Product Partitioned Dirichlet Process Prior Models for Identifying Substantive Clusters and Fitted Subclusters in Social Science Data

Andrew Womack* Jeff Gill† George Casella‡

April 10, 2013

Abstract

We introduce a new model-based clustering design using product partitions. This Bayesian specification simultaneously incorporates substantive clustering and model-fit *sub*clustering on random effects from a Dirichlet process prior. The estimation algorithm directly includes variable context within clusters into a general clustering model that detects latent clustering effects pervasive in social science datasets based on posterior probability. The analysis of terrorist groups shows how this tool reveals important features in a dataset that are otherwise undetectable.

AMS 2000 subject classifications: Primary 62F99; secondary 62P25; secondary 62G99

Keywords and phrases: linear mixed models, generalized linear mixed models, hierarchical models, Dirichlet process priors, clustering models, product partitions, Gibbs sampling. Metropolis-Hastings Algorithm, terrorism data analysis.

*Postdoctoral Fellow, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grants DMS-04-05543, DMS-1026165 and SES-1028329. Email: womack.andrew@gmail.com.

†Professor, Department of Political Science, Division of Biostatistics, and Department of Surgery, Washington University, One Brookings Dr., Seigle Hall, St. Louis, MO. Supported by National Science Foundation Grants DMS-1026165 and SES-1028329. Email: jgill@wustl.edu.

‡Distinguished Professor, Department of Statistics, University of Florida, Gainesville, FL 32611. Supported by National Science Foundation Grants DMS-04-05543, DMS-0631632 and SES-0631588. NOTE: This paper is dedicated to our coauthor, friend, and mentor, George Casella who left us on June 17, 2012 (Fathers' Day). George's considerable effort and thinking are reflected herein.

1 Objectives and Significance

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 withhold information, collection in difficult or even violent areas, embargoed information based on privacy concerns, well-known survey response issues, overlapping explanatory power in model variables, the fluidity of political and social institutions, as well as the willingness of individuals to conceal information from researchers. This has led to many important modeling innovations as a way to meet these challenges. In particular, one problem that is difficult to handle with traditional statistical models is deliberately withheld information that correlates strongly with phenomena of interest. For example, Gill and Casella (2009) used a generalized linear mixed model 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 (Mackenzie and Light ICPSR Study Number 8458, Spring 1987) embargoed key information, such as agency employer, that would have helped researchers but identified these government executives. As a way to draw subtle information out of the data that sheds light on the bureaucratic classification, a Bayesian approach was developed where the random effects are modeled with a Dirichlet process mixture prior. Such information can be thought of as latent clustering in the data, but do not constitute actual clustering in the since the criteria for their creation produces too many categories.

This work addresses the issues of latent clusters in the data, but still including the groupings that result from a *Generalized Linear Mixed Dirichlet Model* (GLMDM), and improves the current state of model-based clustering algorithms. Here we adapt GLMDM models to estimate the probability of alternative posterior cluster arrangements that account for both differing responses to

covariate information as well as non-parametric individual level random effects. Because this approach models two types of latent heterogeneity, it can help us better understand clustering effects that are pervasive in social science datasets, notably with empirical studies of terrorism as shown here. This unique application will improve our understanding underlying commonality and distinctiveness in terrorist groups.

Terrorism has existed since humans first built weapons, and probably before. Naturally, the academic study of terrorism increased after September 11, 2001, and this is now a very active research area. Unfortunately, the empirical analysis of terrorist groups and terrorism events is hampered by the poor quality of the data. Insufficient work has been done using conventional statistical tools, since such data are always more messy and interrelated than in other related areas, such as criminology and international relations. The key underlying problem is that a set of diversely organized covert and violent operatives do not cooperate with data collection efforts. So the resulting information is usually incomplete, sparse, and difficult to model with parametric forms. Normally, such a state of affairs would drive away researchers and shelve projects until the quality of the data changed. However, since the safety of millions depends on understanding terrorist organizations, this is not a viable path. We seek to improve this problem with the model-based clustering tool described here.

2 Background

There are often structures in social science data such as: unexplained clustering effects, unit heterogeneity, autocorrelation, or missingness, that cast doubt on the notion of homogeneous effects of estimated coefficients in a given model. Heterogeneity can be modeled in many ways, and here we describe our approach, and its relation with another popular method.

2.1 Modeling Unobserved Heterogeneity As Clusters

As an example, consider voting in a congressional election where the Democratic candidate favors pulling soldiers out of Afghanistan and the Republican candidate advocates continued military action there. Normally the standard set of explanatory variables from survey research (partisanship, ideology, race, age, education, etc.) are well-justified and powerful determinants of this vote choice, but perhaps not in the same way for a conservative respondent who has a relative in the military based in Afghanistan or a liberal who is a recent veteran of the armed forces. Such issues are not normally covered by questions asked in standard academic or journalistic surveys, but may be an unmeasured strong causal reason for this vote. Here we are concerned with latent clustering that, if not accounted for, adversely affects the quality of the model, since unmeasured explanatory phenomena still affect the modeled relationship. This is a very general problem since these clusters can be described in many ways.

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 voting choice model these are: partisanship, ideology, sex, race, age, and education. The game, according to Leamer, is specifying an additional set of "over the horizon" variables that may provide new knowledge. Often the first type of variables are included in the final specification even if they are not found to be statistically reliable because there is a history of this variable contributing to model specifications in the relevant literature. It is not widely recognized that the effect of such variables can be altered by latent clusters. That is, for some individual cases in the data, a variable could be a strong determinant of the outcome variable, but its effect is sufficiently heterogeneous across individuals that it does not appear statistically reliable in the model. Thus, accounting for latent clusters as proposed here, can affect how an explanatory variable is assessed in model summaries. The model would identify clusters of individuals where the effect of the variable would differ between clusters.

2.2 Dirichlet Process Mixtures Models

We are concerned with how nonparametric priors can enhance the increasing use of Bayesian models in the social sciences, particularly in the (near ubiquitous) presence of latent clusters. As a clarification, we are concerned with accounting for an unknown number of unseen clusters *in the context of building a statistical*

model. We are not working in the general area of spatial clustering with known coordinates.

One effective strategy for dealing with unmeasured grouping phenomena is to use random effect terms, denoted ψ_i here, to capture such underlying clustering information. The distribution of the ψ_i is unknown, by the researcher, but can be determined by custom or intuition. Frequently it is chosen to be a normal distribution, even in the absence of evidence that this provides a good fit. As the random effects, unlike error terms, cannot be checked (there are no corresponding residuals), the normal assumption is justified only as a convenience. A better alternative is a nonparametric Bayesian approach that draws ψ_i from more flexible class of distributions:

$$(Y_1, \ldots, Y_n) \sim f(y_1, \ldots, y_n \mid \beta, \psi_1, \ldots, \psi_n) = \prod_i f(y_i|\beta, \psi_i), \quad \psi_i \sim G, \quad i = 1, \ldots, n, \tag{1}$$

where f is taken as normal in the conventional regression setting, and a popular choice for G is the Dirichlet Process ($\mathcal{DP}$)

$$\psi_i \sim G \sim \mathcal{DP}(\lambda, \phi_0), \quad i = 1, \ldots, n, \tag{2}$$

with base measure ϕ_0 and precision parameter λ. In particular, the observations are modeled as

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \psi_i + \epsilon_i, \quad \psi_i \sim \mathcal{DP}(\phi_0, \lambda), \quad i = 1, \ldots, n, \tag{3}$$

where the ϵ_i are independent normal random variables (note that the addition of a link function, $g^{-1}()$, turns this into a generalized linear mixed model). Since the ψ_i are drawn from a $\mathcal{DP}$ distribution, they are not necessarily unique and thus can be represented by a K-vector, $\boldsymbol{\eta}$, where $K \leq n$. Thus, the model can be succinctly written as

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times K}{\mathbf{A}}\underset{K\times 1}{\boldsymbol{\eta}} + \underset{n\times 1}{\boldsymbol{\epsilon}} \tag{4}$$

where $\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\eta}$ and $\mathbf{A}$ is an $n \times K$ matrix of zeros with a single one in each row which denotes the particular η_k assigned to ψ_i (Kyung *et al.*2010).

