Bayesian Analysis (2017) **12**, Number 3, pp. 831–854

Latent Class Mixture Models of Treatment Effect Heterogeneity

Zach Shahn* and David Madigan†

Abstract. We provide a general Bayesian framework for modeling treatment effect heterogeneity in experiments with non-categorical outcomes. Our modeling approach incorporates latent class mixture components to capture discrete heterogeneity and regression interaction terms to capture continuous heterogeneity. Flexible error distributions allow robust posterior inference on parameters of interest. Hierarchical shrinkage priors on relevant parameters address multiple comparisons concerns. Leave-one-out cross validation estimates of expected posterior predictive density obtained through importance sampling, together with posterior predictive checks, provide a convenient method for model selection and evaluation. We apply our approach to a clinical trial comparing two HIV treatments and to an instrumental variable analysis of a natural experiment on the effect of Medicaid enrollment on emergency department utilization.

Keywords: treatment effect heterogeneity, subgroup analysis, causal inference, latent class mixture model.

1 Introduction

In randomized experiments, it is often of interest to characterize treatment effect heterogeneity in terms of baseline covariates. Usually, the aim is to identify subpopulations likely to have particularly positive or negative (or neutral) responses to treatment. The process of searching for such subpopulations after the completion of an experiment (without pre-specifying which subpopulations will be considered as candidates) is called 'post hoc subgroup analysis'. It is a controversial practice. Concerns about data dredging and multiple comparisons (Rothwell, 2005) have led many authors to advise against reporting results from post hoc subgroup analyses at all. However, it is our view that post hoc analyses can produce informative insights that would be unlikely to arise from limited pre-registered comparisons. Here, we illustrate an approach in which identification of special subgroups is one byproduct of fully modeling treatment effect heterogeneity more generally.

Specifically, we propose to model treatment effect heterogeneity using regularized Bayesian latent class mixture models with treatment interaction terms and flexible error distributions. By placing hierarchical shrinkage priors on relevant parameters, we minimize data dredging concerns (Gelman et al., 2012). Flexible error distributions allow robust posterior inference on parameters of interest. Cross validation based tools for model evaluation and comparison (Gelfand, 1996; Vehtari and Lampinen, 2002),

*Harvard School of Public Health, zshahn@hsph.harvard.edu
†Columbia University, david.madigan@columbia.edu

 DOI: 10.1214/16-BA1022

along with posterior predictive checks (Gelman et al., 1996), provide a mechanism for gauging confidence in the substantive implications of model results.

These models are well suited to illuminate the shape of heterogeneity. The latent class components capture 'discrete heterogeneity' while the treatment interaction terms capture 'continuous heterogeneity'. By continuous heterogeneity we mean variation in subjects' individual treatment effects that is well approximated by a smooth function of underlying covariates. Discrete heterogeneity refers to variation in subjects' individual treatment effects that is associated with latent class membership, where latent class membership may in turn be associated with baseline covariates. Discrete heterogeneity is likely to be present if a treatment works through unobserved causal pathways that may be discretely open or closed. For example, suppose a drug works better in people with a specific phenotype for some protein receptor, but the presence of that phenotype is not recorded as a baseline covariate in a clinical trial evaluating the drug. But suppose that a recorded baseline covariate (say, weight) is associated with the presence of the beneficial phenotype. Then treatment effect variation as a function of weight will be better approximated by a latent class model with weight as a predictor of latent class membership than by any smooth function of weight alone. It can sometimes be useful to understand which type(s) of heterogeneity are present.

Despite our approach being a fairly straightforward application of Bayesian latent class mixture models and existing model comparison and evaluation techniques, we have not seen it in the subgroup analysis literature. Further, our approach offers a different combination of strengths (and weaknesses) from those methods we have seen.

Employing parametric probability models of heterogeneity brings certain automatic advantages. The parameter estimates have interpretable implications about the shape of heterogeneity, and models provide estimates of uncertainty about those parameters. Models also provide estimates of treatment effects for subpopulations and corresponding uncertainty estimates. These are obvious features of parametric probability models and are only worth mentioning because many methods for subgroup analysis are nonparametric or not model based and do not share these features.

One general tactic in the literature is to use nonparametric machine learning algorithms (often based on trees) to predict counterfactual outcomes of future subjects under treatment and control. Examples of works in this vein include Kang et al. (2012); Foster et al. (2011); Su et al. (2009), and others. These methods will frequently have superior predictive accuracy and flexibility to ours. Some of them also use cross validation to mitigate multiple comparisons concerns. However, most do not provide estimates of uncertainty about their predictions and do not characterize the shape of heterogeneity interpretably. Athey and Imbens (2015) recently proposed a machine learning approach that produces valid standard errors for causal effect estimates within nodes of a tree fit to a holdout validation set. Of course, holdout validation sets may not be practical for smaller experiments.

A closely related line of work directly learns optimal treatment assignment rules without first estimating counterfactual outcome response surfaces (Qian and Murphy, 2011; Zhang et al., 2012; Zhao et al., 2012). These methods are not interested in learning about heterogeneity, just assigning the best treatment to each subject. They have similar

strengths and weaknesses relative to our method as the machine learning approaches described above.

Imai and Ratkovic (2013) employ a linear Support Vector Machine (SVM) with interaction terms to model heterogeneity. The output of their model is interpretable, and they place shrinkage penalties on the parameters to discourage overfitting. They use a cross validation measure for model selection. One could replace their SVM with a regression probability model and obtain uncertainty estimates as well. However, they do not directly model discrete heterogeneity and do not consider uncertainty in their model selection criterion.

There have been other examples of latent class mixture models in the literature. In another context, Sobel and Muthen (2012) used a logistic-normal latent class mixture model to reflect the assumption that there exists a subpopulation in which the treatment has zero effect. Shen and He (2015) recently applied a similar model to identify subgroups and developed a corresponding likelihood ratio test for the existence of latent treatment effect classes. Neither of these approaches allows for continuous effect modification, however, and both are very sensitive to the assumption of a normal error distribution. We use very flexible error distributions so that our estimates are robust to departures from normality. Working in a model evaluation and comparison framework as opposed to Shen's and He's hypothesis testing framework allows us to consider more complex models leading to better fits and more reliable results at the expense of theoretical asymptotic guarantees.

The structure of this paper is as follows. In Section 2, we describe our approach in detail, providing specifications of various models that we consider and explaining the cross validation approach to model selection. In Section 3, we provide simulations illustrating the utility of our approach and the importance of some of its features. In Section 4, we reanalyze a clinical trial for an HIV treatment that was used as an example in Shen and He (2015) and Zhang et al. (2012). We conclude that this trial exhibits strong discrete heterogeneity, but not as strong as estimated by Shen and He (2015). In Section 5, we apply our approach to data from the Oregon Health Insurance Experiment (OHIE). The OHIE was a natural experiment that arose when Oregon instituted a lottery to determine who could enroll in a new Medicaid program with limited openings. This experiment allowed researchers to explore various public health and economic effects of Medicaid. One prominent finding was that Medicaid increased emergency department (ED) utilization contrary to many experts' predictions. The researchers performed multiple pre-registered and post hoc comparisons between subgroups and discovered several possible heterogeneities. Because the OHIE study contains lots of noncompliers (i.e. lottery winners who did not enroll in Medicaid and lottery losers who managed to enroll through other channels), we follow the researchers in performing an instrumental variable analysis using the principal stratification framework of Angrist, Imbens and Rubin (1996). Our method extends naturally to this framework because principal strata are themselves latent classes. When we include all covariates and employ hierarchical shrinkage priors to deal with multiple comparisons, we do not see strong evidence of heterogeneity associated with any of the observed covariates. In Section 6, we conclude.

Before proceeding, we prominently note a major limitation. Latent class mixture models are not identifiable for categorical outcome distributions such as the Bernoulli (Titterington, 1985). They are identifiable for the Poisson, negative binomial, or almost any continuous outcome distribution, however (Titterington, 1985).

2 Methods

2.1 Potential Outcomes

We use the potential outcomes framework (Rubin, 1974) which formalizes the notion that each experimental unit (e.g. patient) has a potential outcome for each possible treatment assignment that unit might have received. We get to observe only the potential outcome corresponding to the treatment actually received. We consider the counterfactual potential outcome that would have been observed had the treatment assignment been different to be an unobserved random variable. For the i^{th} unit, let $Z_i \in \{0, 1\}$ denote the treatment assignment and $Y_{z,i}$ the potential outcome corresponding to the possibly counterfactual treatment assignment $Z_i = z$. Then $Y_{z=1,i} - Y_{z=0,i}$ is the effect of treatment on unit i. Note that this individual level causal effect can never be observed because we never observe both potential outcomes. Still, if we specify a model for the observed data it is possible to estimate $E[Y_{z=1,i} - Y_{z=0,i}|X_i]$ – the conditional average treatment effect (or CATE) – and the parameters that govern its value as a function of covariate values X_i.

