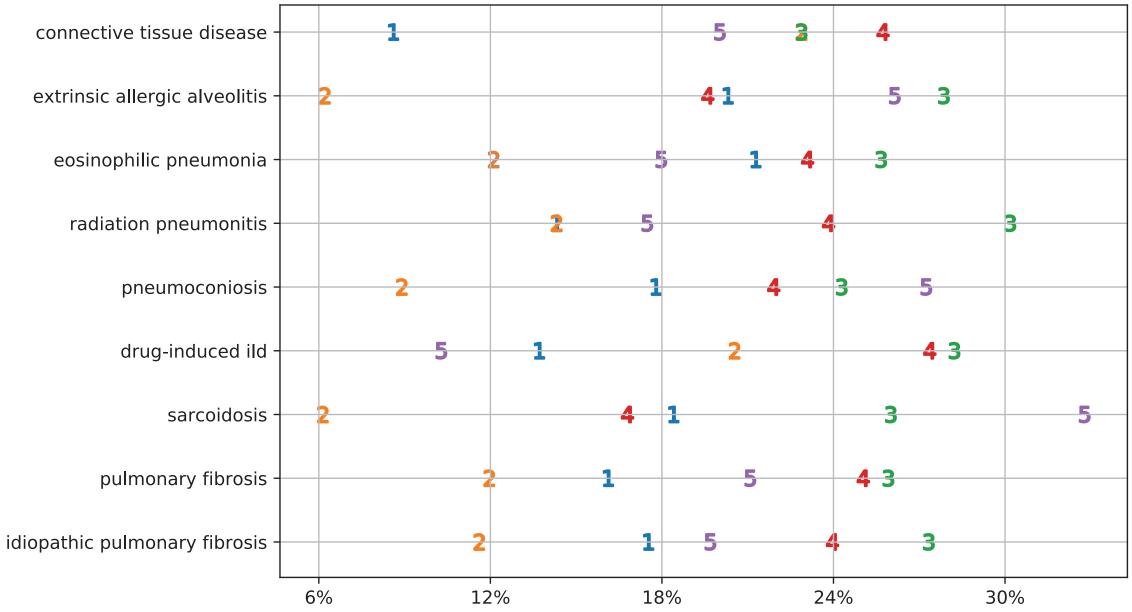

FIGURE 5 Plot for ILD subtypes and the percentage of patients in each cluster (using hard cluster assignments)

appropriate number of clusters from the data using a model selection criterion (ELBO), which is generally preferable to more familiar criteria like BIC (Beal & Ghahramani, 2003). Traditional approaches require manually fitting a range of models with different numbers of clusters and then deciding more subjectively when to stop searching for a better fit (Leroux, 1992; McLachlan & Rathnayake, 2014; Zivkovic & van der Heijden, 2004). Too many clusters may overfit the data and impair model interpretation, while too few clusters limit the flexibility of the mixture to approximate the true underlying data structure. This practice is vulnerable to model-selection bias and also computationally more expensive than the VB approach.

The VB approach also has several advantages over traditional MCMC estimation. Not only is it computationally faster (fitting this model to the AOK data set took only 1 min on mainstream hardware, a comparable MCMC model in the Stan language took over 6 h), but it also avoids some of the identifiability issues (e.g., the label-switching problem) and the convergence problems that MCMC is prone to (Celeux, Hurn, & Robert, 2000; McGrory & Titterington, 2007).

Still, Bayesian mixtures are a highly complex problem that requires estimating a lot of parameters, making these models difficult to fit. Some authors even argued that interpreting the parameters of the mixture coefficients should be avoided (Rodriguez & Dunson, 2011; Ferguson, 1983). VB algorithms are often sensitive to their initialization scheme, even leading to different number of clusters. This may sound as a limitation to applied researchers, but is part of the nature of mixture models, where different fitted models with different number of clusters can be good representatives of the same data. Our initialization of $K = 10$ is a compromise between model complexity and the goal of our analysis to get a reasonable and interpretable number of clusters. In other applications higher values of K may be useful.

There are other limitations to our study. Our simulation covered only a very small subset of possible scenarios, and the data-generating process closely resembled the model specification, making these results hardly conclusive. For example, both finite mixture and VB find the correct number of four clusters in the simulation study. These clusters are very different in both approaches, yet they do not appear wrong in either setting. In real data analysis, unlike in simulations, the concept of “true” clusters depends on the context and analytical goals (Hennig, 2015).

We restricted our consideration to mixtures of Gaussian regression models. Other research recommends more flexible kernels that can better account for the specific distributional properties of health care cost data (Fellingham, Kottas, and Hartman, 2015; Canale and Prünster, 2017). However, the large sample size ensures near-normality of sample means because of the Central Limit Theorem (Mihaylova et al., 2011). Still, distributions like the generalized gamma (Hill & Miller, 2010) or the tweedie (Kurz, 2017) could be examined, among others.