Dirichlet process models were introduced by Ferguson (1973), who defined the process and investigated the basic properties. Blackwell and MacQueen (1973) showed that the marginal distribution is that of the n^{th} step of a Polya urn process. Other work that characterizes the properties of the Dirichlet process includes Korwar and Hollander (1973) and Sethuraman (1994). Work that has particular importance for our development is that of Lo (1984), who derives the analytic form of a Bayesian density estimator, and Liu (1996), who derives an identity for the profile likelihood estimator of λ. The implementation of the Dirichlet process mixture model has been made feasible by modern methods of Bayesian computation and efficient algorithms. The work of Escobar and West (1995) and MacEachern and Müller (1998) developed estimation techniques and sampling algorithms. Neal (2000) provides an extended and more efficient Gibbs sampler.

The model specified in (1) 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 that was developed in Gill and Casella (2009). They found that this approach can model difficult data and produce results that existing alternative methods fail to discover. In that work they were able to account for important latent clustering structures that do not necessarily reflect *confounding variables*, but still provide information about agency environment that was not explicitly available. The Dirichlet process produces clusters due to the fact that any realization of a Dirichlet process is discrete. However, the Dirichlet process is really providing a non-parametric estimation of the distribution of the individual level random effects.

2.3 Cluster Analysis

Model-based cluster analysis has a long and rich history, and recent developments using MCMC methods and hierarchical models are relevant to our work. For recent examples, Zhong and Ghosh (2003) develop a bipartite graph approach that identifies clusters as distinct probabilistic models. Pan and Shen (2007), as well as Xie *et al.*(2008), are concerned with model selection in the presence of unknown clustering and focus on a penalized likelihood approach to get a parsimonious number of elements. Maugis *et al.*(2009) are also concerned with model selection in this context, but they develop a procedure to separate variables that are relevant to clustering, variables that are not relevant to clustering but are dependent of some that are, and variables not relevant at all to clustering. In the next section we describe key papers that inform our approach.

There are two basic approaches to cluster models that have been used in the literature, those based on an underlying mixture model, and those based on a *product partition model* (also called *classification likelihood*). Quintana and Iglesias showed that the $\mathcal{DP}$ model is actually an example of a product partition model. For reasons given in Section 2.3.3 we choose not to use the $\mathcal{DP}$ model for substantive clustering and develop an alternative product partition model in Section 2.4.

2.3.1 Mixtures and Product Partition Models

This section contrasts the finite mixture model for reflecting latent heterogeneity with the product partition model, an alternative approach that we recommend and develop. Commonly applied mixture models begin with the assumption that $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are realizations of n independent and identically distributed (iid) random variables from the m-component mixture density:

$$f(\cdot|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m) = \sum_{\ell=1}^{m} \omega_\ell \, f_\ell(\cdot|\boldsymbol{\theta}_\ell) \, , \tag{5}$$

where m is a fixed positive integer, $0 \leq \omega_\ell \leq 1$, $\sum_{\ell=1}^{m} \omega_\ell = 1$, and $f_\ell(\cdot|\boldsymbol{\theta}_\ell)$ is the density of the ℓth component of the mixture and depends on the parameter $\boldsymbol{\theta}_\ell$. The mixture model does not account for clustering of the data directly. A partition of the data is typically obtained as a byproduct of the use of the EM algorithm. A different way to think about latent heterogeneity, the *product partition model*, starts by conditioning on a given partition, and then determines the posterior probabilities of these partitions using Bayes' rule. Given a partition $\mathcal{C}$ of $\mathbb{N}_n := \{1, 2, \ldots, n\}$ that has m clusters denoted by $C_1, \ldots, C_m$, the data are a realization from

a density of the form:

$$f(\mathbf{Y}|\boldsymbol{\theta}_{\mathcal{C}}, \mathcal{C}) = \prod_{\ell=1}^{m} \prod_{i \in C_\ell} f_\ell(Y_i|\boldsymbol{\theta}_\ell) \; . \tag{6}$$

where $f_\ell(\cdot|\boldsymbol{\theta}_\ell)$ is the density associated with the ℓth cluster and depends on parameter θ_ℓ. Unlike the mixture model (5), model (6) recognizes a parameter, $\mathcal{C}$, that is directly connected to the basic clustering problem. This model was developed by Hartigan (1990) (see also Barry and Hartigan 1992, Crowley 1997) as a product partition model.

McCullaugh and Yang (2008) and Booth *et al.* (2008) argue strongly in favor of the product partition model, the latter noting that not only does the mixture model lack a parameter that defines the clusters, the inference is sometimes confounded with the common application of the EM algorithm. That is, even if the parameters of the model are known, there needs to be some way of generating a latent variable to identify clusters. Park and Dunson (2010) create a related generalized product partition model such that the partition process is predictor-dependent and computationally efficient. The key here is that the mixtures model is also a marginal model, and our perspective is that real interest is in the conditional model as shown in (6).

McCullaugh and Yang (2008) also argue that the mixture model is not appropriate for determining clusters. A further deficiency is that the model needs to be run with a fixed m; the typical strategy is to run a selection of m values and choose the one with the best BIC. Conversely, the product partition model clearly identifies the parameter that determines the cluster, and has no restriction on m, the number of clusters. A stochastic search algorithm, such as one used by Booth *et al.* (2008), can move between different size partitions at each iteration.

Lastly, the mixture model is also prone to a label switching problem, where modes cannot be identified and thus ergodic averages cannot be computed without further processing. See, for example, Stephens (2000), Jasra *et al.*(2005), Sperrin *et al.*(2001). Various solutions include: creating an ordering constraint (Richardson and Green 1997), fixing some of the cases to clusters (Chung *et al.*[2004], as well as the software solution from Grün and Leisch 2009), assigning loss functions that are label-invariant (Celeux 1998), and relabeling using the "maximum a posterior" estimate (Marin *et al.*2005). All of these approaches require additional uncomfortable assumptions or restrictions. Such problems with clustering models are our starting point, and the Section 2.4 describes a solution for recovering estimates of substantive clusters in social science data. In particular, since the product partition model is label-free (the clusters are all defined by unique partitions of $\mathbb{N}_n = \{1, 2, \ldots, n\}$), we can easily identify mappings of cases to clusters.

2.3.2 Classifying Approaches To Model-Based Clustering

We can classify related previous work according to the model used for the clustering, and the model used for the random effects. Clustering based on mixture models was used by Fraley and Raftery (2002), estimating the allocation probabilities and the model parameters with the EM algorithm, and using the Bayesian Information Criteria (BIC) for determination of the number of clusters. Mixture models with Dirichlet random effects (or latent variables) were used by Dahl (2006) for microarray expression data, Kim, *et al.* (2006) for both variable selection and clustering (updating Tadesse *et al.* 2005), and Rodriquez, *et al.* (2008), who used a

nested Dirichlet process structure.

