Demand for Medical Care by the Elderly: A Nonparametric Variational Bayesian Mixture Approach

Christoph F. Kurz[1] **and Rolf Holle**[1]

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

Abstract

Outpatient care is a large share of total health care spending, making analysis of data on outpatient utilization an important part of understanding patterns and drivers of health care spending growth. Common features of outpatient utilization measures include zero-inflation, overdispersion, and skewness, all of which complicate statistical modeling. Mixture modeling is a popular approach because it can accommodate these features of health care utilization data. In this work, we add a nonparametric clustering component to such models. Our fully Bayesian model framework allows for an unknown number of mixing components, so that the data, rather than the researcher, determine the number of mixture components. We apply the modeling framework to data on visits to physicians by elderly individuals and show that each subgroup has different characteristics that allow easy interpretation and new insights.

1998 ACM Subject Classification G.3 Probability and Statistics

Keywords and phrases machine learning, health care utilization, Bayesian statistics

Digital Object Identifier 10.4230/OASIcs.xxx.yyy.p

1 Introduction

Outpatient hospital services account for a large share of health care utilization and therefore of total health care spending. To understand the variation in this major component of health care expenditures, researchers have sought to identify patient subgroups with different utilization and spending patterns.

Health care resource use data are often non-negative, right-skewed, heavy-tailed, and multi-modal with a point mass at zero. Desirable analytical approaches for these data should be sufficiently powerful and flexible to accommodate all of these features. Several authors showed that finite mixture models (FMMs) provide better model fit than single distribution generalized linear models (GLMs) and the hurdle model. [1, 2] In addition, FMMs have two advantages: first, they can easily handle multimodality. This may be important when the outcome distribution suggests decomposing resource use into different components. For example, it may be necessary to fit the tail distribution separately. Second, mixture models allow us to link the prevalence of different mixture components to different covariates. [2] Generally, mixture models distinguish between different groups of users (e.g. low- and high users) and avoid the sharp dichotomy between users and non-users.

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.)

Conference/workshop/symposium title on which this volume is based on.
Editors: Billy Editor and Bill Editors; pp. 1–7

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

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 different user groups can be decided either "ex-ante" by a defined value (two or three groups are common), or "ex-post", i.e. chosen by model fit after calculating different models. While the ex-ante approach is focused on feasibility and is a one-stage decision process, ex-post approaches use information which extends beyond the time at which the actual model is prepared and involves a second decision process. Both approaches introduce a decision and model selection bias.

In this paper, we present a fully variational Bayesian (VB) hierarchical mixture model, where the optimal number of components is evaluated during model fit. This one-stage process yields both the ideal number of components and allows interpretation of each component. In this Bayesian nonparametric mixture model, we let the data determine both the number and the form of the local mean functions. In contrast to frequentist nonparametric regression methods, this Bayesian approach creates a model that is only as complex as the data require. [3] In models with a fixed, finite number of parameters, there may be misfit between the complexity of the model and the amount of data available. By contrast, Bayesian nonparametric models are less subject to over- or under-fitting: the unbounded complexity of the infinite mixture mitigates under-fitting, while the Bayesian approach of computing the full posterior over parameters mitigates over-fitting. [4]

Our model uses a Dirichlet process (DP) prior for the mixing component and comprises a fully VB regression scheme. VB is an alternative to Markov chain Monte Carlo (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.

In the following, we define a VB regression mixture model for counts and apply it on a data set to analyze outpatient health care utilization. The data set has already been used in [1] where Deb and Trivedi showed that a FMM with two components provides better model fit than a simple GLM. In this paper, we apply our proposed VB mixture model on the DebTrivedi data set and demonstrate that this model has good clustering and inference properties that allow new insights.

2 Model Definition

2.1 Dirichlet Process Mixtures for Generalized Linear Models

The DP is a *measure on measures* [5] or a *distribution over distributions* [6] parameterized by a base distribution G_0 and a concentration parameter α. Each draw from a DP is a distribution that is discrete with countably infinite parameters, making this a nonparametric model. Using a DP prior for the distribution of component means in mixture models does not require one to specify the number of components. Instead, a concentration parameter controls it implicitly. Suppose the sample space Ω is partitioned into measurable subsets $U_1, ..., U_k$. Let $\mathfrak{U}$ be the collection of all possible subsets of Ω. If G is a random probability measure over $(\Omega, \mathfrak{U})$ that assigns probabilities to all subsets, then $G \sim \mathrm{DP}(\alpha, G_0)$ is a measure with property

$$G(U_1), ..., G(U_k) \sim \mathrm{Dir}(\alpha G_0(U_1), ..., \alpha G_0(U_k)).$$