For the AOK data set, the VB model finds five clusters of individuals with strikingly different distributions of health care costs and ILD subtypes. All five clusters show an equal distribution of drug-related ILD and radiotherapy–induced ILD, that is, if one of these subtypes has a high proportion, the other has too. Both subtypes are indicators for the presence of lung cancer, causing high costs because of lung cancer treatment. However, the regression coefficients do not always point in the same direction (e.g., in clusters 2 and 4). An explanation could be that if standard chemotherapy causes ILD, a change to non-standard chemotherapy could be more expensive. Sarcoidosis can occur in two forms: the slow-progressing symptom-poor onset chronic course and the acute form (Löfgren syndrome). This could explain the high presence of sarcoidosis in both low-cost cluster 1 and high-cost cluster 5. Cluster 3 might capture all the "average" cases with relatively stable disease onset and progress. The same goes for cluster 4 but with higher average costs and less cases of sarcoidosis. Maybe focus in this cluster is more on palliative care. Cluster 5 includes the high-cost cases where there might by rapid disease progression and expensive treatment of comorbidities (e.g., exacerbations) or due to end of life care costs.

Identifying ILD subgroups with different spending and treatment patterns could help policymakers and clinicians seeking to improve care and reduce spending design interventions tailored to specific subpopulations.

7 | CONCLUSION

This work presents a VB approach for mixture regression models that can be used to find interpretable subgroups of patients. It is easy-to-use, flexible, and avoids bias introduced by model specifications. In contrast to MCMC methods, it is extremely fast. In our application to health care costs of ILD patients, we find subgroups with specific properties that correspond well to the different spending pattern in each component. Our findings indicate that our model is ideal to represent the underlying heterogeneity of ILD and their association with costs.

ACKNOWLEDGMENTS

The authors would like to thank Michael Hughes for valuable guidance on using the BNPY package, and Phillen Maqhuzu and Larissa Schwarzkopf for help with interpreting the clusters.

Open access funding enabled and organized by Projekt DEAL.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

FUNDING

No funding was received for this study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from AOK. Restrictions apply to the availability of these data, which were used under license for this study. The code is provided on Github at https://github.com/krz/ild_bnpy.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially for data confidentiality reasons.

ORCID

Christoph F. Kurz https://orcid.org/0000-0001-9498-8002

REFERENCES

Agostini, C., Albera, C., Bariffi, F., De Palma, M., Harari, S., Lusuardi, M., … Tinelli, C. (2001). First report of the Italian register for diffuse infiltrative lung disorders (ripid). *Monaldi Archives for Chest Disease*, *56*(4), 364–368.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267–281.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 21–30.

Bates, D. W., Kuperman, G. J., Wang, S., Gandhi, T., Kittler, A., Volk, L., … Middleton, B. (2003). Ten commandments for effective clinical decision support: Making the practice of evidence-based medicine a reality. *Journal of the American Medical Informatics Association*, *10*(6), 523–530.

Beal, M., & Ghahramani, Z. (2003). The variational Bayesian em algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian Statistics*, *7*, 453–464.

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, *1*(1), 121–143.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(1), 859–877.

Böhning, D., & Seidel, W. (2003). Editorial: Recent developments in mixture models. *Computational Statistics & Data Analysis*, *41*(3–4), 349–357.

Canale, A., & Prünster, I. (2017). Robustifying Bayesian nonparametric mixtures for count data. *Biometrics*, *73*(1), 174–184.

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*(451), 957–970.

Corduneanu, A., & Bishop, C. M. (2001). Variational Bayesian model selection for mixture distributions. *Artificial Intelligence and Statistics*, *2001*, 27–34.

Coultas, D. B., Zumwalt, R. E., Black, W. C., & Sobonya, R. E. (1994). The epidemiology of interstitial lung diseases. *American Journal of Respiratory and Critical Care Medicine*, *150*(4), 967–972.

Deb, P., & Trivedi, P. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, *12*(3), 313–336.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, *39*(1), 1–38.

Fellingham, G. W., Kottas, A., & Hartman, B. M. (2015). Bayesian nonparametric predictive modeling of group health claims. *Insurance: Mathematics and Economics*, *60*, 1–10.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In Rizvi, M. H., Rustagi, J. S., & Siegmund, D. (Eds.), *Recent advances in statistics: Papers in honor of Herman Chernoff on his sixtieth birthday* (pp. 287–302). New York: Academic Press.

Fraley, C., & Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, *24*(2), 155–181.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, *96*(453), 194–209.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62.

Hill, S. C., & Miller, G. E. (2010). Health expenditure estimation and functional form: Applications of the generalized gamma and extended estimating equations models. *Health Economics*, *19*(5), 608–627.

Hingorani, A. D., van der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G., Steyerberg, E. W., … Hemingway, H. (2013). Prognosis research strategy (progress) 4: Stratified medicine research. *British Medical Journal*, *346*, e5793.

