Advance Access publication November 27, 2014

Political Analysis (2015) 23:1–20
doi:10.1093/pan/mpu018

Modeling Latent Information in Voting Data with Dirichlet Process Priors[†]

Richard Traunmüller
Department of Social Sciences, Goethe University Frankfurt, Frankfurt am Main, D-60323, Germany
e-mail: traunmueller@soz.uni-frankfurt.de

Andreas Murr
Department of Politics and International Relations, University of Oxford, Oxford, OX1 3UQ, United Kingdom
e-mail: andreas.murr@politics.ox.ac.uk

Jeff Gill
Departments of Political Science, Biostatistics, and Surgery, Washington University, St. Louis, MO 63130-4899, United States
e-mail: jgill@wustl.edu (corresponding author)

Edited by Jonathan Katz

We apply a specialized Bayesian method that helps us deal with the methodological challenge of unobserved heterogeneity among immigrant voters. Our approach is based on *generalized linear mixed Dirichlet models* (GLMDMs) where random effects are specified semiparametrically using a Dirichlet process mixture prior that has been shown to account for unobserved grouping in the data. Such models are drawn from Bayesian nonparametrics to help overcome objections handling latent effects with strongly informed prior distributions. Using 2009 German voting data of immigrants, we show that for difficult problems of missing key covariates and unexplained heterogeneity this approach provides (1) overall improved model fit, (2) smaller standard errors on average, and (3) less bias from omitted variables. As a result, the GLMDM changed our substantive understanding of the factors affecting immigrants' turnout and vote choice. Once we account for unobserved heterogeneity among immigrant voters, whether a voter belongs to the first immigrant generation or not is much less important than the extant literature suggests. When looking at vote choice, we also found that an immigrant's degree of structural integration does not affect the vote in favor of the CDU/CSU, a party that is traditionally associated with restrictive immigration policy.

1 Introduction

The analysis of social science data is often difficult for reasons that tend to affect other fields less substantially. Such problems include: high levels of measurement error, governments that falsify or hide information, collection in difficult or even violent areas, embargoed information for privacy reasons, well-known survey response issues, overlapping explanatory power in model variables, the fluidity of political and social institutions, as well as the willingness of humans to conceal information from researchers. This has led to many important modeling innovations as a way to compensate for these issues. One problem in particular is difficult to handle with traditional statistical models: unobserved grouping effects that correlate strongly with phenomena of

Authors' note: The authors would like to thank the participants at this event, two anonymous referees, and the editors for helpful comments and remarks. Full replication materials for this study are available on the *Political Analysis* Web site at http://dx.doi.org/10.7910/DVN/27564.

[†]A previous version of this article was presented at the 3rd Annual General Conference of the European Political Science Association 2013 in Barcelona.

© The Author 2014. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

interest. In a survey context, perhaps there are subnational ethnic groups, gender differences that differ by culture, heterogeneous family structures, or important questions eliminated for privacy concerns. Such unobserved phenomena may lead to different ways that the set of observed variables interact with each other in a statistical model.

In this article, we apply a specialized Bayesian method that helps us deal with the methodological challenge of unobserved heterogeneity. Our approach is based on *generalized linear mixed Dirichlet models* (GLMDMs) where random effects are specified semiparametrically using a Dirichlet process mixture prior that has been shown to account for unobserved grouping in the data (Dorazio et al. 2007; Gill and Casella 2009). This approach does not provide estimates of the unknown clustering, which is a very difficult problem; instead, it accounts for it in the context of the Bayesian stochastic simulation process so that this clustering does not harm the estimation of regular regression quantities. Such models are drawn from Bayesian nonparametrics, a general term that includes Dirichlet process models, to help overcome objections handling latent effects with strongly informed prior distributions. We show that for difficult problems of missing key covariates and unexplained heterogeneity this approach gives: (1) overall improved model fit, (2) smaller standard errors on average, and (3) less bias from omitted variables.

We encountered the latent heterogeneity problem in electoral studies when evaluating both the turnout and the votes of immigrants who have recently become citizens and may now vote. Immigration is one of the most prominent structural features of Western democracies, and this growing social group is likely to affect national political life in important ways for many nations (Bird, Saalfeld, and Wüst 2011). However, as has been noted by many scholars, "social scientists attempting to address immigrant political behavior face the challenge of limited data" (Dancygier and Saunders 2006, 963). The most obvious concern is that minorities typically make up only a small proportion of the population being studied and can be hard to reach for data-collection purposes (but see Heath et al. 2013). Therefore, immigrant sample sizes in standard surveys available to electoral researchers tend to be small and the resulting estimates are less precise than desired.

Importantly for our purposes, immigrants are a heterogeneous group in most countries, and this heterogeneity, expressed in cultural traits and social preferences related to socialization in foreign countries or conditions of integration into the host society, is not often directly measured. So even if we had survey data with a *large proportion* of immigrants, we would want a tool that accounts for such heterogeneity. Failing to account for this obvious diversity means that some systematic component of the data falls to an error term, exacerbating efforts to find parsimonious models with good fit and potentially leading to biased results. Lacking direct covariate information, we seek here to find help in the data itself by specifying a nonparametric prior that reflects information in the data to help account for underlying structure in the context of the model. Thus, this unmeasured heterogeneity of immigrants is actually the motivation for our methodological approach.

The remainder of this article is structured as follows. In the following section, we first briefly review the basic ideas of Bayesian inference. We then give the essential background on Bayesian nonparametric models designed to account for unobserved heterogeneity (Section 3). Section 4 provides a more technical description of Dirichlet process prior (DPP) models. We then illustrate the workings and benefits of these models using data on the voting of immigrants in the 2009 German federal election (Section 5). In the final section, we conclude with a brief discussion.

2 Background on Bayesian Political Science

Key features of Bayesian inference are well documented in political science (Western and Jackman 1994; Bartels 1997; Gill 2007) but warrant a brief repeating. Because all unknown quantities are described probabilistically in Bayesian models, the full mechanism of inference is also probabilistic based on updating with observed data: the joint posterior distribution for the coefficients of interest is obtained by multiplying the prior by the likelihood function. So the specification of prior distributions is required. This process can also be iterative where current posteriors become priors in future work as new data arrive, and inferences are therefore refreshed in a "learning" process that can continue indefinitely. The latter point is an appealing feature of the Bayesian approach since most political phenomena change over time.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Of course the real product of interest is the set of marginal posterior distributions, which can be summarized in convenient forms for readers. When this marginalization of the joint posterior is difficult, researchers employ an increasingly powerful set of computational tools, including Markov chain Monte Carlo (MCMC) algorithms, generally referred to as Bayesian stochastic simulation. MCMC techniques solved a lingering problem in Bayesian analysis by producing empirical draws for the parameters of interest from complicated joint posteriors.

Most Bayesian work to date in the social sciences avoids putting substantial information into prior specifications with so-called "uninformed" forms (Rubin and Schenker 1987; Western 1998; Quinn, Martin, and Whitford 1999; Smith 1999; Hill and Kriesi 2001; Schweinberger and Snijders 2003). Thus, one finds evangelical preaching of inference through Bayes's law using prior information and the likelihood in political science, $\pi(\theta|\mathbf{x}) \propto p(\theta)L(\theta|\mathbf{x})$, but forms of the prior are specifically designed to minimize prior effects such as a diffuse uniform, $p(\theta) = 1$, $-\infty < \theta < \infty$, a Gaussian normal with huge variance, $p(\theta) \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 \gg 0$, and Jeffreys' prior, $p(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right]$. Early critics of the Bayesian paradigm (Pearson 1920a, 1920b; Fisher 1922; Neyman 1937) focused on the almost exclusive use of uniform (flat) priors at the time as a method for expressing prior ignorance or uncertainty. Their concern was the influential effect that this prior has with small samples (since large enough samples produce standard likelihood analysis results), and the fact that uniform shape does not represent a genuine lack of information about a parameter.

Bayesian work in political science is mostly centered on applying the associated simulation tools (i.e., MCMC) to solve difficult computational problems. This approach is not very Bayesian in the philosophical sense and they could be termed "Bayesians of convenience" since part of the Bayesian model is defaulted to diffuse (often uniform) forms (Martin and Quinn 2002). A few authors have argued for priors that are actually informed by current scientific knowledge in the social sciences (Leamer 1978; Berk, Bruce, and Weiss 1995; Western 1996), but most authors are simply more comfortable with less informed prior information. The reasons for this are two fold: a lack of attention to available information, and a reluctance to sell priors given their historical controversy.

We argue here for nonparametric priors, which exist in the middle of the spectrum because they retain some researcher intuition but allow the data to drive the analysis. Note that the prior distribution in this context is actually a *stochastic process*, rather than some typically applied parametric distribution (e.g., "normal" or "uniform"), and has thus been labeled as *nonparametric* in the relevant literature. The use of *nonparametric* here is pervasive but unfortunate, and a better term would be *semiparametric* since other parts of such models include specified distributions. Here, we only apply the nonparametric priors to random effects components as a way to capture *semi-informed* prior information that improves the analysis of difficult social science data containing latent effects by using information in such data that is not directly expressed in the likelihood function. This is done by allowing the data to indirectly inform a logical grouping in the data. Standard regression parameters are given conventional priors since DPPs can bias these terms (Kyung, Gill, and Casella 2010), so it is not reasonable to be *fully* nonparametric in the regression setting.

