PA

Modeling Context-Dependent Latent Effect Heterogeneity

Diogo Ferrari

Department of Political Science, University of Michigan, 505 South State Street, 5700 Haven Hall, Ann Arbor, MI 48104, USA.
Email: diogoferrari@gmail.com

Abstract

Classical generalized linear models assume that marginal effects are homogeneous in the population given the observed covariates. Researchers can never be sure *a priori* if that assumption is adequate. Recent literature in statistics and political science have proposed models that use Dirichlet process priors to deal with the possibility of latent heterogeneity in the covariate effects. In this paper, we extend and generalize those approaches and propose a hierarchical Dirichlet process of generalized linear models in which the latent heterogeneity can depend on context-level features. Such a model is important in comparative analyses when the data comes from different countries and the latent heterogeneity can be a function of country-level features. We provide a Gibbs sampler for the general model, a special Gibbs sampler for gaussian outcome variables, and a Hamiltonian Monte Carlo within Gibbs to handle discrete outcome variables. We demonstrate the importance of accounting for latent heterogeneity with a Monte Carlo exercise and with two applications that replicate recent scholarly work. We show how Simpson's paradox can emerge in the empirical analysis if latent heterogeneity is ignored and how the proposed model can be used to estimate heterogeneity in the effect of covariates.

Keywords: bayesian nonparametric model, latent variables, heterogeneous effects, generalized linear models, semiparametric mixture modeling, Dirichlet regression

1 Introduction

This paper proposes a model to deal with context-dependent latent heterogeneity in the effect of covariates in generalized linear models (GLMs). Generalized linear models, including those with mixed effects, are still one of the most used tools for multivariate analyses in political science. Among many assumptions required by such models, e.g., the conditional independence, researchers need to assume that important covariates were not left out. In that regard, much has been said in political science, statistics, and econometrics about the problems caused by omitting additive covariates in the model, but much less about the issues surrounding unobserved confounders that condition the effects of observed covariates. Conditioning features can lead to a well-known phenomenon in statistics called Simpson's paradox (a.k.a. aggregation paradox): an effect found when data are aggregated can be completely different or even reversed when data are separated into groups (Pearson, Lee, and Bramley-Moore 1899; Yule 1903; Simpson 1951; Blyth 1972). The crucial point connecting the paradox and omitting variables is that, in the typical situation, researchers can never be sure *a priori* that there are not latent or unobserved groups—a.k.a. clusters—with heterogeneous effect nor how many of them exist.

Consider for example the study of voter's preferences for redistribution. It is well known that features such as income and race can affect support for redistributive policies (Alesina and Angeletos 2005; Rehm 2009; Shayo 2009; Alesina and Giuliano 2010), but the effects of

Author's note: The author is thankful to Robert Franzese, Walter Mebane, Kevin Quinn, Long Nguyen, as well as participants of 2018 Polmeth and 2018 APSA Annual meeting for helpful comments on previous versions of this manuscript. The author also thanks the editor Jeff Gill and two anonymous reviewers for their invaluable suggestions. Replication materials are publicly available on the *Political Analysis* Harvard Dataverse (Ferrari 2018) as well as author's website.

Political Analysis (2020)
vol. 28:20–46
DOI: 10.1017/pan.2019.13

Published
20 May 2019

Corresponding author
Diogo Ferrari

Edited by
Jeff Gill

such observed factors like income and race can be heterogeneous among subpopulations due to unobserved factors such as motivation, personal history, and ability (Stegmueller 2013). Consequently, the estimated effect of income, e.g., found when data are aggregated can be very different from the effect that would be estimated if we had observed motivation and considered low- and high-motivation groups separately or considered the income effect as conditional upon motivation.

Although that problem occurs in all scientific disciplines, perhaps it is more salient in the social sciences because, to mention a few reasons, problems have high dimensionality and often many dimensions remain unmeasured; data are often difficult to collect or are unavailable for privacy or other reasons; culture-specific aspects are not well measured; some subjects may conceal information from researchers purposefully; or researchers may simply be unaware of possible latent interactive factors (Stegmueller 2013; Traunmuller, Murr, and Gill 2015).

Especially in comparative politics, an additional layer of complication seems likely: the latent heterogeneity can depend on context-level features. For instance, some researchers have shown that the effect of income is conditional on country-level variables such as the progressivity of the tax system (Beramendi and Rehm 2016), the levels of inequality and crime rates (Rueda and Stegmueller 2016), national identity (Johnston *et al.* 2010), and the existing levels of redistribution (Svallfors 1997; Arts and Gelissen 2001). If there is latent effect heterogeneity due to unobserved factors like motivation or personal experiential history among the population from a given context, say a country, it is very likely that a different heterogeneity manifests in other countries. In other words, suppose the effect of income is heterogeneous between two groups of voters (clusters) in the United States (the context) and we do not know the group membership of individual voters. We would expect to see heterogeneity among the population in another context, say Italy, but we should not expect to have the same two latent groups in Italy (or in other contexts). Maybe there are more or fewer latent groups in Italy, or maybe some latent subpopulations are similar in Italy and the United States. For instance, high- and low-motivation Italians and Americans may have welfare opinions similarly affected by their income. But Italians' personal experiences of crime modify income effects on welfare support very differently than Americans' personal experiences of crime modify their income effects on welfare support. In sum, the characteristics of the within-context heterogeneity (clustering) can vary from one context (e.g., country) to another, and that within-context heterogeneity may depend on the characteristics of the context itself.

Practitioners in political science have long recognized these challenges. The possibility of omitting relevant conditioning factors, in conjunction with cross-context differences, have been stressed as an important source of an attitude of *radical skepticism* regarding the results of observational and experimental empirical investigation in the social sciences in general, and in comparative analysis in particular (Przeworski 2007; Stokes 2014).

The literature has proposed different approaches to address effect heterogeneity. The approaches depend on whether the grouping features are known and measured. When the groups are observed, classical approaches include mixed models with gaussian distributed random effects (e.g., hierarchical linear models (HLMs)). Suppose, for example, that we are analyzing data from many countries (contexts) and in each country there are different subpopulations with heterogeneous effects. If we knew the subpopulation to which each individual belongs, we could use a classical mixed-effects model at country and subpopulation levels. However, the distributional assumption on the random effect in such an approach is often criticized because of the single modality, light tails, and symmetry of the normal distribution, which imposes unnecessary and often unjustifiable constraints to the analysis in the empirical modeling stage (Verbeke and Lesaffre 1997; Heinzl and Tutz 2013). In addition, such approach only works if the heterogeneous groups within each context are known and observed, but researchers have to assume that there is no other latent or unobserved feature that can cause effect heterogeneity.

There are some modeling approaches that work for single-context cases in which subpopulation membership is unknown or unobserved. When one wants to investigate subpopulations with latent heterogeneity within a given context and the number of subpopulations are known or are assumed to be finite and fixed, finite mixture models (FMM) are often used (Ng *et al.* 2006; De la Cruz-Mesía, Quintana, and Marshall 2008; Villarroel, Marshall, and Barón 2009). More commonly, however, researchers do not know if or how many latent heterogeneous subpopulations exist within a given context. Recent contributions in statistics literature have proposed models that use Dirichlet process prior (DPP) to deal with these single-context cases with unknown subpopulation heterogeneity/clustering (Mukhopadhyay and Gelfand 1997; Kleinman and Ibrahim 1998b; Hannah, Blei, and Powell 2011; Heinzl and Tutz 2013). Models using DPP have been used in marketing literature to model the error term with a flexible distribution, the heterogeneity of consumer's demand in discrete choice models (Rossi, Allenby, and McCulloch 2006; Rossi 2014), and in latent instrumental variable (LIV) models to deal with endogeneity of covariates (Ebbes *et al.* 2005; Ebbes, Wedel, and Böckenholt 2009). Related work has also been developed in econometrics and program evaluation literature to study effect heterogeneity of training programs (Aakvik, Heckman, and Vytlacil 2005; Chen 2007; Heckman and Vytlacil 2007; Ichimura and Todd 2007; Matzkin 2007). DPP models have been applied in political science to study lengths of time political appointees stay in their appointed position (Gill and Casella 2009), political priorities of senators (Grimmer 2009), intraparty voting (Spirling and Quinn 2010), immigrant turnout in elections (Traunmuller, Murr, and Gill 2015), and dynamic aspects of preferences for redistribution (Stegmueller 2013).

Those DPP approaches, however, have three limitations. First, they are usually designed to be used with specific types of dependent variables, e.g., with outcome variables measured on an ordered scale. Second, particularly in political science literature, previous works have used DPP mostly as a prior only for the intercept (or error) term. Third and more importantly, previous works were not designed to study cases in which the latent heterogeneity is context-dependent.

To redress these limitations, this paper proposes a Dirichlet mixture of generalized linear models in which the within-context effect heterogeneity (clustering) can be context-dependent. The proposed model is a generalization, from the point of view of the expectation of the dependent variable, of usual generalized linear model (GLM), classical generalized linear mixed models (GLMM), finite mixture models (FMMs), and current single-context Dirichlet mixtures of generalized linear models. The proposed model has several advantages over those special cases.

First, when there are multiple contexts, for instance in cross-country comparative analysis, the model can be used to investigate if country features are associated with latent heterogeneity in the covariate effects; that is, if country-level features affect the number and the characteristics of the subpopulation clusters.

Second, the proposed Dirichlet mixture of generalized linear models is developed in its full generality to handle Dirichlet mixtures of any distribution in the exponential family, investigate heterogeneity not only in the error term but in the effect of any observed covariates, and, as mentioned, study how such heterogeneity varies with context-level features. This paper implements two special cases: binary and continuous outcomes, modeled using Bernoulli and gaussian distributions, respectively. The algorithms for estimation of these special cases are presented, but an MCMC algorithm with a Gibbs sampler is derived for the more general model, so it can easily be extended to other outcome variable distributions.

Third, as a generalization of the other models, it can be used in situations in which any of the more specialized models are well justified. If, in fact, one believes that a single GLM can be used across contexts and there is no latent heterogeneity in the population, the proposed model can be estimated and it will produce similar results for the conditional expectation of the dependent variable as those estimated using a GLM. If there is just one context, but unknown clusters, it can be

used instead of the single-context Dirichlet mixture of GLMs. The analogous situation is true for the other special cases, i.e., whenever the researcher is estimating a GLMM or a finite mixture model (FMM) the proposed model can be used, and it has two additional advantages: the number of latent clusters, whose number is allowed to grow with the size of the data, is being simultaneously estimated. As already mentioned, if the data comes from different contexts, the effect of the context on the characteristics of the clusters are also being investigated.

Fourth, the model estimates cluster memberships, so we can classify the data points into (latent) groups. The clusters differ in terms of the vector of linear coefficients that connect covariates to the outcome variable. So it can be used to study and characterize the heterogeneity in the effect of the covariates within and across contexts.

Fifth, the statistics and epidemiology literature has proposed approaches for dealing with Simpson's paradox based on domain knowledge (Hernán, Clayton, and Keiding 2011) and estimation diagnostics (Kievit *et al.* 2013). Its formal aspects and its connection to other problems have also been studied (Samuels 1993; Hernán, Clayton, and Keiding 2011; Pearl 2011, 2014). However, to the best of our knowledge, the literature has not proposed any modeling solution. We connect the model proposed here with Simpson's paradox in the context of generalized linear models and show how it can be used to detect the occurrence of the paradox and to deal with such problems by estimating the cluster-specific effects.

The rest of the paper is organized as follows. The next section presents the model. Then, the following section demonstrates how the proposed model is connected to classical GLM, mixed models, FMM, and the econometric models mentioned above that use DPP to deal with heterogeneity in single-context analyses. Section 4 develops MCMC algorithms to estimate the model in its full generality and for two special cases of outcome variables. In the Section 5 we conduct a Monte Carlo exercise to study the frequentist properties of the estimation. The estimation is tested against a large variety of scenarios with and without latent heterogeneity. The section also illustrates how the model can be used to deal with Simpson's paradox in the context of generalized linear models. It also compares the estimated results of GLM using a maximum likelihood estimator (MLE) with those produced by the proposed model using the MCMC developed in the Section 4. Section 6 uses the model to analyze real data sets. It replicates some studies and shows how it uncovers latent heterogeneity and Simpson's paradox. Finally, the conclusions are presented.

2 The Model

To restate the problem, we want to use a generalized linear model to estimate the effect of the covariates X_i on y_i. Second, we want to take into account the possibility that the effect of the covariates is heterogeneous across different subgroups whose defining features are latent or were not observed. In other words, there might be latent subpopulations of individuals for which the covariates have a different relationship with the outcome. Finally, we want to allow this latent subpopulation heterogeneity or clustering to be investigated both for data that comes from a single context or from multiple contexts. Finally, when the observed population comes from different contexts (e.g., different countries and different years), we want to investigate if context-level features change not only the effect of observed covariates on the outcome but also the existence and the characteristics of latent subpopulations in which the observed covariates have different effects.