The product partition model, which explicitly searches for the best cluster, was used by Heard *et al.* (2006) and Booth *et al.* (2008), the latter using a substantive objective function to drive the search. Specifically, they evaluated the posterior probability of each partition $\mathcal{C}_m$, using this probability as the target in a Metropolis-Hasting search algorithm. With the product partition model and Dirichlet random effects, Quintana and Iglesias (2003) proposed a Bayesian clustering algorithm that minimizes a posterior loss, which is similar to the approach of Lau and Green (2007), who minimized a misclassification loss.

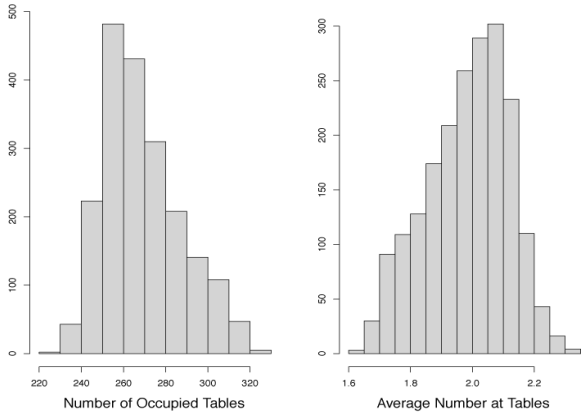
2.3.3 Bayesian Nonparametric Cluster Strategies

It is important to understand the clustering strategy that has previously been used in applications of the Dirichlet random effects model. In particular, consider a standard linear model, where a subject is modeled with covariates and a random effect. A typical strategy is to use the Dirichlet to generate a very large number of candidate "clusters," which are actually *sub*clusters (fractions of clusters), then choose the best of these by a post-hoc scheme that processes the MCMC output through some objective function to find the best grouping.[1] Therefore, the supposed-clusters that are produced only by the MCMC repeated realizations in each iteration of the Dirichlet process are:

- not substantive in any way,
- not able to reflect any real cluster structure driven by the covariates,
- temporary random effect assignments to make the model fit better in the context of the sampler.

Furthermore, since there is *no over-fitting penalty in the Dirichlet process*, we can expect there to always be more *sub*clusters than actual substantive clusters in the data. For example, in the analysis of the tenure of political appointees (Gill and Casella 2009), there were 512 cases and therefore a maximum of 512 "restaurant tables" or *sub*clusters from the Dirichlet process. A iteration of the Gibbs sampler reportedly had an average of 278 of these *sub*clusters, and no scholar of the Federal bureaucracy would make the claim that there are 278 fundamentally different agency environments, so there is evidence that these are not the substantive clusters desired. Furthermore, Figure 1 is a replication of

Figure 1: Dirichlet Subclustering from the Political Executive Data (Gill and Casella, 2009). Histogram Description of the Last 3,000 Iterations



[1] We have used the term *subcluster* in previous work to differentiate these from actual clusters in the data, since the Dirichlet random effects model merely uses these imposed categories to improve the fit of the model rather than to imply actual underlying clustering in the data.

their sampler that shows in the left panel a histogram of the *sub*cluster assignments across $3,000$ draws where the mean is obvious. The right panel histogram shows that the average number of cases per *sub*cluster is around two, across these 3,000 draws, further supporting the point that these are not substantive clusters. This is because the Dirichlet *sub*clustering produces partitions with a large number of clusters *with the objective of reducing model variation, not finding meaningful covariate-based groupings.* This distinction is the motivation for the model produced in this paper.

2.4 Substantive Clustering Strategy

We introduce a model that combines clustering using a product partition model and non-parametric estimation of random effects through the Dirichlet process in the context of Gaussian linear models. Conditioned on a particular partition $\mathcal{C}$ with clusters $C_\ell, \ell = 1, \ldots, m$ and random effect *sub*cluster assignment matrix $\mathbf{A}$ with k rows (as well as the necessary parameter values), the data $\mathbf{Y}$ are assumed to be normally distributed within clusters. In particular let $\mathbf{Y}_\ell$ be a vector of length n_ℓ containing the Y_i in cluster C_ℓ. Informally, the data in cluster C_ℓ are described by,

$$\mathbf{Y}_\ell = \mathbf{X}_\ell \boldsymbol{\beta}_\ell + \mathbf{A}_\ell \boldsymbol{\eta} + \boldsymbol{\epsilon}_\ell \tag{7}$$

where $\mathbf{X}_\ell$ and $\mathbf{A}_\ell$ are composed of the rows corresponding to the Y_i in cluster C_ℓ and $\boldsymbol{\epsilon}_\ell \sim \mathcal{N}\left(0, \sigma_\ell^2 I_{n_\ell}\right)$. The parameters $\boldsymbol{\beta}_\ell$ and σ_ℓ^2 are specific to cluster C_ℓ, and we assume σ_ℓ^2 is unknown. The parameter $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ is common between the clusters and the η_j are iid draws from ϕ_0.

This model incorporates clustering in the data in two distinct ways. First, it utilizes $\mathcal{DP}$ random effects to model latent clustering in the data that relates to model fit. Second, the product partition model, using $\mathcal{C}$, provides substantive clusters to the data that serve to provide insights into how the data can be broken into groups that have different behavior (e.g. different reactions to the covariates). Note that these groupings do not need to nest, and so observations in the same cluster C_ℓ can belong to different *sub*cluster defined by the columns of $\mathbf{A}$, and vice-versa since observations in different clusters can be in the same *sub*cluster. In contrast to the usual conditional independence assumption of the product partition model (Hartigan and Barry 1992), the introduction of the $\mathcal{DP}$ random effects produces a correlation between individuals both within the same cluster and in different clusters.

The model is formally defined by providing a hierarchical Bayesian specification, including priors for the cluster specific parameters as well as the parameters of the Dirichlet process random effects. Formally, the model is defined by

$$\mathbf{Y}|\mathcal{C}, \boldsymbol{\beta}_\mathcal{C}, \boldsymbol{\sigma}^2_\mathcal{C}, \mathbf{A}, \boldsymbol{\eta} \sim \prod_{\ell=1}^{m} \mathcal{N}(\mathbf{Y}_\ell | \mathbf{X}_\ell \boldsymbol{\beta}_\ell + \mathbf{A}_\ell \boldsymbol{\eta}, \sigma_\ell^2 I_{n_\ell}) \tag{8}$$

where we assume the priors

$$\boldsymbol{\beta}_\ell | \sigma_\ell^2 \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \sigma_\ell^2 \mathbf{S}^{-1}\right) \quad \text{and} \quad \sigma_\ell^2 \sim \mathcal{IG}\left(\frac{a_{\sigma^2}}{2}, \frac{b_{\sigma^2}}{2}\right)$$

and $\mathbf{A}\eta = \boldsymbol{\psi} \sim \mathcal{DP}(\phi_0, \lambda)$. The model is completed by providing priors for the hyper parameters, which are assumed to be

$$\begin{array}{ccc} \phi_0 = N\left(0, \tau^2\right) & \tau^2 \sim \mathcal{IG}\left(\frac{a_{\tau^2}}{2}, \frac{b_{\tau^2}}{2}\right) & \lambda \sim G\left(\frac{a_\lambda}{2}, \frac{b_\lambda}{2}\right) \\ \boldsymbol{\beta}_0 \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\beta}}\mathbf{S}^{-1}) & \sigma^2_{\boldsymbol{\beta}} \sim \mathcal{IG}\left(\frac{a_{\sigma^2_{\boldsymbol{\beta}}}}{2}, \frac{b_{\sigma^2_{\boldsymbol{\beta}}}}{2}\right) & \mathbf{S} \sim \mathcal{W}(V^{-1}, a_{\mathbf{S}}) \\ V = \mathrm{Diag}(v_1, \ldots, v_p) & v_i \sim G\left(\frac{a_v}{2}, \frac{b_v}{2}\right) & \end{array}$$