More precisely, if $\alpha > 0$ and G is an instantiation of a DP with base measure G_0, then each component k has mixture weight c_k sampled as follows:

$$G = \sum_{k=1}^{\infty} c_k \delta(\theta = \zeta_k), \quad \text{where } \zeta_k \overset{\text{iid}}{\sim} G_0, \ k = 1, ..., \infty,$$
$$v_k = \Phi(\alpha), \quad c_k = v_k \prod_{j=1}^{k-1}(1 - v_j), \quad \sum_{k=1}^{\infty} c_k = 1, \tag{1}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution and δ is the delta function. The base measure G_0 provides an initial guess at G, and α controls how close samples from the Dirichlet process are to G_0. The DP serves as a nonparametric prior on the mixture components. As the ζ's are drawn from a (discrete) DP-distributed distribution, it is very likely that they will be the same in each draw. The distinct number of ζ's defines the number of components. The representation in Equation 2.1 is called a *stick-breaking process* and yields an infinite mixture model representation:

$$f_{\text{mix}}(x|\alpha, G_0) = \sum_{k=1}^{\infty} c_k f(x|\phi_k),$$

where f is the density function with parameters ϕ_k. Note that we define the stick-breaking process according to the probit representation [7] instead of using Beta random variables.

In addition to the usual regression parameters, these nonparametric mixture models produce several additional parameters of interest. For each mixture component k, we want to estimate the relative prevalence of the mixture component in the data and parameters of the mixture component's distribution, such as the mean, variance, and regression coefficients. The mixture weights c_k are the probabilities associated with each component and come directly from the stick-breaking proportions v_k. The features of the mixture component are in ϕ_k. In addition, for each observation, we want to estimate the mixture component from which it was most likely drawn, also called the component assignment.

2.2 The Negative Binomial Regression Model

The Negative Binomial distribution is a flexible alternative to the Poisson model for counts that accommodates over-dispersion with a longer, fatter tail. [8] Hilbe identified more than 12 different parameterizations of the Negative Binomial in the literature; [9] here, we use the definition in [1]. For $i = 1, \ldots, N$ observations and $d = 1, \ldots, D$ covariates, the data comprise an (N, D)-dimensional covariate matrix $\mathbf{X}$ with rows $\mathbf{x}_i$ and an N-vector of outcomes $\mathbf{y} = (y_1, \ldots, y_N)'$. For simplicity, we omit the subscript i in what follows. The density function for the $y \sim \text{NegBin}(\mu, \psi)$ distribution is

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

where we specify a regression model (with regression coefficients β) for the mean parameter

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

and ψ is a precision parameter. 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. [10]

2.3 Variational Inference Scheme

We assume a mixture distributions with K components, each following a negative binomial 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 component assignment. The conditional distribution of the observed data vectors given the latent variables and the component parameters can be defined as:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\psi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \text{NegBin}(y_n|\boldsymbol{x}\boldsymbol{\beta}, \psi)^{z_{nk}}.$$

We define a Dirichlet prior over the mixing proportions c:

$$p(\mathbf{c}) = \text{Dir}(\mathbf{c}|\boldsymbol{\alpha_0})$$

and introduce a Gaussian-Wishart prior over the mean and dispersion component:

$$p(\boldsymbol{\beta}, \boldsymbol{\psi}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\beta}_k|\hat{\boldsymbol{\beta}}_k, \psi_k^{-1}\hat{P}_k^{-1})\mathcal{W}(\psi_k|\hat{\nu}_k, \hat{\tau}_k).$$

The joint distribution over all random variables is:

$$p(\mathbf{y}, \mathbf{x}, \mathbf{z}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\psi}) = p(\mathbf{y}|\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\psi})p(\boldsymbol{z}|\boldsymbol{\beta})p(\boldsymbol{c})p(\boldsymbol{\beta}|\boldsymbol{\psi})p(\boldsymbol{\psi}).$$

The goal of variational inference is to optimize the parameters of a fully factorized variational distribution q that minimizes the Kullback-Leibler divergence from the true intractable posterior. The optimal q maximizes the *evidence lower bound* objective. Because of intractable integrals in the variational distribution

$$q(\mathbf{z}, \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\psi}) = q(\mathbf{z})q(\mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\psi}),$$

We define

$$q(\mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\psi}) = q(\mathbf{c}) \prod_{k=1}^{K} q(\boldsymbol{\beta}_k, \boldsymbol{\psi}_k) = q(\mathbf{c}) \prod_{k=1}^{K} q(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k^{-1}),$$

where $q(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k^{-1}) = \mathcal{N}(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k^{-1})$.