The model that deals with such problems can be developed as follows. For each observation i, suppose we have a set of observed covariates $X_i' \in \mathbb{R}^{D_x}$ and an outcome variable y_i. Denote $X_i = (1, X_i')$. Let K denote the number of heterogeneous groups in the population such that it can be bigger if the population is bigger, and let Z_i indicate the group of i. Z_i and K may or may not be known or observed. When Z_i is not observed, we use the term "clusters" instead of "groups."

Denote C_i the context of i, so $C_i = j$ indicates that the observation i comes from context indexed by $j \in \{1, \ldots, J\}$, were J is the number of contexts.

For the purpose of illustration and as a toy example, suppose we want to investigate the effect of income and race on voters' support for welfare policies in different countries. Then X_i are measures of income and race of individual i, and y_i is his degree of support for welfare policies. The variable $C_i = j$ indicates the country where i lives, and data are collected in J countries. Suppose further that in each country the population is divided into *types* of individuals with different personal experiences with class and racial conflict. The types are not observed but we suspect the effect of income and race is conditional on the type. The latent variable $Z_i = k$ indicates that i is type k and K denotes the number of different types.

If $p(\,)$ denotes a distribution in the exponential family, g a link function, and $\theta = (\beta, \sigma)$, then the group- and context-specific GLM is given by:

$$y_i \mid Z_i, X_i, C_i, \theta_{C_i Z_i} \sim p(y_i \mid X_i, \theta_{C_i Z_i}) \ni \mathbb{E}[y_i \mid \cdot] = \mu_i = g^{-1}(X_i^T \beta_{C_i Z_i}), \quad Z_i = 1, \ldots, K. \tag{1}$$

If Z_i was observed and K was therefore known, one could use classical mixed-effects models to estimate groups and context-specific heterogeneous effects. If K was known, but Z_i was latent or unobserved, one option would be to use finite mixture models for the estimation (Gaffney 2003; Ng *et al.* 2006; De la Cruz-Mesía, Quintana, and Marshall 2008; Villarroel, Marshall, and Barón 2009). When Z_i is latent and K is unknown, some authors have proposed models that use DPP on θ in order to estimate cluster-specific effects[1] (Mukhopadhyay and Gelfand 1997; Kleinman and Ibrahim 1998a,b; Dorazio *et al.* 2008; Gill and Casella 2009; Heinzl and Tutz 2013; Stegmueller 2013; Traunmuller, Murr, and Gill 2015). We refer here to such models as Dirichlet process generalized linear model (dpGLM), as adopted by Hannah, Blei, and Powell (2011). Contrary to their formulation, however, we assume that X_i is given. If we denote by $\mathcal{DP}(\alpha, G)$ the Dirichlet process with location parameter α and base measure G, the GLM is modified in the following way to produce the dpGLM:

$$\begin{aligned} G \mid \alpha_o, G_o &\sim \mathcal{DP}(\alpha_o, G_o) \\ \theta_i \mid G &\sim G \\ y_i \mid X_i, \theta_i, &\sim p(y_i \mid X_i, \theta_i), \quad \mathbb{E}[y_i \mid \cdot] = \mu_i = g^{-1}(X_i^T \beta_i). \end{aligned} \tag{2}$$

Authors have warned that using DPP can lead to biased estimators, and it is known that neither weak consistency nor asymptotic unbiasedness are guaranteed in general in DPP models (Diaconis and Freedman 1986; Ghosal, Ghosh, and Ramamoorthi 1999; Tokdar 2006; Kyung *et al.* 2010). Although bias will always be present due to the bayesian priors, Hannah, Blei, and Powell (2011) demonstrated that the dpGLM satisfies the conditions that guarantee weak consistency of the joint posterior distribution and consistency of the regression estimates (see also Tokdar 2006).

The dpGLMs lacks the hierarchical clustering approach that we would like to have in the model, that is, that the clusters can be a function of higher level context features in a multi-context analysis. We want to preserve the structure of the dpGLM and the DPP—because there might be unknown clusters with heterogeneous effect and unknown cluster membership—but include such context dependency—because the heterogeneity may depend on the context characteristics.

Some authors have proposed different approaches to model hierarchical clustering and to create dependencies among multiple Dirichlet processes (Mallick and Walker 1997; Carota and Parmigiani 2002; De Iorio *et al.* 2004; Müller, Quintana, and Rosner 2004; Teh *et al.* 2006). We can

1 The clustering property of the DPP will not be revised here because there are already good sources explaining such feature of the Dirichlet process prior. The reader interested in a review can check Teh *et al.* (2006) and Müller and Mitra (2013) and the references therein.

generalize and combine these approaches with the dpGLM in the following way. Let $W_j' \in \mathbb{R}^{D_w}$ denote the context-level features of context j and J the total number of contexts as before. Let $W_j = (1, W_j')$.

In our toy example, W_j can be thought of as the level of economic development and inequality of the country (context) j. It means we want to investigate if the effect of income and race (the observed covariates X_i) on support for redistribution (y_i) varies with the degree of economic development and inequality of the country (W_j). Moreover, we also want to investigate if the effect of those observed covariates (income and race) is different among within-country subpopulations whose membership (Z_i) is unobserved. Finally, we want to verify if those subpopulations vary from one country (context) to another due to a country's level of inequality and economic development (context-level features W_j).

The model can be modified in the following way to introduce context-level dependency among DPP:

$$\begin{aligned} G_j \mid \alpha_o, G_o, W_j &\sim \mathcal{DP}(\alpha_o, G_o(W_j)) \\ \theta_{ji} \mid G_j &\sim G_j \\ y_i \mid X_i, C_i, \theta_{ji}, &\sim p(y_i \mid X_i, \theta_{ji}), \quad \mathbb{E}[y_i \mid \cdot] = \mu_{ji} = g^{-1}(X_i^T \beta_{ji}). \end{aligned} \tag{3}$$

We refer to the model (3) as hierarchical Dirichlet process generalized linear model (hdpGLM). It generalizes the GLM and the dpGLM and provides a hierarchical clustering structure that is context-dependent. Because it generalizes GLMs, it can be used even if there is neither heterogeneity nor multiple contexts. The advantage of using hdpGLM is that clusters can be uncovered if the researcher is uncertain about the existence of heterogeneous effects. A model selection procedure can be adopted to decide if either the results of GLM or hdpGLM is adequate for the data at hand (Mukhopadhyay and Gelfand 1997).

To complete the formulation of the model and connect (3) and (1), denote Z_{ik} the case in which individual i belongs to the subpopulation indexed by k, that is, $Z_i = k$. Let C_{ij} indicates that individual i belongs to the context (country) j, that is, $C_i = j$. We can parameterize the effect of the context-level covariates W with $\tau \in \mathbb{R}^{(D_w+1)\times(D_x+1)}$ and rewrite the model (3) using the stick-breaking construction (Sethuraman 1994; Teh *et al.* 2006). The resulting model in its full generality is the following:

$$\begin{aligned} &V_l \mid \alpha_o \sim \text{Beta}(1, \alpha_o) \\ &\pi_k = \begin{cases} V_1, & k = 1, \\ V_k \prod_{l=1}^{k-1}(1 - V_l), & k > 1, \end{cases} \\ &Z_i \mid \pi \sim \text{Cat}(\pi), && \pi \in \Delta^{\infty} \\ &\tau_d \sim p(\tau_d), && d = 1, \ldots, D_x + 1 \\ &\theta_{kj} \mid Z_{ik}, \tau, C_{ij}, W \sim p(\theta_{jk} \mid W, \tau), && j = 1, \ldots, J \\ &y_i \mid Z_{ik}, \theta_{kj}, X_i, C_{ij} \sim p(y_i \mid Z_{ik}, C_{ij}, X_i, \theta_{kj}) \ni \mathbb{E}[y_i \mid Z_{ik}, \theta_{kj}, X_i, C_{ij}] = g^{-1}(X_i^T \beta_{kj}), \\ &\qquad\qquad p(y_i \mid Z_{ik}, C_{ij}, X_i, \theta_{kj}) \text{ from exponential family.} \end{aligned} \tag{4}$$

We further assume that, for $\theta = (\beta, \sigma)$

$$\begin{aligned} \tau_d \mid \mu_\tau, \Sigma_\tau &\sim N_{D_w+1}(\mu_\tau, \Sigma_\tau), && d = 1, \ldots, D_x + 1 \\ \beta_{kj} \mid Z_{ik}, \tau, C_{ij}, W &\sim N_{D_x+1}([W_j^T \tau]^T, \Sigma_\beta), && j = 1, \ldots, J. \end{aligned}$$

So, the variable τ_d is a vector of linear coefficients of the country-level features. It determines the average effect of the individual-level features X_{id} on the outcome y_i in the cluster k.

In our toy example, $\tau_1 = (\tau_{11}, \tau_{21})$ would be the linear effect of the inequality and economic development (the context-level features) on β_{1k}, the linear effect of income (observed covariate) on support for redistribution among people with similar history of social and racial conflict, which are unobserved, indexed by k. The parameter $\tau_2 = (\tau_{12}, \tau_{22})$ would be the linear effect of inequality and economic development on β_{2k}, which is the effect of race on support for redistribution among people of type k. Therefore, we have a DPP clustering model that is context-dependent because the linear coefficients β of the outcome variable depend on the cluster probability π, (thought Z_i) and on the context-level feature W (through its linear effect τ).

3 Generalized Linear Models, Finite Mixture Models, and hdpGLM

This section shows the relationship between the hdpGLM and the classical GLMs, GLMM, FMMs, and dpGLM in terms of the structure of the average parameters of the outcome variable y_i. In that sense, the hdpGLM can be viewed as a generalization of the other models. That generalization allows us to estimate latent or unobserved heterogeneity in the population in terms of how the covariates and the outcome are linearly related when the number of heterogeneous groups is not known in advance. The section also explores some connections between hdpGLM, latent instrumental variable (LIV) and latent-index models, which are approaches that use DPP with regression models.

As before, we denote $X_i = (1, X_i') \in \mathbb{R}^{(D_x+1)\times 1}$ the observed characteristics of unit i, $Z_i \in \{0,1\}^\kappa$ the design variable indicating the group (or cluster) of i. The parameter κ represents the number of clusters. Let $\gamma_i \in \mathbb{R}^{(D_x+1)\times\kappa}$ be the cluster-specific matrix of linear coefficients such that $\gamma_{ik} \in \mathbb{R}^{(D_x+1)\times 1}$ is the k^{th} column of γ_i with linear coefficients of cluster k. The most general formulation of the GLM in which every individual and groups have their own set of linear coefficients is:

$$y_i \mid X_i, \beta_i, \gamma_i \sim p(y_i \mid X_i, \beta_i, \gamma_i) \ni \mathbb{E}[y_i \mid \cdot] = \mu_i = g^{-1}(X_i^T \beta_i + X_i^T \gamma_i Z_i). \tag{5}$$

Define $\eta_i = X_i^T \beta_i + X_i^T \gamma_i Z_i$ and for simplicity let $D_x = 1$. We can write

$$\eta_i = (\beta_{oi} + \gamma_{oi} Z_i) + (\beta_{1i} + \gamma_{1i} Z_i) X_i'$$
$$\eta_{ik} = (\beta_{oi} + \gamma_{oik}) + (\beta_{1i} + \gamma_{1ik}) X_i'.$$

Classical GLMs, GLMMs, FMMs, or hdpGLMs emerge from model (5) depending on what we know or believe about κ, Z_i, γ_i and β_i. More precisely, it depends on the structural assumptions we impose on those parameters.

Classical GLM can be interpreted in two ways. Either one assumes $\gamma_i = 0$ for all i and $\beta_i = \beta$, which gives

$$\eta_i = \beta_o + \beta_1 X_i' \tag{6}$$

or one assumes $\kappa = 1$, which gives

$$\eta_i = (\beta_o + \gamma_o) + (\beta_1 + \gamma_1) X_i = \theta_o + \theta_1 X_i'. \tag{7}$$

Clearly, (6) and (7) are structurally equivalent, and treating either θ in (7) of β in (6) as the parameter to be estimated should produce the same results.

When $\kappa > 1$, Z_i is observed, and one believes $\gamma \neq 0$, the common approach is to use fixed, random, or mixed-effects models. For fixed effects, one either assumes that each group k has its own fixed intercept term θ_{ok}, or both its own fixed intercept and slope $(\theta_{ok}, \theta_{1k})$. Classical models with random effects similarly assume that each observed group has its own intercept (and/or slope) but also that, instead of being fixed, they are drawn from a common

distribution. A gaussian distribution with zero mean is the standard choice for the random effects (Hayashi 2000; Woodridge 2002), but one can also have group-specific averages (Gelman and Hill 2007) (see Table 1). Mixed models use a combination of random and fixed effects.

