Bayesian Nonparametric Clustering and Inference for Health Care Utilization of Interstitial Lung Disease Patients

Christoph F. Kurz
Helmholtz Zentrum München
Institute of Health Economics and Health Care Management
Neuherberg, Germany
christoph.kurz@helmholtz-muenchen.de

Abstract

Interstitial lung disease is a very heterogeneous disease with various subtypes, making analysis of health care spending patterns for this disease very difficult. Common features of health care utilization measures include zero inflation, overdispersion, and skewness, all of which complicate statistical modeling. Mixture modeling is a popular approach that can accommodate these features of health care utilization data. In this work, we add a nonparametric clustering component to such models. Our fully variational Bayesian model framework allows for an unknown number of mixing components, so that the data determine the number of mixture components. When we apply the modeling framework 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.

1 Introduction

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. [11] That makes ILD a very heterogeneous group of disorders that includes at least two dozens of different types. Some forms of ILD are short-lived, while others are chronic and reversible. A US study reported incidences of 31.5 per 100,000 among men and 26.1 per 100,000 among women. [3] European studies have reported slightly lower ILD incidences between 4.6 and 7.6 per 100,000 inhabitants/year. [17, 1, 18, 16] The economic burden of ILD is very high. ILD not only causes indirect costs (e.g., loss of productivity), [6] but also high direct costs for hospitalization and exacerbation. [12, 14] 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. [5]

Models based on mixtures of parametric models are commonly used as a flexible way to accommodate excess zeros, overdispersion and heavy tails that often define health care utilization cost data. [10] These finite mixture models (FMMs), also known as latent class models, are motivated by the concern that different parts of the response distribution could be differently affected by covariates (i.e., low cost users, high cost users). FMMs identify groups of observations with similar outcomes using unsupervised clustering, and often perform better than standard generalized linear models and the hurdle model. [4]

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

A key question in mixture models is the optimal number of components. (Note that we use *component*, rather than *cluster*, to describe the subpopulations identified by FMMs.) Too many components may overfit the data and impair model interpretation, while too few components limit the flexibility of the mixture to approximate the true underlying data structure. 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. In practice, most analysts use information criteria such as AIC or BIC or bootstrap resampling of the likelihood ratio test statistic values. [8, 9] In the Bayesian literature, many authors use the maximum *a posteriori* estimator for model selection. [19]

Our model uses a Dirichlet process (DP) prior for the mixing component, where the optimal number of components is determined simultaneously with the model fit. This one-stage process yields the ideal number of components and allows interpretation of each component. This potentially unbounded infinite mixture model avoids both over- and under-fitting by allowing the data to determine the optimal number of components. [13] Because the model is fully Bayesian, prior information about the number of components can be incorporated. Previous work has shown that DP mixtures can more accurately find the true number of components than AIC or BIC model selection. [7]

In the last decades, the development of advanced Markov chain Monte Carlo (MCMC) methods were a key step in making it possible to compute large hierarchical models that require integrations over hundreds of unknown parameters. [15] Still, MCMC sampling can be prohibitively slow for large data sets or complicated likelihood functions. [2] Recently, variational Bayesian (VB) methods were developed for fast approximation of 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. In particular, whereas MCMC techniques provide a numerical approximation to the exact posterior using a set of samples, VB provides a locally-optimal, exact analytical solution to an approximation of the posterior. VB inference algorithms are usually faster than MCMC and suitable for large scale data sets, which are becoming more and more prevalent through the analysis of claims data and electronic health records.

We are motivated by the desire to understand variation in total health care expenditures among patients diagnosed with ILD. To understand patient subpopulations defined by health care costs and effects of covariates on costs, we develop and apply Gaussian regression models that use a DP prior for nonparametric mixing of the components using VB inference.