and the definition of the prior for $\mathcal{C}$ is treated later. We have defined the model in this way so that the distribution of $\mathbf{Y}|\mathcal{C}, \boldsymbol{\beta}_0, \mathbf{S}, \mathbf{A}, \boldsymbol{\eta}$, obtained by integrating out $\boldsymbol{\beta}_\mathcal{C}$ and $\boldsymbol{\sigma}^2_\mathcal{C}$, is given by a product of multivariate t distributions with the same form. In particular

$$\mathbf{Y}_\ell|X_\ell, \boldsymbol{\beta}_0, \mathbf{S}, \mathbf{A}_\ell, \boldsymbol{\eta} \sim \mathcal{MVT}_{n_\ell}\left(\mathbf{X}_\ell\boldsymbol{\beta}_0 + \mathbf{A}_\ell\boldsymbol{\eta}, \left(I_{n_\ell} + \mathbf{X}_\ell\mathbf{S}^{-1}\mathbf{X}'_\ell\right)\frac{b_{\sigma^2}}{a_{\sigma^2}}\right) \tag{9}$$

This provides a sampling distribution for $\mathbf{Y}$, marginalized over $(\boldsymbol{\beta}_\mathcal{C}, \boldsymbol{\sigma}^2_\mathcal{C})$, of

$$f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}_0, \mathbf{S}, \mathbf{A}, \boldsymbol{\eta}, \mathcal{C}) = \prod_{\ell=1}^{m} f(\mathbf{Y}_\ell|\mathbf{X}_\ell, \boldsymbol{\beta}_0, \mathbf{S}, \mathbf{A}_\ell, \boldsymbol{\eta}, \mathcal{C}) \tag{10}$$

where each $f(\mathbf{Y}_\ell|\mathbf{X}_\ell, \boldsymbol{\beta}_0, \mathbf{S}, \mathbf{A}_\ell, \boldsymbol{\eta}, \mathcal{C})$ is the appropriate multivariate t density.

2.4.1 Cluster Prior Probabilities

The specification of a prior for $\mathcal{C}$ requires special attention in order to obtain reasonable posterior inference for the partitions of the data. Because we are adjudicating partitions $\mathcal{C}$ of N_n using their posterior probabilities, it is important to develop the prior on the set of partitions so that it distributes mass in a reasonable way. Numerous priors have been defined for the clustering problem, and we discuss two of them here; the Ewens-Pitman prior (EPP) and the hierarchical uniform prior (HUP). We aim to understand the priors in two ways. One, we want to determine the prior probability of the set of partitions that have m clusters as $n \to \infty$. Two, we consider the random variable $F = (F_1, \ldots, F_m) = \left(\frac{n_1}{n}, \ldots, \frac{n_m}{n}\right)$ conditioned on the set of partitions that have m clusters. We want to determine the distribution of F as $n \to \infty$.

First, we fix some language and notation. We will refer to the number of clusters in the partition $\mathcal{C}$ as the *size* of $\mathcal{C}$. For each choice of m, the choice of $(n_1, \ldots, n_m)$ such that $0 < n_m \leq n_{m-1} \leq n_2 \leq n_1 \leq n$ and $n_1 + \cdots + n_m = n$ represents a specific *type* of partition. The function $b(n, m)$ counts the number of partition types of N_n of size m and $N(n_1, \ldots, n_m)$ counts the number of partitions of a given type.

The Ewens-Pitman prior arises as the marginal prior distribution over the set of partitions from a Dirichlet process and depends on one parameter, which we shall call ρ. The probability of a partition depends only on its type. A partition $\mathcal{C}$ with type $(n_1, \ldots, n_m)$ has prior probability

$$p^{EPP}(\mathcal{C}|n) = \frac{1}{\Gamma(n+\rho)}\prod_{\ell=1}^{m}(\rho\Gamma(n_\ell)) \tag{11}$$

Fix an integer $m > 0$; the ratio of prior probabilities

$$\frac{p^{EPP}(\{\mathcal{C} \ : \ size(\mathcal{C}) = m+1\}|n)}{p^{EPP}(\{\mathcal{C} \ : \ size(\mathcal{C}) = m\}|n)} \approx \rho \log(n) \tag{12}$$

shows that the EPP places mass on sets of partitions with larger and larger *size* as $n \to \infty$. The random variable F converges in a locally in the weak sense to the Haldane measure on the m simplex, restricted to $F_1 \geq F_2 \geq \cdots \geq F_m$. Thus, the EPP both places mass on partitions with greater size as n increases and converges weakly to an improper measure when restricted to partitions of a particular size, both undesirable properties.

The hierarchical uniform prior, on the other hand, accounts for the *size* and *type* of the partition in the prior construction. First, a prior is placed on m. For convenience, we assume that this prior is a truncated Poisson distribution whose (non-truncated) mean is 1, but any distribution which sums to 1 as $n \to \infty$ is appropriate. Conditioned on $size(\mathcal{C}) = m$, the prior over types is uniform. Conditioned on both the *size* and *type* of partition, the partitions are uniformly distributed. Thus, a partition $\mathcal{C}$ with $size(\mathcal{C}) = m$ and $type(\mathcal{C}) = (n_1, n_2, \ldots, n_m)$ has prior probability

$$p^{HUP}(\mathcal{C}|n) = p^{HUP}(m|n)\frac{1}{b(n,m)N(n_1,\ldots,n_m)}$$

The number of types with fixed size m increases as $b(n,m) \approx \frac{n^{m-1}}{m!(m-1)!}$ as $n \to \infty$ and

$$N(n_1,\ldots,n_m) = \binom{n}{n_1,\ldots,n_m}\frac{1}{R(n_1,\ldots,n_m)}$$

where $R(n_1,\ldots,n_m) = \prod_{j=1}^{n}\left([\sum_{\ell=1}^{m} I(n_\ell = 1)]!\right)$ is a redundancy factor that accounts for reordering subsets of the partition that have the same number of elements. By its construction, the HUP places finite mass on each choice of m. When restricted to partitions of *size* m, the random variable F converges weakly to the uniform measure restricted to the set $F_1 \geq F_2 \geq \cdots \geq F_m$. Thus, the HUP alleviates the issues encountered with the EPP and we choose to focus on the HUP as the prior for the partitions $\mathcal{C}$.