When Z_i is not observed, obviously it is not possible to use classical mixed models for the group heterogeneity. When one does not observe Z_i, but κ is known or it is assumed to be finite and fixed, then a finite mixture model is usually used (Lenk and DeSarbo 2000). If we let $Z_i \sim \text{Cat}(\pi), \pi \in \Delta^\kappa$ then

$$\eta_i' = \mathbb{E}[\eta_i \mid X_i] = (\beta_o + \gamma_o \pi) + (\beta_1 + \gamma_1 \pi) X_i' = \theta_o + \theta_1 X_i' \tag{8}$$

and

$$\eta_{ik}' = \mathbb{E}[\eta_i \mid X_i] = (\beta_o + \gamma_o \pi_k) + (\beta_1 + \gamma_1 \pi_k) X_i = \theta_{ok} + \theta_{1k} X_i'$$

which implies a finite mixture distribution for y_i, that is,

$$\begin{aligned} Z_i \mid \pi &\sim \text{Cat}(\pi), \pi \in \Delta^\kappa \\ y_i \mid X_i, Z_{ik}, \theta_k &\sim p(y_i \mid \mu_{ik}). \end{aligned} \tag{9}$$

By averaging over κ we get

$$\overline{\eta}_i' = \overline{\theta}_o + \overline{\theta}_1 X_i'.$$

Again, it has the same basic structure of the classical GLM.

The dpGLM generalizes that structure by allowing κ to be undetermined. It emerges naturally from finite mixtures when there might be clusters in the populations that are latent or that were not measured and, in addition, we do not know exactly the number of clusters. By letting $\kappa \to \infty$ in the finite mixture model in (9), and by putting a prior on θ and a stick-breaking prior on π we have the dpGLM, as described in (10) (Teh *et al.* 2006; Hannah, Blei, and Powell 2011).

$$\begin{aligned} V_l \mid \alpha_o &\sim \text{Beta}(1, \alpha_o) \\ \pi_k &= \begin{cases} V_1, & k = 1, \\ V_k \prod_{l=1}^{k-1} (1 - V_l), & k > 1, \end{cases} \\ Z_i \mid \pi &\sim \text{Cat}(\pi), \quad \pi \in \Delta^\infty \\ \theta_{Zi} \mid Z_i &\sim p_\theta \\ y_i \mid X_i, Z_{ik}, \theta_k &\sim p(y_i \mid \mu_{ik}). \end{aligned} \tag{10}$$

Starting with the dpGLM, by restricting the possible number of clusters to be finite ($\kappa < \infty$), and treating π and θ as fixed we are again back to the FMM. If, in addition, we either average out the clusters and treat those averaged elements as the fixed parameters to estimate or if $\kappa = 1$, we have the classical GLM. In sum, the GLM can be viewed a special case of the dpGLM.

Finally, the hdpGLM proposed here generalizes that structure to account for the possibility of context-dependent clustering. It does that by letting the linear coefficients of the clusters be a function of context-level covariates. We modify the model (10) by adding the parameter τ and context-level information W. Given J different contexts, the context-level covariates $W \in \mathbb{R}^{J \times (D_w + 1)}$, and the variable C_i that indicates the context to which i belongs, we have the hdpGLM model by modifying the dpGLM and adding the following structure to it:

$$\begin{aligned} \tau_d &\sim p(\tau_d), && d = 1, \ldots, D_x + 1 \\ \theta_{Z_i C_i} \mid Z_{ik}, \tau, C_{ij}, W &\sim p(\theta_{jk} \mid W, \tau), && j = 1, \ldots, J. \end{aligned} \tag{11}$$

https://doi.org/10.1017/pan.2019.13 Published online by Cambridge University Press

Table 1. Relationship between GLM, GLMM, FMM and hdpGLM based on structural assumptions on κ, Z_i and φ_i.

Model*	κ **(# of groups)**	Z_i **(group indicator)**	γ_i **(group effect)**	η_i **(linear predictors)**
GLM	Known ($\kappa = 1$)	Observed ($Z_i = 1$)	$\gamma_i = \gamma = 0$	$X_i\beta_1$
FE (I)	Known ($\kappa = K \in \mathbb{N}$)	Observed ($Z_i \in \{1, \ldots, K\}$)	$\gamma_{oi} = \gamma_{ok}, \gamma_{1ki} = 0$	$\beta_o + \gamma_{ok} + X_i'\beta_1$
FE (I + S)	Idem	Idem	$\gamma_{oi} = \gamma_{ok}, \gamma_{1ki} = \gamma_{1k}$	$\beta_o + \gamma_{ok} + (\beta_1 + \gamma_{1k})X_i'$
RE (I)	Idem	Idem	$\gamma_{oi} = \gamma_{ok}, \gamma_{1ki} = 0 \ni \gamma_{ok} \sim N(\mu_{\gamma_o}, \sigma_{\gamma_o})$	$\beta_o + \gamma_{ok} + X_i'\beta_1$
RE (I + S)	Idem	Idem	$\gamma_{oi} = \gamma_{ok}, \gamma_{1ki} = \gamma_{1k} \ni \gamma_{dk} \sim N(\mu_{\gamma_d}, \sigma_{\gamma_d})$	$\beta_o + \gamma_{ok} + (\beta_1 + \gamma_{1k})X_i'$
FMM	Idem	Unobserved/latent	$\gamma_{oi} = \gamma_{ok}, \gamma_{1ki} = \gamma_{1k}$	$\beta_o + \sum_{k=1}^{K} Z_{ik}\gamma_{ok} + (\beta_1 + \sum_{k=1}^{K} Z_{ik}\gamma_{1k})X_i'$
hdpGLM	Unknown ($\kappa \in \mathbb{N} \cup \{\infty\}$)	Unobserved/latent	Idem	$\beta_o + \sum_{k=1}^{\kappa} Z_{ik}\gamma_{ok} + (\beta_1 + \sum_{k=1}^{\kappa} Z_{ik}\gamma_{1k})X_i'$

* GLM: Generalized Linear Models; FE: Fixed Effect in the intercept (FE (I)) and both in the intercept and slope (FE (I + S)); RE: Random Effect in the intercept (FE (I)) and both in the intercept and slope (RE (I + S)); FMM: finite mixture models.

Hence, if there is just one context ($J = 1$) we have the dpGLM again, and it demonstrates the connection between hdpGLM and the other models. Table 1 summarizes the connection between them.

The hdpGLM model is also structurally connected to latent instrumental variable (LIV) models (Ebbes, Böckenholt, and Wedel 2004; Ebbes *et al.* 2005; Ebbes, Wedel, and Böckenholt 2009). Such models can be used to deal with endogenous covariates. The main feature of the LIV is the introduction of a latent categorical instrumental variable, which turns the instrumental variable (IV) regression model into a FMM. To see this, consider this simple example of a classical IV model with endogenous covariate x_1, a gaussian outcome, and the instrumental variable z (index i omitted for simplicity):

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \\ x_1 = \gamma_0 + \gamma_2 x_2 + \gamma_3 z + \nu. \end{cases} \tag{12}$$

For $\phi_0 = \beta_0 + \beta_i \gamma_0, \phi_2 = \beta_1 \gamma_2 + \beta_2, \phi_3 = \beta_1 \phi_3$, and $\varepsilon' = \beta_1 \nu + \varepsilon$ we have the reduced form:

$$y = \varphi_0 + \varphi_2 x_2 + \varphi_3 z + \varepsilon'. \tag{13}$$

The LIV approach defines a latent K-level categorical random variable Z_i to be used instead of the instrument z_i. Each group k is assumed to have its is own mean value φ_{0k}. It leads to a FMM with K latent groups such that the outcome is given by:

$$y = (\varphi_0 + \varphi_{0k}) + \varphi_2 x_2 + \varepsilon' = \theta_{0k} + \theta_2 x_2 + \varepsilon' \tag{14}$$

or equivalently, and assuming group-specific errors:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_{1k} + \beta_2 x_2 + \varepsilon \\ x_{1k} &= \gamma_0 + \gamma_2 x_2 + \varphi_{0k} + \nu_i. \end{aligned} \tag{15}$$

Ebbes *et al.* (2005) and Ebbes, Wedel, and Böckenholt (2009) have proposed (15), called LIV, to deal with endogeneity in the regressors. Obviously, by (14) and (15) we can see how it is structurally connected to the hdpGLM. Some differences between LIV and the model presented here is that in the former the number of latent groups needed to be selected in advance before the parameters are estimated, which is a feature of any FMM. Moreover, LIV is designed for a single-context estimation, that is, the endogeneity and the instrument are not context-dependent. The main difference, however, is that the LIV approach uses the joint distribution of the endogenous covariates and the outcome due to its goal of dealing with endogeneity of the covariate, while here the covariates are assumed to be exogeneous. The hdpGLM leads to a LIV model if we truncate the DPP, restrict the hdpGLM to its non-hierarchical version (dpGLM) with gaussian outcome, and model the distribution of the endogenous covariates as in the equations above.

Finally, the hdpGLM also has a close connection with the latent-index model that has been designed to deal with single-context heterogeneity (Aakvik, Heckman, and Vytlacil 2005; Rossi 2014). Aakvik, Heckman, and Vytlacil (2005) propose a model in which the marginal effects are heterogeneous in the population and indexed by a continuous latent random variable. They also provide a special case with two latent groups by using a binary transformation of that (gaussian distributed) latent index, which produces a mixture model with two latent components. The model here generalizes that continuous latent-index approach in two ways. First, it imposes a much more flexible distribution on the latent index and allows us to estimate effect heterogeneity when there are unknown number of finite or countably infinite groups. Their gaussian index model can be approximated by a countably infinity partition of the real line and a symmetric

unimodal discrete distribution on that partition. It is embedded in the structure of the model and the DPP can naturally be used to estimate such a distribution of the indexes. Second, the model here generalizes their approach by adding a context-dependent structure to the latent heterogeneity.

4 Estimation

There are many options in the literature to estimate models that use DPP (Ishwaran and Zarepour 2000; Neal 2000; Blei *et al.* 2006; Walker 2007). Here we extend the approach proposed by Ishwaran and James (2001). In order to implement a (blocked) Gibbs sampler for a DPP model, one of the algorithms they propose uses a truncated version of the stick-breaking construction in conjunction with the generalized Dirichlet distribution. We extend their basic algorithm in two ways. First, we incorporate the hierarchical structure of the model proposed here and develop a Gibbs sampler in its full generality. Second, we derive the sampler for two special cases: continuous outcome variable y_i, modeled using a gaussian distribution, and a binary outcome variable, modeled using a Bernoulli distribution with a logistic transformation of the average parameter. For the gaussian outcome, the Gibbs update can be used for all parameters. Therefore, in practice for that special case the estimation shows good convergence diagnostics within thousand iterations and it can be performed in a relatively short time depending on the size of the data set. For the binary outcome, the Gibbs update is available for all parameters but the linear coefficients of the generalized linear model. So we extend the algorithm and implement a Metropolis–Hasting update within Gibbs to sample the linear coefficients (β) using Riemann manifold Hamiltonian Monte Carlo (Neal 2000; Shahbaba and Neal 2009; Neal *et al.* 2011). The R package hdpGLM contains the implementation of the model with the algorithms presented here.

The truncation of the DPP used in the MCMC algorithm restricts the mixing probability parameter $\pi \in \Delta^{\infty}$ described in (4) to $\pi \in \Delta^{K}$. To estimate the model properly, we set a large value for K and monitor the estimation to check the maximum number of clusters the sampler used to allocate the data points during the iterations. If it reached K at any point we increase its value and repeat the process. By selecting a K much larger than the number of clusters the sampler activates during the estimation, we make sure the truncation is not changing the estimated results.

As before, let $\mathbf{X} \in \mathbb{R}^{n\times(D_x+1)}$ denote the individual-level covariates including a column with ones for the intercept term, and n the number of data points including all contexts. Denote $C = (C_1, \ldots, C_n)$ and $C_i \in \{1, \ldots, J\}$ the variable that indicates the context to which i belongs, and let $\mathbf{W} \in \mathbb{R}^{J\times(D_w+1)}$ be the $(D_w + 1)$-dimensional context-level features of the contexts J. Finally, let $Z = (Z_1, \ldots, Z_n)$.

The following additional notation is used to derive the algorithm: Z^* denotes the unique values of Z, and Z^{*C} the values between 1 and K that are not in Z^*. We denote by Z_j^* the unique values of Z in the context j, and Z_j^{*C} its complement in j, I_k is the set of indexes i of the data points assigned to the cluster k, N_k the total number of data points in k, and X_{jk} (or y_{jk}) the covariates (outcome variable) of the observations i in context j and assigned to the cluster k.