2.2 Model Specification

In Figure 1, we provide a graphical specification that describes all the models we consider in this paper. We elaborate on the interpretation of the graph in the following subsections. Note that this is not a *causal graph* in the sense of Pearl (1995), but rather a distributional graph that contains counterfactual variables at some nodes.

Intent to Treat Analyses

An Intent to Treat (ITT) analysis is concerned with the effect of the random treatment assignment Z and ignores whether subjects actually complied with their treatment assignments. An interpretation of the nodes of the graph in Figure 1 in the context of a standard ITT analysis is as follows. The potential outcomes for the i^{th} unit under treatment assignment $Z = z$ for z in $\{0, 1\}$ are denoted $Y_{z,i}$. Again, for each patient we only observe one of these and the other is treated as missing. The expected potential outcome for a unit with covariates X_i and latent class G_i under treatment assignment $Z = z$ is denoted by $\mu_{z,i}$. That is, the $\mu_{z,i}$ are the marginal expectations of the potential outcomes conditional on covariates and latent class. ϕ_z are parameters governing the marginal distributions of the potential outcomes $Y_{z,i}$ apart from their means $\mu_{z,i}$. For example, in a normal model, the ϕ_z would be standard deviations. ρ governs the dependence between the two potential outcomes but is completely unidentified because we never observe both potential outcomes for any one unit. $\mu_{z=0}$ is a function of covariates X_i, latent class G_i, and parameters β_C. $\mu_{z=1,i} = \mu_{z=0,i} + \Delta_i$, where Δ_i denotes the

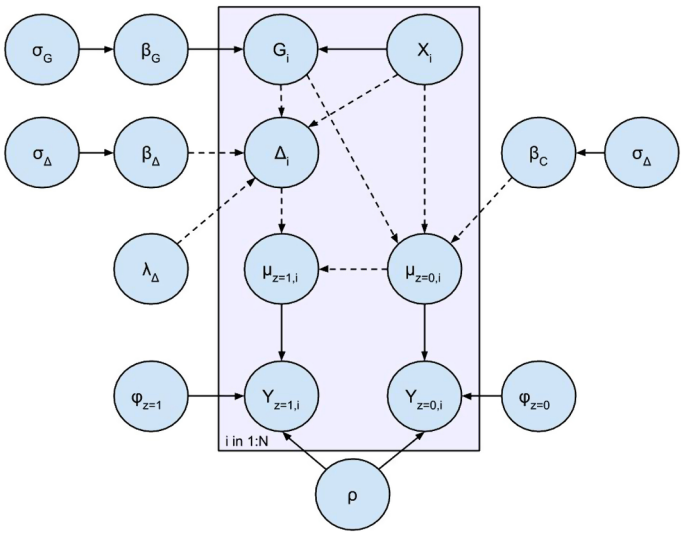

Figure 1: Graphical model specification. The dashed lines indicate deterministic relationships, and the solid lines indicate stochastic relationships.

average treatment effect for units with covariates X_i and latent class G_i. Δ_i is a function of X_i, G_i, and parameters β_Δ and λ_Δ. β_Δ determines how treatment effect varies continuously as a function of covariates, and λ_Δ determines the magnitudes of the discrete differences in treatment effect between latent classes. The probability distribution that generates G_i is a function of X_i and parameters β_G. β_Δ, λ_Δ, and β_G are the parameters of interest as together they describe how treatment effect heterogeneity is related to covariates. The parameters σ_G, σ_Δ, and σ_C are variances for shrinkage priors that we place on relevant parameters to avoid overfitting. We put weakly informative priors on the shrinkage variances themselves so that the appropriate level of shrinkage is learned from the data.

We make a few remarks on identifiability. We have included the dependence parameter ρ in the graphical model even though it is completely unidentified by data. This is because we do not wish to assert that the potential outcomes are conditionally independent given X and G, as would be implied if ρ were not in the graph. An important consequence of the unidentifiability of ρ is that it is impossible to obtain an informative posterior or posterior predictive distribution for an individual level causal effect without making unverifiable assumptions about the dependence between potential outcomes. We are limited to inferences and predictions involving only parameters governing the marginal distributions of the potential outcomes. These parameters are identified by the data, and their posterior distributions are not impacted by ρ (Chib, 2007), which we therefore do not include in the fit of our model. Suppose, for example, we want to predict the causal effect of a drug on a new patient using a heterogeneity model of the sort sketched above fit to data from the clinical trial for the drug. The most we can extract from this (or any) model without making assumptions about ρ is

a posterior predictive distribution of the average treatment effect for patients with the same covariates and (redundantly, unobserved) latent class as our new patient (i.e. the posterior predictive distribution of that new patient's Δ parameter). We can also obtain marginal posterior predictive distributions for each of that patient's potential outcomes but not their difference (i.e. the patient's treatment effect).

Also, to avoid aliasing issues in parameters of interest, in all models of this form we require that $\lambda_\Delta^{G_i}$ increases with the latent class label G_i.

The framework of the graphical model in Figure 1 allows for flexibility in the selection of functional forms, distributions, and number of latent classes. In this paper, we only consider linear models for the potential outcomes and logistic regression models for latent class membership. For potential outcomes, we consider models with two different error distributions – M_{Norm} with a normal error distribution and M_{Flex} with a more flexible three component Gaussian mixture error distribution.

$\mathbf{M}_{Norm}$ specification:

$$
\begin{aligned}
&Y_{z=0,i} \sim N(\mu_{z=0,i}, \sigma_{z=0}) \qquad Y_{z=1,i} \sim N(\mu_{z=1,i}, \sigma_{z=1}),\\
&G_i \sim Bernoulli(p_i),\\
&logit(p_i) = \alpha_G + X_i\beta_G,\\
&\mu_{z=0,i} = \alpha_C^{G_i} + X_i\beta_C^{G_i}, \qquad \Delta_i = \lambda_\Delta^{G_i} + X_i\beta_\Delta, \qquad \mu_{z=1,i} = \mu_{z=0,i} + \Delta_i,\\
&\alpha_C^G \sim N(0, 1000), \qquad \beta_C \sim N(0, \sigma_C),\\
&\alpha_G \sim N(0, 1000), \qquad \beta_G \sim N(0, \sigma_{\beta_G}), \qquad \beta_\Delta \sim N(0, \sigma_\Delta),\\
&\sigma_C \sim Uniform(0, 10), \qquad \sigma_G \sim Uniform(0, 5),\\
&\sigma_\Delta \sim U(0, 10), \qquad \sigma_1 \sim U(0, 10), \qquad \sigma_0 \sim U(0, 10),\\
&\lambda_\Delta^0 \sim N(0, 1000), \qquad \lambda_\Delta^1 - \lambda_\Delta^0 \sim Truncated\_Normal(0, 1000; 0+),\\
&\text{(where } \lambda_\Delta^1 - \lambda_\Delta^0 \text{ is restricted to be positive to prevent aliasing).}
\end{aligned}
\tag{1}
$$

$\mathbf{M}_{Flex}$ is the same as M_{Norm} except the M_{Flex} error distributions are each a mean 0 mixture of three Gaussian components. In $\mathbf{M}_{Flex}$,

$$
\begin{aligned}
&Y_{z,i} \sim q_1^z N(\mu_{z,i} + d_1^z, \sigma_1^z) + q_2^z N(\mu_{z,i} + d_2^z, \sigma_2^z) + q_3^z N(\mu_{z,i} + d_3^z, \sigma_3^z),\\
&\sum q_i^z = 1, \qquad \sum q_i^z d_i^z = 0,\\
&d_1 \sim N(0, 1000), \qquad d_2 \sim N(0, 1000),\\
&d_3 \text{ is determined by the constraint that the error distribution has mean 0,}\\
&\mathbf{q} \sim Dirichlet(1).
\end{aligned}
\tag{2}
$$

We demonstrate through simulations in Section 3 that violations of normality in M_{Norm} can lead to biased estimation of parameters of interest, but using flexible error distributions as in M_{flex} solves this problem. M_{Norm} and M_{Flex} are specified above for the case of two latent classes and include application specific weakly informative priors. Extension to multinomial logistic regression latent class models is straightforward.

Aliasing can arise in the estimation of the parameters governing the flexible error distribution, but that is not a problem because we are not interested in interpreting those parameters.