2.4.2 Cluster Posterior Probabilities

Our goal is to find the best partition $\mathcal{C} = (C_1, \ldots, C_m)$ in the sense of finding the $\mathcal{C}$ which maximizes $\pi(\mathcal{C}|\mathbf{Y}, \mathbf{X})$. The posterior probability of $\mathcal{C}$ is given (up to a proportion) by:

$$\begin{aligned}\pi(\mathcal{C}|\mathbf{Y},\mathbf{X}) &\propto f(\mathbf{Y}|\mathcal{C},\mathbf{X})P(\mathcal{C}) \\ &= \Bigg(\sum_{\mathbf{A}}\int f(\mathbf{Y}|\boldsymbol{\beta}_0,\boldsymbol{\eta},\tau^2,\mathbf{A},\sigma^2_{\boldsymbol{\beta}},V,\mathbf{S},\lambda,\mathcal{C},\mathbf{X})d\boldsymbol{\beta}_0 d\boldsymbol{\eta} d\tau^2 d\mathbf{A} d\sigma^2_{\boldsymbol{\beta}} dV d\mathbf{S} d\lambda \\ &\qquad \times p(\boldsymbol{\beta}_0,\boldsymbol{\eta},\tau^2,\mathbf{A},\sigma^2_{\boldsymbol{\beta}},V,\mathbf{S},\lambda)\Bigg)P(\mathcal{C})\end{aligned} \tag{13}$$

Computing this posterior probability requires not only integrating out the cluster specific coefficients and the corresponding hyper parameters, but also marginalizing over the random effects, which requires integration over $\boldsymbol{\eta}$, integrating out the priors for τ^2 and λ, and summation over the $\mathbf{A}$ matrices. These tasks make direct computation intractable. In Section 3 we develop an RJMCMC algorithm for generating samples from the posterior distribution of the parameters of the model, including $\mathcal{C}$ and $\mathbf{A}$.

3 Overview of Posterior Estimation

We use a *Metropolis within Gibbs* procedure to drive a Reversible Jump MCMC algorithm over the set of clusters $\{\mathcal{C}\}$, *sub*clusters $\{\mathbf{A}\}$ and the continuous parameters.. Specifically, we create a Markov chain by dividing the parameters into the sets $\{(\mathbf{A}, \boldsymbol{\eta}), (\mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}), \xi'\}$, where $\xi' = (\boldsymbol{\beta}_0, \sigma^2_{\boldsymbol{\beta}}, \mathbf{S}, \boldsymbol{v}, \tau, \lambda)$. Its stationary distribution is the joint posterior distribution of the parameter sets. The RJMCMC will therefore explore through the space of clusters in a manner that maintains detailed balance. The sampling procedure utilizes Metropolis corrections for sampling $(\mathbf{A}, \boldsymbol{\eta})$ and $(\mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}})$ and Gibbs sampling for ξ'. At each iteration, the sampling of any one component is performed while conditioning on all other components.

The iterative sampling produces model estimation of potentially different dimension on each iteration since changing the number of partitions changes the dimension of the model. Therefore there is a direct analogy with reversible jump MCMC processes (Green 1995), although there is not a bijection function specified in the literal sense. A unique model $\mathcal{C}$ is given by a partition from the sampler, and the likelihood function for the data given $\mathcal{C}$ is conditioned on model-specific parameters, $\theta_{\mathcal{C}} = (\boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}})$, as well as the other parameters whose prior forms do not depend on the partition status, $\xi = (\boldsymbol{\beta}_0, \sigma^2_{\boldsymbol{\beta}}, \mathbf{S}, \boldsymbol{v}, \tau, \boldsymbol{\eta}, A, \lambda)$. So $\boldsymbol{\theta}_{\mathcal{C}}$ and $\boldsymbol{\eta}$ potentially differ in structure on each iteration and ξ retains its original dimension since it depends only on the class of models specified independent of the product partitioning and Dirichlet process random effects.

The sampling of $(\mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}})$ and $(\mathbf{A}, \boldsymbol{\eta})$ are RJMCMC steps, with the dimensions of $(\boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}})$ and $\boldsymbol{\eta}$ possibly changing at each iteration of the sampler due to the changing clusters and *sub*clusters. These parameters are sampled directly from their full conditional distributions after proposing $\mathcal{C}$ or $\mathbf{A}$ and then $(\mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}})$ or $(\mathbf{A}, \boldsymbol{\eta})$ are accepted or rejected as blocks. The Metropolis corrections for these blocks do not depend on the particular values of $(\boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}})$ or $\boldsymbol{\eta}$ sampled because they are drawn from their full conditional distributions. For example, if we let $K(\mathbf{A}'|\mathbf{A}, \xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y})$ be a proposal kernel for $\mathbf{A}$ then the proposal for $(\mathbf{A}', \boldsymbol{\eta}')$ is given by:

$$K(\boldsymbol{\eta}', \mathbf{A}'|\mathbf{A}, \xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y}) = \pi(\boldsymbol{\eta}'|\mathbf{A}', \xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y}) K(\mathbf{A}'|\mathbf{A}, \xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y}) \tag{14}$$

and the target is given by:

$$\pi(\boldsymbol{\eta}', \mathbf{A}'|\xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y}, \mathbf{y}) = \pi(\boldsymbol{\eta}'|\mathbf{A}', \xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}) \pi(\mathbf{A}'|\xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y}) \tag{15}$$

Therefore, the ratio of the two is given by:

$$\frac{K(\boldsymbol{\eta}', \mathbf{A}'|\mathbf{A}, \xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y})}{\pi(\boldsymbol{\eta}', \mathbf{A}'|\xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y})} = \frac{K(\mathbf{A}'|\mathbf{A}, \xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y})}{\pi(\mathbf{A}'|\xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y})} \tag{16}$$

which is free of $\boldsymbol{\eta}$. A similar calculation is easily carried out for $\mathcal{C}$. Because rejecting $\mathcal{C}$ or $\mathbf{A}$ also rejects $\boldsymbol{\theta}_{\mathcal{C}}$ and $\boldsymbol{\eta}$, we add an extra sampling step of $\boldsymbol{\theta}_{\mathcal{C}}$ and $\boldsymbol{\eta}$ after sampling $\mathcal{C}$ and $\mathbf{A}$. This additional sampling step promotes mixing whenever the discrete parameters are rejected and does not affect detailed balance. We discuss specific distributions used for sampling and detailed balance after discussing the use of samples from the Markov chain to produce estimates of the posterior probabilities of the partitions.

3.1 Estimation of Partition Posterior Probabilities

This setup is useful because the computation of marginal partition probabilities can be produced using essentially the same process as RJMCMC estimates. In particular, we take advantage of (9) to form an estimator of the probability of each cluster conditioned on the draws of $(\mathbf{A}, \boldsymbol{\eta})$ and ξ.

First, calculate the posterior probabilities of $\mathcal{C}$ conditionally on $[\mathbf{y}, \mathbf{X}]$ and $(\mathbf{A}, \boldsymbol{\eta}, \xi)$, $f(\mathbf{y}|\mathcal{C}, \mathbf{A}, \boldsymbol{\eta}, \xi, \mathbf{X})$ given in (10). This can be directly transformed into an estimate of the conditional posterior of $\mathcal{C}$ through

$$\hat{P}(\mathcal{C}|\xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)}, \mathbf{y}, \mathbf{X}) = \frac{f(\mathbf{y}|\mathcal{C}, \xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)}, \mathbf{X})p(\mathcal{C})}{\sum_{\mathcal{C}'} f(\mathbf{y}|\mathcal{C}', \xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)}, \mathbf{X})p(\mathcal{C}')}. \tag{17}$$

where the sum is taken over the $\mathcal{C}'$ visited during the stochastic search and (k) represents the value from the kth iteration of the Markov chain. Restricting to the set of partitions visited during the stochastic search is a convenience that is used because the size of the partition space is so large. In theory, one can include all $\mathcal{C}$ in the denominator or some other — restricted — set of partitions. Thus, $\hat{P}(\mathcal{C}|\xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)}, \mathbf{y}, \mathbf{X})$ given in (17) is really an estimate of the conditional posterior probability of $\mathcal{C}$ restricted to the class of partitions used in the denominator of (17).