The optimization problem is therefore

$$q^*(z) = \arg\min_{q(z)\in\mathcal{D}} \text{KL}(q(\mathbf{z}, \mathbf{c}, \boldsymbol{\beta}, \boldsymbol{\Sigma})\|p(\mathbf{y}, \mathbf{x}, \mathbf{z}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\psi})).$$

and we solve this by memoized online variational inference as in [11].

3 Data

We explore the model on the data set from Deb and Trivedi. [1] It contains 4406 individuals, aged 66 and over, who are covered by Medicare, a public insurance program. Originally obtained from the US National Medical Expenditure Survey (NMES) for 1987/88, the data are available in the `R` package `MixAll`. The objective is to model the demand for medical care–as captured by the number of physician/non-physician office and hospital outpatient visits–by the covariates available for the patients. Here, we adopt the number of physician office visits `ofp` as the dependent variable and use the health status variables `hosp` (number of hospital stays), `health` (self-perceived health status), `numchron` (number of chronic conditions: cancer, heart attack, gall bladder problems, emphysema, arthritis, diabetes, other heart disease), as well as the socioeconomic variables `gender`, `age`, `black` (race), `faminc` (family income), `school` (number of years of education), and `privins` (private insurance indicator) as regressors.

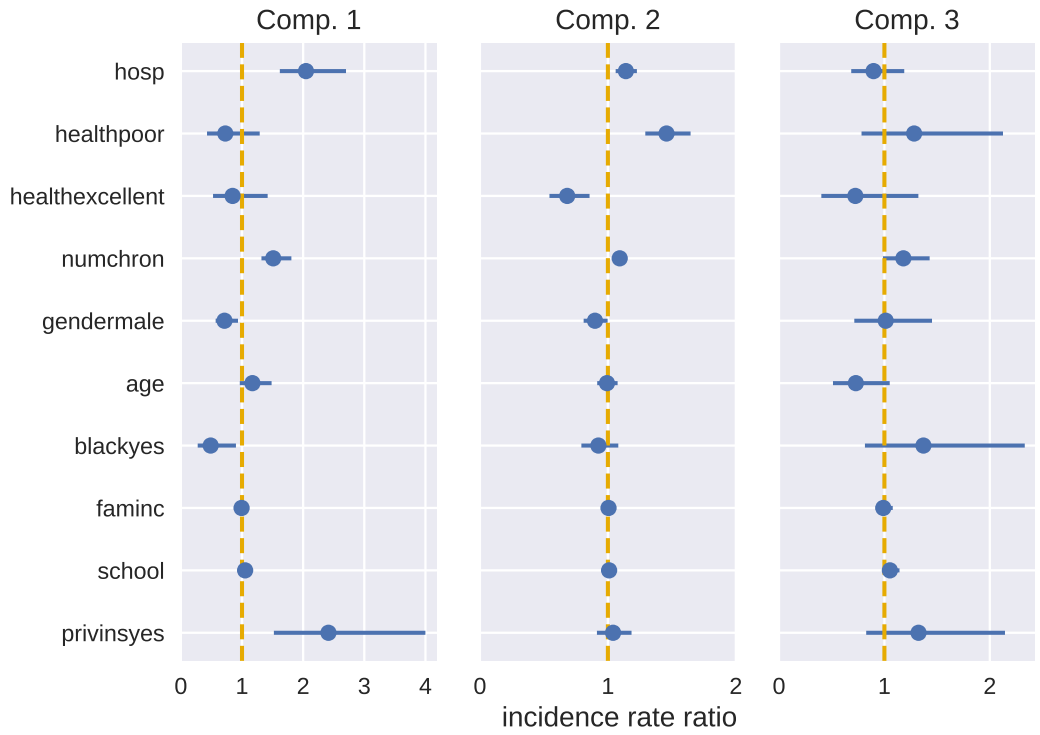

Figure 1 Parameter estimates for all three components on the DebTrivedi data set based on the negative binomial VB regression mixture model. Parameter estimates are presented as incidence rate ratios and 95% high probability density intervals. Intercept is not shown. The yellow dashed line at one marks no effect.