Instrumental Variable Analyses

Sometimes a situation arises in which treatment is not randomly assigned, but an encouragement to take treatment is randomly assigned. If the random encouragement only affects the outcome through the treatment and is indeed effective at inducing some people to take the treatment, then the encouragement is referred to as an 'instrument' and an Instrumental Variable (IV) analysis may be performed. IV analyses estimate the treatment effect in the subpopulation of units that would take the treatment if and only if encouraged by their value of the instrument. Such units are referred to as 'compliers' and the causal estimand in an IV analysis is referred to as the Complier Average Causal Effect (CACE). The canonical example of an IV setting is a randomized clinical trial with noncompliance. It is frequently the case in clinical trials that participants do not comply with their treatment assignments. Patients in the treatment arm may fail to take the treatment, and those in the control arm may find a way to take the treatment anyway. Thus, a simple ITT analysis comparing the two arms of the trial estimates the effect of treatment assignment rather than the effect of the treatment itself. However, random assignment to the treatment arm can be viewed as an instrument that encourages patients to take the treatment. An IV analysis with assignment as the instrument then estimates the effect of treatment on those patients who would comply with whatever random treatment assignment they happened to receive.

We follow Sobel and Muthen (2012) in extending latent class heterogeneity to an instrumental variable (IV) setting. We consider the case of a randomly assigned binary instrument Z that encourages a binary treatment D. Suppose without loss of generality that $Z = 1$ encourages $D = 1$. The outcome is denoted by Y. We use a potential outcomes framework modified for the IV setting. As before, each subject is assumed to have a potential outcome for each possible treatment (i.e. $Y_{D=1}$ and $Y_{D=0}$). Each subject is also assumed to have a potential treatment for each possible value of the instrument (i.e. $D_{Z=1}$ and $D_{Z=0}$). Further, each subject has a potential outcome for each possible instrument value (i.e. $Y_{Z=1}$ and $Y_{Z=0}$). We assume that Z only affects the outcome Y through D, so

$$Y_{Z=z} = Y_{D_{Z=z}}. \tag{3}$$

Under this assumption, the CACE is equivalent to the average causal effect of the instrument Z on Y among compliers. That is,

$$CACE = E[Y_{D=1} - Y_{D=0} | D_{Z=1} = 1, D_{Z=0} = 0] = E[Y_{Z=1} - Y_{Z=0} | D_{Z=1} = 1, D_{Z=0} = 0]. \tag{4}$$

In other words, estimating the CACE amounts to estimating the causal effect of the instrument among a subgroup (compliers). We are therefore interested in modeling the heterogeneity of the effect of the instrument within the (latent) subgroup of compliers. This places us back in a similar position to the ITT case. Indeed, models of treatment effect heterogeneity in the instrumental variable setting can be represented by the same graphical model (Figure 1) as the standard ITT case. However, the interpretation of cer-

tain nodes changes, and there are certain added constraints on parameter values. Latent class (the G_i node in Figure 1) now encodes compliance status as well as treatment effect class. We follow Angrist, Imbens and Rubin (1996) in defining four types of subjects or 'principal strata': always takers, never takers, compliers, and defiers. Their definitions are as follows: always takers would take the treatment regardless of their instrument value; never takers would not take the treatment regardless of their instrument value; compliers would take the treatment if and only if encouraged by their instrument; and defiers would take the treatment if and only if discouraged by their instrument. We make the common assumption that there are no defiers. We get to observe the principal strata of some subjects, but other subjects' principal strata are latent. Units with $Z = 1$ and $D = 0$ are definitely never takers, and units with $Z = 0$ and $D = 1$ are definitely always takers. But units with $Z = 1$ and $D = 1$ could either be compliers or always takers, and units with $Z = 0$ and $D = 0$ could either be compliers or never takers. G_i takes one value for never takers, one value for always takers, and one value for each treatment effect class for compliers to allow for discrete heterogeneity in the CACE. Because we assume that the instrument only affects the outcome through the treatment, the instrument effect (represented by Δ_i in Figure 1) must be 0 whenever latent class G_i indicates a never-taker or always-taker.

In the IV application we consider in this paper, the outcome (number of visits to the emergency department) is a count variable which we model as negative binomial. We parameterize the outcome in terms of its mean and allow the log of the mean to vary discretely with latent class and continuously with covariates. We call the resulting model M_{IV}, and it is specified below (including application specific weakly informative priors).

$\mathbf{M}_{IV}$ specification:

$$
\begin{aligned}
&Y_{z=0,i} \sim NegBinom(p_{z=0,i}, r_{z=0}), \qquad Y_{z=1,i} \sim NegBinom(p_{z=1,i}, r_{z=1}),\\
&G_i \sim \text{Multinomial\_Logistic\_Regression}(X_i; \beta_G),\\
&p_{z=0,i} = r_{G_i}/(r_{G_i} + \mu_{z=0,i}), \qquad p_{z=1,i} = r_{G_i}/(r_{G_i} + \mu_{z=1,i}),\\
&log(\mu_{z=0,i}) = \alpha_C^{G_i} + X_i\beta_C^{G_i},\\
&\Delta_i = \lambda_\Delta^{G_i} + X_i\beta_\Delta^{G_i}, \text{ where } \lambda_\Delta^{G_i} \text{ and } \beta_\Delta^{G_i} \text{ are set to 0}\\
&\qquad\qquad \text{when } G_i \text{ is Never Taker or Always Taker,}\\
&log(\mu_{z=1,i}) = log(\mu_{z=0,i}) + \Delta_i,\\
&\alpha_C^G \sim N(0, 100), \qquad \beta_C \sim N(0, \sigma_C), \qquad \beta_G \sim N(0, \sigma_{\beta_G}), \qquad \beta_\Delta \sim N(0, \sigma_\Delta),\\
&\sigma_C \sim Uniform(0, 10), \qquad \sigma_G \sim Uniform(0, 5), \qquad \sigma_\Delta \sim U(0, 10),\\
&r_1, \ldots, r_M \sim U(0, 10) \text{ where } M \text{ denotes the number of latent classes,}\\
&\lambda_\Delta^{complier,1} \sim N(0, 5), \qquad \lambda_\Delta^{complier,2} - \lambda_\Delta^{complier,1} \sim Unif(0, 5).
\end{aligned}
\tag{5}
$$

Identifiability Issues

Identifiability issues can arise for any of the models discussed above when in reality there are no latent classes. In this case, if the error distributions are properly specified, there

can be negligible or no difference in the likelihood between different parameter settings. For instance, the value of α_G (which determines probability of class membership) is irrelevant if there is no difference between classes. The estimates of β_G will still converge to 0 if there are no latent classes, though, so there is no danger of wrongly concluding that there is discrete heterogeneity *associated with observed covariates*.

If the error distributions are misspecified, the latent class component of the model might help to better model them. If there are no latent classes in reality, a Markov Chain Monte Carlo (MCMC) may still converge to unique parameter values that best model the misspecified error distributions. Despite convergence, it is still not correct to interpret latent class component parameters in terms of heterogeneity in this scenario. But, again, if there are no latent classes in reality then the estimates of β_G should be near 0 and there is no danger of wrongly concluding that heterogeneity is associated with observed covariates.

Even if there are latent classes in reality, improvements in likelihood from better modeling misspecified error distributions can pull estimates away from their 'correct' values (that is, the values with correct implications about heterogeneity if interpreted as intended). That is why it is important to include flexible error distributions in the model. Simulations in Section 3.2 illustrate this phenomenon.

Model Evaluation

We follow the framework for model comparison by Bayesian cross validation laid out in Vehtari and Lampinen (2002). A sensible measure of a model M's value is the expected utility of using M to make predictions about future observations generated by the same process that generated the training data. A Bayesian model produces a posterior predictive distribution for future outcome y_{new} given future covariates x_{new} and the data D that the model M was fit to:

$$p(y|x_{new}, D, M) = \int p(y|x_{new}, \theta, D, M)p(\theta|D, M)d\theta, \tag{6}$$

where θ denotes the model parameters. The utility of M for predicting a new outcome is some function $u[y_{new}, x_{new}, p(y|x_{new}, D, M)]$ of the outcome and the posterior predictive distribution that measures how well the posterior predictive distribution predicted the outcome. We can estimate the expected utility of a model on populations similar to the training data as the average

$$\frac{1}{N}\sum_{i=1}^{N} u[y_i, x_i, p(y_i|x_i, D^{-i}, M)], \tag{7}$$

where D^{-i} denotes the data with the i^{th} observation removed and N is the number of observations in D. This is the Leave One Out Cross Cross Validation (LOO-CV) estimate of the expected utility of a model M. To estimate the expected utility on a population whose covariates differ from the training data in known ways, a weighted average can be used. Because it is computationally prohibitive to fit the model once for each data point, we approximate the LOO-CV estimate using an importance sampling

scheme proposed by (Gelfand, 1996; Vehtari and Lampinen, 2002). To compare two models M_1 and M_2, interest lies in their expected difference in utility, which can be estimated as:

$$\bar{u}_{M_1-M_2} = \frac{1}{N}\sum_{i=1}^{N} u[y_i, x_i, p(y_i|x_i, D^{-i}, M_1)] - u[y_i, x_i, p(y_i|x_i, D^{-i}, M_2)]. \quad (8)$$

Generally, the choice of utility function depends on the application. For many applications, the posterior predictive mean is taken as the forecast and an appropriate utility is a monotonic function of the distance of the posterior predictive mean from the actual outcome. For example, the squared error utility function would be:

$$u_{se}[y_{new}, x_{new}, p(y|x_{new}, D, M)] = (y_{new} - \int yp(y|x_{new}, D, M)dy)^2. \quad (9)$$

Such utilities are problematic for the purpose of distinguishing models that contain discrete latent class heterogeneity from those that contain only continuous heterogeneity because they ignore the shape of the posterior predictive distribution. If there really is heterogeneity, the posterior predictive distribution for a latent class model will be multimodal and its mean will lie somewhere between the modes. The posterior predictive distribution for a continuous effect modification model will usually be unimodal but have a similar mean, so utilities based on the accuracy of the posterior predictive mean will have low power to distinguish these potentially quite different models.