3 Dirichlet Process Prior Modeling Background

This section gives the essential background on the type of Bayesian nonparametric models designed to account for unobserved heterogeneity in a variety of circumstances. We then proceed to a detailed description of this family of model specifications.

3.1 *Types and Terms*

Generalized linear models (GLMs) have enjoyed considerable attention over many years, providing a flexible framework for modeling discrete responses using a variety of error structures. If we have observations that are discrete or categorical, $\mathbf{Y} = (Y_1, \ldots, Y_n)$, such data can often be assumed to be independent and from a distribution in the exponential family, where the likelihood function has

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

components of the model such as the form of the link function $g()$ and the type of error structures that result. This links the expected value of the outcome variable with a linear additive component that includes an estimated regression vector β and a matrix of explanatory variables $\mathbf{X}$: $E[\mathbf{Y}] = g^{-1}(\mathbf{X}\beta)$. The classic book by McCullagh and Nelder (1989) describes these models in detail; see also the more recent developments in Dey, Ghosh, and Mallick (2000) or Fahrmeir and Tutz (2001).

Generalized linear mixed models (GLMM) are an extension of a GLM that allows for random effects, and can give us flexibility in developing a more suitable model when the observations are correlated, or where there may be other underlying phenomena that contribute to the resulting variability. Note that the researcher has to pick one or more factors for the random effects. The standard GLMM begins with the assumption that $Y_1, \ldots, Y_n$ are realizations of n independent and identically distributed (i.i.d.) random variables within m-component mixtures, giving the density $f(y|\beta_\ell, \omega_\ell) = \sum_{\ell=1}^{m} \omega_\ell f(\cdot|\beta_\ell)$, where $m < n$ is a fixed positive integer set in advance, $0 \leq \omega_\ell \leq 1$ and $\sum_{\ell=1}^{m} \omega_\ell = 1$. Details of such models, covering both statistical inferences and computational methods, can be found in the texts by McCulloch and Searle (2001) and Jiang (2007).

GLMDMs are variations of a GLMM, where the random effects are modeled with a DPP, resulting in a GLMDM. We give the technical description below in Section 4. For now it is enough to note that for the GLMDM, the researcher does not have to specify a fixed value (m) for the number of unique random effects modeled in the mixture, as is necessary with the GLMM. Although this is a newly developed model, there are revealing implementations. Dorazio et al. (2007) used a GLMDM with a log link for spatial heterogeneity in animal abundance. They proposed an empirical Bayesian approach with the Dirichlet process, instead of the regular assumption of a normally distributed random effect, because they argued that for some species, the source of heterogeneity in abundance is poorly understood or unobservable. In a political science setting, Gill and Casella (2009) used a GLMDM with an ordered probit link to estimate levels of stress in presidential political appointees as a means of understanding their surprisingly short tenures. In order to obtain open and honest responses, the collectors of these data embargoed key information such as agency employer that would have helped researchers but identified these government executives. The random effects modeled with a Dirichlet process mixture prior produced an enhanced understanding of the agency environment that was not directly available by conventional means. Spirling and Quinn (2010) similarly modeled latent voting blocks within a party in the UK House of Commons; Kyung, Gill, and Casella (2012) used DPPs to produce new findings in terrorism data; and Stegmueller (2013) was able to capture individual preference heterogeneity in hierarchical latent dynamic panel models with this approach.

3.2 *Nonparametric Priors and Latent Clustering*

In the non-Bayesian context, nonparametric data analysis has proven useful in finding important and useful features in observations that researchers are unwilling to assign an assumed standard parametric data-generating process. There is a large literature spanning many fields on the use of smoothing, outlier analysis, and more structurally, generalized additive models. These ideas are relatively new to Bayesian modeling, having been introduced first by Ferguson (1973), but not fully developed until the advent of better computing resources for estimation after 1990.

The key idea is the use of DPPs (described in detail below), which stipulate that the data are produced by a mixture distribution wherein the Bayesian prior specifications are produced by a Dirichlet process, which constitutes a distribution of distributions since each produced parameter defines a particular distribution. Subsequent realizations of the Dirichlet process are discrete (with probability 1), even given support over the full real line, and are thus treated like *countably infinite mixtures*. Note that this approach is substantially different from the use of the Dirichlet distribution in conventional Bayesian models as a conjugate prior distribution for the multinomial likelihood.

What can nonparametric priors add to the emerging Bayesian paradigm in the social sciences? Consider the question of modeling dichotomous individual choices, Y_i, like turning out to vote, voting for a specific candidate, joining a social group, discontinuing education, and so on. The most

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

common "regression-style" modeling approach is to assume that an underlying smooth utility curve dictates such preferences, providing the unobserved but estimated threshold $\theta \in [0, 1]$. The individual's threshold along this curve then determines the zero or one outcome conditional on an additive right-hand specification, $\mathbf{X}\beta$. Realistically, we should treat this threshold differently for each individual, but we can apply the reasonable Bayesian approach of assuming that these are different thresholds, but still generated from a single distribution G, such that $E[nG(\theta_i|\mathbf{X}\beta, \alpha)]$ is the expected number of affirmative outcomes. Suppose there were structures in the data such as unexplained clustering effects, unit heterogeneity, autocorrelation, or missingness that cast doubt on the notion of G as a single model.[1] The choice of G is unknown by the researcher but determined by custom or intuition. We suggest, instead, a semiparametric Bayesian approach that draws θ from a mixture of appropriate prior distributions conditional on data and parameters where the mixture information is derived from latent clustering in the data. In this way G can be informed by the data, relying less on customary practices.

Unknown clustering also has an effect on variable selection. Most literatures in the social sciences have a collection of explanatory phenomena that need to be included because the theories supporting them are very strong. In many cases, the resulting decision is simply which measured version of the phenomenon should be used as a right-hand-side variable. Leamer (1978) called these "inside the horizon" variables since their value is so well established. In the above case of a vote choice model, these are: partisanship, ideology, race, age, and education. The game, according to Leamer, is specifying an additional set of "over the horizon" variables that may provide new knowledge if supported by the data and the model. Often the first type of variables is included in the final specification even if they are not found to be statistically reliable because the norms of the discipline are strong motivators. It is not widely recognized that the effect of these variables can be confounded by latent clusters. That is, for some individual cases in the data this variable is a strong determinant of the outcome variable, but its effect is sufficiently heterogeneous across clusters that it does not appear statistically reliable in the model and may be excluded in the final specification. Thus, accounting for latent clusters, as proposed here, affects both variable selection and variable importance in estimated clusters.

Finally, it is important to realize that we are relying on *proven* statistical theory in this GLMDM application. Kyung, Gill, and Casella (2010) derived the critical profile likelihood based on forms given by Liu (1996) and Lo (1984). Burr and Doss (2005) showed that a random effects specification of the Dirichlet process can account for latent heterogeneity. Escobar and West (1995), Neal (2000), and MacEachern and Müller (1998) produced the necessary computational approaches to estimate these specifications. We now explain this technology in detail.

4 Technical Description of Dirichlet Process Prior Models

Dirichlet process mixtures were introduced by Ferguson (1973) and Antoniak (1974), and Blackwell and MacQueen (1973) showed that the marginal distribution of the Dirichlet process is equal to the distribution of the nth step of a Pólya Urn Process (explained below). Korwar and Hollander (1973) characterized the joint distribution and looked at nonparametric empirical Bayes estimation of the distribution function based on DPPs. Sethuraman (1994) later showed that the Dirichlet measure is a distribution on the space of all probability measures, giving probability one to the subset of discrete probability measures, and Doss (1994) showed how to sample this distribution exactly. Lo (1984) derived the analytic form of a Bayesian density estimation that is generated by convoluting a known density kernel with a Dirichlet process, and Liu (1996) derived an identity for the profile likelihood estimator of precision parameter λ. However, Kyung, Gill, and Casella (2010) looked at the properties of the maximum-likelihood estimate (MLE) of the precision parameter and found that the likelihood function can be ill behaved. They noted that incorporating a gamma prior, and using posterior mode estimation, results in a more stable solution.

[1] Note that this can happen in a Bayesian or non-Bayesian setting, the difference being the distributional or deterministic interpretation of θ.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Models with DPPs are treated as hierarchical models in a Bayesian framework, and the implementation of these models through Bayesian computation and efficient algorithms has had much attention. Escobar and West (1995) produced a Gibbs sampling algorithm for the estimation of marginal posterior distributions for all model parameters plus the direct evaluation of predictive distributions, and also discussed inference about the precision parameter using a gamma prior. MacEachern and Müller (1998) presented a Gibbs sampler with non-conjugate priors by using auxiliary parameters, and Neal (2000) provided an extended and more efficient Gibbs sampler to handle general Dirichlet process mixture models with non-conjugate priors by using a set of auxiliary parameters. Teh et al. (2006) also extended the auxiliary variable method of Escobar and West (1995) for posterior sampling of the precision parameter with a gamma prior. Kyung, Gill, and Casella (2012) developed algorithms for estimation of the precision parameter and new MCMC algorithms for a linear mixed Dirichlet random effects model. Also, they showed how to extend the results to a GLMDM with a probit link function. They used a new parameterization of the hierarchical model to derive a Gibbs sampler for the model parameters and the subclusters of the Dirichlet process that more fully exploits the structure of the model and mixes very well. They adapted the results of Hobert and Marchev (2008) to establish that the proposed sampler is an improvement, in terms of operator norm and efficiency, over other commonly used algorithms.