Given the most general formulation of the hdpGLM in (4) and the truncation used for the sampler we have the following proposition (see proof in the appendix A):

PROPOSITION 1 (Blocked Gibbs sampler for hdpGLM). *A Blocked Gibbs sampler for the model described in (4) with $\pi \in \Delta^{K}$ is given by the Algorithm 1.*

A special case of the model described in (4) occurs when y_i is gaussian distributed. Let $N_d(\mu, \Sigma)$ denote a d-dimensional multivariate gaussian distribution. Then, for $\theta = (\beta, \sigma)$, we can have a Gibbs sampler for all parameters if we use the following distribution for τ, β and σ

Algorithm 1 Gibbs Sampler for hdpGLM

Require: $Z^{(t)} = (Z_1^{(t)}, \ldots, Z_n^{(t)}), \theta_{Z_i}^{(t)}, \tau^{(t)}, \pi^{(t)}$

1: For $d \in \{1, \ldots, D_x + 1\}$, sample $\tau_d^{(t+1)} \mid \theta^{(t)}, \mathbf{W} \sim p(\theta_d^{(t)} \mid \mathbf{W}, \tau_d^{(t)}) p(\tau_d)$

2: For $j = 1, \ldots, J$

For all $k \in Z_j^*$ sample $\theta_{kj}^{(t+1)} \mid Z^{(t)}, \theta^{(t)}, \tau^{(t+1)}, \mathbf{X}, \mathbf{W}, C,$

$y \sim p(\theta_{kj} \mid \tau^{(t+1)}, \mathbf{W}) \prod_{i \in I_k} p(y_i \mid Z_{ik}^{(t)}, C_{ij}, X_i, \theta_{kj}^{(t)})$

For all $k \in Z_j^{*C}$ sample $\theta_{kj}^{(t+1)} \mid \tau^{(t+1)}, \mathbf{W} \sim p(\theta_{kj} \mid \tau^{(t+1)}, \mathbf{W})$

3: For $i = 1, \ldots, n$, sample $Z_i^{(t+1)} \mid \theta^{(t+1)}, \pi^{(t)}, X_i, y \sim \sum_{k=1}^{K} p_{ik} \delta(Z_{ik}) \ni p_{ik}$

$\propto \pi_k^{(t)} p(y_i \mid X_i, Z_{ik}^{(t)}, C_{ij}, \theta_{kj}^{(t+1)})$

4: For $k = 1, \ldots, K-1$ sample $v_k^{(t+1)} \overset{iid}{\sim} \text{Beta}\left(1 + N_k^{(t+1)}, \alpha + \sum_{l=k+1}^{K} N_l^{(t+1)}\right) \ni N_k^{(t+1)}$

$= \sum_{i=1}^{n} I(Z_{ik}^{(t+1)})$

Set $v_K^{(t+1)} = 1$ and compute $\pi_k^{(t+1)} = \begin{cases} v_1^{(t+1)}, & k = 1 \\ v_k^{(t+1)} \prod_{l=1}^{k-1} (1 - v_l^{(t+1)}), & k = 2, \ldots, K. \end{cases}$

(see proof in the appendix A):

$$
\begin{aligned}
&\tau_d \mid \mu_\tau, \Sigma_\tau \sim N_{D_w+1}(0, \Sigma_\tau), && d = 1, \ldots, D_x + 1 \\
&\beta_{kj} \mid Z_{ik}, \tau, C_{ij}, \mathbf{W} \sim N_{D_x+1}([W_j^T \tau]^T, \sigma_\beta I), && j = 1, \ldots, J, k = 1, \ldots, K \\
&\sigma_k^2 \mid Z_{ik} \sim \text{Scale-inv-}\chi^2(\nu, s^2) \\
&\varepsilon_i \mid \sigma_k, Z_{ik} \sim N(0, \sigma_k) \\
&y_i = X_i^T \beta_{Z_i C_i} + \varepsilon_i
\end{aligned}
\tag{16}
$$

PROPOSITION 2 (Gibbs for hdpGLM with gaussian mixtures). *The Gibbs sampler for the model described in (16) is given by the Algorithm 2.*

When the outcome variable y_i in the model (4) is binomial distributed, or in general has a distribution that does not have a conjugate prior for the linear coefficients, the full conditional of the parameters θ (or β) is not standard and we cannot sample from it directly. To deal with such cases we use a Riemman manifold Hamiltonian Monte Carlo (RMHMC) update (Girolami and Calderhead 2011) within Gibbs to sample the β coefficients. We can still sample all the other parameters as before. For the sake of completeness, the RMHMC algorithm is presented in the supplementary material.

The random variable of interest is $\beta_{kj} \in \mathbb{R}^{D_x+1}$, called the position variable of the Hamiltonian Monte Carlo (HMC) algorithm (Neal *et al.* 2011), and we denote by $v \in \mathbb{R}^{D_x+1}$ the ancillary variable (momentum) such that $v \sim N_{D_x+1}(0, G(\beta_{kj}))$. The Hamiltonian for our model is defined by

$$
\begin{aligned}
H(\beta_{kj}, v) &= U(\beta_{kj}, v) + K(\beta_{kj}, v) \\
&= -\ln p(\beta_{kj} \mid \cdot) + \frac{D_x + 1}{2} \ln(2\pi) + \frac{1}{2}[\ln(\det[G(\beta_{kj})]) + v^T G(\beta_{kj})^{-1} v]
\end{aligned}
\tag{17}
$$

Algorithm 2 Gibbs Sampler for the hdpGLM with gaussian mixtures

Require: $Z^{(t)} = (Z_1^{(t)}, \ldots, Z_n^{(t)}), \theta_{Z_i}^{(t)}, \tau^{(t)}, \pi^{(t)}$

1: For all $d \in \{1, \ldots, D_x + 1\}$ sample $\tau_d^{(t+1)} \mid \beta^{(t)}, \mathbf{W} \sim N(\overline{\mu}_{\tau_{dj}}, \overline{\Sigma}_{\tau_d}) \ni$

$$\overline{\mu}_{\tau_{dj}} = \frac{1}{K}\sum_{k=1}^{K}\mu_A^{(k)}; \quad \overline{\Sigma}_{\tau_d} = \frac{1}{K}\Sigma_A; \quad S_A = (\Sigma_\tau^{-1}\sigma_\beta^2 + \mathbf{W}^T\mathbf{W})^{-1}; \quad \mu_A^{(k)} = S_A\mathbf{W}^T\beta_{dk}^{(t)}; \quad \Sigma_A = S_A\sigma_\beta^2$$

2: For $j = 1, \ldots, J$

For all $k \in Z_j^*$ sample $\beta_{kj}^{(t+1)} \mid Z^{(t)}, \sigma^{2(t)}, \tau^{(t+1)}, \mathbf{X}, \mathbf{W}, C, y \sim N_{D+1}(\overline{\mu}_\beta, \overline{\Sigma}_\beta) \ni$

$$S_\beta = (\Sigma_\beta^{-1}\sigma_k^2 + \mathbf{X}_{kj}^T\mathbf{X}_{kj})^{-1}, \quad \overline{\mu}_\beta = S_\beta\left[\Sigma_\beta^{-1}(\mathbf{W}^T\tau^{(t+1)})^T + \frac{\mathbf{X}_{kj}^T y_{kj}}{\sigma_k^{2(t)}}\right]\sigma_k^2; \quad \overline{\Sigma}_\beta = S_\beta\sigma_k^{2(t)}$$

For all $k \in Z_j^{*C}$ sample $\beta_{kj}^{(t+1)} \mid \tau^{(t+1)}, \mathbf{W} \sim N_{D+1}((\mathbf{W}^T\tau^{(t+1)})^T, \Sigma_\beta)$

3: For all $k \in Z^*$ sample $\sigma_k^{2(t+1)} \mid Z^{(t)}, \beta^{(t+1)}, \tau^{(t+1)}, \mathbf{X}, \mathbf{W}, C, y \sim$ Scale-inv-$\chi_2(\overline{\nu}, \overline{s}^2) \ni$

$$\overline{\nu} = \nu + N_k^{(t)}; \quad \overline{s}^2 = \frac{\nu s^2 + N_k^{(t)}\hat{s}^2}{\nu + N_k^{(t)}}; \quad \hat{s}^2 = \frac{1}{N_k^{(t)}}(y_k - \mathbf{X}_k\beta_k^{(t+1)})^T(y_k - \mathbf{X}_k\beta_k^{(t+1)})$$

For all $k \in Z^{*C}$ sample $\sigma_k^{2(t+1)} \mid Z_i = k \sim$ Scale-inv-$\chi^2(\nu, s^2)$

4: For $i = 1, \ldots, n$, sample $Z_i^{(t+1)} \mid \theta^{(t+1)}, \pi^{(t)}, X_i, y \sim \sum_{k=1}^{K} p_{ik}\delta(Z_i = k) \ni$

$$p_{ik} \propto \pi_k^{(t)} p(y_i \mid X_i, Z_{ik}^{(t)}, C_{ij}, \theta_{kj}^{(t+1)})$$

5: For $k = 1, \ldots, K-1$ sample $v_k^{(t+1)} \overset{iid}{\sim} \text{Beta}\left(1 + N_k^{(t+1)}, \alpha + \sum_{l=k+1}^{K} N_l^{(t+1)}\right) \ni$

$$N_k^{(t+1)} = \sum_{i=1}^{n} I(Z_{ik}^{(t+1)})$$

Set $v_K^{(t+1)} = 1$ and compute $\pi_k^{(t+1)} = \begin{cases} v_1^{(t+1)}, & k = 1 \\ v_k^{(t+1)}\prod_{l=1}^{k-1}(1 - v_l^{(t+1)}), & k = 2, \ldots, K. \end{cases}$

whose solution is

$$\begin{aligned} \nabla_v H(\beta_{kj}, v) &= G(\beta_{kj})^{-1}v \\ \nabla_{\beta_{kj}} H(\beta_{kj}, v) &= -\left[\nabla_{\beta_{kj}} U(\beta_{kj}, v) - \tfrac{1}{2}\text{tr}\{G(\beta_{kj})^{-1}\nabla_{\beta_{kj}} G(\beta_{kj})\}\right. \\ &\quad \left. + \tfrac{1}{2}(v^T G(\beta_{kj})^{-1} G(\beta_{kj})^{-1} v)\nabla_{\beta_{kj}} G(\beta_{kj})\right]. \end{aligned} \tag{18}$$

The Hamiltonian equations are solved using the generalized Stormer–Verlet leapfrog integrator (Calin and Chang 2006; Girolami and Calderhead 2011). For L leapfrog steps with size ε, and $l = 1, \ldots, L$, it is given by:

$$\begin{aligned} v^{l+\varepsilon/2} &= v^l - \frac{\varepsilon}{2}\nabla_{\beta_{kj}} H(\beta_{kj}^l, v^{l+\varepsilon/2}) \\ \beta_{kj}^{l+\varepsilon} &= \beta_{kj}^l + \frac{\varepsilon}{2}[\nabla_v H(\beta_{kj}^l, v^{l+\varepsilon/2}) + \nabla_v H(\beta_{kj}^{l+\varepsilon}, v^{l+\varepsilon/2})] \\ v^{l+\varepsilon} &= v^{l+\varepsilon/2} - \frac{\varepsilon}{2}\nabla_{\beta_{kj}} H(\beta_{kj}^{l+\varepsilon}, v^{l+\varepsilon/2}). \end{aligned} \tag{19}$$

When y_i is binomial, that is, the distribution of y_i in the model (4) is defined by

$$y_i \sim \text{Bin}(p_{kj}), \quad p_{kj} = \frac{1}{1 + e^{-X_i^T\beta_{kj}}}$$

then, given the Equation (A 4), the elements of the RMHMC for the model hdpGLM when $k \in Z_j^*$ are defined by the following equations:

$$
\begin{aligned}
U(\beta_{kj}) &= -\ln p(\beta_{kj} \mid \cdot) \\
&\propto -\left[-\frac{D_x + 1}{2} \ln 2\pi - \frac{1}{2} \ln(\det(\Sigma_\beta)) - \frac{1}{2} (\beta_{kj} - (W_j^T \tau)^T)^T \Sigma_\beta^{-1} (\beta_{kj} - (W_j^T \tau)^T) \right. \\
&\quad \left. - \sum_{i \in I_k} y_i \ln(1 + e^{-X_i^T \beta_{kj}}) - \sum_{i \in I_k} (1 - y_i) \ln(1 + e^{X_i^T \beta_{kj}}) \right] \\
\nabla_{\beta_{kj}} U(\beta_{kj}) &= -\left[-(\beta_{kj} - (W_j^T \tau)^T)^T \Sigma_\beta^{-1} + \sum_{i \in I_k} X_i y_i p(y_i = 0 \mid \cdot) - \sum_{i \in i_k} X_i (1 - y_i) p(y_i = 1 \mid \cdot) \right].
\end{aligned}
$$

In practice we use $G(\beta_{kj}) = I_{(D_x+1)\times(D_x+1)}$, which is the most widely used approach in applications (Liu 2008; Neal *et al.* 2011). It also simplifies the Equations (17), (18), and (19) substantially. Using $v \sim N_{D_x+1}(0, I)$, the integrator reduces to the standard Stormer–Verlet leapfrog integrator (Duane *et al.* 1987; Neal *et al.* 2011). We follow that approach in this paper.