A commonly used utility function that does not suffer from this problem is the posterior predictive density (ppd):

$$u_{ppd}[y_{new}, x_{new}, p(y|x_{new}, D, M)] = p(y_{new}|x_{new}, D, M). \quad (10)$$

This utility rewards models that place lots of posterior predictive probability mass near future outcome values. A model with a multimodal posterior predictive distribution would be rewarded for outcomes that lie near any mode and penalized for outcomes that lie in the low density regions between modes. This utility has several nice theoretical properties as well. The model with the highest mean posterior predictive density minimizes Kullback Leibler distance to the true model. Posterior predictive density is also a proper scoring rule (Dawid and Musio, 2014). A drawback of this utility is that it is sensitive to our choice of error distribution, and we do not directly care about modeling the error distribution for our application. If we use very flexible error distributions for all candidate models, though, this should not be a serious problem.

(Many authors prefer to use $log(ppd)$ as a utility because it is more Bayesian in spirit. If an observation has a ppd of 0 under a given model, this barely impacts the mean ppd over all observations and the model is hardly penalized. Under standard Bayesian reasoning, however, if a single observation has ppd equal to zero this should effectively disqualify the model. Such an observation would have $log(ppd)$ equal to $-\infty$. We performed all analyses using both ppd and $log(ppd)$ as utilities, and there was no substantive difference in results. We present just the ppd comparisons.)

In practice, for any given experiment we might consider many candidate models $M_1, \ldots, M_K$ with varying numbers of latent classes and functional forms. We prefer

the model with the highest LOO-CV estimated expected posterior predictive density. However, we want to be mindful of the possibility that, due to sampling variability, the model with the highest estimated expected utility is not the model with the highest true expected utility. Each model's estimated expected utility is the sample mean of the LOO-CV posterior predictive densities of all the observations from the experiment. Since the samples of LOO-CV ppds produced by each model are based on the same observations, they are dependent and their centers can be compared using classical methods for dependent samples such as paired t-tests or Wilcoxon signed rank tests. Suppose that M_i has the highest estimated expected utility. We can obtain a conservative p-value for the null hypothesis that M_i has the highest true expected utility of all models considered by taking the p-value of the comparison between M_i and the next best model and adjusting for K multiple comparisons using Holm's method (Holm, 1979). There are K possible comparisons we might have made because we would have tested this null hypothesis for whichever model had the best estimated utility. If the p-value we obtain in this way is very low, we would weight the implications of the top model highly compared to the other candidates. If the p-value is high, we would not dismiss the implications of other models with comparable utilities and would accept uncertainty where those implications conflicted with our chosen model.

Of course, just because a model is the best of those we considered does not mean it is a good model. We perform posterior predictive checks (Gelman et al., 1996) to try to identify deviations of our chosen model from the data. If we fail to identify any serious lack of fit, this improves confidence in the conclusions we draw from our model. If we do identify lack of fit, we can address them with new models and repeat the process described above.

3 Simulations

We apply our approach in several simulated examples demonstrating its capabilities and the importance of some of its features. First, we demonstrate the ability of cross validation to distinguish between discrete and continuous heterogeneity. Next, we illustrate the necessity of flexible error distributions. All code for simulations discussed in this section is available in the supplementary materials.

In Sim_1, we generated data from M_{Norm} with the following settings:

$$Y_{z=0,i} \sim N(\mu_{z=0,i}, 1), \qquad Y_{z=1,i} \sim N(\mu_{z=1,i}, 1), \qquad G_i \sim Bernoulli(p_i),$$
$$logit(p_i) = \alpha_G + X_i \cdot \beta_G,$$
$$\mu_{z=0,i} = \alpha_C^{G_i} + X_i \cdot \beta_C^{G_i}, \qquad \Delta_i = \lambda_\Delta^{G_i} + X_i \cdot \beta_\Delta, \qquad \mu_{z=1,i} = \mu_{z=0,i} + \Delta_i,$$

where

$$\alpha_G = -1, \qquad \beta_G = (-2, -1, 0, 1, 2),$$
$$\alpha_C^{G_i} = \begin{cases} 5 & \text{if } G_i = 1, \\ 0 & \text{if } G_i = 1, \end{cases} \qquad \beta_C^{G_i} = \begin{cases} (-2, -1, 0, 1, 2) & \text{if } G_i = 0, \\ (1, 2, 3, 4, 5) & \text{if } G_i = 1, \end{cases}$$
$$\beta_\Delta = (-2, -1, 0, 1, 2), \qquad \lambda_\Delta^{G_i} = \begin{cases} 5 & \text{if } G_i = 0, \\ 15 & \text{if } G_i = 1. \end{cases}$$

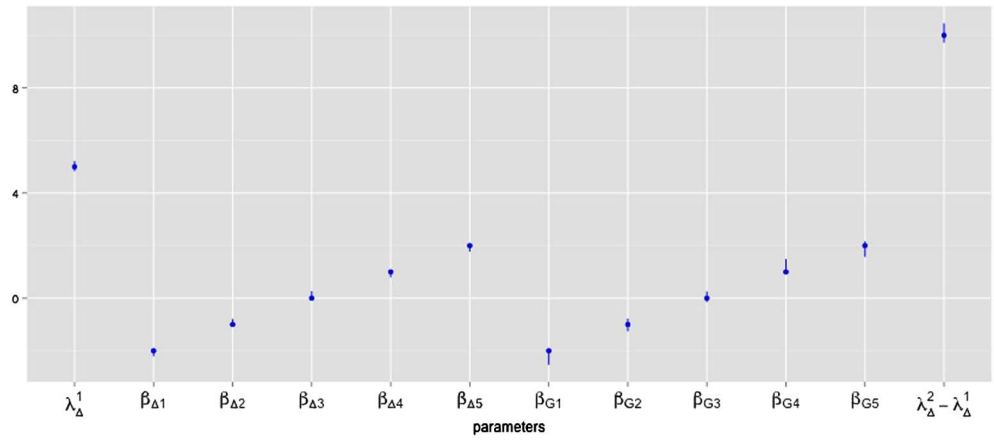

Figure 2: The results of fitting M_{Flex} to Sim_1. The dots are the true parameter values and the lines are 95% credible intervals.

We simulated a two armed clinical trial with 500 patients in each arm. X consisted of 5 predictor variables generated from a standard normal distribution. Both continuous and discrete heterogeneity was present. We then fit three models to this data: M_{Flex}, $M_{Flex}^{Continuous}$ (which is identical to M_{Flex} but without a discrete heterogeneity component), and $M_{Flex}^{Constant}$ (which is identical to M_{Flex} but without discrete or continuous heterogeneity). In Figure 2, we see that the posterior of M_{Flex} is accurate and clearly does not sacrifice too much precision for the robustness gained from flexible error distributions.

The LOO-CV estimated expected posterior predictive densities for M_{Flex}, $M_{Flex}^{Continuous}$, and $M_{Flex}^{Constant}$ were 0.22, 0.12, and 0.12 respectively. A paired t-test comparing M_{Flex} and $M_{Flex}^{Continuous}$ rejected the null hypothesis that $E[u_{M_{Flex}} - u_{M_{Flex}^{Continuous}}] \leq 0$ with p-value numerically 0. Hence, cross validation decisively favored the correct heterogeneity model M_{Flex}.