4.1 *Generalized Linear Mixed Dirichlet Model Specification*

Define $\mathbf{X}$ as an $n \times p$ matrix containing p covariates for each of n cases, β as an estimated p-length coefficient vector, and ψ_i as a non-unique random effect ($\psi_i \in \psi_1, \ldots, \psi_k, k < n$) accounting for subject-specific deviation from the underlying model. Assume that the set of $\mathbf{Y}|\psi$ are conditionally independent, each y_i with a density from the exponential family form. Then the GLMDM can be expressed as follows:

$$Y_i|\psi_i \sim f(y_i|\psi_i), i = 1, \ldots, n, \tag{1}$$

where $E[Y_i|\psi] = \mu_i$. Using a link function $g(\cdot)$, we can express the transformed mean of Y_i, $\mathrm{E}[Y_i|\psi]$, as a linear function of the data

$$g(\mu_i) = \mathbf{X}_i\beta + \psi_i, \tag{2}$$

where

$$\psi_i \sim G. \tag{3}$$

The key part is the assumed form of G to describe the variation in ψ_i, which in generalized linear mixed models is often taken to be a normal distribution with zero mean and estimated variance $\hat{\sigma}^2$, that is, $G = N(0, \sigma^2)$. The typical restriction of G to a single normal distribution is sometimes thought to be limiting, as it may not be very realistic in a given application. For instance, what if there were distinct groups in the population that would suggest a multimodal distribution? Unfortunately, the normal assumption also cannot be formally checked as it refers to random effects and not error terms (Burr and Doss 2005).

With Dirichlet process mixture models, we relax this assumption of a single Gaussian form and instead model G semiparametrically as a mixture distribution of weighted point masses with

$$G \sim \mathcal{DP}(\lambda G_0), \tag{4}$$

where $\mathcal{DP}$ is the Dirichlet process with base measure G_0, which serves as a prior for unknown G, and precision parameter λ. Informally speaking, the $\mathcal{DP}$ is really just a distribution over distributions—that is, each draw from a $\mathcal{DP}$ is itself a discrete distribution with probability 1. It thus allows us to express our uncertainty about the form of G. Intuitively, G_0 can be thought of as the mean of the DPP and thus our best prior guess about G and λ as the degree of certainty: with higher values of λ, G is closer to G_0. Figure 1 illustrates this with three simulated draws from a $\mathcal{DP}$ with $G_0 = N(0, 1)$ and three different precision parameters λ.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

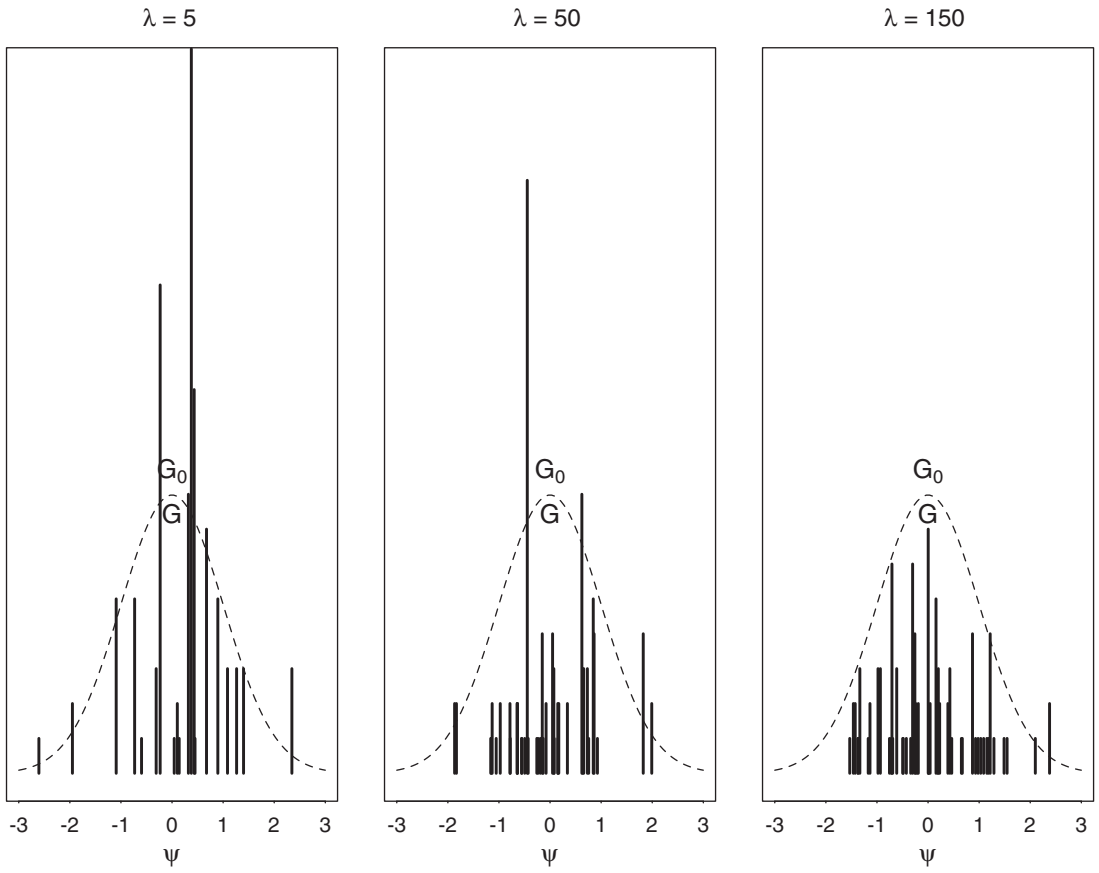


Fig. 1 Simulated draws from a Dirichlet process with base measure $G_0 = N(0, 1)$ and three different precision parameters λ.

By moving to this model, we not only relax the single-form normal assumption, but also provide a richer model for the random effects. Such a model has the potential for capturing more types of variability in those effects with the possible end result of more precise estimates of the fixed effects β regression terms. Note that whereas the random effect ψ_i is indexed by i, it retains the usual hierarchical interpretation because individuals will share ψ values by being grouped together. In other words, the ψ_i are not necessarily unique. Now the data are hierarchical with respect to this latent group, which gives the structure of the DPP.

This clustering property of the Dirichlet process can be illustrated by referring to the result of Blackwell and MacQueen (1973), who proved that for $\psi_1, \ldots, \psi_n$ i.i.d. from $G \sim \mathcal{DP}$, the joint distribution of ψ is a product of successive conditional distributions of the form

$$\psi_i|\psi_1, \ldots, \psi_{i-1}, \lambda \sim \frac{\lambda}{i-1+\lambda} g_0(\psi_i) + \frac{1}{i-1+\lambda} \sum_{l=1}^{i-1} \delta(\psi_l = \psi_i), \tag{5}$$

where $\delta()$ denotes the Dirac delta function and $g_0()$ is the density function of the base measure. The Pólya urn process for sampling ψ is equivalent to the following permutation scheme, sometimes called the restaurant algorithm:

- a restaurant has many large circular tables,
- n diners enter one at a time to be seated, where the first person sits at the first table,
- for a given weight, λ, the ith person sits at the unoccupied ith table with probability $\lambda/(i-1+\lambda)$,
- otherwise this diner selects the jth $(j < i)$ *previously occupied* table with probability $n_j/(i-1+\lambda)$, where n_j is the number seated at that table already.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Now the table locations of the seated diners, $\psi_1, \ldots, \psi_n$, is a *dependent exchangeable sequence*, where $\mathcal{N}(n)$ is the number of non-empty tables. Surprisingly this process is not particularly demanding of the model or the data in the sense that weakly defined or absent subclusters mean that the model results will look like the corresponding GLM (a good reason for the comparison in Section 5). Small numbers of cases in each group, including just one case, simply indicate that the data do not have important latent heterogeneity of this type. Furthermore, large-n data sets in political science with thousands of cases can be handled with the GLMDM, where the cost is simply more time for the MCMC process.

This representation of the Dirichlet process illustrates that the ψ_i come from a mixture between the base distribution and the distribution of previously realized random effect parameters. With a probability proportional to λ, ψ_i will either be newly drawn from the base distribution G_0 or assigned a value equal to one of the previously realized groups, revealing the clustering property of the Dirichlet process. Greater values of λ and thus greater precision of the prior make it more likely that ψ_i is drawn from G_0. Recall that the standard GLMM assumes that the random effects ψ_i are i.i.d. with a mixture of normal distributions, $N(0, \sigma^2_\psi)$. However, the Dirichlet process clustering assigns different normal parameters across groups, and the same parameters within groups and cases are only i.i.d. if they are assigned to the same cluster.