5 Monte Carlo Simulation

In this section, we conduct a Monte Carlo exercise[2] to demonstrate the properties of the estimates of the model produced by the algorithms developed in Section 4. The exercise is divided into three parts. First, we reproduce a particular situation that often occurs in practice if one omits factors that condition the association between the variable of interest and the outcome. In order to do that, we compare results produced by hdpGLM with those produced by GLM when there is no latent heterogeneity in the population and when there are latent clusters. We show how Simpson's paradox can happen in the latter case and how it is uncovered by the proposed model. In the second part of the MC exercise we simulate a large variety of possible scenarios, each with different types of heterogeneity and numbers of observed covariates to show that the model has good performance in a large variety of situations. We evaluate the frequentist properties of the estimators in each case, particularly their coverage probability (Carlin and Louis 2000; Little *et al.* 2011). Lastly, we compare the predictive performance of GLM and hdpGLM for different possible number of clusters in terms of root-mean-squared error (RMSE).

We start by comparing the estimates produced by GLM and by hdpGLM with and without latent heterogeneity in the population. We generated data sets from two parameter configurations with 3 continuous covariates sampled from a gaussian distribution. In the first data set, there is no effect heterogeneity. Hence, a single GLM would be appropriate because the effect of those three covariates are homogeneous in the population. In the second data set, we let the effect of one covariate to be conditional on a latent factor such that it has opposite signs and similar magnitude for half of the population. The effect of the other two covariates is homogeneous. Data sets used in this exercise contain 2000 observations.

As a toy example, we can think that the first covariate represents income, the second age, the third the degree of racial fragmentation in the neighborhood, and the outcome the degree of support for redistributive policies. The effect of income on support for redistribution depends on a latent feature, let's say, if the individuals have experienced economic reward due to their effort and hard work, as opposed to luck or family monetary heritage. The latent heterogeneous effect of income can occur, for instance, if more income means less support for redistribution only for those that believe upward mobility can be achieved through effort and hard work.

2 See Ferrari (2018) for replication.

Table 2. Comparing estimates of GLM (estimated using MLE) and hdpGLM (estimated using MCMC) with and without latent heterogeneity in the population.

				hdpGLM with MCMC estimates		GLM with MLE estimates	
Cluster	**Covariate**	**Parameter**	**True**	**MCMC Mean**	**95% HPD**	**MLE estimate**	**95% CI**
No latent heterogeneity in the population ($K = 1$)							
1	(Intercept)	β_0	−0.15	−0.16	(−0.21, −0.12)	−0.16	(−0.21, −0.12)
1	X_1	β_1	−3.09	−3.11	(−3.15, −3.07)	−3.11	(−3.15, −3.07)
1	X_2	β_2	9.90	9.91	(9.86, 9.95)	9.91	(9.86, 9.95)
1	X_3	β_3	3.90	3.87	(3.83, 3.92)	3.87	(3.83, 3.92)
Two subpopulations ($K = 2$) **with heterogeneous effect on** X_1							
1	(Intercept)	β_0	−3.30	−3.23	(−3.30, −3.14)	−3.25	(−3.34, −3.17)
1	X_1	β_1	2.00	2.00	(1.93, 2.08)	0.16	(0.08, 0.25)
1	X_2	β_2	−5.29	−5.31	(−5.39, −5.23)	−5.33	(−5.41, −5.24)
1	X_3	β_3	2.25	2.29	(2.21, 2.36)	2.23	(2.14, 2.32)
2	(Intercept)	β_0	−3.30	−3.28	(−3.35, −3.21)	—	—
2	X_1	β_1	−1.50	−1.52	(−1.58, −1.45)	—	—
2	X_2	β_2	−5.29	−5.29	(−5.37, −5.24)	—	—
2	X_3	β_3	2.25	2.19	(2.11, 2.26)	—	—

Table 2 compares the point estimates and their confidence intervals produced by estimating a GLM using MLE, with the posterior average and the 95% HPD interval produced by estimating the hdpGLM with the MCMC proposed here. We can compare the estimates with the true value, which is displayed in the fourth column of the table. After estimating the hdpGLM, we classified the data into clusters using the estimated cluster probability of each data point. We assigned each observation to the cluster they have the highest probability to belong to. The indexes of the clusters occupied by data points are displayed in the first column of the table. We can see in the upper half of the Table 2 the estimates when the data comes from a population in which there is no heterogeneity. All data points were classified into the same single cluster by the hdpGLM. The estimates of the two models are very similar. In fact, they are identical up to two significant figures, as shown in the table. The lower half of the table presents the results of the estimation using GLM and hdpGLM for the second data set with heterogeneous effects in the first covariate X_1. We used the same procedure just described to classify the data into clusters. The hdpGLM estimated two clusters in virtually all repetitions of the procedure. The results produced by the GLM and the hdpGLM are very similar for the covariates 2 and 3 (β_3 and β_4), whose effects are homogeneous in the population. For the hdpGLM, the values of the linear effect of those covariates with homogeneous effects are indistinguishable in the two clusters estimated, as expected. However, for the heterogeneous effect β_1 (e.g., income) the GLM estimated a positive effect when in fact there are two subpopulations, one with a positive and another with a negative effect. The hdpGLM, on the other hand, estimated the marginal effect of X_1 correctly for both clusters.

Table 2 contains an example of Simpson's paradox: the aggregate effect found for X_1 when one uses GLM and ignores the clusters is quite different from the effect found when the clusters are considered. We can see it clearer in Figure 1. The lines represent the fitted values using the MLE estimate for the GLM model and the fitted values using the posterior average for the hdpGLM. In the left panel, we can compare the estimated marginal effects produced by each model. In the right panels, we see the data points after they were clustered by the hdpGLM. The right panels also display the fitted values. We would have reached incomplete conclusions using GLM in such situation: the effect is positive and significant for the MLE estimates of the GLM but, in fact, it is

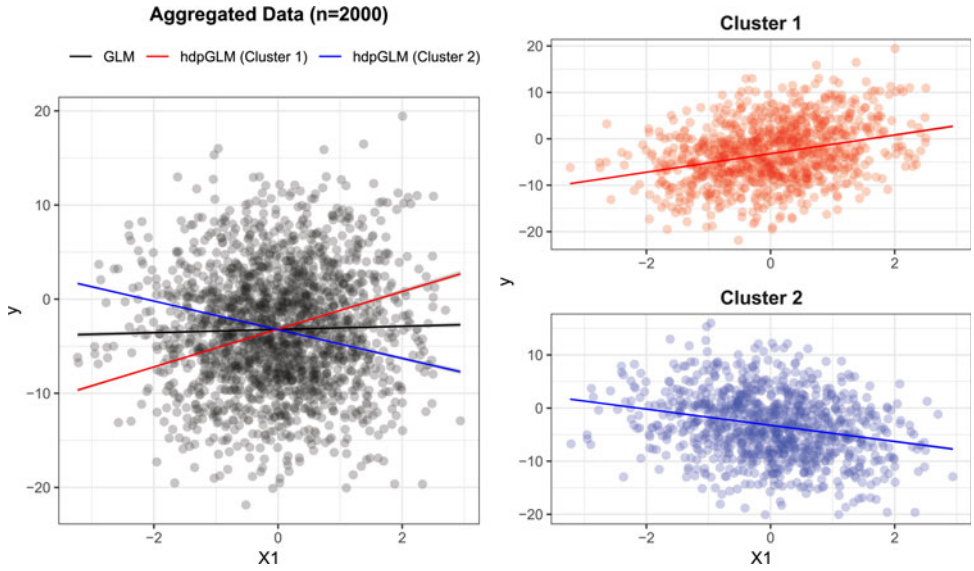

Figure 1. Comparing marginal effect estimated using GLM (MLE estimator) and hdpGLM (MCMC posterior average) when there are 3 clusters in the population.

negative for half of the population, and it has a larger positive effect than estimated by the GLM for the other half.

The general take away from these results is that when GLM is well specified and there is no effect heterogeneity due to latent features, using hdpGLM will not harm the estimation. When there are clusters with heterogeneous effects, GLM will produce accurate aggregate results but can nevertheless be incorrect for each one of the subpopulations. Table 2 and Figure 1 demonstrate that the hdpGLM reduces to GLM when there is no heterogeneity (see Section 3). When the assumptions that justify the adoption of the GLM holds, the hdpGLM can still be used and it estimates the mean value of the linear parameters quite close to the ones produced by MLE estimates of GLM. When there was heterogeneity, the hdpGLM classified the data correctly into two clusters, the marginal effects were correctly estimated, and Simpson's paradox was uncovered.

Next, in order to evaluate the performance of the hdpGLM in a wide range of possible scenarios, we randomly generated 10 different sets of parameters. To make the Monte Carlo exercise faster and easy to visualize, we simulate data for a single context ($J = 1$) with a continuous outcome variable. An example with context-dependent heterogeneity ($J > 1$) are presented in the sequel. Examples with binary outcome variables are provided in the supplementary material.

For each parameter set, we randomly generated 100 data sets. The number of clusters K and the number of covariates in each case was also randomly generated. We allowed the heterogeneity to occur in the effect of all covariates. Values of the linear coefficients range from -20 to 20. We estimated the hdpGLM for each one of the 1,000 data sets (10 parameter sets *times* 100 data sets for each parameter set) and all the usual convergence diagnostics were conducted (Geweke 1992; Cowles and Carlin 1996; Brooks and Gelman 1998; Flegal 2008; Flegal, Haran, and Jones 2008). The high posterior density (HPD) intervals were computed across data sets generated by each parameter set and so was the posterior average.

Table 3 summarizes the coverage probability of the linear coefficients β for each one of the 10 parameter sets along with the cluster estimation. The first column indicates the number of covariates in each parameter set, and the second indicates the true number of clusters in the population. The third through fifth columns display the summaries of the estimation across the 100 data sets generated by each parameter set. It shows the mean, minimum, and maximum number of clusters the data points were assigned to after the estimation. As before,

Table 3. Summary of the performance of the hdpGLM when estimating number of clusters (K) and linear coefficients (β) across 100 replications generated by 10 different parameter sets.

	Number of Clusters (K)							
		Estimates across replications				Coverage and HPD of linear coefficients (β)		
Number of Covariates	True	Mean	Minimum	Maximum	Correct (%)	Minimum	Average	95% HPD (largest average)
0	1	1.05	1	2	95	99.05	99.05	(−0.49, 0.04)
5	2	2.04	2	3	96	90.00	93.20	(−7.74, −7.1)
2	3	3.15	3	4	85	95.00	97.66	(2.02, 4.05)
3	4	4.14	4	5	86	91.00	95.44	(−3.04, −1.09)
2	4	4.12	4	5	88	93.00	96.28	(−5.97, −4.30)
3	5	5.02	5	6	98	93.00	96.27	(7.69, 8.20)
4	7	7.07	7	9	95	91.59	96.54	(−3.69, −2.69)
5	7	7.08	7	8	92	92.00	96.30	(−2.87, −1.77)
3	10	10.11	10	12	91	93.00	96.54	(0.76, 3.34)
2	10	10.25	10	12	76	92.31	96.73	(−2.8, 2.03)

we assigned the data points to the clusters based on the maximum estimated probability of cluster membership. The sixth column shows the proportion of the time the data was classified into correct number of clusters across the replications. The table also displays the minimum and the average coverage across linear coefficients for each parameter set. For instance, the second line displays a case in which there are two latent clusters in the population and five covariates. The sixth column indicates that the data points were classified into two clusters in 96% of the estimations performed using the 100 data sets generated by that parameter set. There are 10 linear coefficients across clusters for that case (5 linear coefficients per cluster). Among those 10 linear coefficients, the minimum coverage probability was 90%. It means that the linear parameter whose estimation had the worst coverage still was correctly estimated 90% of the time. By correct estimation we mean the true value was within the 95% HPDI. So in at least 90 out of 100 cases, the true values were contained in the 95% HPD interval for all the linear parameters. One may argue that such good coverage probability occurs because the posterior intervals are too wide. So, we display in the last column of the table the maximum average of the HPD intervals among the linear coefficients in each case. As the intervals are generally small we can be confident that the model and the estimation procedure proposed here have good coverage probability and such results are not due to the large variance of the posterior distribution. Another possible objection is that the number of replications is too small. In the supplementary material we provide a much larger MC exercise for two additional parameter sets with 1,000 replications each. The supplementary material also contains tables with the MC standard error for all simulations and for all linear parameter β. The results are similar to those presented here.