In Sim_2, we generated data from a model we will call $M_{Norm}^{Continuous}$, which is identical to M_{Norm} but without any discrete components:

$$Y_{z=0,i} \sim N(\mu_{z=0,i}, 1), \qquad Y_{z=1,i} \sim N(\mu_{z=1,i}, 1),$$
$$\mu_{z=0,i} = X_i \cdot \beta_C, \qquad \Delta_i = X_i \cdot \beta_\Delta, \qquad \mu_{z=1,i} = \mu_{z=0,i} + \Delta_i,$$

where

$$\beta_C = (-2, -1, 0, 1, 2), \qquad \beta_\Delta = (-2, -1, 0, 1, 2).$$

Again, we simulated a clinical trial with 500 patients in each arm. X again consisted of 5 predictor variables generated from a standard normal distribution. We then fit the same three models to this data that we fit to Sim_1. M_{Flex} exhibited the identifiability issues discussed in the previous section that can arise when there are no latent

classes in the true data generating process and the error distributions are correctly (over-)specified. Different MCMC chains got stuck at very high or low values of α_G, but all chains converged to 0 for β_G and the correct values for β_Δ. The LOO-CV estimated expected posterior predictive densities for M_{Flex}, $M_{Flex}^{Continuous}$, and $M_{Flex}^{Constant}$ were .2799, .2801, and 0.133 respectively. The p-value from a paired t test comparing the samples from M_{Flex} and $M_{Flex}^{Continuous}$ was 0.001. So cross validation selected the simplest correct model $M_{Flex}^{Constant}$.

3.1 The Importance of Flexible Error Distributions

We simulated data from a model similar to M_{Norm} from Sim_1 but with highly skewed error distributions:

$$
\begin{aligned}
&Y_{z=0,i} \sim Gamma(\mu_{z=0,i}, shape_0, scale_0),\\
&Y_{z=1,i} \sim Gamma(\mu_{z=1,i}, shape_1, scale_1),\\
&G_i \sim Bernoulli(p_i),\\
&logit(p_i) = \alpha_G + X_i \cdot \beta_G,\\
&\mu_{z=0,i} = \alpha_C^{G_i} + X_i \cdot \beta_C^{G_i}, \qquad \Delta_i = \lambda_\Delta^{G_i} + X_i \cdot \beta_\Delta, \qquad \mu_{z=1,i} = \mu_{z=0,i} + \Delta_i,
\end{aligned}
$$

where

$$
\begin{aligned}
&\alpha_G = -1, \qquad \beta_G = (-2,-1,0,1,2),\\
&\alpha_C^{G_i} = \begin{cases} 5 & \text{if } G_i = 1, \\ 0 & \text{if } G_i = 1, \end{cases} \qquad \beta_C^{G_i} = \begin{cases} (-2,-1,0,1,2) & \text{if } G_i = 0, \\ (1,2,3,4,5) & \text{if } G_i = 1, \end{cases}\\
&\beta_\Delta = (-2,-1,0,1,2), \qquad \lambda_\Delta^{G_i} = \begin{cases} 5 & \text{if } G_i = 0, \\ 15 & \text{if } G_i = 1. \end{cases}
\end{aligned}
$$

X consisted of 5 predictor variables generated from a standard normal distribution. $Gamma(\mu, shape, rate)$ denotes a Gamma distribution shifted to have mean μ. We chose $shape_0 = shape_1 = 1$ and $scale_0 = scale_1 = 10$ so that the error distributions were

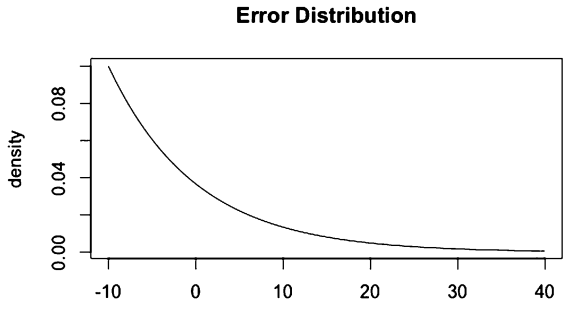

Figure 3: Skewed Error Distribution.

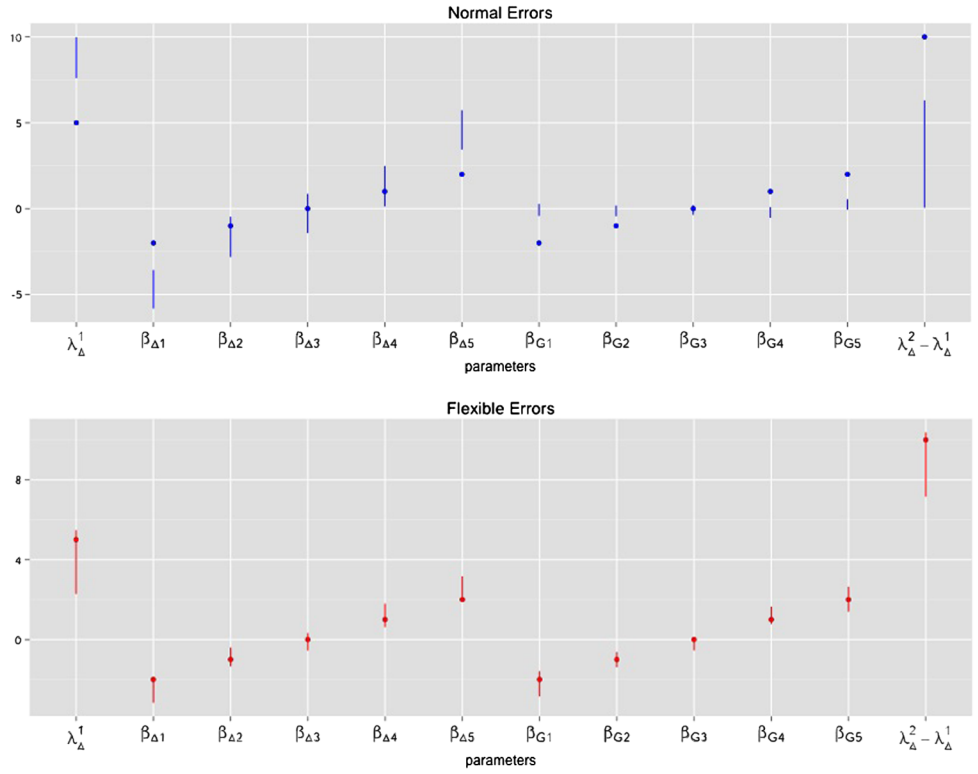

Figure 4: Comparison of M_{Norm} and M_{Flex}.

highly skewed as in Figure 3. We fit the models M_{Norm} and M_{Flex} described in the previous section to the data simulated from the above process. The two models only differed in their error distributions and were correctly specified in all other respects. Figure 4 compares the models' estimates of certain parameters of interest. We see that the M_{Norm} estimates are off target for some parameters, including the $\lambda_\Delta^2 - \lambda_\Delta^1$ parameter that represents the magnitude of discrete heterogeneity between latent classes. The M_{Flex} estimates are fairly accurate for all parameters. These results illustrate sensitivity to misspecification of the error distribution and reassure us that the strategy of employing a flexible (mixture of normals) error distribution is sufficient to handle the problem.

To illustrate model comparison in this setting, we also consider $M_{Flex}^{Continuous}$. We use LOO-CV to compare $M_{Flex}^{Continuous}$ to M_{Flex}, which we know to be the superior model. In our simulated example, the LOO-CV estimated expected posterior predictive density of M_{Flex} was .04 compared to .03 for $M_{Flex}^{Continuous}$. A paired t-test comparing the samples of LOO-CV ppd's from the two models rejected the null hypothesis that $E[u_{M_{Flex}} - u_{M_{Flex}^{Continuous}}] \leq 0$ with p-value numerically 0.

4 Re-Analysis of Data from the ACTG 320 Clinical Trial

The ACTG 320 trial compared two AIDS treatments – a combination of indinavir, zidovudine, and lamivudine versus just zidovudine and lamivudine. Following Shen and He (2015), who themselves follow Hammer et al. (1997) and Zhao et al. (2012), we take change in CD4 count at the 24th week of treatment as the response variable, exclude patients with missing outcome values or extreme CD4 counts, and ignore any bias that we may induce by these exclusions. We are left with a dataset of 800 patients. A summary of the data is included in the Supplementary Appendix (Shahn and Madigan, 2016).

Before fitting any models, we test the null hypothesis of a constant treatment effect using Rosenbaum's covariance adjustment test (Rosenbaum, 2002). The test produces a p-value that is approximately 0, so we are quite certain that there is heterogeneity. The question remains whether it is related to observed covariates and whether we can effectively model it.

We fit multiple models and compare them using LOO-CV with posterior predictive density as the utility function. In some models, we use just the 3 covariates that Shen and He considered (baseline CD4, baseline RNA, and age), while in others we take advantage of regularization to include the 9 other variables that were available. All continuous effect modification was specified as linear, and all latent class membership models were specified as logistic or, in the case of M_9, multinomial logistic regressions. The table at the top of Figure 5 summarizes key attributes of the models. A summary of parameter estimates from select models is in the Supplementary Appendix. Figure 5 depicts the LOO-CV estimated expected posterior predictive densities of each model. The first thing that jumps out in this plot is that M_1, which is Shen and He's model with a normal error distribution, performs far worse than the other models which all use flexible error distributions. This is not necessarily meaningful, however. Our parameters of interest do not govern the error distribution, so the utility could be rewarding these models purely for better modeling an aspect of the data that is not important to us. But comparing the parameter estimates of M_3 (which is Shen and He's model with flexible error distributions) to M_1, we see that there are substantive differences. The estimated difference in treatment effects between latent classes is significantly smaller in M_3 than in M_1. Observing that the residuals are highly skewed and recalling the lessons learned from the simulation in Section 3.1, we suspect that the misspecified error distribution of M_1 biased the estimates of parameters of interest. However, it is still true that much of the difference in expected utility could be due to error distribution alone.