Another way to understand the grouping achieved by the DPP is to first define C to be a partition of the sample of size n into k groups, $k = 1, \ldots, n$. We call these groups "subclusters," as done by Gill and Casella (2009), since the grouping is done nonparametrically rather than on substantive criteria. That is, it is likely that any real underlying clusters would be broken up into multiple subclusters by the nonparametric fit since there is no penalty for over-separation. Each partition C can now be associated with an $n \times k$ matrix $\mathbf{A}$ defined by

$$\mathbf{A} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \tag{6}$$

where a_i is a $1 \times k$ vector of all zeros except for a 1 in one position that indicates which group the observation is from. In other words, each column of matrix $\mathbf{A}$ now represents a subcluster or group, and column sums are $(n_1, n_2, \ldots, n_k)$, the number of observations in each of the k groups. If partition C has the groups $\{S_1, \ldots, S_k\}$, then if $i \in S_j$, $\psi_i = \eta_j$ and the random effect can be rewritten as

$$\psi = \mathbf{A}\eta, \tag{7}$$

where $\eta = (\eta_1, \ldots, \eta_k)$ and $\eta_j \sim G_0$ for $j = 1, \ldots, k$. Now all observations in the same cluster or group share a common random effect parameter η_j. For example, if $S_1 = \{1, 4, 6\}$, $S_2 = \{2, 3\}$, and $S_3 = \{5\}$, then

$$\begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}. \tag{8}$$

Note that both the number of columns and the placement of 1s is random, representing the subclustering process. See Kyung, Gill, and Casella (2009, 2012) for further details of the Dirichlet process mixture models with the indicator matrix $\mathbf{A}$ and Kyung, Gill, and Casella (2011) for an application of this approach.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Importantly for our purposes, Kyung, Gill, and Casella (2012) also observed that in every instance the Dirichlet model produces shorter intervals on the coefficient estimates, as seen in the examples below. For example, when Kyung, Gill, and Casella (2010) analyzed data from the British General Election Study, Scottish Election Survey, 1997, they observed that the widths of 90% credible intervals from the Dirichlet random effects model were all shorter than those from a standard Bayesian linear model. Such results were so persistent that it seemed unlikely to be a coincidence, and Kyung, Gill, and Casella (2009) were able to prove that for all $\mathbf{y}$ not containing a within-subcluster contrast, the mean of the posterior distribution of the variance from the Dirichlet random effects model is smaller than that from the normal random effects model, and therefore the standard Bayesian linear model as well. Note that this theorem *does not* guarantee that all the credible intervals from the Dirichlet random effects model will always be smaller. Instead, the theorem says that the posterior means of the variance parameters are smaller, which then implies that on average the distribution of the fixed effects will be more concentrated. This will typically result in shorter credible intervals for the fixed effects (the β parameters in the regression).

4.2 *Generalized Linear Mixed Dirichlet Model Estimation*

The likelihood function, which by definition is integrated over the random effects, is given definitionally by

$$L(\theta|\mathbf{y}) = \int f(\mathbf{y}\,|\theta, \psi)\pi(\psi)\,\mathrm{d}\psi. \tag{9}$$

This model is actually a classical semiparametric random effects model, and with further Bayesian modeling of the parameters, lends itself to a Gibbs sampler. Unfortunately, the presence of the Dirichlet term makes the use of the Gibbs sampler somewhat complicated in non-conjugate situations, which is the algorithm developed in Gill and Casella (2009). In addition, the estimation of the precision parameter λ needs more attention. Gill and Casella also found that the fit of the model was relatively insensitive to the value of the precision parameter, λ, but only with one particular data set looking at stress in presidential political appointees. Kyung, Gill, and Casella (2010) looked at the maximum-likelihood estimation of λ, and found that the standard approach to finding the MLE, given in Liu (1996), may have problems. Likelihood estimation is not reliable, and they proved that by introducing a prior distribution on λ, one could guarantee an interior mode, stabilizing the estimation procedure.

The central problem comes from the awkward form of the likelihood function, as produced by Lo (1984) and Liu (1996):

$$L(\theta \mid \mathbf{y}) = \frac{\Gamma(\lambda)}{\Gamma(\lambda + n)} \sum_{k=1}^{n} \lambda^k \sum_{C:|C|=k} \prod_{j=1}^{k} \Gamma(n_j) \int f(\mathbf{y}_{(j)}\,|\theta, \psi_j) g_0(\psi_j)\,\mathrm{d}\psi_j, \tag{10}$$

where C defines the subclusters, $\mathbf{y}_{(j)}$ is the vector of y_is that are in subcluster j, ψ_j is the common random effects parameter for that subcluster, and vector θ contains all model parameters. Note that there are $\mathcal{S}_{n,k} = \left\{ {n \atop k} \right\} = \frac{1}{k!}\sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n$ different partitions C, the Stirling number of the second kind (which increases rapidly with increasing n).

To remedy the likelihood function problem, we again turn to the matrix representation of the Dirichlet process (Kyung, Gill, and Casella 2009) and define

$$f(\mathbf{y} \mid \theta, \mathbf{A}) = \int f(\mathbf{y} \mid \theta, \mathbf{A}\eta) g_0(\eta) \mathrm{d}\eta. \tag{11}$$

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

The likelihood function then simplifies to

$$L(\theta \mid \mathbf{y}) = \frac{\Gamma(\lambda)}{\Gamma(\lambda + n)} \sum_{k=1}^{n} \lambda^k \sum_{A \in \mathcal{A}_k} \prod_{j=1}^{k} \Gamma(n_j) f(\mathbf{y} \mid \theta, \mathbf{A}), \tag{12}$$

where $\mathcal{A}_k$ is the set of all k matrices $\mathbf{A}$. Note that the random effects have been marginalized from the likelihood and replaced with the $\mathbf{A}$ matrices, which group the observations into the subclusters. The MCMC estimation process is now straightforward, as the Dirichlet process is aided by the binary $\mathbf{A}$ matrix.

The model parameters are given the following prior distributions:

$$\begin{aligned} \beta|\sigma^2 &\sim N(\mathbf{0}, d^* \sigma^2 \mathcal{I}) \\ \tau^2 &\sim \text{Inverted Gamma}(a, b), \end{aligned} \tag{13}$$

where $d^* > 1$ and the hyperparameter values, (a, b), are fixed in order to make the inverse gamma diffuse ($a = 1$, b very small). We assume a normal distribution with mean 0 and variance τ^2, $N(0, \tau^2)$ for the base measure of the Dirichlet process. With a flat prior on $\mathbf{A}$ and $\pi(\theta)$, the joint posterior is given by

$$\pi(\theta, \mathbf{A}|\mathbf{y}) = \frac{\lambda^k f(\mathbf{y}|\theta, \mathbf{A}) \pi(\theta)}{\int_\Theta \sum_A \lambda^k f(\mathbf{y} \mid \theta, \mathbf{A})\, \pi(\theta)\, d\theta}. \tag{14}$$

The choice of a normal distribution as the base measure here is an arbitrary choice for computational convenience. In general, this model is not particularly sensitive to this choice and a Student's t or some other choice may be reasonable as well, even some skewed form if there was a strong theoretical reason. In addition, replacing the normal with some other form is an easy modification of the sampler. However, the normal with mean zero works well since it matches up with the traditional distributional choice for random effects, although we are still more general than this restriction since $\mathcal{DP}$ is a distribution over such distributions.

An overview of the general sampling scheme is as follows. We identify three groups of parameters: (a) λ, the precision parameter of the Dirichlet process, (b) $\mathbf{A}$, the indicator matrix of the partition defining the subclusters, and (c) $\theta = (\eta, \beta, \tau^2)$, the standard model parameters. In the Gibbs sampler, we iterate between these three groups until convergence:

1. Conditional on λ and $\mathbf{A}$, generate (η, β, τ^2);
2. Conditional on (η, β, τ^2) and λ, generate $\mathbf{A}$, a new partition matrix;
3. Conditional on (η, β, τ^2) and $\mathbf{A}$, generate λ, the new precision parameter.

Kyung, Gill, and Casella (2012) provide a new Gibbs sampler implementing these steps that was proven to be in excess of 50% more efficient than previous algorithms. So we make use of their software here for the MCMC process (the `glmdm` package in `R`). As with any reasonably complicated MCMC implementation, the `glmdm` function can take up to several hours to converge to the stationary distribution and then produce a useful number of samples. In our example in Section 5, the number of cases is $N = 197$, which takes much less than this time to produce 5000 draws.

5 Immigrant Voting Data Analysis

In the previous section, we provided the history of Dirichlet process models, explained their technical underpinnings, and noted that they are more efficient than alternatives in the presence of unobserved heterogeneity. Now we turn our attention to an application that demonstrates that such approaches can account for latent information that is not available to alternatives. By "account for," we mean that the resulting model fits better by incorporating the latent heterogeneity into the estimation process. This does not provide actual *substantive clusters*, which is a different and more difficult problem (Womack, Gill, and Casella 2014).