Now we turn to a full example of an estimation with context-dependent latent heterogeneity. For this example, we used ten contexts ($J = 10$) and two covariates ($D_x = 2$). We let the expectation of the effect β_1 of first covariate X_1 be a function of the context-level feature W_1, but the expectation of the linear effect β_2 of the second covariate X_2 is not a function of context-level features. In other words, we randomly sampled τ_{11} (the effect of W_1 on the expectation of β_1) from its prior distribution and set τ_{12} (the effect of W_1 on the expectation of β_2) to zero. We set the number of clusters to two ($K = 2$). Figure 2 shows the result of the estimation. On the left panel of the figure we see the posterior distribution of the linear coefficients in each context. The vertical lines indicate the true values. We clearly see the posterior concentrated around the true values of the clusters. On the top right of the figure, we see the estimated posterior averages for β_1 and β_2 for each cluster, in each context, as a function of the context feature W_1. We clearly see that

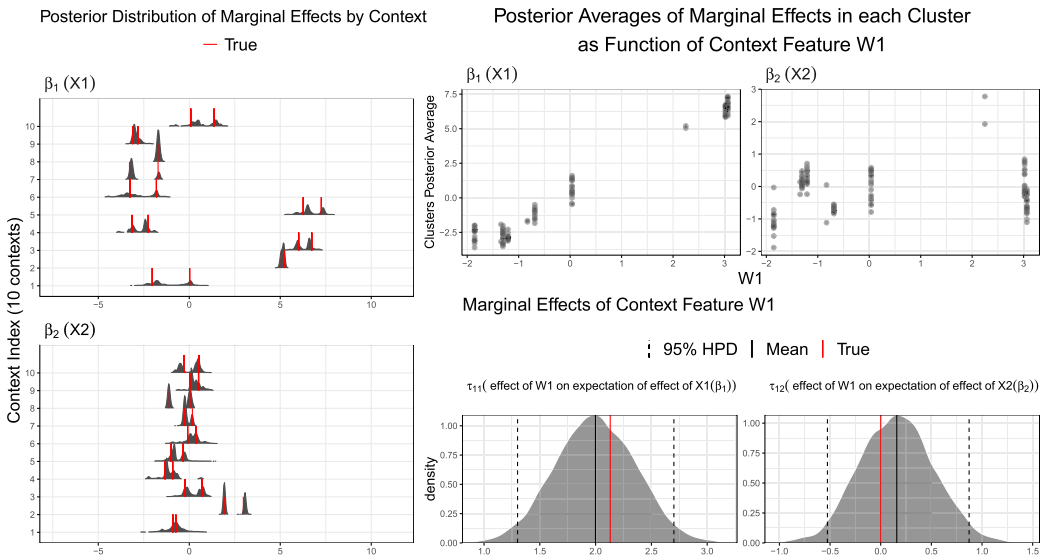

Figure 2. Output of estimation of hdpGLM model for a data set with 10 contexts and positive effect of context-level feature W_1 on the marginal effect β_1 of X_1.

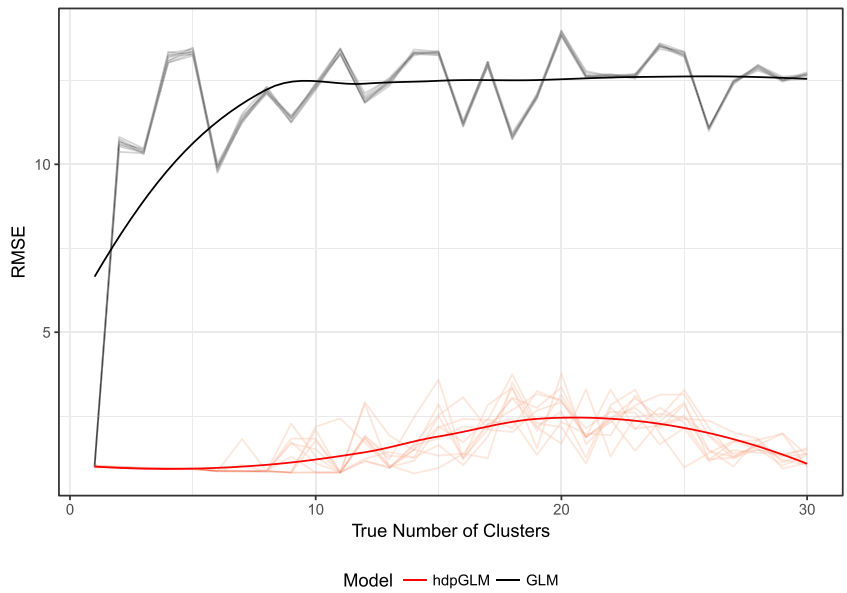

Figure 3. Comparing performance of hdpGLM and GLM using root-mean-squared error (RMSE) as a function of the number of clusters in the data.

the expectation of β_1 and the clusters are positive functions of W_1, but that is not the case for β_2. Finally, in the bottom right we see the posterior expectation of τ. The estimated values are quite close to the true values and within a small 95% HPD interval.

To complete this section, we compared the predictive performance of the GLM and the hdpGLM in terms of RMSE for different numbers of latent clusters. We randomly generated 30 parameter sets, each one with the number of clusters ranging from 1 to 30. For each case, we generated 10 data sets and estimated both the GLM and the hdpGLM. The RMSE was computed in each case. The Figure 3 compares the predictive performance of the GLM and the hdpGLM. The value of the RMSE stays always low for the hdpGLM, as expected.

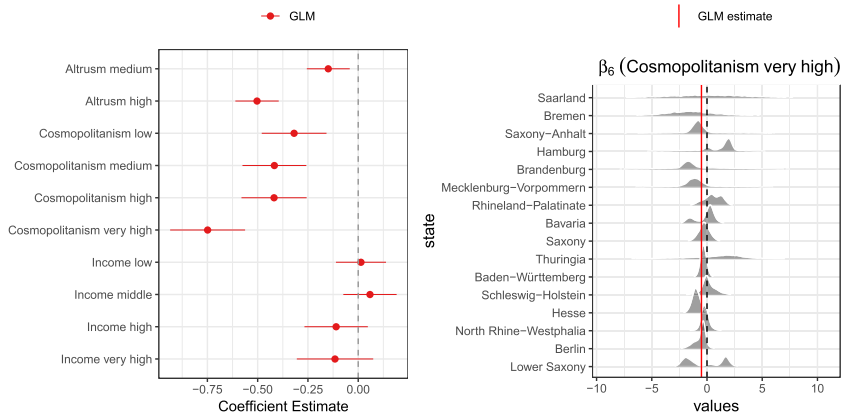

Figure 4. Left panel shows the effect of altruism, cosmopolitanism, and income on support for bailout estimated by GLM, reproducing Table 3 of Bechtel, Hainmueller, and Margalit (2014). Right panel shows latent heterogeneity in the marginal effect of very high cosmopolitanism as function of German states.

All estimations in this section and in the following use $(\mu_{\tau_d}, \sigma_{\tau_d} I, \sigma_{\beta_{kj}} I, s^2, \nu, \alpha_0) = (0, 10I, 10I, 10, 10, 1)$ as prior parametrization, where I represents the identity matrix. Those values give a reasonably large variation for the underlying random variables, and the simulation results have shown that they produce good coverage and small 95% HPD intervals in a large variety of situations. The supplementary material contains details of a prior perturbation study. Briefly, it shows that on average the model is not very sensitive to different prior settings, but in the worst case for certain combinations of prior parameters the model can demand very large data sets to escape the influence of the prior specification. This is true specially for extreme values of the concentration parameter α and values that produce highly dispersed inverse-scaled-χ^2 distribution, which can be generated by low values (below five) of the scale parameter s^2. For details, see supplementary material.

6 Empirical Application

In this section, we illustrate some applications of the model by replicating empirical studies and comparing the original results with the ones produced by the hdpGLM estimates.

We start with Bechtel, Hainmueller, and Margalit (2014), who present a study in Germany using online and telephone survey data. The paper investigates why some voters agree with bailout payments for other countries. The dependent variable is a dichotomous measure coded as 1 if the person is against bailout payments for over-indebted EU countries and 0 otherwise. They find that social dispositions, in particular feelings of cosmopolitanism and altruism, are the strongest predictors of attitudes toward providing financial help to other countries. The left panel of Figure 4 reproduces their results and displays the marginal effects of their three main variables. The right panel shows the estimation of the hdpGLM using an indicator variable for German states. The panel shows the effect of (very high) cosmopolitanism on support for bailout payments in each region. We see that for most of the states there is no latent heterogeneity. Moreover, the aggregate average effect estimated using a GLM is similar, for most cases, to the posterior average effect found by hdpGLM in each state. One exception is Lower Saxony, in which we see Simpson's paradox: there are two clusters with opposite effects of very high cosmopolitanism on support for bailout. Although we would need further investigation to provide a substantive account of these results, we can see how the hdpGLM can be used to estimate context-dependent heterogeneity.

For the second empirical application, we replicate Newman, Johnston, and Lown (2015). Using national surveys conducted in the USA, they investigate if residential proximity to inequality affect US citizens' beliefs in meritocracy, defined as the idea that the economic system rewards

Table 4. GLM vs hdpGLM estimates for counties with no latent heterogeneity.

Covariate	Parameter	GLM estimate	GLM Std Error	Average of posterior expectation across counties	Std Dev of posterior expectation across counties
(Intercept)	β_0	0.54	0.06	0.50	0.32
Income	β_1	−0.01	0.07	−0.08	0.33
educ_i	β_2	−0.10	0.02	−0.07	0.29
age_i	β_3	−0.00	0.00	0.00	0.01
gender_i	β_4	0.00	0.01	−0.00	0.19
unemp_i	β_5	0.01	0.01	0.02	0.22
union_i	β_6	0.02	0.01	0.06	0.20
partyid_i	β_7	−0.12	0.01	−0.13	0.24
ideo_i	β_8	−0.07	0.02	−0.06	0.29
attend_i	β_9	−0.03	0.01	−0.02	0.26

individuals based on their hard work and ability. The data set contains individual- and county-level covariates. They show that the association between the individual's income and the probability of rejecting meritocracy is conditional on the levels of inequality in the county: low-income individuals become more likely to reject meritocracy when inequality increases. We reproduce their results for white residents, as they present in Table 1 of their paper, using a linear probability model. In their results, income and the percentage of blacks in the county do not matter alone, but the interaction between income and inequality is significant. We focus here on that result. We estimate the hdpGLM using the same individual- and county-level covariates they included in their model. County covariates are inequality, county income, percentage of black, percentage of votes for Bush in 2004, and county population. The estimation of the hdpGLM found no latent heterogeneity in 1,633 out of 1,688 counties. Two latent clusters were estimated in 54 counties, and three latent clusters in one of them. Table 4 compares the MLE estimates of the GLM with the posterior expectation of the hdpGLM, averaged across counties with no latent heterogeneity. We can see in that table that those values are similar.

The left panel of Figure 5 shows the posterior distribution of the income effect in 20 randomly sampled counties. We see that the GLM and the hdpGLM estimates agree in many cases, but for some counties, there are latent heterogeneous groups and the estimates of the two models disagree. In the county with index 543, for instance, there are three latent groups, one in which the income plays no role, and two with opposite income effects. That case represents an example of Simpson's paradox in the effect of income in that county.

As discussed, one of the advantages of using hdpGLM is that we can evaluate if there is any effect of context(county)-levels variables after we take into account the latent heterogeneity in the effect of observed individual-level covariates. Newman, Johnston, and Lown (2015) found that inequality conditions the effect of income on the probability of rejecting meritocracy. However, when we take into account the latent heterogeneity of the income effect in each county, that conditional effect disappears. It can be seen in Figure 5. In the top-right panel of the figure, we see the posterior expectation of each cluster within each one of the 1,688 counties. In the bottom right, we see the posterior distribution of τ_{11}, the effect of inequality on the expectation of the effect of income for each county and cluster. The results indicate that inequality does not change the effect of income when we consider latent heterogeneity in the effect of covariates.

As we can see, the hdpGLM model can be used to investigate latent heterogeneity in the effect of observed covariates in generalized linear models. When there is no heterogeneity, the results of GLM and hdpGLM are similar. When there is latent heterogeneity, the GLM can produce estimates that are incorrect for all or some subpopulations. By using GLM, one is simply assuming that

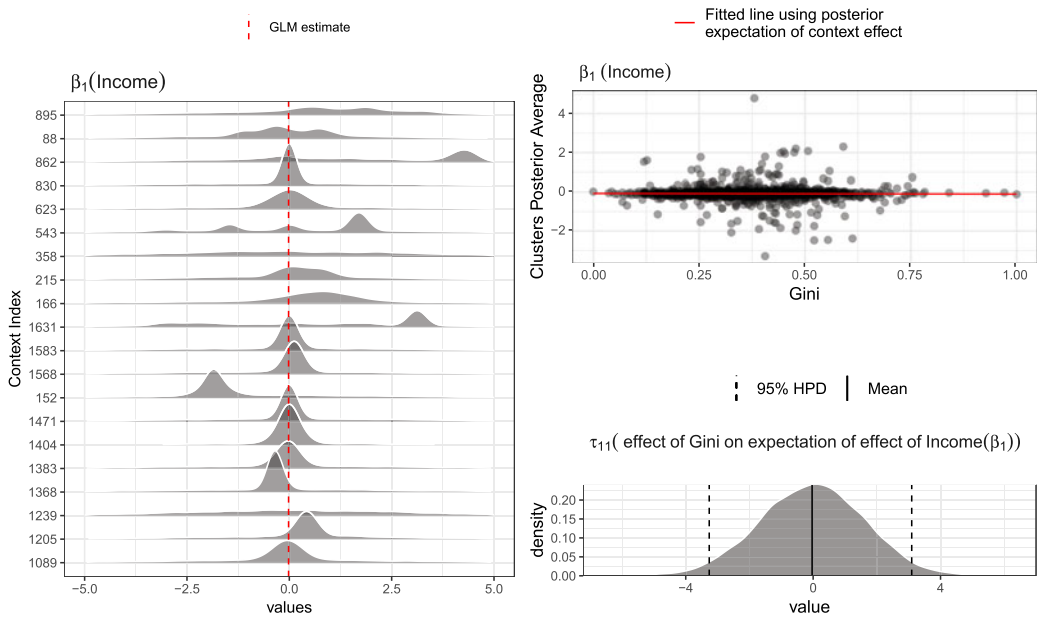

Figure 5. Posterior distribution of income effect for 20 selected counties (left panel), posterior expectation of income effect in each cluster as function of inequality (top-right panel), and posterior distribution of the effect of inequality on the income effect (bottom right).