Next we turn our attention to the models with flexible error distributions. The most complex model, M_9, has the highest expected utility. When comparing models, we note that the estimated utilities are the sample means of the LOO-CV posterior predictive densities of the 800 observations. Since the samples of LOO-CV ppds produced by each model are based on the same observations, they are dependent and can be compared using classical methods for dependent samples such as paired t-tests or Wilcoxon signed rank tests. Paired t-tests indicate that the sample mean LOO-CV utility for M_9 is statistically significantly greater than the sample mean LOO-CV utilities of every other model. We can obtain a conservative p-value for the null hypothesis that M_9

	M1	M2	M3	M4	M5	M6	M7	M8	M9
Flexible Error Distributions		✔	✔	✔	✔	✔	✔	✔	✔
Continuous Heterogeneity					✔	✔	✔	✔	✔
Discrete Classes	2	1	2	2	1	1	2	2	3
# Covariates	3	0	3	12	3	12	3	12	12

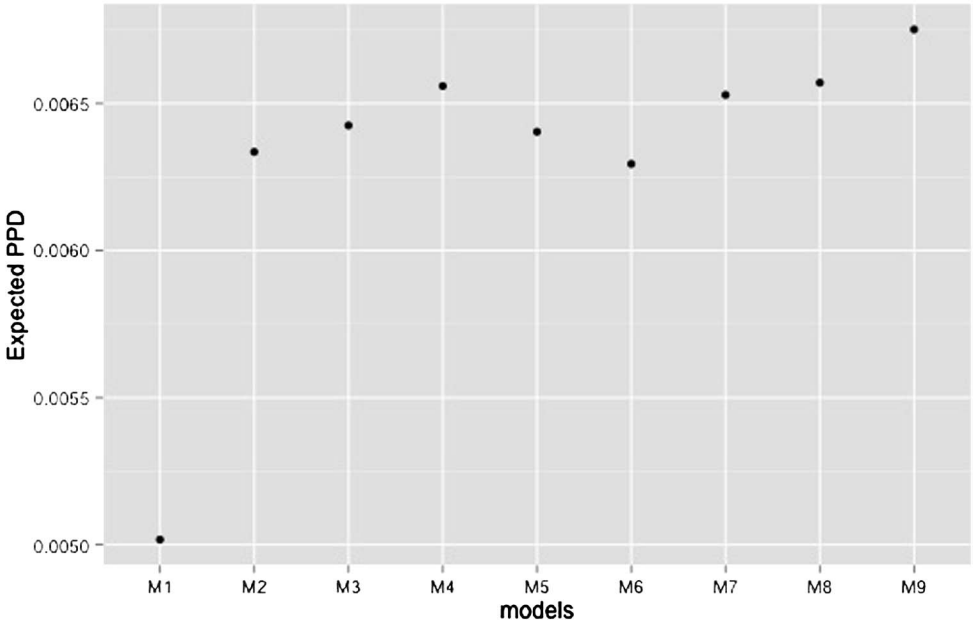

Figure 5: LOO-CV estimated expected posterior predictive densities of each candidate model fit to the ACTG data. M_1 is Shen and He's model with two latent classes, a common normal error distribution for all patients, no continuous effect modification, and just 3 covariates. M_2 is a constant effect model with separate flexible error distributions for each treatment group. Note that 'constant effect' is a misnomer, since the distinct error distributions for the two treatment arms allow for heterogeneity, just not associated with the covariates. M_3 is the same as M_1 but with separate flexible error distributions for each treatment group. M_4 is the same as M_3 but includes all 12 covariates. M_5 is a continuous effect modification model with separate flexible error distributions for each treatment group and only 3 covariates. M_6 is the same as M_5 but includes all 12 covariates. M_7 is a 2 latent class mixture model with continuous effect modification and flexible error distributions and only 3 covariates. M_8 is the same as M_7 but includes all 12 available covariates. M_9 is the same as M_8 except that it has 3 latent classes instead of 2.

has the highest true expected utility of all models considered by taking the p-value of the comparison between M_9 and the next best model (M_8) and adjusting for multiple comparisons using Holm's method. The paired t-test comparing M_9 to M_8 had p-value .006, and adjusting for the other 7 similar comparisons we might have made (i.e. testing

whether models 2 through 8 were the best) yields $p \approx .04$. That is, the probability of M_9 having such a superior estimated utility due to sampling variation alone if any of the other models had true expected utilities as good as M_9's is less than .04 ('less than' because our p-value is conservative). This indicates strong but not necessarily overwhelming support for M_9, so we would not completely dismiss other models with similar utilities such as M_8, M_7, and M_4. We definitely prefer M_9 but would take its implications with a grain of salt if they contradicted one of the other models with fairly similar utility.

We now take a closer look at what the models said about heterogeneity. Where comparisons between models could be made, the models were generally in agreement. First, every model found that heterogeneity was associated with the covariates (apart from the constant model $M2$, obviously). Of the three covariates that were included in every model (except M_2), baseline CD4 count and RNA levels, which we will denote $cd4_0$ and rna_0, were unanimously positively associated with treatment effect after adjusting for other covariates. The models that contained all 12 covariates also agreed that weight was positively associated and prior zidovudine exposure negatively associated with treatment effect adjusting for other covariates. (We will omit 'adjusting for other covariates' for the remainder of this discussion, but it should be understood that all associations might depend on which other covariates were included in the model.)

Every model that contained both continuous and discrete heterogeneity components (M_9, M_8, and M_7) attributed most covariate associated heterogeneity to discrete differences between latent classes. Every model that included latent classes agreed that there was a substantial difference of about 60-80 CD4 count between the highest treatment effect class and the lowest. The three class model, M_9, also included a middle class with estimated treatment effect approximately 10 higher than in the lowest class. The treatment effect in the low class for an average patient was about 45-55 CD4 count in all models. All latent class models agreed that $cd4_0$ and rna_0 were positively associated with membership in the highest class. The models with 12 covariates also agreed that weight was positively associated and prior zidovudine negatively associated with membership in the highest class. In the two class models (M_3,M_4,M_7, and M_8), strength of association with class membership is easily discerned from the posterior distributions of the relevant logistic regression coefficients from the β_G parameter. In the preferred three class model, in which class is determined by a multinomial logistic regression, the relationship between β_G and the nature of the association is more subtle. Figure 6 compares the M_9 posterior distributions of probability of membership in the highest effect class for three hypothetical patients – one with the maximum observed value of $cd4_0$, one with the median observed value of $cd4_0$, and one with the minimum observed value of $cd4_0$. All three hypothetical patients were assigned median values for all other covariates. Comparing the high and medium patients, we see that the high $cd4_0$ value makes low probabilities of membership in the highest treatment effect class less likely but does not alter the mode. For very low values of $cd4_0$, membership in the high treatment effect class is virtually impossible. So M_9 appears to pick up on a nonlinear aspect to the association between $cd4_0$ and highest treatment effect class membership probability. Low values of $cd4_0$ are also strongly associated with membership in the middle class. Figure 7 depicts how the M_9 posterior predictive distribution of Δ varies with

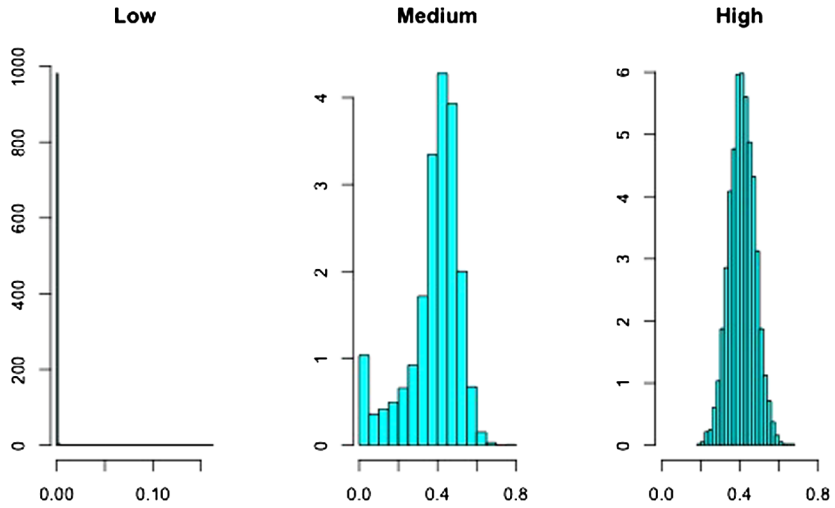

Figure 6: Posterior predictive distributions from model M_9 of probability of membership in the highest treatment effect class for hypothetical patients with low, median, and high $cd4_0$ values.

rna_0. The models that contained both continuous and discrete heterogeneity components (M_9, M_8, and M_7) also all detected a possible moderate linear association with rna_0 and no other significant linear associations. So our preferred model M_9 and all the other credible models together imply mostly discrete heterogeneity associated with $cd4_0$, rna_0, weight, and zidovudine exposure along with possible modest continuous effect modification by rna_0. The strong performance of M_9 might be attributed to the more flexible relationships it allows between covariates and high effect class membership. A more thorough analysis would explore models with nonlinear regressions, more than three latent classes, interactions among covariates, and distinct error distributions for each latent class instead of just for each treatment group.