2 Model Definition

In our model, we assume a Mixture distributions with K components, each following a Gaussian linear regression model. The data set consists of pairs $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$ where x_n is a vector of length D and y_n is scalar. Therefore, for each pair of observations there exists a latent variable $\mathbf{z}_n$ indicating the cluster assignment. The latent regression coefficients are $\boldsymbol{\beta}$, and the precision latent variables are τ_k. The conditional distribution of the observed data vectors given the latent variables and the component parameters is defined as:

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\prod_{d=1}^{D} \mathcal{N}(y_{nd} \mid \mathbf{x}_{nd}\boldsymbol{\beta}_d, \tau_{kd}^{-1}) \right)^{z_{nk}} .$$

The conditional distribution of $\mathbf{z}$, given the mixing coefficients $\mathbf{c}$ can be written as a categorical distribution:

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

We define a Dirichlet prior over the mixing proportions c with fixed hyperparameters α_0:

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

where $\Gamma(\cdot)$ is the Gamma function, and introduce a Gaussian-Gamma prior over the mean and precision parameters:

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}) = p(\boldsymbol{\beta} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau}),$$

with

$$p(\boldsymbol{\beta} \mid \boldsymbol{\tau}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\beta}_k \mid \mathbf{b}_0, (\tau_k \boldsymbol{\Lambda}_0)^{-1}),$$

$$p(\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathrm{Gam}(\tau_k \mid a_0, b_0),$$

We use the following values for the hyperparameters:

$$b_0 = 0, \quad \mathrm{diag}(\Lambda_0) = 0.01, \quad a_0 = 0.01, \quad b_0 = 0.01.$$

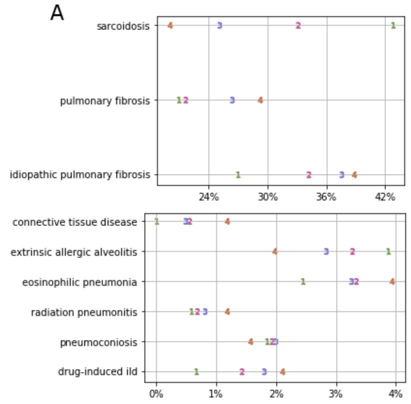
3 AOK Data Set

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 2010 to 2013, as well as service utilization. We used a study population previously derived from this data set consisting of patients with interstitial lung disease 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 $N = 9010$ individual observations.

We included the following covariates in the model: age, sex, ILD type (9 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.

4 Results



B

variables	Comp.1	Comp.2	Comp.3	Comp.4
Intercept	64.60	847.40	7465.50	50534.90
age	11.80	24.00	-1.60	-298.30
sex (female)	167.50	235.30	107.70	-2812.80
connective tissue disease	-25.00	234.70	1527.30	6364.20
extrinsic allergic alveolitis	58.40	323.30	1906.10	6693.60
eosinophilic pneumonia	109.80	-30.70	-515.80	245.10
radiation pneumonitis	-43.90	888.60	5708.60	8853.30
pneumoconiosis	-246.70	-255.40	-1184.10	-6307.50
drug-induced ild	157.30	716.80	488.20	7520.90
sarcoidosis	-23.30	-149.70	1263.90	6334.40
pulmonary fibrosis	78.00	-229.60	-994.50	12785.60
idiopathic pulmonary fibrosis	0.00	-650.60	-734.20	8045.20
Charlson index	178.40	507.70	1090.00	1740.40
lung cancer	-112.00	-803.10	1287.60	-3657.00
GERD	197.90	398.10	726.00	-1954.20
nursing home	44.30	1659.20	5184.80	15115.40

Figure 1: (A) Plots for ILD subtypes as percentages after using hard component assignments based on the VB model for the AOK data set. Numbers mark the four components. Common subtypes are shown in the upper plot, rare subtypes in the lower plot. (B) Parameter estimates for each component of the VB mixture model.

For the AOK data set, the VB model finds 4 components as having the highest expected posterior mixture weights. In the following, we only show results based on this final model with 4 components.

Component 1 is the largest components and contains 34.4% (3134/9110) of all observations and corresponds to individuals with the lowest health care costs, on average, but with high variance. Component 1 is the only component where connective tissue disease and radiation pneumonitis have a negative effect on the costs. Component 2 is slightly smaller, with 28.6% (2605/9110) of individuals; they have higher health care costs with less variance. Component 3 comprises 26.6% (2423/9110) of the population, and these patients have even higher health costs, on average. Component 4 has the individuals with the highest costs, containing 10.4% (948/9110) of individuals.

When we use hard assignments to classify individuals into components according to the highest posterior probability, we see that ILD subtype patterns are very different across components (see Figure 1). Sarcoidosis, pulmonary fibrosis, and idopathic pulmonary fibrosis are the most common subtypes. Sarcoidosis is by far more common in component 1, associated with lower costs, while pulmonary fibrosis and idopathic pulmonary fibrosis are more present with the high cost individuals in component 4.