The estimator averaged over the sampled values is given by:

$$\hat{P}(\mathcal{C}|\mathbf{y}, \mathbf{X}) = \frac{1}{N_{sim}} \sum_{k=1}^{N_{sim}} \hat{P}(\mathcal{C}|\xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)}, \mathbf{y}, \mathbf{X}). \tag{18}$$

and is a direct estimator of (13) through a Rao-Blackwell style estimator. Even though the draw of each $(\xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)})$ is obtained by conditioning on a specific model and the model specific parameters that have been sampled, the distribution of the marginal chain $(\xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)})$ follows $f(\xi^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\eta}^{(k)}|\mathbf{y}, \mathbf{X})$. Thus (18) is a consistent estimator of $P(\mathcal{C}|\mathbf{y}, \mathbf{X})$ as $N_{sim} \to \infty$. Also, since the chain maintains detailed balance, each $\mathcal{C}$ is visited by the chain *eventually* and (17) and (18) can — theoretically — be estimated over the entire partition space.

3.2 Transition Kernels

A variety of proposal kernels are used for $\mathcal{C}$ and $\mathbf{A}$. The kernel used on any specific iteration of the MCMC algorithm is chosen at random, where the multinomial probability vector for this choice is an MCMC tuning parameter. The need for a variety of kernels for clustering problems is discussed in Jain and Neal (2004). Both large and small steps are needed in the sampling algorithm in order to effectively search the model space. Large steps are needed in order to make the sampler move between cluster spaces of different sizes. Small steps are needed to provide a fine adjustment to existing clusters, moving single observations at a time.

For clusters and *sub*clusters, we implement different versions of Multinomial-Dirichlet sampling, and so we describe it for *sub*clusters. Let $\mathbf{A}^{(t)}$ be the current *sub*cluster configuration. Define the Dirichlet probability parameter vector as $\boldsymbol{\alpha} = (m_1 + r, \ldots, m_K + r, r, \ldots, r)$ for fixed $r > 0$ where $\boldsymbol{\alpha}$ has length n and K is the number of *sub*clusters in $\mathbf{A}^{(t)}$ with *sub*cluster lengths $m_1, \ldots, m_K$. A proposal $\mathbf{A}'$ is generated according to a Multinomial distribution with parameter $\mathbf{q} \sim \text{Dirichlet}(\boldsymbol{\alpha})$. The larger r is, the more uniform the sampling from the Multinomial-Dirichlet and thus for large r many small *sub*clusters will be sampled. For small r (say $r = 1/n$), the sampling produces *sub*cluster sizes in $\mathbf{A}'$ which are more like those from $\mathbf{A}$. Since this sampling is indifferent to the actual *sub*clusters in $\mathbf{A}^{(t)}$ and only uses the *sub*cluster sizes in the sampler, this produces samples which are nearly independent of $\mathbf{A}$.

The sampler for clusters has two levels performed on each cycle. The top *macro*-level chooses a move algorithm. Since clustering searches need big steps to mix through the sample space as well as small steps to refine high probability clustering outcomes, we mix two algorithms large moves and two algorithms with small moves from:

- Jain-Neal sampling scheme (large moves)
- Pitman Jump (large moves)
- random walk (small moves)
- restricted Gibbs sampling (small moves).

These choices dictate the *micro*-level algorithm steps contained within. The result of each lower level process is a candidate position for a Metropolis-Hastings accept/reject step at the higher level. Since this decision nests the micro-level decisions, the Markov chain is ergodic at the macro-level regardless of the decisions made within these inner steps.

The **Jain-Neal sampling scheme** for splits and merges (see also Frühwirth-Schnatter 2006, Chapter 5, and Viallefont *et al.*2002) starts with the latest interaction of the cluster assignment, C denoted `C.Vec` in the software and sampling uniformly at random sample cases, labeled X_i and X_j. If these are in the same cluster C_ℓ, then create a new cluster, C_{L+1}, and assign one of them to this new cluster, arbitrarily X_j here. For each of the remaining cases in C_ℓ assign them to remain with probability 0.5 and to move to C_{L+1} with probability 0.5. This is not a key part of the stochastic process, it is simply done to create a "launch state," ($\mathcal{C}_{launch}$) for a small set of Gibbs draws. Now for k in 1 to K_ℓ units in the first group and k in 1 to K_{L+1} in the second group do:

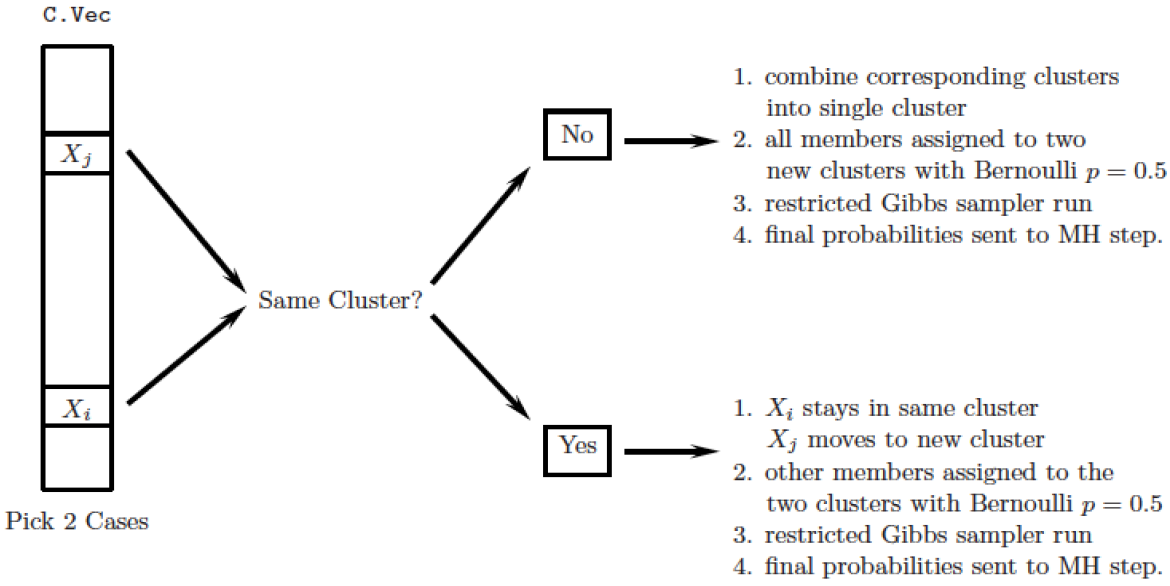
1. draw a cluster move between the groups for unit k based on $p(X_k|X_{-k})$ where $-k$ denotes the other cases in that case's group,

2. make a Gibbs between groups move based on this draw, save $p(X_k|X_{-k})$ and its complement $1 - p(X_k|X_{-k})$,

3. repeat steps 1. and 2. m^* times (depending on hardware capability),

4. final configuration provides a candidate position for the enveloping Metropolis-Hastings algorithm to accept or reject this split of the original cluster.

Figure 2: Illustration of Jains-Neal



The initial transition probability is $(0.5)^{j-2}$ where j is the size of the *sub*cluster being split. The result of the intermediate Gibbs sampling steps is to produce $\mathcal{C}_{launch}$ and the randomly drawn proposal kernel is $K(\mathcal{C}'|\mathcal{C}, \mathcal{C}_{launch})$. The sampling of $\mathcal{C}'$ is achieved by one more iteration of the restricted Gibbs sampler and the overall transition probability is computed using the product of the individual transition probabilities of the sampled states. This not only produces a transition probability that is larger than $(0.5)^{j-2}$, it also produces sampled $\mathcal{C}'$ with larger (conditional) posterior probability, making $\mathcal{C}'$ more likely to be accepted by a Metropolis correction.