As a last step, we performed some basic posterior predictive checks (Gelman et al., 1996) to affirm that M_9 not only outperforms the other candidates but also fits the data reasonably well. First, Figure 8 compares a histogram of the outcomes from the real trial to a histogram of fake outcomes that were simulated from the posterior mean values of the parameters from M_9 and the observed covariate values. The distributions are remarkably similar. We then simulated 1000 fake data sets from the posterior predictive distribution of M_9, computed summary statistics of each fake data set, and checked whether the corresponding summary statistics of the true data fall within the range of the simulations. The exercise is summarized in Figure 9, in which the red dots indicate the summary statistic values in the real data. The model appears to fit well by these criteria. While it is possible that a four latent class model could have superior expected utility to M_9, we stopped at three classes because the substantive implications of the three class model are very similar to the implications of similar models with two classes and all of these models fit the data well. Interpretability would also suffer from fur-

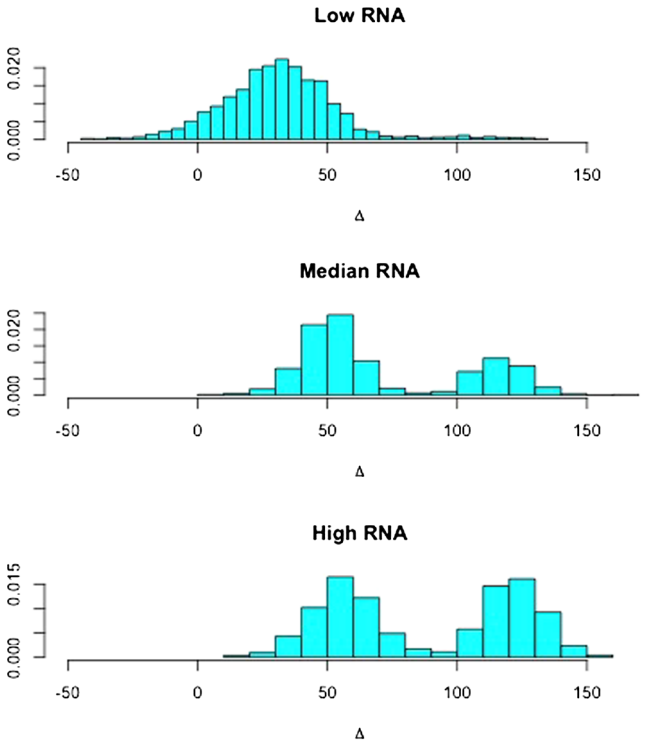

Figure 7: Posterior predictive distributions from model M_9 of the Δ parameter for hypothetical patients with low, median, and high rna_0 values.

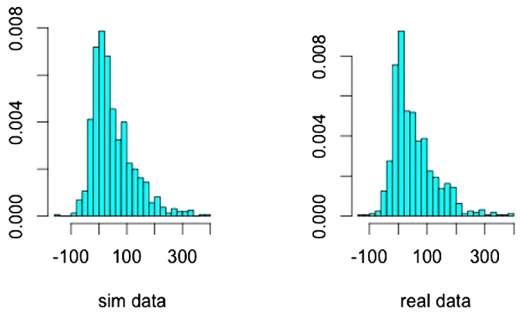

Figure 8: The distribution of the outcome variable in the trial (right) and a draw from the M_9 posterior predictive distribution of the outcome variable (left).

ther complexity, as we have seen that interpretation of the parameters governing class membership in M_9 is already subtle. Further, it is well known that LOO-CV exhibits a tendency toward overfitting, so it can be good practice to cap model complexity when substantive implications are constant and fit is adequate.

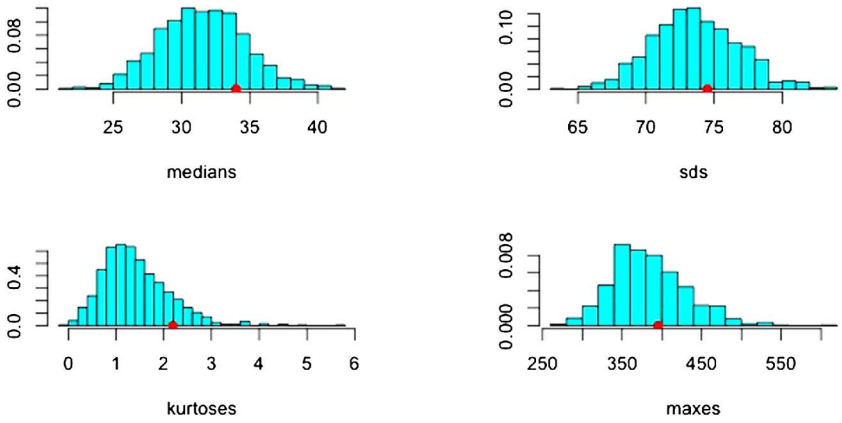

Figure 9: Posterior Predictive Checks of M_9. Posterior predictive distributions of the median, standard deviation, kurtosis, and maximum of the outcomes assign high probability to the values observed in the actual data.

5 The Oregon Health Insurance Experiment (OHIE) Study

5.1 Description of the Study and the Data

In 2008, Oregon instituted a lottery to determine who could enroll in a new Medicaid program with limited openings. The randomness of the lottery induced a natural experiment that has allowed researchers to explore various public health and economic effects of Medicaid (Taubman et al., 2014). One important health economic question was what effect if any Medicaid might have on emergency department (ED) utilization. The intuitive and naive guess would be that health insurance would increase utilization by decreasing cost. However, many experts had predicted that expanding health insurance would actually decrease ED utilization for two main reasons. First, uninsured patients sometimes go to EDs for problems that could be addressed in a primary care setting because, unlike primary care physicians, EDs cannot turn patients away for being unable to afford treatment. Second, assuming Medicaid coverage would increase primary care utilization, regular monitoring of chronic conditions at primary care visits might prevent flareups that necessitate trips to the ED.

Taubman et al. addressed this question by looking at ED utilization among the 24,000 lottery participants who lived in Portland. They matched these lottery participants to medical records from 12 hospitals that accounted for almost all ED visits for Portland residents over the period of the study. Unfortunately, because many lottery winners did not go on to actually enroll in Medicaid and some lottery losers managed to enroll through other channels, Taubman et al. could not simply compare lottery winners to losers to directly estimate the average causal effect (ACE) of Medicaid coverage. In these situations, the best one can do is to estimate the 'Complier Average Causal Effect' (CACE) using an instrumental variable analysis. The CACE is the average causal effect

of Medicaid coverage on those lottery participants who would enroll in Medicaid if and only if they won the lottery (Angrist, Imbens and Rubin, 1996; Frangakis and Rubin, 2002), i.e. the compliers. Taubman et al estimated the CACE to be positive with high confidence, supporting the naive and intuitive prediction that Medicaid coverage would increase ED utilization.

Restricting their data to approximately 10,000 lottery participants who filled out a survey containing questions pertaining to pre-treatment covariates, Taubman et al. explored heterogeneity in the CACE by performing both pre-registered and post hoc subgroup comparisons. They discovered several possible disparities in treatment effect (e.g. between smokers and non-smokers and between people with and without a prior serious chronic disease) but did not adjust either the pre-registered or the post-hoc comparisons for multiple testing. See the Supplementary Appendix for tables summarizing the data and covariates.

5.2 Results of Application to OHIE

We applied model M_{IV} defined in Section 2 to the OHIE data. In the context of the OHIE, the definitions of the principal strata are as follows: always takers would enroll in Medicaid regardless of whether they won the lottery; never takers would not enroll in Medicaid regardless of whether they won the lottery; compliers would enroll if they won the lottery and not enroll if they lost; and defiers would enroll if they lost and not enroll if they won. We make the common assumption that there are no defiers. We get to observe the principal strata of some subjects, but other subjects' principal strata are latent. Lottery winners who don't enroll in Medicaid are definitely never takers, and lottery losers who do enroll are definitely always takers. But winners who enroll could either be compliers or always takers, and losers who do not enroll could either be compliers or never takers. G_i takes one value for never takers, one value for always takers, and one value for each treatment effect class for compliers to allow for discrete heterogeneity in the CACE. The model specified two treatment effect classes within the subgroup of compliers. Because we assume that the instrument only affects the outcome through the treatment, the instrument effect (represented by Δ_i in Figure 1) must be 0 whenever latent class G_i indicates a never-taker or always-taker.