For the more rare subtypes, we see only extrinsic allergic alveolitis as being more prevalent for low cost individuals in component 1. All other rare subtypes are more prevalent in component 4. Components 2 and 3 are always in between. The only exception is pneumoconiosis where the spending pattern is more heterogeneous across the components.

5 Discussion

This paper explores a VB regression model for that combines nonparametric clustering with the advantages of finite mixture models. It avoids model selection and decision bias because all parameters, including the ideal number of mixture components, are estimated from the data.

For the AOK data set, the VB models find four components of individuals with strikingly different distributions of costs and ILD subtypes. Sarcoidosis is a disease with often no, or only mild, symptoms. It is therefore not surprising to have the highest prevalence in component 1. On the other hand, (idiopathic) pulmonary fibrosis is an uncurable disease where patients suffer from perpetual shortness of breath. This leads to high health care costs, making it most common in component 4.

Pneumoconiosis is caused by the inhalation of dust and has very different form, leading to no consistent spending pattern. Extrinsic allergic alveolitis is caused by hypersensitivity to inhaled organic dusts. It is relatively easy to treat, which is represented by being most common among the low health care spenders in component 1.

The health care costs received by people in components 2 and 3 are quite similar, with only the proportion of having sarcoidosis being significantly higher in 2 than in 3. However, the coefficients governing the relationships between ILD subtype and costs in these components are quite different. While lung cancer is associated with significantly higher costs in component 2, it has the opposite association in component 3. The same applies to pneumoconiosis.

There are several limitations to this study. DP mixture models present computational challenges and interpretation is necessarily more difficult in complex models such as these. In the case of our presented model, with four mixture components, the number of regression coefficients is four 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.

Future work will perform simulation studies on the accuracy of the parameter estimates compared with standard finite mixture approaches.

References

[1] C Agostini, C Albera, F Bariffi, M De Palma, S Harari, M Lusuardi, A Pesci, V Poletti, L Richeldi, G Rizzato, et al. First report of the italian register for diffuse infiltrative lung disorders (ripid). *Monaldi Archives for Chest Disease*, 56(4):364–368, 2001.

[2] David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

[3] David B Coultas, Ross E Zumwalt, William C Black, and Richard E Sobonya. The epidemiology of interstitial lung diseases. *American journal of respiratory and critical care medicine*, 150(4):967–972, 1994.

[4] Partha Deb, Pravin K Trivedi, et al. Demand for medical care by the elderly: a finite mixture approach. *Journal of applied Econometrics*, 12(3):313–336, 1997.

[5] Aroon D Hingorani, Daniëlle 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.

[6] Paweł P Kawalec and Krzysztof P Malinowski. 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, 2015.

[7] Christoph F Kurz and Laura A Hatfield. Bayesian nonparametric clustering and inference for inpatient health care utilization. *under review*.

[8] Brian G Leroux et al. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3):1350–1360, 1992.

[9] 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.

[10] 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.

[11] Ganesh Raghu and Kevin K Brown. Interstitial lung disease: clinical evaluation and keys to an accurate diagnosis. *Clinics in chest medicine*, 25(3):409–419, 2004.

[12] Karina Raimundo, Eunice Chang, Michael S Broder, Kimberly Alexander, James Zazzali, and Jeffrey J Swigris. Clinical and economic burden of idiopathic pulmonary fibrosis: a retrospective cohort study. *BMC pulmonary medicine*, 16(1):2, 2016.

[13] Carl Edward Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[14] J Bradford Rice, Alan White, Andrea Lopez, Alexandra Conway, Aneesha Wagh, Winnie W Nelson, Michael Philbin, and George J Wan. Economic burden of sarcoidosis in a commercially-insured population in the united states. *Journal of Medical Economics*, (just-accepted):1–16, 2017.

[15] Christian P Robert. *Monte carlo methods*. Wiley Online Library.

[16] Eulogio Rodríguez-Becerra et al. Incidence of interstitial lung diseases in the south of spain 1998–2000: the renia study. *European journal of epidemiology*, 19(2):155–161, 2004.

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

[18] M Thomeer, M Demedts, K Vandeurzen, and VRGT Working Group on Interstitial Lung Diseases. Registration of interstitial lung diseases by 20 centres of respiratory medicine in flanders. *Acta Clinica Belgica*, 56(3):163–172, 2001.

[19] 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.