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

5.1 *Challenges in Modeling Immigrant Voting Behavior*

In this section, we provide a real-data example that highlights the workings of generalized linear Dirichlet random effects models. We are interested in the voting behavior of immigrants in Germany, a topic of great concern to politicians and political scientists alike. The data come from the post-election cross-section of the 2009 *German Longitudinal Election Study* (GLES),[2] which includes face-to-face interviews of $N = 2095$ citizens eligible to vote. Based on the official definition of the National Statistics Office, we define immigrants as citizens who either migrated to Germany after 1949 (i.e., first-generation immigrants, `firstgeneration`) or who were born in Germany with at least one parent who migrated to Germany (second-generation immigrants). Among the respondents, $N = 197$ or 9.4% have a migration background, which roughly corresponds to official statistics. This illustrates the typical case that general surveys may only include a limited number of observations of interest and thus lead to imprecise coefficient estimates. However, in the present example we are less worried about the low N for immigrants than we are with the challenge of capturing the heterogeneity among them (see below).

We model both turnout and vote choice among these immigrants. The turnout variable scores as 1 if the respondent stated that they had voted in the 2009 federal election and as 0 if they did not vote. We also code the vote choice variable as binary, where one indicates a vote for the CDU/CSU—Germany's conservative party, which is generally associated with restrictive policies toward immigrants and minorities—and zero all other options. Our model of turnout replicates Wüst (2011), who using the same data includes as predictors the political and economic attitudes of respondents together with their socio-demographic characteristics. In particular, we include respondents' democratic satisfaction (`satsif`, 1 = *not at all satisfied* to 5 = *very satisfied*), their agreement with the statement that voting is a civic duty (`civicduty`, 1 = *strongly disagree* to 5 = *strongly agree*), their interest in politics (`interest`, 1 = *not at all interested* to 5 = *very interested*), their strength of partisanship (`strengthpid`, 0 = *no PID* to 5 = *strong*), their absolute distance in evaluations of CDU and SPD (`absdieval`, 0 = *none* to 10 = *maximum*), and their subjective class membership (`class`, 1 = *lower class* to 6 = *upper class*). Besides these political and economic attitudes, we also consider respondent's age (`age`), whether the respondent is female (`female`), currently employed (`employ`), member of a trade union or other club (`member`), and has at least intermediary secondary school qualification (`educ`).

For the vote choice model, which also follows Wüst (2011), we keep the basic socio-demographic variables of the turnout model and add variables capturing the employment type and religiosity as well as political ideology and policy preferences of the respondent. We include information about the type of previous or current employment by adding two dummy variables, one for workers (`worker`) and the other for the self-employed (`selfempl`). The religiosity variable indicates how frequently the respondent attends a religious service (`attendance`, 0 = *never* to 7 = *more than once a week*). Besides these socio-demographics, the vote choice model includes a respondent's position on several 11-point scales, indicating their stance on the left–right scale (`leftright`, 1 = *left* to 11 = *right*) as well as on three political issues. The issue scales cover taxation and spending (`spending`, 1 = *lower taxes, even at the cost of less government spending on health, education, and social benefits* to 11 = *more government spending on health, education, and social benefits, even at the cost of higher taxes*), immigration (`immigration`, 1 = *laws on immigration should be relaxed* to 11 = *laws on immigration should be tougher*), and nuclear energy (`nuclear`, 1 = *more nuclear power stations should be built* to 11 = *all nuclear power stations should be closed down today*). In contrast to Wüst (2011), we include party identification with the CDU/CSU as a predictor of vote choice (`partisanship`, 0 = *no*, 1 *yes*). Otherwise one could argue that our method picks up heterogeneity that we could have accounted for in the first place. All (quasi-)continuous variables were standardized by subtracting their mean and dividing by two standard deviations, making them roughly comparable to the dummy variables (Gelman and Hill 2007).

[2]The data can be downloaded from the original source at http://www.gesis.org/wahlen/gles/.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

The main challenge in modeling immigrants' voting behavior is that they are a highly heterogeneous social group in Germany, and that we lack variables in the data set to account for this heterogeneity. This heterogeneity among immigrants is very likely to affect their political behavior. Political participation and party preferences may depend on the political socialization in the country of origin and the exact conditions of immigration and integration in the host society. In Germany, there are three main types of immigrants. The first type, *Aussiedler*, are resettlers, that is, ethnic Germans from Eastern Europe. The second type are *Gastarbeiter*, guest workers that were recruited as cheap labor forces from Southern Europe and Turkey and expected to stay for a fixed amount of time. The final type are *asylum seekers and displaced people* from civil war countries such as the former Yugoslavia.

Similar to Wüst (2011), we code people as *Aussiedler* (`aussiedler`) if at least one of the three family members (i.e., the respondent and her two parents) was born in the former Soviet Union and its successor states, as well as other Central and Eastern European states, and German territories in Eastern Europe. Classifying *Gastarbeiter* is even more challenging. During the 1950s and 1960s, West Germany signed bilateral recruitment agreements with Italy, Greece, Turkey, Morocco, Portugal, Tunisia, and Yugoslavia. Again, if at least one of the three family members was born in one of these countries, we code this person as a guest worker (`guestworker`). With the data at hand, however, we are likely to misclassify some of the respondents. Although the GLES asked about the specific country of origin for each member of the family, the provided data set often groups countries together: for instance, instead of "Tunisia," we would only know that the person was born in "Africa" and thus miss crucial information.[3] Although we also know whether respondents speak another language at home (`other language`, $0 = no$, $1 = yes$), we may still group together immigrants who came as guest workers with other groups. It is simply not possible to classify displaced persons or other immigrant groups with the data at hand, a problem that frequently occurs in social science settings.

There remain many further cultural or social characteristics that are likely to impact immigrants' voting behavior but were not measured and/or included in the data set. For instance, while much debate about immigrants' political attitudes and behaviors in Germany focuses on the role of Islam, the religious denomination variable lacks a response category for Muslims. Although it provides an `other` category, there is no information of what other religion respondents may be. According to the codebook, the survey asked a question about mosque attendance, but the responses were combined with the responses to questions about church and synagogue attendance. The reason for this combination is privacy concerns: "[b]ecause of data protection the variables 'Church attendance,' 'Synagogue attendance,' and 'Mosque attendance' were condensed under 'Church Synagogue/Mosque attendance.'" It is impossible to get the raw data. This, again, illustrates a typical challenge in analyzing social science data that we are addressing with this work.

In summary, we *know* that there are latent groups in our immigrant sample that should, to a great extent, affect how the explanatory variables relate to our outcomes of interest, turnout, and vote choice. We also know that there is information in the data beyond the likelihood function that can inform these unseen measures. Our objective is to use a GLMDM to categorize respondents by estimating subclusters: data-determined temporary "bins" on each iteration of the Gibbs sampler that serve to make the model fit better, increase efficiency, and un-bias coefficient estimates.

5.2 *Comparing Generalized Linear Mixed Dirichlet Model and Generalized Linear Model*

To estimate the GLMDM probit models for immigrants' turnout and vote choice, we use the `R` package `glmdm` that implements the MCMC algorithm by Kyung, Gill, and Casella (2012). For each model, one chain was run for 5000 iterations. Inspection of graphical diagnostics as well as formal convergence tests implemented in the `R` package `superdiag` show no signs of

[3]The only single-country response category we have is Turkey. The others are "Iberian Peninsula," "Italy/Malta/San Marino," "Other Central/East Europe," "Balkan States (not EU)," and "Africa." The codebook states that the GLES provides the country coding scheme online, but it could not be found.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

non-convergence (see Gill [2008] for a discussion of MCMC diagnostics). We then compare the results to the results from a standard GLM probit estimated with the `glm` function in `R`. This is also essentially a Bayesian probit model with highly diffuse (flat) priors (Gill 2007), only with slightly smaller coefficient standard errors for finite samples. The comparison with GLM is more appropriate here than with GLMM since the latter requires a priori determination of the number of mixture elements (m) by the researcher and neither of the other two have this restriction.

What do we gain from estimating GLMDMs compared with regular GLMs? Before turning to the interpretation of the substantive results, we first assess model fit. One simple way to compare the model fit between the two models is to assess their predictive performance. Since over-fitting could be a concern here, we compare the accuracy in out-of-sample predictions in a cross-validation exercise (Geisser 1975; Efron 1983). Cross-validation randomly divides the data into two sets, a "training" and a "test" set. Using the training set, we estimate the coefficients using both procedures, and then compare their predictive power using the test set. More specifically, we performed a 4-fold cross-validation. For each outcome variable, we randomly assigned respondents to one of four equally sized groups. Three of those groups form the training set to estimate the coefficients of both models, whereas the remaining group represents the test set for the out-of-sample predictions. We repeat this procedure four times, such that each of the groups serves in turn as the test set while the other three make up the training set. Because the results of this process depend on how the data were divided in the first place, the 4-fold cross-validation was repeated 10 times.[4]

Figure 2 shows the respective ROC curves (Dorfman and Alf 1969) that plot the "true positive rate," which is predicted turnout when the respondent actually voted divided by the sum of correctly predicted turnout and falsely predicted non-turnout, versus the proportion of "false positives," which is the predicted turnout when the respondent actually did not vote divided by all actual non-voters for different thresholds of the probit models. Curves that are further toward the upper left corner indicate superior model fit (give consistently higher rates of true positives than false positives). Apparently, both turnout and vote choice among immigrants are predicted satisfactorily by the chosen model specification. However, comparing the predictive performance of GLMDM and GLM, we find that the ROC curves for the GLMDM clearly dominate the curves for the GLM. In the turnout models, the area under the curve is 0.88 for the GLMDM and 0.83 for the GLM. In the vote choice models, it is 0.95 and 0.91, respectively. This indicates a somewhat better model fit for the GLMDM that accounts for unobserved groupings in the immigrant population.