When we included all recorded baseline covariates and used hierarchical shrinkage priors, we did not find evidence that treatment effect among compliers was associated with any of the recorded covariates. (Or, assuming that all covariates are associated with treatment effect at least a little bit, we did not find strong evidence of the direction of any of the associations.) The posterior distributions of all components of β_G and β_Δ were centered near zero with substantial probability mass on either side. A summary of the results is in the Supplementary Appendix.

6 Conclusion

We have illustrated a general Bayesian framework for modeling treatment effect heterogeneity in experiments with non-categorical outcomes. Our modeling approach incorpo-

rates latent class mixture components to capture discrete heterogeneity and regression interaction terms to capture continuous heterogeneity. Flexible error distributions allow robust posterior inference on parameters of interest. Hierarchical shrinkage priors on relevant parameters address multiple comparisons concerns. Leave-one-out cross validation estimates of expected posterior predictive density obtained through importance sampling, together with posterior predictive checks, provide a convenient method for model selection and evaluation.

Simulated and real examples demonstrate the utility of this framework and the importance of its various features. The method provides convincing evidence that the heterogeneity in the ACTG HIV trial is truly discrete and characterizes potential subgroups in terms of baseline covariates. Parameter estimates differ substantially from a prior analysis using a similar method (though the subjective interpretation of the output remains the same) as a result of using flexible error distributions, the importance of which is illustrated in simulations. In the IV analysis of the OHIE data, shrinkage priors serve their purpose and prevent premature identification of heterogeneities that may be due to multiple comparisons.

We see five immediate opportunities for future work. First, it should be relatively straightforward to develop implementations of this approach for other specialized outcome models, in particular for survival analyses. Second, if one could obtain stable estimates of Bayes factors, possibly using the method of Chib and Jeliazkov (2005), more formal methods for model comparison with certain desirable properties would be available, and model averaged estimates of some relevant quantities could be computed. Third, variational Bayes approximations to the posteriors of these models would enable applications to experiments with very large numbers of covariates. Fourth, nonparametric implementations could mitigate concerns about model misspecification. And finally, we have focused on the context of randomized experiments in this paper as opposed to observational studies so that issues surrounding adjustment for confounding did not distract from a direct emphasis on treatment effect heterogeneity. In observational studies, it is usually difficult enough to obtain reliable estimates of average treatment effects without fitting elaborate models for heterogeneity. However, it is conceptually straightforward if perhaps overly ambitious to extend this method to observational settings.

Supplementary Material

Supplementary Appendices of "Latent Class Mixture Models of Treatment Effect Heterogeneity" (DOI: 10.1214/16-BA1022SUPP; .pdf).

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. "Identification of causal effects using instrumental variables." Journal of the American Statistical Association 91.434 (1996): 444–455. 833, 838, 851

Athey, Susan, and Guido Imbens. "Machine learning methods for estimating heterogeneous causal effects." arXiv:1504.01132 (2015). 832

Chib, Siddhartha. "Analysis of treatment response data without the joint distribution of potential outcomes." Journal of Econometrics 140.2 (2007): 401–412. MR2408912. doi: http://dx.doi.org/10.1016/j.jeconom.2006.07.009. 835

Chib, Siddhartha, and Ivan Jeliazkov. "Accept-reject Metropolis–Hastings sampling and marginal likelihood estimation." Statistica Neerlandica 59.1 (2005): 30–44. MR2137380. doi: http://dx.doi.org/10.1111/j.1467-9574.2005.00277.x. 852

Dawid, Alexander Philip, and Monica Musio. "Theory and applications of proper scoring rules." Metron 72.2 (2014): 169–183. MR3233147. doi: http://dx.doi.org/10.1007/s40300-014-0039-y. 840

Foster, Jared C., Jeremy M. G. Taylor, and Stephen J. Ruberg. "Subgroup identification from randomized clinical trial data." Statistics in Medicine 30.24 (2011): 2867–2880. MR2844689. doi: http://dx.doi.org/10.1002/sim.4322. 832

Frangakis, Constantine E., and Donald B. Rubin. "Principal stratification in causal inference." Biometrics 58.1 (2002): 21–29. MR1891039. doi: http://dx.doi.org/10.1111/j.0006-341X.2002.00021.x. 851

Gelfand, Alan E. "Model determination using sampling-based methods." Markov chain Monte Carlo in Practice (1996): 145–161. MR1397969. 831, 840

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. "Why we (usually) don't have to worry about multiple comparisons." Journal of Research on Educational Effectiveness 5.2 (2012): 189–211. 831

Gelman, Andrew, Xiao-Li Meng, and Hal Stern. "Posterior predictive assessment of model fitness via realized discrepancies." Statistica Sinica 6.4 (1996): 733–760. MR1422404. 832, 841, 848

Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., Eron Jr, J. J., Feinberg, J. E., Balfour Jr, H. H., Deyton, L. R. and Chodakewitz, J. A. "A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less." New England Journal of Medicine 337.11 (1997): 725–733. 845

Holm, Sture. "A simple sequentially rejective multiple test procedure." Scandinavian Journal of Statistics (1979): 65–70. MR0538597. 841

Imai, Kosuke, and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation." The Annals of Applied Statistics 7.1 (2013): 443–470. MR3086426. doi: http://dx.doi.org/10.1214/12-AOAS593. 833

Kang, Joseph, et al. "Tree-structured analysis of treatment effects with large observational data." Journal of Applied Statistics 39.3 (2012): 513–529. MR2880431. doi: http://dx.doi.org/10.1080/02664763.2011.602056. 832

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669–688. MR1380809. doi: http://dx.doi.org/10.1093/biomet/82.4.669. 834

Qian, Min, and Susan A. Murphy. "Performance guarantees for individualized treatment

rules." Annals of Statistics 39.2 (2011): 1180. MR2816351. doi: http://dx.doi.org/10.1214/10-AOS864. 832

Rosenbaum, Paul R. "Covariance adjustment in randomized experiments and observational studies." Statistical Science 17.3 (2002): 286–327. MR1962487. doi: http://dx.doi.org/10.1214/ss/1042727942. 845

Rothwell, Peter M. "Subgroup analysis in randomised controlled trials: importance, indications, and interpretation." The Lancet 365.9454 (2005): 176–186. 831

Rubin, Donald B. "Estimating causal effects of treatments in randomized and nonrandomized studies." Journal of Educational Psychology 66.5 (1974): 688. 834

Shahn, Z. and D. Madigan. "Supplementary Appendices of "Latent Class Mixture Models of Treatment Effect Heterogeneity"." Bayesian Analysis (2016). doi: http://dx.doi.org/10.1214/16-BA1022SUPP. 845

Shen, Juan, and Xuming He. "Inference for subgroup analysis with a structured logistic-normal mixture model." Journal of the American Statistical Association 110.509 (2015): 303–312. MR3338504. doi: http://dx.doi.org/10.1080/01621459.2014.894763. 833, 845

Sobel, Michael E., and Bengt Muthen. "Compliance mixture modelling with a zero-effect complier class and missing data." Biometrics 68.4 (2012): 1037–1045. MR3040010. doi: http://dx.doi.org/10.1111/j.1541-0420.2012.01791.x. 833, 837

Su, Xiaogang, et al. "Subgroup analysis via recursive partitioning." The Journal of Machine Learning Research 10 (2009): 141–158. 832

Sarah Taubman, Heidi Allen, Bill Wright, Katherine Baicker, Amy Finkelstein, and the Oregon Health Study Group, "Medicaid increases emergency department use: evidence from Oregon's health insurance experiment." Science 343.6168 (2014 Jan 17): 263–268. 850

Titterington, D. Michael, Adrian F. M. Smith, and Udi E. Makov. Statistical Analysis of Finite Mixture Distributions, vol. 7. New York: Wiley, 1985. MR0838090. 834

Vehtari, Aki, and Jouko Lampinen. "Bayesian model assessment and comparison using cross-validation predictive densities." Neural Computation 14.10 (2002): 2439–2468. 831, 839, 840

Zhang, Baqun, et al. "A robust method for estimating optimal treatment regimes." Biometrics 68.4 (2012): 1010–1018. MR3040007. doi: http://dx.doi.org/10.1111/j.1541-0420.2012.01763.x. 832, 833

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). "Estimating individualized treatment rules using outcome weighted learning." Journal of the American Statistical Association 107.499, 1106–1118. MR3010898. doi: http://dx.doi.org/10.1080/01621459.2012.695674. 832, 845