Simpson's paradox does not occur in the analysis. The hdpGLM can be used instead to estimate the heterogeneous effect, cluster the data into groups, uncover Simpson's paradox, and evaluate if the effect of context-level features remains relevant after latent heterogeneity is considered.

7 Final Discussion

Researchers in any academic discipline can never be sure *a priori* that there is no effect heterogeneity caused by latent or omitted variables in their investigation. In other words, we are never sure if there are latent subpopulations in which the average effect found using the aggregated data is different or even reversed. When there are such subpopulations, using GLM or GLMM can produce an incomplete picture and in the worst case scenario a completely misleading conclusion. This is true in analyses using either observation or experimental data. It is desirable to use a method that is robust to latent heterogeneity. Moreover, when data comes from different contexts, for instance, different states or different countries, it is common to assume that the effect of observed covariates varies from context to context due to context-level features. Likewise, it is also desirable to consider that the latent heterogeneity of the covariate effects within each context (e.g., country) can vary from context to context (from country to country) due to context-level features.

We have provided a model to deal with those issues. The model is designed to estimate marginal effects in linear models and consider if there are latent subpopulations in which the marginal effects differ. If data comes from different contexts, the model also estimates if the existence of such subpopulations and their specific marginal effects are functions of context-level features. We have shown that the proposed model causes no harm when the GLM is correct—that is, when there are no latent heterogeneous effects—but it correctly estimates the heterogeneous effects when they exist.

Similar to GLM, however, the proposed model requires specifying which and how observed covariates are included in the model. This is not a trivial task. It can affect the estimation as much as it does for GLMs. Further research is needed to develop methods to compare and select which

observed covariates should be used and how. However, if for any reason one believes a specific set of covariates is adequate for a GLM, the proposed model can be used instead with the advantage that it will be robust to subpopulation heterogeneous effects.

Appendix A. Markov Chain Monte Carlo Algorithms

The proof of the Proposition 1 is the following.

PROOF (Blocked Gibbs sampler for hdpGLM). The full conditional of τ is given by

$$p(\tau \mid \theta, \pi, Z, y, X, W, C) \propto p(\theta \mid W, \tau)p(\tau) \propto \prod_{d=1}^{D_x+1} p(\theta_d \mid W, \tau_d)p(\tau_d).$$

For each $d = 1, \ldots, D_x + 1$ we have

$$p(\tau_d \mid \cdot) \propto p(\theta_d \mid W, \tau_d)P(\tau_d).$$

The full conditional for θ is

$$\begin{aligned} p(\theta \mid \tau, \pi, y, X, W, Z, C) &\propto p(y \mid \theta, X, W, Z, C)p(\theta \mid W, \tau) \\ &= \prod_{j=1}^{J} \prod_{i:Z_i \in Z_j^*} p(y_i \mid X_i, Z_i, C_i, \theta_{C_i,Z_i})p(\theta_{C_iZ_i} \mid \tau, W_j) \prod_{i:Z_i \in Z_j^{*C}} p(\theta_{C_iZ_i} \mid W_j, \tau) \\ &= \prod_{j=1}^{J} \left[\left(\prod_{k \in Z_j^*} p(\theta_{jk} \mid W_j, \tau) \prod_{i:Z_i \in Z_j^*} p(y_i \mid X_i, \theta_{jk}) \right) \left(\prod_{k \in Z_j^{*C}} p(\theta_{jk} \mid W_j, \tau) \right) \right]. \end{aligned}$$

Therefore, for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$, we have

$$p(\theta_{jk} \mid \cdot) \propto \begin{cases} p(\theta_{jk} \mid W_j, \tau) \prod_{i:Z_i=k} p(y_i \mid X_i, \theta_{jk}), & \text{if } k \in Z_j^*, \\ p(\theta_{jk} \mid W_j, \tau), & \text{if } k \in Z_j^{*C}. \end{cases} \tag{A 1}$$

For the variable Z, the full conditional is given by

$$\begin{aligned} p(Z \mid \tau, \theta, \pi, y, X, W, C) &\propto p(y \mid \theta, Z, X, W, C)p(Z \mid \pi) \\ &= \prod_{i=1}^{n} p(y_i \mid \theta_{C_iZ_i}, X_i, C_i, Z_i)p(Z_i \mid \pi). \end{aligned}$$

Therefore for all $i = 1, \ldots, n$ we have

$$p(Z_i = k \mid \cdot) \propto \pi_k p(y_i \mid \theta_{C_ik}, X_i, C_i)$$

or similarly

$$p(Z_i \mid \cdot) \propto \sum_{k=1}^{K} p_{ik} I(Z_i = k) \ni p_{ik} = \pi_k p(y_i \mid \theta_{C_ik}, X_i, C_i). \tag{A 2}$$

Finally, for π the full conditional is

$$p(\pi \mid \tau, \theta, Z, y, X, W, C) \propto p(Z \mid \pi)p(\pi) = \prod_{i=1}^{n} p(Z_i \mid \pi)p(\pi).$$

Now, for simplicity, let $\pi \sim \text{Dir}(\alpha/K)$. The connection between this distribution and the stick-breaking process described in (4) can be found in Ishwaran and James (2001). Then we have

$$p(\pi \mid \tau, \theta, Z, y, X, W, C) \propto \prod_{i=1}^{n}\left(\prod_{k=1}^{K}\pi_k^{I(Z_i=k)}\right)\prod_{k+1}^{K}\pi_k^{\alpha/K-1} = \prod_{k=1}^{K}\pi_k^{N_k+(\alpha/K)-1}.$$

Therefore,

$$p(\pi \mid \cdot) \propto \text{Dir}\left(N_1 + \frac{\alpha}{K}, \ldots, N_K + \frac{\alpha}{K}\right). \qquad \square \quad \text{(A 3)}$$

The proof of the Proposition 2 is the following.

Proof (Gibbs for hdpGLM with gaussian mixtures). Considering the results in Proposition 1 and the model described in (16), we have the following. For τ, for each $d = 1, \ldots, D_x + 1$

$$p(\tau_d \mid \cdot) \propto p(\beta_d \mid W, \tau_d)P(\tau_d) = \prod_{k=1}^{K}\left[\prod_{j=1}^{J} p(\beta_{dkj} \mid W, \tau_d)p(\tau_d)\right].$$

But by conjugacy of the gaussian distributions, we have $\prod_{j=1}^{J} p(\beta_{dkj} \mid W, \tau_d)p(\tau_d) \propto N_{D_w+1}(\mu_A^{(k)}, \Sigma_A)$ where

$$\begin{aligned} S_A &= (\Sigma_\tau^{-1}\sigma_\beta^2 + W^T W)^{-1} \\ \Sigma_A &= S_A\sigma_\beta^2 \\ \mu_k^{(k)} &= S_A W^T \beta_{dk}. \end{aligned}$$

Therefore

$$p(\tau_d \mid \cdot) \propto \prod_{k=1}^{K} N_{D_w+1}(\mu_A^{(k)}, \Sigma_A) \propto \exp\left\{-\frac{1}{2}\left[\tau_d^T(k\Sigma_A^{-1})\tau_d - 2\tau_d^T\Sigma_A^{-1}\left(\sum_{k=1}^{K}\mu_A^{(k)}\right)\right]\right\}.$$

If we denote $\overline{\Sigma}_{\tau_d} = \frac{1}{K}\Sigma_A$ and $\overline{\mu}_{\tau_d} = \frac{1}{K}\sum_{k=1}^{K}\mu_A^{(k)}$ then

$$\tau_d \mid \cdot \propto N_{D_w+1}(\overline{\mu}_{\tau_d}, \overline{\Sigma}_{\tau_d}).$$

The full conditional for β is

$$\begin{aligned} p(\beta \mid \tau, \sigma^2, \pi, y, X, W, Z, C) &\propto p(y \mid \beta, \sigma^2, X, W, Z, C)p(\beta \mid W, \tau) \\ &= \prod_{j=1}^{J}\prod_{i:Z_i\in Z_j^*} p(y_i \mid X_i, Z_i, C_i, \beta_{C_i,Z_i}, \sigma_{Z_i}^2)p(\beta_{C_iZ_i} \mid \tau, W_j)\prod_{i:Z_i\in Z_j^{*C}} p(\beta_{C_iZ_i} \mid W_j, \tau) \\ &= \prod_{j=1}^{J}\left[\left(\prod_{k\in Z_j^*} p(\beta_{jk} \mid W_j, \tau)\prod_{i:Z_i\in Z_j^*} p(y_i \mid X_i, \beta_{jk}, \sigma_k^2)\right)\left(\prod_{k\in Z_j^{*C}} p(\beta_{jk} \mid W_j, \tau)\right)\right]. \end{aligned}$$

Therefore, for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$, we have

$$p(\beta_{jk} \mid \cdot) \propto \begin{cases} p(\beta_{jk} \mid W_j, \tau)\prod_{i:Z_i=k} p(y_i \mid X_i, \beta_{jk}, \sigma_k^2), & \text{if } k \in Z_j^*, \\ p(\beta_{jk} \mid W_j, \tau), & \text{if } k \in Z_j^{*C}. \end{cases} \quad \text{(A 4)}$$

Denote $X_{kj} = \{X_i \mid C_i = j, Z_i = k\}$, $y_{kj} = \{y_i \mid C_i = j, Z_i = k\}$, it is clear from (A 4), (16), and the conjugacy of the normal distribution that for $k \in Z_j^*$

$$\beta_{jk} \mid \cdot \propto N_{D_x+1}(\overline{\mu}_\beta, \overline{\Sigma}_\beta) \quad \text{where } S_\beta = (\Sigma_\beta^{-1}\sigma_k^2 + X_{kj}^T X_{kj})^{-1}, \quad \overline{\Sigma}_\beta = S_\beta \sigma_k^2$$
$$\overline{\mu}_\beta = S_\beta \left[\Sigma_\beta^{-1}(W_j^T \tau)^T + \frac{X_{kj}^T y_{kj}}{\sigma_k^2}\right] \sigma_k^2.$$

The full conditional for σ^2 is

$$\begin{aligned} p(\sigma^2 \mid \tau, \beta, \pi, Z, y, X, W, C) &\propto p(y \mid \beta, \sigma^2, Z, X, W, C) p(\sigma^2) \\ &= \prod_{i=1}^{n} p(y_i \mid \beta_{C_i Z_i}, \sigma_{Z_i}^2, X_i, Z_i, C_i) p(\sigma_{Z_i}^2) \\ &= \left(\prod_{k \in Z^*} p(\sigma_k^2) \prod_{i: Z_i = k} p(y_i \mid \beta_{C_i k}, X_i, C_i)\right)\left(\prod_{k \in Z^{*C}} p(\sigma_k^2)\right). \end{aligned}$$

Therefore, for all $k = 1, \ldots, K$ we have

$$p(\sigma_k^2 \mid \cdot) \propto \begin{cases} p(\sigma_k^2) \prod_{i: Z_i = k} p(y_i \mid \beta_{C_i k}, X_i, C_i), & \text{if } k \in Z^* \\ p(\sigma_k^2), & \text{if } k \in Z^{*C}. \end{cases} \tag{A 5}$$

Given the full conditional of σ^2 in (A 5), the distributions in (16), and the fact that the scaled inverse χ^2 distribution is a conjugate prior for a gaussian likelihood with known mean, which is the case for the full conditional, it is straightforward to see that for $k \in Z^*$, $X_k = \{X_i \mid Z_i = k\}$, and $y_k = \{y_k \mid Z_i = k\}$

$$\sigma_k^2 \mid \cdot \propto \text{Scale-inv-}\chi^2(\overline{\nu}, \overline{s}^2)$$

where

$$\overline{\nu} = \nu + N_k, \quad \overline{s}^2 = \frac{\nu s^2 + N_k \hat{s}^2}{\nu + N_k}, \quad \hat{s}^2 = \frac{1}{N_k}(y_k - X_k \beta_k)^T (y_k - X_k \beta_k).$$

The full conditionals for Z and π are as in (A 2) and (A 3), respectively. □

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2019.13.