Next, Fig. 3 shows a comparison of the standard errors for the fixed effects coefficients obtained by the GLMDM and GLM using the full sample, respectively. These would fall on the diagonal line if they were the same under both model specifications. Clearly, they are concentrated in the lower right of the diagonal and thus—in accordance with the theorem of Kyung, Gill, and Casella (2009)—the standard errors of the GLMDM are *smaller* than in the GLM. Interestingly, this is the case for virtually *all* of the coefficients and for both the turnout and the vote choice models. This result suggests that the GLMDM indeed accounts for additional latent variation in the data and thus increases the efficiency of the fixed effects estimates.

Looking at the substantive results of the *turnout* model presented in Table 1 and Fig. 4, it is obvious that both socio-demographic characteristics as well as political attitudes of immigrants matter for vote turnout. Perhaps not surprisingly, higher political interest and stronger identification with a particular party is related to higher turnout among immigrants. Moreover, older immigrants and those with employment—a crucial indicator of structural integration—tend to have a higher probability to have voted in the 2009 federal election. With regard to the characteristics more directly related to immigration, members of the first generation are less likely to have turned

[4]According to Ward, Greenhill, and Bakke (2010, 8, footnote 5), "[t]here is no evidence that the choice of [the number of folds] makes much difference in most applications, as long as the subsample in each fold is large enough to calculate the usual statistics." In the present case, the turnout data with 158 respondents have the smallest fold size of 39 respondents, and the vote choice data with 129 respondents have the smallest fold size of 32 respondents. We should also note that the actual number of iterations is lower than the theoretical value of 40 because GLM faced computational problems. In some iterations, GLM showed perfected separation, failed to converge, or both. We excluded these iterations as computational failures from the cross-validation exercise.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

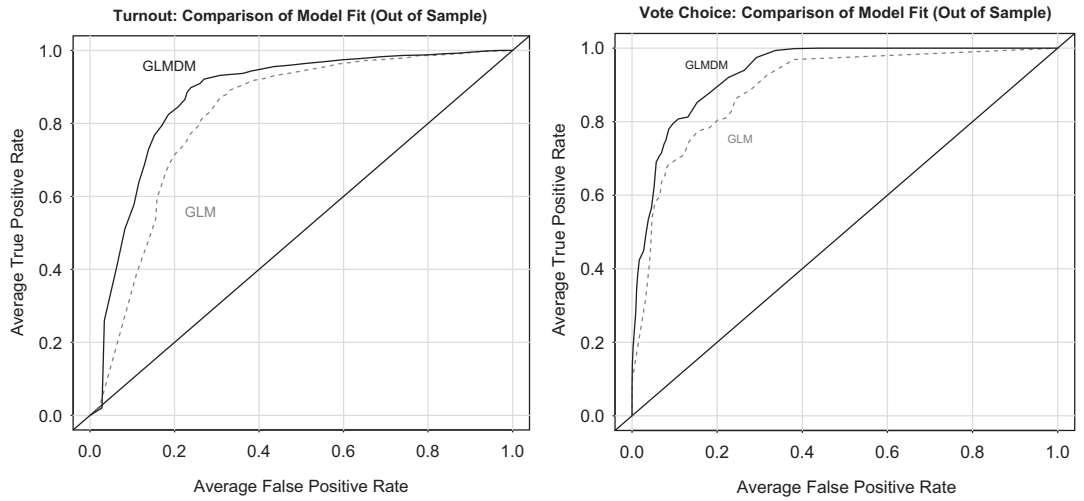


Fig. 2 ROC curves of out-of-sample predictive performance of GLMDM probit and GLM probit models. Results from a 4-fold cross-validation with 10 iterations.

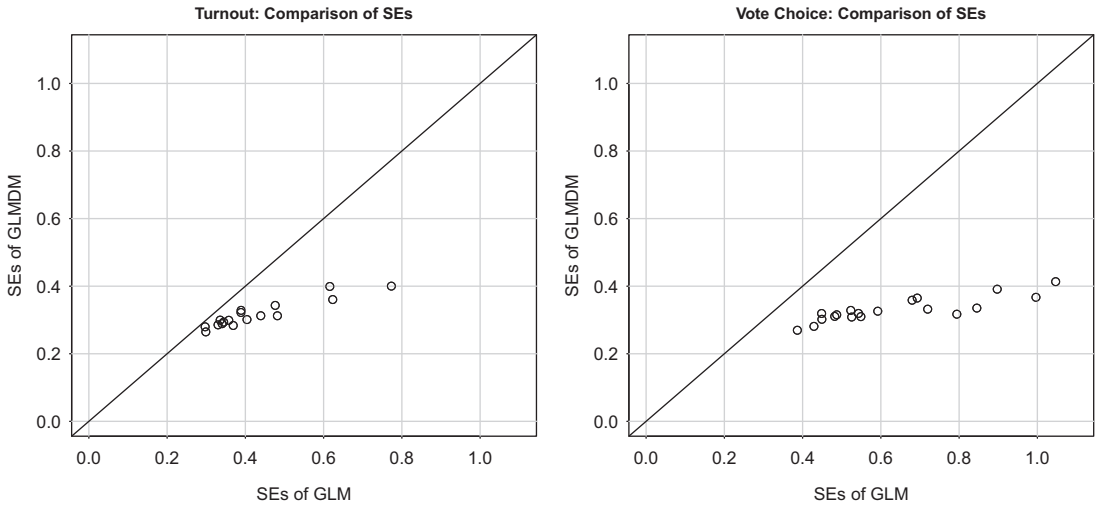


Fig. 3 Comparison of the size of the fixed effect coefficients' standard errors from GLMDM probit and GLM probit models.

out to vote. This is in accordance with the general theoretical view that due to socialization in the host country, the participation deficit of immigrants decreases over the generations (Wüst 2011). For the remaining explanatory variables we found no reliable coefficients.

How do the results of the GLMDM and GLM differ? In contrast to the smaller 95% highest probability densities (HPDs) that we already noted above, the posterior means (i.e., the coefficients) do not differ dramatically between the two specifications. The models come to similar substantive conclusions, but note that for the key socio-demographic variables `firstgeneration`, `age`, and `employ`, the posterior means of the GLMDM tend to be markedly smaller than those in the GLM. This indicates that there is some unobserved heterogeneity in the data that, when ignored, biases these coefficients upward. Note that these differences are with the explanatory variables referring to important features of immigrants' situation—in particular belonging to the first immigrant generation and whether one is employed in Germany—suggesting that the new model is successful because the binning process is picking up heterogeneity between immigrant groups, which we know to exist but cannot directly observe due to data restrictions. In our case, the posterior mean of the number of latent groups produced by the Dirichlet process is $k = 61.8$ with a

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Table 1 Posterior means and 95% HPDs of GLM and GLMDM probit models for immigrants' turnout in the 2009 German federal election

	GLM probit		*GLMDM probit*	
	Mean	*95% HPD*	*Mean*	*95% HPD*
(Intercept)	0.23	[−0.99: 1.45]	0.47	[−0.23: 1.17]
Firstgeneration	−1.09	[−1.81: −0.37]	−0.52	[−1.06: 0.06]
Aussiedler	0.37	[−0.57: 1.32]	0.28	[−0.33: 0.89]
Guestworker	0.35	[−0.58: 1.28]	0.24	[−0.42: 0.93]
Otherlanguage	0.61	[−0.60: 1.81]	0.18	[−0.59: 0.98]
Other	0.86	[−0.66: 2.38]	0.30	[−0.49: 1.08]
Female	0.27	[−0.40: 0.94]	0.06	[−0.51: 0.62]
Age	1.31	[0.45: 2.17]	0.65	[0.02: 1.25]
Educ	0.30	[−0.40: 1.00]	0.25	[−0.32: 0.85]
Employ	1.08	[0.28: 1.87]	0.67	[0.09: 1.26]
Member	0.04	[−0.55: 0.62]	0.12	[−0.40: 0.65]
Class	−0.09	[−0.85: 0.67]	−0.04	[−0.68: 0.58]
Interest	0.89	[0.12: 1.65]	0.72	[0.06: 1.36]
Strengthpid	0.84	[0.18: 1.50]	0.84	[0.26: 1.42]
Civicduty	0.40	[−0.19: 0.98]	0.47	[−0.07: 1.03]
Absdiffeval	0.27	[−0.38: 0.91]	0.36	[−0.18: 0.93]
Satisf	0.03	[−0.64: 0.71]	0.04	[−0.53: 0.62]