This process is shown graphically in Figure 2. If the X_i and X_j are in different clusters then these two clusters are potentially eliminated (depending final acceptance in the Metropolis-Hastings step), and a new larger cluster is created from their union. Then the above process is performed with the final decision offered to the enveloping algorithm based on a new split of this new cluster. Second, we employ a "Pitman Jump," (Pitman 1976) which is simply a multinomial-Dirichlet draw for the clusters. This simple step is also designed to make large adjustments in order to move between modes efficiently.

We also need to make small moves to fully explore high density areas of the parameter space. First, the simple biased random walk randomly samples one of the existing clusters at some iteration of the Markov chain and then samples a single observation from that cluster. This case is then uniformly randomly assigned to one of the existing clusters (including the original cluster it came from). This done for *sub*clusters as well. For clusters, we also implement restricted Gibbs sampling. At random, two clusters are chosen and then two observations from those clusters. These observations are fixed as remaining in distinct clusters. The rest of the observations in these two clusters are then resampled according to a random scan Gibbs sampler, restricted to being in one of the two clusters.

3.3 Sampling ξ'

Since sampling ξ' is done in a Gibbs sampler fashion, we list the full conditional distributions of all of the parameters in ξ' here. This is done in six steps:

Step 1: Sampling $\boldsymbol{\beta}_0$.

The full conditional is given by:

$$\boldsymbol{\beta}_0|\boldsymbol{\beta}_C, \boldsymbol{\sigma}^2_C, \mathbf{S}, \sigma^2_{\boldsymbol{\beta}}, \mathbf{y} \sim \mathcal{N}\left(\frac{\sum_\ell \sigma_\ell^{-2}\boldsymbol{\beta}_\ell}{\sigma_{\boldsymbol{\beta}}^{-2} + \sum_\ell \sigma_\ell^{-2}}, \mathbf{S}\left(\sigma_{\boldsymbol{\beta}}^{-2} + \sum_\ell \sigma_\ell^{-2}\right)\right). \tag{19}$$

Step 2: Sampling $\sigma^2_{\boldsymbol{\beta}}$.

The full conditional is given by:

$$\sigma^2_{\boldsymbol{\beta}}|\boldsymbol{\beta}_0, \mathbf{S}, a_{\sigma^2_{\boldsymbol{\beta}}}, b_{\sigma^2_{\boldsymbol{\beta}}}, \mathbf{y} \sim \mathcal{IG}\left(\frac{p + a_{\sigma^2_{\boldsymbol{\beta}}}}{2}, \frac{b_{\sigma^2_{\boldsymbol{\beta}}} + \boldsymbol{\beta}_0'\mathbf{S}\boldsymbol{\beta}_0}{2}\right). \tag{20}$$

Step 3: Sampling $\mathbf{S}$.

The scaled precision matrix, $\mathbf{S}$, has a Wishart full conditional distribution according to:

$$\mathbf{S}|C, \boldsymbol{\beta}_C, \sigma_C, \boldsymbol{\beta}_0, \sigma^2_{\boldsymbol{\beta}}, V, \mathbf{y} \sim \mathcal{W}\left(\left(V^{-1} + \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{\beta}_0\boldsymbol{\beta}_0' + \sum_\ell \sigma_\ell^{-2}(\boldsymbol{\beta}_\ell - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_\ell - \boldsymbol{\beta}_0)'\right)^{-1}, (m + 1 + a_{\mathbf{S}})\right). \tag{21}$$

Step 4: Sampling $\mathbf{v}$.

To update V we draw p independent v_i from a gamma full conditional distribution using $\mathbf{S}_{ii}$:

$$v_i|\mathbf{S}, a_{\mathbf{S}}, a_{\mathbf{v}}, b_{\mathbf{v}}, \mathbf{y} \sim \mathcal{G}\left(\frac{a_{\mathbf{S}} + a_{\mathbf{v}}}{2}, \frac{\mathbf{S}_{ii} + b_{\mathbf{v}}}{2}\right). \tag{22}$$

Step 5: Sampling λ.

Sampling the Dirichlet process precision term is aided by a parameter expansion process simplified from that

of Escobar and West (1995). First draw a value γ from the distribution:

$$\gamma|\lambda, \mathbf{A}, \mathbf{y} \sim \text{Beta}(\lambda, n). \tag{23}$$

The full conditional distribution of λ is now given by:

$$\lambda|\gamma, \mathbf{A}, \mathbf{y} \sim \mathcal{G}(a_\lambda + k, b_\lambda - \log(\gamma)). \tag{24}$$

Step 6: Sampling τ.

Sampling τ (the constant variance term in the distribution of $\boldsymbol{\eta}$) is also a Gibbs step with the full conditional:

$$\tau|\mathbf{A}, \boldsymbol{\eta}, \mathbf{y} \sim \mathcal{IG}\left(\frac{k + a_\tau}{2}, \frac{b_\tau + \sum_{j=1}^{k} \eta_j^2}{2}\right) \tag{25}$$

At this point the Gibbs sampler has a complete draw for ξ', for a single step.

3.4 Sampling $(\mathcal{C}, \boldsymbol{\beta}_\mathcal{C}, \boldsymbol{\sigma}^2_\mathcal{C})$

$(\mathcal{C}, \boldsymbol{\beta}_\mathcal{C}, \boldsymbol{\sigma}^2_\mathcal{C})$ is sampled in a RJMCMC step conditioned on $\{\xi', (\mathbf{A}, \boldsymbol{\eta})\}$. Here we use a Metropolis step with target distribution given from Bayes' Law (up to proportion) by:

$$\pi(\mathcal{C}|\xi', \boldsymbol{\eta}, \mathbf{A}, \lambda, \mathbf{y}) \propto \left(\prod_{\ell=1}^{m} f_\ell\left(\mathbf{y}_\ell|\xi', \boldsymbol{\eta}, \mathbf{A}, \lambda\right)\right) P(\mathcal{C}), \tag{26}$$

with $p(\mathcal{C})$ indicating the hierarchical uniform prior on $\mathcal{C}$. Within each $\ell = 1, \ldots, m$ cluster, the n_ℓ outcomes, $\mathbf{y}_\ell$ have a Students-t distribution unique to that cluster:

$$f_\ell(\mathbf{y}_\ell|\xi', \boldsymbol{\eta}, A, \lambda) = \mathcal{MVT}_{n_\ell}\left(\mathbf{y}\middle| a_{\sigma^2}, \hat{\mu}_{\mathbf{y}_\ell}, \hat{\boldsymbol{\Sigma}}_{\mathbf{y}_\ell}\right) \tag{27}$$

with:

$$\hat{\mu}_{\mathbf{y}_\ell} = \mathbf{A}_\ell\boldsymbol{\eta} + \mathbf{X}_\ell\boldsymbol{\beta}_0 \qquad\qquad \hat{\boldsymbol{\Sigma}}_{\mathbf{y}_\ell} = \frac{b_{\sigma^2}}{a_{\sigma^2}}\left(\mathbf{I}_{n_\ell} + \mathbf{X}_\ell\mathbf{S}^{-1}\mathbf{X}_\ell'\right).$$

Notice that this draw incorporates both *sub*cluster information from the Dirichlet process as well as substantive clustering information from the ℓ grouping.