References

Aakvik, A., J. J. Heckman, and E. J. Vytlacil. 2005. "Estimating Treatment Effects for Discrete Outcomes When Responses to Treatment Vary: an Application to Norwegian Vocational Rehabilitation Programs." *Journal of Econometrics* 125(1–2):15–51.

Alesina, A., and G.-M. Angeletos. 2005. "Fairness and Redistribution." *The American Economic Review* 95(4):960–980.

Alesina, A., and P. Giuliano. 2010. "Preferences for Redistribution." In *Handbook of Social Economics*, edited by J. Benhabib, A. Bisin, and M. O. Jackson, 93–131. Amsterdam: Elsevier.

Arts, W., and J. Gelissen. 2001. "Welfare States, Solidarity and Justice Principles: Does the Type Really Matter? *Acta Sociologica* 44(4):283–299.

Bechtel, M. M., J. Hainmueller, and Y. Margalit. 2014. "Preferences for International Redistribution: The Divide Over the Eurozone Bailouts." *American Journal of Political Science* 58(4):835–856.

Beramendi, P., and P. Rehm. 2016. "Who gives, who gains? Progressivity and Preferences." *Comparative Political Studies* 49(4):529–563.
Blei, D. M, and M. I. Jordan et al. 2006. "Variational inference for Dirichlet process mixtures." *Bayesian Analysis* 1(1):121–143.
Blyth, C. R. 1972. "On Simpson's Paradox and the Sure-Thing Principle." *Journal of the American Statistical Association* 67(338):364–366.
Brooks, S. P., and A. Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7(4):434–455.
Calin, O., and D.-C. Chang. 2006. *Geometric Mechanics on Riemannian Manifolds: Applications to Partial Differential Equations*. Birkhauser: Springer Science & Business Media.
Carlin, B. P., and T. A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Boca Raton, FL: Chapman & Hall/CRC.
Carota, C., and G. Parmigiani. 2002. "Semiparametric Regression for Count Data." *Biometrika* 89(2):265–281.
Chen, X. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." *Handbook of Econometrics* 6:5549–5632.
Cowles, M. K., and B. P. Carlin. 1996. "Markov chain Monte Carlo convergence diagnostics: a comparative review." *Journal of the American Statistical Association* 91(434):883–904.
De Iorio, M., P. Müller, G. L. Rosner, and S. N. MacEachern. 2004. "An ANOVA Model for Dependent Random Measures." *Journal of the American Statistical Association* 99(465):205–215.
De la Cruz-Mesía, R., F. A. Quintana, and G. Marshall. 2008. "Model-Based Clustering for Longitudinal Data." *Computational Statistics & Data Analysis* 52(3):1441–1457.
Diaconis, P., and D. Freedman. 1986. "On the Consistency of Bayes Estimates." *Annals of Statistics* 14(1):1–26.
Dorazio, R. M., B. Mukherjee, L. Zhang, M. Ghosh, H. L. Jelks, and F. Jordan. 2008. "Modeling Unobserved Sources of Heterogeneity in Animal Abundance Using a Dirichlet Process Prior." *Biometrics* 64(2):635–644.
Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth. 1987. "Hybrid monte carlo." *Physics Letters B* 195(2):216–222.
Ebbes, P., M. Wedel, and U. Böckenholt. 2009. "Frugal IV Alternatives to Identify the Parameter for an Endogenous Regressor." *Journal of Applied Econometrics* 24(3):446–468.
Ebbes, P., M. Wedel, U. Böckenholt, and T. Steerneman. 2005. "Solving and Testing for Regressor-Error (in) Dependence When No Instrumental Variables are Available: With New Evidence for the Effect of Education on Income." *Quantitative Marketing and Economics* 3(4):365–392.
Ebbes, P., U. Böckenholt, and M. Wedel. 2004. "Regressor and random-effects dependencies in multilevel models." *Statistica Neerlandica* 58(2):161–178.
Ferrari, D. 2018. "Replication Data for: Modeling Context-Dependent Latent Effect Heterogeneity." https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WB9XLZ.
Flegal, J. M., M. Haran, and G. L. Jones. 2008. "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?" *Statistical Science* 23(2):250–260.
Flegal, J. M. 2008. "Monte Carlo Standard Errors for Markov Chain Monte Carlo." PhD thesis, University of Minnesota.
Gaffney, S. 2003. "Curve Clustering with Random Effects Regression Mixtures." In *Ninth International Workshop on Artificial Intelligence and Statistics, AISTATS*. Key West, Florida.
Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
Geweke, J. 1992. "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." In *Bayesian Statistics*, 4th edn. 169–193. Oxford: Oxford University Press.
Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi. 1999. "Consistent Semiparametric Bayesian Inference about a Location Parameter." *Journal of Statistical Planning and Inference* 77(2):181–193.
Gill, J., and G. Casella. 2009. "Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation." *Journal of the American Statistical Association* 104(486):453–454.
Girolami, M., and B. Calderhead. 2011. "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(2):123–214.
Grimmer, J. 2009. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1):1–35.
Hannah, L. A., D. M. Blei, and W. B. Powell. 2011. "Dirichlet Process Mixtures of Generalized Linear Models." *Journal of Machine Learning Research* 12(Jun):1923–1953.
Hayashi, F. 2000. *Econometrics,* vol. 1. Princeton, NJ: Princeton University Press.
Heckman, J. J., and E. J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments." *Handbook of Econometrics* 6:4875–5143.
Heinzl, F., and G. Tutz. 2013. "Clustering in Linear Mixed Models with Approximate Dirichlet Process Mixtures Using EM Algorithm." *Statistical Modelling* 13(1):41–67.

Hernán, M. A., D. Clayton, and N. Keiding. 2011. "The Simpson's Paradox Unraveled." *International Journal of Epidemiology* 40(3):780–785.
Ichimura, H., and P. E. Todd. 2007. "Implementing Nonparametric and Semiparametric Estimators." *Handbook of Econometrics* 6:5369–5468.
Ishwaran, H., and L. F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association* 96(453):161–173.
Ishwaran, H., and M. Zarepour. 2000. "Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models." *Biometrika* 87(2):371–390.
Johnston, R., K. Banting, W. Kymlicka, and S. Soroka. 2010. "National Identity and Support for the Welfare State." *Canadian Journal of Political Science* 43(02):349–377.
Kievit, R., W. E. Frankenhuis, L. Waldorp, and D. Borsboom. 2013. "Simpson's Paradox in Psychological Science: A Practical Guide." *Frontiers in Psychology* 4(513):1–14.
Kleinman, K. P., and J. G. Ibrahim. 1998a. "A Semi-Parametric Bayesian Approach to Generalized Linear Mixed Models." *Statistics in Medicine* 17(22):2579–2596.
Kleinman, K. P., and J. G. Ibrahim. 1998b. "A Semiparametric Bayesian Approach to the Random Effects Model." *Biometrics* 54(3):921–938.
Kyung, M., J. Gill, and G. Casella et al. 2010. "Estimation in Dirichlet Random Effects Models." *The Annals of Statistics* 38(2):979–1009.
Lenk, P. J., and W. S. DeSarbo. 2000. "Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects." *Psychometrika* 65(1):93–119.
Little, R. et al. 2011. "Calibrated Bayes, for Statistics in General, and Missing Data in Particular." *Statistical Science* 26(2):162–174.
Liu, J. S. 2008. *Monte Carlo Strategies in Scientific Computing*. New York: Springer Science & Business Media.
Mallick, B. K., and S. G. Walker. 1997. "Combining Information from Several Experiments with Nonparametric Priors." *Biometrika* 84(3):697–706.
Matzkin, R. L. 2007. "Nonparametric Identification." *Handbook of Econometrics* 6:5307–5368.
Mukhopadhyay, S., and A. E. Gelfand. 1997. "Dirichlet Process Mixed Generalized Linear Models." *Journal of the American Statistical Association* 92(438):633–639.
Müller, P., and R. Mitra. 2013. "Bayesian Nonparametric Inference-Why and How." *Bayesian Analysis* 8(2):269–302.
Müller, P., F. Quintana, and G. Rosner. 2004. "A Method for Combining Inference Across Related Nonparametric Bayesian Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(3):735–749.
Neal, R. M. 2000. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics* 9(2):249–265.
Neal, R. M. et al. 2011. *MCMC Using Hamiltonian Dynamics*, vol. 2. New York, NY: CRC Press.
Newman, B. J., C. D. Johnston, and P. L. Lown. 2015. "False Consciousness or Class Awareness? Local Income Inequality, Personal Economic Position, and Belief in American Meritocracy." *American Journal of Political Science* 59(2):326–340.
Ng, S.-K., G. J. McLachlan, K. Wang, L. Ben-Tovim Jones, and S.-W. Ng. 2006. "A Mixture Model with Random-Effects Components for Clustering Correlated Gene-Expression Profiles." *Bioinformatics* 22(14):1745–1752.
Pearl, J. 2011. "Simpson's Paradox: An Anatomy." Technical Report UCLA: Department of Statistics Los Angeles, California. https://escholarship.org/uc/item/3s62r0d6.
Pearl, J. 2014. "Comment: Understanding Simpson's Paradox." *The American Statistician* 68(1):8–13.
Pearson, K., A. Lee, and L. Bramley-Moore. 1899. "Mathematical Contributions to the Theory of Evolution. VI. Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses." *Philosophical Transactions of the Royal Society of London Series A* 192:257–330.
Przeworski, A. 2007. "Is the Science of Comparative Politics Possible? In *The Oxford Handbook of Comparative Politics*, edited by C. Boix Boix and S. C. Stokes, Oxford Handbooks Online.
Rehm, P. 2009. "Risks and Redistribution an Individual-Level Analysis." *Comparative Political Studies* 42(7):855–881.
Rossi, P. 2014. *Bayesian Non- and Semi-Parametric Methods and Applications*. Princeton, NJ: Princeton University Press.
Rossi, P. E., G. M. Allenby, and R. McCulloch. 2006. *Bayesian Statistics and Marketing*. Chichester: John Wiley & Sons.
Rueda, D., and D. Stegmueller. 2016. "The Externalities of Inequality: Fear of Crime and Preferences for Redistribution in Western Europe." *American Journal of Political Science* 60(2):472–489.
Samuels, M. L. 1993. "Simpson's Paradox and Related Phenomena." *Journal of the American Statistical Association* 88(421):81–88.
Sethuraman, J. 1994. "A Constructive Definition of Dirichlet Priors." *Statistica Sinica* 4:639–650.

Shahbaba, B., and N. Radford. 2009. “Nonlinear Models Using Dirichlet Process Mixtures.” *Journal of Machine Learning Research* 10(Aug):1829–1850.
Shayo, M. 2009. “A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution.” *American Political Science Review* 103(02):147–174.
Simpson, E. H. 1951. “The Interpretation of Interaction in Contingency Tables.” *Journal of the Royal Statistical Society. Series B (Methodological)* 13(2):238–241.
Spirling, A., and K. Quinn. 2010. “Identifying Intraparty Voting Blocs in the UK House of Commons.” *Journal of the American Statistical Association* 105(490):447–457.
Stegmueller, D. 2013. “Modeling Dynamic Preferences: A Bayesian Robust Dynamic Latent Ordered Probit Model.” *Political Analysis* 21(3):314–333.
Stokes, S. C. 2014. “A Defense of Observational Research.” In *Field Experiments and their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, edited by D. L. Teele, 33–57. New Haven, CT: Yale University Press.
Svallfors, S. 1997. “Worlds of Welfare and Attitudes to Redistribution: A Comparison of Eight Western Nations.” *European Sociological Review* 13(3):283–304.
Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. “Hierarchical Dirichlet Processes.” *Journal of the American Statistical Association* 101:1566–1581.
Tokdar, S. T. 2006. “Posterior Consistency of Dirichlet Location-Scale Mixture of Normals in Density Estimation and Regression.” *Sankhyā: The Indian Journal of Statistics* 68(1):90–110.
Traunmuller, R., A. Murr, and J. Gill. 2015. “Modeling Latent Information in Voting Data with Dirichlet Process Priors.” *Political Analysis* 23(1):1, http://dx.doi.org/10.1093/pan/mpu018.
Verbeke, G., and E. Lesaffre. 1997. “The Effect of Misspecifying the Random-Effects Distribution in Linear Mixed Models for Longitudinal Data.” *Computational Statistics & Data Analysis* 23(4):541–556.
Villarroel, L., G. Marshall, and A. E. Barón. 2009. “Cluster Analysis Using Multivariate Mixed Effects Models.” *Statistics in Medicine* 28(20):2552–2565.
Walker, S. G. 2007. “Sampling the Dirichlet Mixture Model with Slices.” *Communications in Statistics - Simulation and Computation* 36(1):45–54.
Woodridge, J. M. 2002. *Econometric Analysis of Cross-Sectional and Panel Data*. Cambridge and London: MIT Press.
Yule, G. U. 1903. “Notes on the Theory of Association of Attributes in Statistics.” *Biometrika* 2(2):121–134.