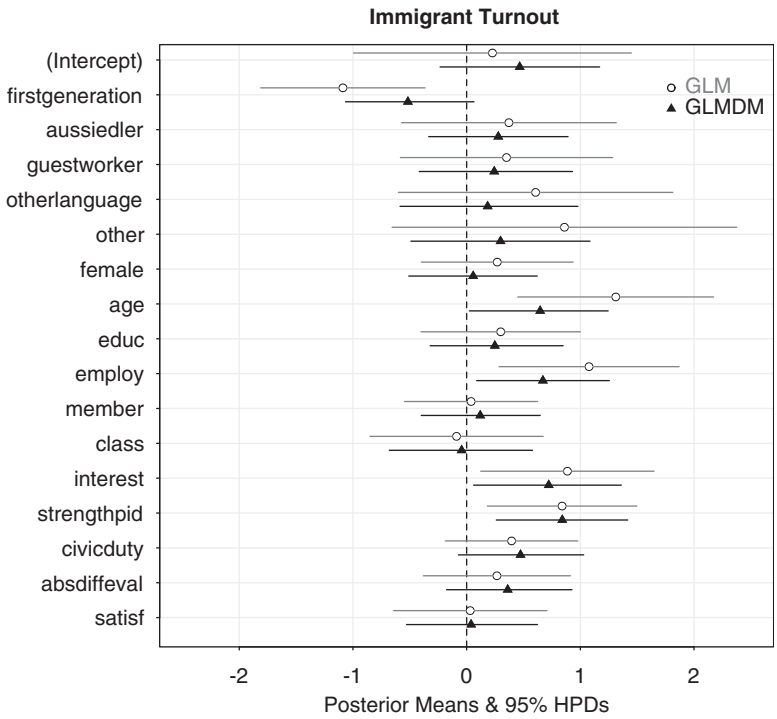


Fig. 4 Posterior means and 95% highest probability densities (HPDs) of GLMDM probit and GLM probit models for immigrants' turnout in the 2009 German federal election.

standard deviation of 4.5. It is important to remember that this binning process is there to provide better model fit without regard to the parsimony or interpretability of the number of groups. Therefore, these are not groups in the substantive sense.

Substantively, once we account for these latent group differences in the GLMDM, we find that the difference in turnout between first-generation immigrants who were born outside the country

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

and immigrants whose parents or ancestors already came to Germany is far less pronounced than often believed. Indeed, not only is immigrant generation now less important than the general political predictors of political interest and party identification, the respective 95% HPD now even includes zero. Under strict adherence to conventional cutoffs, the effect of belonging to the first immigrant generation would not be considered reliable anymore.

Of course, we can only speculate about *what* immigrant features are being picked up by the GLMDM and are thus driving this result. Given that the voting age in Germany is 18 years and the immigration waves for *Aussiedler* and refugees from former Yugoslavia occurred only in the beginning of the nineties, only a few individuals from these groups born in Germany (i.e., belonging to the second generation) would have been eligible to vote in 2009. Therefore, immigrant generation and immigrant type are likely confounded. The lower turnout rate for the first generation then results from differences between types which, as we discussed, are hard to classify given the lack of relevant covariates. One plausible explanation for the lower propensity to turn out to vote is that the political socialization of these groups occurred under socialist regimes. The GLMDM accounts for this unobserved source of heterogeneity and un-biases the coefficient for first-generation immigrants.

Turning to *vote choice*, we can report that in both models, as expected, those immigrants who identify with the CDU/CSU are more likely to have voted for them in the 2009 election (see Table 2 and Fig. 5). Moreover, none of the variables capturing different immigrant groups have reliable coefficients.

Again, the main difference between the GLMDM and the GLM specification lies in the markedly shorter 95% HPDs. But we also observe some notable shifts in the posterior means of the intercept and some variables. The GLMDM generally produces smaller coefficient estimates, as is particularly evident for `educ` and `employ`. For education, the HPD now clearly excludes the null, making the coefficient reliable under conventional levels; for the employment status, the HPD now clearly includes the null, making the coefficient unreliable under conventional levels. In other words, once we account for the unobserved grouping of immigrants in the GLMDM ($k = 52.66$ with a standard deviation of 3.9), our substantive conclusions about the role of education and employment status for conservative vote choice change. The highly educated immigrants were less likely to have voted for the CDU/CSU than the less educated, whereas those who are employed do not favor the CDU/CSU any more than immigrants without employment. Both facts could have been easily missed in a model ignoring the heterogeneity of immigrants.

6 Conclusion

In this article, we apply recently developed Bayesian nonparametric tools, some of which by one of the authors in the statistics literature, to a difficult problem in political science. We have argued that the GLMDM can account for additional variability in the data, which strongly affects parameter estimates. This is because there exists information in the data *that is not expressed directly through the likelihood function* that produces GLM results. The Likelihood Principle (Birnbaum 1962) states that once the data are observed, and therefore treated as given, all of the available evidence for producing the maximum-likelihood estimate of $\hat{\theta}$ is contained in the (log) likelihood function, $\ell(\theta|X)$. This is a very handy data-reduction tool because it tells us exactly what treatment of the data is important to us and allows us to ignore an infinite number of alternatives (Poirer 1988, 127). However, *all* Bayesian inference is a combination of prior information and likelihood information, and therefore it adds another component to the Likelihood Principle for inference rather than violating it. Recall that for very large data sets the Bayesian result matches the likelihood result, meaning that the likelihood function plays the same role in both approaches. We simply find that additional information, beyond the joint distribution of the observed data, can be incorporated into an algorithmically produced prior. The measure of success for this enhanced prior process is the extent to which we capture unobserved heterogeneity that may exist in the data.

We have demonstrated the benefits of GLMDM in the study of voting behavior of immigrants in Germany. Not only did the GLMDM fit the vote data better than the GLM, it also produced consistently smaller standard errors for the regression coefficients. Most importantly, the GLMDM

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Table 2 Posterior means and 95% HPDs of the GLM and GLMDM probit models for immigrants' vote choice in the 2009 German federal election

	GLM probit		*GLMDM probit*	
	Mean	*95% HPD*	*Mean*	*95% HPD*
(Intercept)	−2.48	[−4.43: −0.53]	−0.86	[−1.59: −0.14]
Firstgeneration	0.13	[−0.71: 0.97]	−0.11	[−0.68: 0.46]
Aussiedler	−0.37	[−1.78: 1.04]	−0.33	[−1.00: 0.34]
Guestworker	−0.49	[−1.82: 0.85]	−0.45	[−1.13: 0.25]
Otherlanguage	−0.55	[−2.60: 1.50]	−0.30	[−1.13: 0.53]
Other	−0.19	[−1.95: 1.57]	−0.19	[−1.00: 0.59]
Female	0.10	[−0.98: 1.18]	0.08	[−0.53: 0.68]
Age	1.62	[−0.04: 3.28]	0.33	[−0.32: 0.99]
Educ	−0.92	[−1.86: 0.03]	−0.79	[−1.40: −0.18]
Employ	1.80	[0.24: 3.36]	0.34	[−0.29: 0.96]
Member	−0.35	[−1.23: 0.53]	−0.21	[−0.80: 0.39]
Attendance	0.75	[−0.01: 1.50]	0.55	[0.02: 1.07]
Worker	−0.12	[−1.07: 0.84]	−0.12	[−0.75: 0.47]
Selfempl	0.34	[−1.02: 1.70]	0.11	[−0.58: 0.82]
Leftright	0.55	[−0.61: 1.71]	0.67	[0.01: 1.31]
Spending	−0.64	[−1.66: 0.39]	−0.36	[−1.03: 0.31]
Immigration	−0.12	[−1.15: 0.91]	0.07	[−0.59: 0.69]
Nuclear	0.15	[−0.73: 1.03]	−0.17	[−0.80: 0.46]
Partisanship	2.26	[1.20: 3.33]	1.24	[0.60: 1.88]

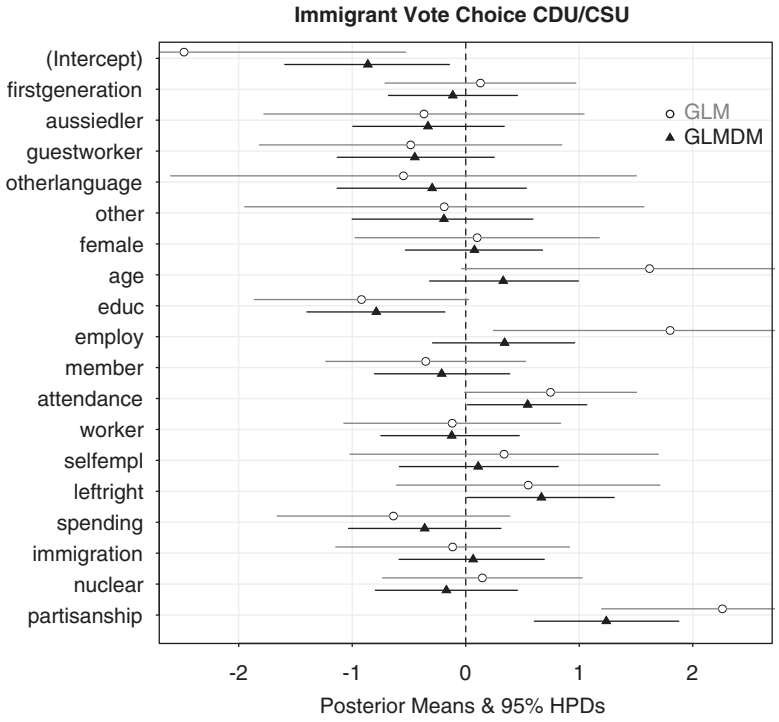


Fig. 5 Posterior means and 95% HPDs of GLMDM probit and GLM probit models for immigrants' vote choice in the 2009 German federal election.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

produced smaller posterior means of the effects of many important variables related to voters' immigrant status and structural integration. It thus changed our substantive understanding of the factors which affect immigrants' turnout and vote choice. With regard to turnout, for instance, we found that once we account for unobserved heterogeneity among immigrant voters, whether a voter belongs to the first immigrant generation or not is much less important than the extant literature suggests. When looking at vote choice, we also found that an immigrant's employment status and thus degree of structural integration does not affect the vote in favor of the CDU/CSU. Instead there exists a general reluctance on the side of immigrants to support this party, which is traditionally associated with restrictive immigration policy.