4 Results and Conclusion

For the DebTrivedi data set, the VB regression mixture model finds three components. The first component contains only 4.2% (186/4406) of all observations and corresponds to individuals who, on average, utilize less health care, but with higher variance. The second component corresponds to the largest proportion of individuals, 62.3%, (2745/4406) with medium health care utilization. The third component captures 33.5% (1475/4406) of individuals, with high utilization counts and again high variance. Figure 1 presents the parameter estimates from the regression model as incidence rate ratios (IRRs). In the first component, insurance status has the largest influence with an IRR of 2.41. This means that individuals with private insurance in the first component visit the doctor more than twice as often as those without private insurance. A similar explanation can be made for the number of hospital stays in the first component: one hospital stay accounts for 2.05 times more doctor visits, on average. This effect diminishes in the other components who contain individuals with higher utilization.

In the second component, a self perceived excellent health condition reduces the doctor visits by a factor or 0.68, while a poor health condition increases them by 1.46. This trend is slightly reduced in the third component. Most other variables show only slight effects on the number of doctor visits in the second component. In the third component, age has a protective influence on utilization, one additional year of age represents 0.73 times the utilization, on average. This seems counter-intuitive, but may be explained when comparing

the age of the individuals in each cluster: Figure 2 shows that age is increasing over the components. In addition, the number of chronic diseases is also increasing in each component. That explains the highest number of doctor visits in component 3. Interestingly, the years of education also increase slightly in each component. This should be subject of further investigation as, for example, Fiscella et al. [12] found that the time spent for physical examination is lower for more educated individuals.

Regarding computational speed, the VB inference method only takes 2 seconds to analyze the data set. A comparable MCMC approach took about 45 minutes on a 2016 Core i7 CPU with 32 GB RAM. This difference is mainly due to VB providing only a solution to an approximation of the posterior, while MCMC estimates the exact posterior. While we did not find great differences in the estimates in the present case, future research should investigate this difference, for example, in simulation studies.

In conclusion, the defined VB regression mixture model provides an interesting alternative with good accuracy and speed, especially suited for large data sets.

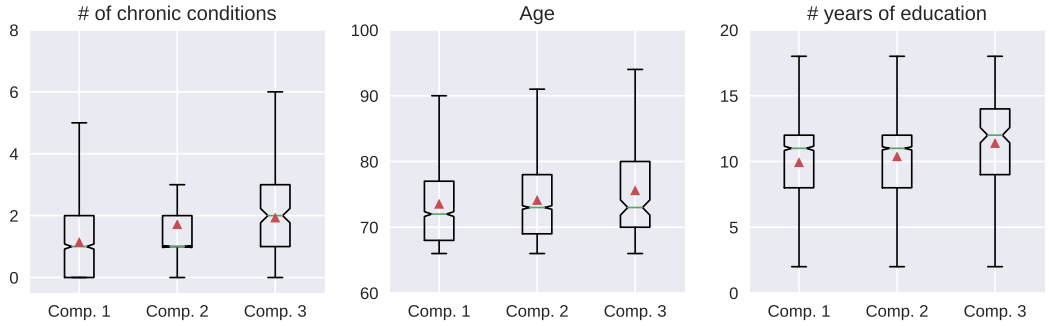

Figure 2 Boxplots for number of chronic conditions, age, and years of education for each component as modeled by the VB mixture model for the DebTrivedi data set. The red triangle marks the mean.

Acknowledgements We thank Laura Hatfield for improving the manuscript.

References

1 Partha Deb, Pravin K. Trivedi. *Demand for Medical Care by the Elderly: A Finite Mixture Approach*, Journal of Applied Econometrics, 12, 313–336, 1997

2 Borislava Mihaylova, et al. *Review of statistical methods for analysing healthcare resources and costs*, Health economics 20.8, 897-916 2011.

3 Lauren A. Hannah, David M. Blei, and Warren B. Powell, *Dirichlet process mixtures of generalized linear models*, Journal of Machine Learning Research, 1923-1953, 2011.

4 Carl Edward Rasmussen, *The infinite Gaussian mixture model*, NIPS. Vol. 12, 1999.

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

6 Claude Sammut and Geoffrey I. Webb, *Encyclopedia of Machine Learning*, Springer Publishing Company, Incorporated, 1st edition, 2011.

7 Abel Rodriguez and David B. Dunson, *Nonparametric bayesian models through probit stick- breaking processes*, Bayesian Analysis, 6(1), 2011.

8 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.

9 Joseph Hilbe, *Negative Binomial Regression*, Cambridge University Press, 2011.

10 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.

11 Michael C. Hughes, and Erik Sudderth, *Memoized online variational inference for Dirichlet process mixture models*, Advances in Neural Information Processing Systems, 2013.

12 Kevin Fiscella, Meredith A. Goodwin, and Kurt C. Stange, *Does patient educational level affect office visits to family physicians?*, Journal of the National Medical Association 94.3, 157, 2002.