Hughes, M. C. (2016). *Reliable and scalable variational inference for nonparametric mixtures, topics, and sequences*. PhD thesis, Brown University. Retrieved from https://cs.brown.edu/research/pubs/theses/phd/2016/hughes.michael.pdf.

Hughes, M. C. (2018). *bnpy documentation, release 0.1.6*. Retrieved from https://media.readthedocs.org/pdf/bnpy/latest/bnpy.pdf.

Hughes, M. C., & Sudderth, E. (2013). Memoized online variational inference for Dirichlet process mixture models. *Advances in Neural Information Processing Systems*, *26*, 1133–1141.

Kawalec, P. P., & Malinowski, K. P. (2015). The indirect costs of systemic autoimmune diseases, systemic lupus erythematosus, systemic sclerosis and sarcoidosis: A summary of 2012 real-life data from the social insurance institution in poland. *Expert Review of Pharmacoeconomics & Outcomes Research*, *15*(4), 667–673.

Kurz, C. F. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, *17*(1), 171.

Kurz, C. F. (2018). *Github repository*. Retrieved from https://github.com/krz/ild_bnpy.

Kurz, C. F., & Hatfield, L. A. (2019). Identifying and interpreting subgroups in health care utilization data with count mixture regression models. *Statistics in Medicine*, *38*(22), 4423–4435.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, *20*(3), 1350–1360.

López-Campos, J., & Rodríguez-Becerra, E. (2004). Incidence of interstitial lung diseases in the south of Spain 1998–2000: The Renia Study. *European Journal of Epidemiology*, *19*(2), 155–161.

McGrory, C. A., & Titterington, D. M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, *51*(11), 5352–5367.

McLachlan, G., & Peel, D. (2004). *Finite mixture models. Wiley series in probability and statistics*. New York: Wiley.

McLachlan, G., & Rathnayake, S. (2014). On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(5), 341–355.

Mihaylova, B., Briggs, A., O'hagan, A., & Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, *20*(8), 897–916.

Müller, P., Erkanli, A., & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, *83*(1), 67–79.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, *55*(2), 463–469.

Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, *11*, 125–139.

Oganisian, A., Mitra, N., & Roy, J. A. (2020). A Bayesian nonparametric model for zero-inflated outcomes: Prediction, clustering, and causal estimation. *Biometrics*, *1*(1), 1–11.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*(2), 140–153.

Raghu, G., & Brown, K. K. (2004). Interstitial lung disease: Clinical evaluation and keys to an accurate diagnosis. *Clinics in Chest Medicine*, *25*(3), 409–419.

Raimundo, K., Chang, E., Broder, M. S., Alexander, K., Zazzali, J., & Swigris, J. J. (2016). Clinical and economic burden of idiopathic pulmonary fibrosis: A retrospective cohort study. *BMC Pulmonary Medicine*, *16*(1), 2.

Rice, J. B., White, A., Lopez, A., Conway, A., Wagh, A., Nelson, W. W., Philbin, M., & Wan, G. J. (2017). Economic burden of sarcoidosis in a commercially-insured population in the United States. *Journal of Medical Economics*, *20*(10), 1048–1055.

Robert, C., & Casella, G. (2005). *Monte Carlo statistical methods. Springer texts in statistics*. New York: Springer.

Rodriguez, A., & Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, *6*(1), 145–177.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Schweisfurth, H. (1996). Mitteilung der wissenschaftlichen arbeitsgemeinschaft fur die therapie vonlungenkrankheiten (watl): Deutsches fibroseregister mit ersten ergebnissen. *Pneumologie*, *50*(12), 899–901.

Sundararajan, V., Henderson, T., Perry, C., Muggivan, A., Quan, H., & Ghali, W. A. (2004). New ICD-10 version of the Charlson Comorbidity Index predicted in-hospital mortality. *Journal of Clinical Epidemiology*, *57*(12), 1288–1294.

Tan, S. L., & Nott, D. J. (2014). Variational approximation for mixtures of linear mixed models. *Journal of Computational and Graphical Statistics*, *23*(2), 564–585.

Tennekes, M., de Jonge, E., & Daas, P. J. (2013). Visualizing and inspecting large datasets with tableplots. *Journal of Data Science*, *11*(1), 43–58.

Thomeer, M., Demedts, M., & Vandeurzen, K. (2001). Registration of interstitial lung diseases by 20 centres of respiratory medicine in flanders. *Acta Clinica Belgica*, *56*(3), 163–172.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, UK: Wiley.

Ueda, N., & Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, *15*(10), 1223–1241.

Wu, B., McGrory, C. A., & Pettitt, A. N. (2012). A new variational Bayesian algorithm with application to human mobility pattern modeling. *Statistics and Computing*, *22*(1), 185–203.

Zivkovic, Z., & van der Heijden, F. (2004). Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(5), 651–656.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Kurz CF, Stafford S. Isolating cost drivers in interstitial lung disease treatment using nonparametric Bayesian methods. *Biometrical Journal*. 2020;1–13. https://doi.org/10.1002/bimj.202000076