Note that the GLMDM specification does not *always* produce smaller posterior means. In the presence of unobserved, and otherwise unmodeled, heterogeneity, they will produce *different* posterior means unless the sample size is quite large. When latent information exists and is important to the relationship between the explanatory variables and the outcome, then the posterior means will be different between a standard GLM specification and GLMDM. For example, Gill and Casella (2009) found that a group of distinctly bureaucratic variable coefficients *increased* relative to other nonbureaucratic variable coefficients when an ordered probit GLMDM was applied to explanations for leadership turnover in U.S. public agencies.

As presented in this work, the methodological argument for using the more involved Dirichlet process modeling approach is direct and clear. These are tools for pulling additional information from the estimation process that is guaranteed to be missed in the presence of unobserved heterogeneity with standard tools. Since an `R` package (`glmdm`) already exists to perform the difficult MCMC steps, the primary hurdle for more widespread use in political science is the requirement to be explicitly Bayesian in one's work. Empirical workers are shifting dramatically in this direction in a number of related fields, and we hope that this applied work contributes to such progress in political science.

Funding

Jeff Gill was supported by National Science Foundation Grants DMS-0631632 and SES-0631588.

Conflict of interest statement. None declared.

References

Antoniak, Charles E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* 2:1152–74.

Bartels, L. 1997. Specification uncertainty and model averaging. *American Journal of Political Science* 41:641–74.

Berk, R. A., W. Bruce, and Robert E. Weiss. 1995. Statistical inference for apparent populations. *Sociological Methodology* 25:421–58.

Bird, K., T. L. Saalfeld, and A. Wüst, eds. 2011. *The political representation of immigrants and minorities*. London: Routledge.

Birnbaum, A. 1962. On the foundations of statistical inference. *Journal of the American Statistical Association* 57:269–306.

Blackwell, D., and J. B. MacQueen. 1973. Discreteness of Ferguson selections. *Annals of Statistics* 1:365–58.

Burr, D., and H. Doss. 2005. A Bayesian semi-parametric model for random effects meta-analysis. *Journal of the American Statistical Association* 100:242–51.

Dancygier, R., and E. N. Saunders. 2006. A new electorate? Comparing preferences and partisanship between immigrants and natives. *American Journal of Political Science* 50:962–81.

Dey, D. K., S. K. Ghosh, and B. K. Mallick. 2000. *Generalized linear models: A Bayesian perspective*. New York: Marcel Dekker.

Dorazio, R. M., B. Mukherjee, L. Zhang, M. Ghosh, H. L. Jelks, and F. Jordan. 2007. Modelling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* 64:635–44.

Dorfman, D. D., and E. Alf Jr. 1969. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals rating-method data. *Journal of Mathematical Psychology* 6:487–96.

Doss, Hani. 1994. Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics* 22:1763–86.

Efron, Bradley. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 78:316–31.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Escobar, M. D., and M. West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90:577–88.
Fahrmeir, L., and G. Tutz. 2001. *Multivariate statistical modelling based on generalized linear models*. 2nd ed. New York: Springer.
Ferguson, T. S. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1:209–30.
Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transmissions of the Royal Statistical Society A* 222:309–68.
Geisser, Seymour. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70:320–28.
Gelman, A., and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
Gill, Jeff. 2007. *Bayesian methods for the social and behavioral sciences*. 2nd ed. New York: Chapman & Hall.
———. 2008. Is partial-dimension convergence a problem for inferences from MCMC algorithms? *Political Analysis* 16:153–78.
Gill, Jeff, and George Casella. 2009. Nonparametric priors for ordinal Bayesian social science models: Specification and estimation. *Journal of the American Statistical Association* 104:453–64.
Heath, A. F., S. D. Fisher, G. Rosenblatt, D. Sanders, and M. Sobolewska. 2013. *The political integration of ethnic minorities in Britain*. Oxford: Oxford University Press.
Hill, Jennifer L., and Hanspeter Kriesi. 2001. Classification by opinion-changing behavior: A mixture model approach. *Political Analysis* 9:301–24.
Hobert, J. P., and D. Marchev. 2008. A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *Annals of Statistics* 36:532–54.
Jiang, Jiming. 2007. *Linear and generalized linear mixed models and their applications*. New York: Springer-Verlag.
Korwar, R. M., and M. Hollander. 1973. Contributions to the theory of Dirichlet processes. *Annals of Probability* 1:705–11.
Kyung, M., J. Gill, and G. Casella. 2009. Characterizing the variance improvement in linear Dirichlet random effects models. *Statistics and Probability Letters* 79:2343–50.
———. 2010. Estimation in Dirichlet random effects models. *Annals of Statistics* 38:979–1009.
———. 2011. New findings from terrorism data: Dirichlet process random effects models for latent groups. *Journal of the Royal Statistical Society, Series C* 60:701–21.
———. 2012. Sampling schemes for generalized linear Dirichlet process random effects models. *Statistical Methods and Applications* 20:259–90.
Leamer, E. E. 1978. *Specification searches: Ad hoc inference with nonexperimental data*. New York: John Wiley & Sons.
Liu, J. S. 1996. Nonparametric hierarchical Bayes via sequential imputations. *Annals of Statistics* 24:911–30.
Lo, A. Y. 1984. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics* 12:351–57.
MacEachern, S. N., and P. Müller. 1998. Estimating mixture of Dirichlet process model. *Journal of Computational and Graphical Statistics* 7:223–38.
Martin, Andrew, and Kevin Quinn. 2002. Dynamic ideal point estimation via Markov chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis* 10:134–53.
McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models*. 2nd ed. New York: Chapman & Hall.
McCulloch, C. E., and S. R. Searle. 2001. *Generalized, linear, and mixed models*. New York: John Wiley & Sons.
Neal, R. M. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 9:249–65.
Neyman, Jerzy. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. In *A selection of early statistical papers*, ed. J. Neyman, 250–90. Berkeley: University of California Press.
Pearson, Karl. 1920a. The fundamental problem of practical statistics. *Biometrika* 13(1):1–16.
———. 1920b. Note on the fundamental problem of practical statistics. *Biometrika* 13:300–301.
Poirer, D. J. 1988. Frequentist and subjectivist perspectives on the problems of model building in economics. *Journal of Economic Perspectives* 2:121–44.
Quinn, Kevin M., Andrew Martin, and Andrew B. Whitford. 1999. Voter choice in multi-party democracies: A test of competing theories and models. *American Journal of Political Science* 43:1231–47.
Rubin, Donald B., and Nathaniel Schenker. 1987. Logit-based interval estimation for binomial data using the Jeffreys prior. *Sociological Methodology* 17:131–44.
Schweinberger, Michael, and Tom A. B. Snijders. 2003. Settings in social networks: A measurement model. *Sociological Methodology* 33:307–41.
Sethuraman, J. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica* 4:639–50.
Smith, Alastair. 1999. Testing theories of strategic choice: The example of crisis escalation. *American Journal of Political Science* 43:1254–83.
Spirling, Arthur, and Kevin, Quinn. 2010. Identifying intraparty voting blocs in the U.K. House of Commons. *Journal of the American Statistical Association* 105:447–57.
Stegmueller, Daniel. 2013. Modeling dynamic preferences: A Bayesian robust dynamic latent ordered probit model. *Political Analysis* 21:314–33.
Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101:1566–81.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press

Ward, Michael D., Brian D. Greenhill, and Kristin M. Bakke. 2010. The perils of policy by *p*-value: Predicting civil conflicts. *Journal of Peace Research* 47:1–13.

Western, Bruce. 1998. Causal heterogeneity in comparative research: A Bayesian hierarchical modelling approach. *American Journal of Political Science* 42:1233–59.

———. 1996. Vague theory and model uncertainty in macrosociology. *Sociological Methodology* 26:165–92.

Western, Bruce, and Simon Jackman. 1994. Bayesian inference for comparative research. *American Political Science Review* 88:412–23.

Womack, Andrew, Jeff Gill, and George Casella. 2014. *Product partitioned Dirichlet process prior models for identifying substantive clusters and fitted subclusters in social science data*. Washington University Technical Paper.

Wüst, Andreas M. 2011. Dauerhaft oder temporär? Zur Bedeutung des Migrationshintergrunds für Wahlbeteiligung und Parteiwahl bei der Bundestagswahl 2009. *Politische Vierteljahresschrift* 45:157–78.

https://doi.org/10.1093/pan/mpu018 Published online by Cambridge University Press