The cluster specific parameters $(\boldsymbol{\beta}_\mathcal{C}, \boldsymbol{\sigma}^2_\mathcal{C})$ are directly drawn from their joint conditional distribution with each $(\boldsymbol{\beta}_\ell, \sigma^2_\ell)$ being distributed as

$$\begin{aligned}
\sigma^2_\ell|\mathbf{y}_\ell, \mathcal{C}, \mathbf{A}_\ell, \boldsymbol{\eta}, \boldsymbol{\beta}_0, \mathbf{S} \sim \mathcal{IG}\Bigg(&\frac{n_\ell + a_{\sigma^2}}{2}, \\
&\frac{b_{\sigma^2}}{2} + \frac{1}{2}\left(\mathbf{y}_\ell - \mathbf{A}_\ell\boldsymbol{\eta} - \mathbf{X}_\ell\boldsymbol{\beta}_0\right)'\left(\mathbf{I}_{n_\ell} - \mathbf{X}_\ell(\mathbf{X}_\ell'\mathbf{X}_\ell + \mathbf{S})^{-1}\mathbf{X}_\ell'\right)\left(\mathbf{y}_\ell - \mathbf{A}_\ell\boldsymbol{\eta} - \mathbf{X}_\ell\boldsymbol{\beta}_0\right)\Bigg) \\
\boldsymbol{\beta}_\ell|\mathbf{y}_\ell, \mathcal{C}, \mathbf{A}_\ell, \boldsymbol{\eta}, \boldsymbol{\beta}_0, \mathbf{S}, \sigma^2_\ell \sim \mathcal{N}&\left(\left(\mathbf{X}_\ell'\mathbf{X}_\ell + \mathbf{S}\right)^{-1}\left(\mathbf{X}_\ell'\left(\mathbf{y}_\ell - \mathbf{A}_\ell\boldsymbol{\eta}\right) + \mathbf{S}\boldsymbol{\beta}_0\right), \sigma^2_\ell\left(\mathbf{X}_\ell'\mathbf{X}_\ell + \mathbf{S}\right)^{-1}\right)
\end{aligned}$$

At this point the Gibbs sampler has a drawn complete version of the vector ξ', as well as $(\mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}})$ during this single step of the Markov chain.

3.5 Sampling $(\mathbf{A}, \boldsymbol{\eta})$

Sampling of $(\mathbf{A}, \boldsymbol{\eta})$ is also achieved using a RJMCMC step. The key to this update is determining $\pi(\mathbf{A}'|\xi', \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \mathbf{y})$ in order to compute the Metropolis correction. The prior for $\mathbf{A}$ is the Ewens-Pitman prior:

$$p(\mathbf{A}|\lambda) = \lambda^K \frac{\Gamma(\lambda)}{\Gamma(\lambda + n)} \prod_{k=1}^{K} \Gamma(\#\{i \ : \ y_i \text{ is in } \textit{sub}\text{cluster } k\}) \tag{28}$$

The likelihood for $\mathbf{y}$ can be expressed in the form of an integral over $\boldsymbol{\eta}$:

$$f(\mathbf{y}|\mathbf{A}, \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \xi') = \int \prod_{\ell=1}^{m} f_\ell(\mathbf{y}_\ell|\mathbf{A}_\ell, \mathbf{X}_\ell, \mu_\ell, \boldsymbol{\beta}_\ell, \sigma^2_\ell, \tau^2, \boldsymbol{\eta}) \pi(\boldsymbol{\eta}|\tau^2) d\boldsymbol{\eta}.$$

Now the acceptance ratio can be computed as soon as we determine the value of $f(\mathbf{y}|\mathbf{A}, \mathcal{C}, \boldsymbol{\beta}_{\mathcal{C}}, \boldsymbol{\sigma}^2_{\mathcal{C}}, \xi')$. This is relatively easy in the regular GLMDM as there are only the *sub*clusters to consider within each cluster. In order to do this for the substantively clustering GLMDM, we can recognize that it will be normal and determine the appropriate mean and covariance structure. In fact, the mean vector is $\mathbf{X}_\ell \boldsymbol{\beta}_\ell$ in cluster ℓ. The precision structure (using clusters as blocks) is given by:

$$\Omega_{\ell,\ell'} = \frac{\mathbf{I}_{n_\ell} \delta_\ell(\ell')}{\sigma^2_\ell} - \frac{\mathbf{A}_\ell}{\sigma^2_\ell} \left(\frac{\mathbf{I}_k}{\tau^2} + \sum_{\iota=1}^{m} \frac{\mathbf{A}'_\iota \mathbf{A}_\iota}{\sigma^2_\iota} \right)^{-1} \frac{\mathbf{A}'_{\ell'}}{\sigma^2_{\ell'}} \tag{29}$$

where $\delta_\ell(\ell')$ denotes the Kronecker delta function (1 if the variables are equal and 0 otherwise). This is simply the rows one gets from the precision for y after integrating out η (note that we have conditioned here on the cluster specific parameters). If we define $m_{j,\ell}$ to be the number of observations from cluster ℓ with random effect η_j from the Dirichlet process, then:

$$\begin{aligned}
\left(\frac{\mathbf{I}_k}{\tau^2} + \sum_{\iota=1}^{m} \frac{\mathbf{A}'_\iota \mathbf{A}_\iota}{\sigma^2_\iota} \right)^{-1} &= \left(\frac{\mathbf{I}_k}{\tau^2} + \sum_{\iota=1}^{m} \frac{\text{diag}(m_{j,\iota})}{\sigma^2_\iota} \right)^{-1} \\
&= \left(\text{diag} \left(\frac{1}{\tau^2} + \sum_{\iota=1}^{m} \frac{m_{j,\iota}}{\sigma^2_\iota} \right) \right)^{-1} \\
&= \text{diag} \left(\left(\frac{1}{\tau^2} + \sum_{\iota=1}^{m} \frac{m_{j,\iota}}{\sigma^2_\iota} \right)^{-1} \right)
\end{aligned}$$

which we label as $\text{diag}(\phi_j)$. Therefore the variance of a particular y_i is

$$\frac{\sigma^4_{\ell[i]}}{\sigma^2_{\ell[i]} - \phi_{j[i]}} \tag{30}$$

and the cross-precision between two observations y_i and $y_{i'}$ is

$$-\frac{\phi_{j[i]}\delta_{j[i]}(j[i'])}{\sigma^2_{\ell[i]}\sigma^2_{\ell[i']}} \tag{31}$$

where again $\delta_{j[i]}(j[i'])$ denotes the Kronecker delta function. Unfortunately, there is no easy analytical way to invert this matrix and so we are left with inverting it numerically, which is computationally costly due to sparsity of the matrix whenever there are a large number of *sub*clusters. However, we can also notice that all we need to compute the normal random variable is the determinant of this matrix and an appropriate inner product. The inner product is not too difficult to compute, but the determinant still requires an arduous computation.

We can first sample $\mathbf{A}$ (since the acceptance ratio does not depend on $\boldsymbol{\eta}$). Then we can sample $\boldsymbol{\eta}$ given the accepted $\mathbf{A}$. Its full conditional is given by:

$$\boldsymbol{\eta}|\mathbf{A},\mathbf{q}^{(t+1)},\boldsymbol{\theta}^{(t)},\mathbf{y} \sim \mathcal{N}_k\left(\left(\sum_{\ell=1}^{m}\frac{\mathbf{A}'_\ell\mathbf{A}_\ell}{\sigma^2_\ell}+\frac{\mathbf{I}_k}{\tau^2}\right)^{-1}\left(\sum_{\ell=1}^{m}\frac{\mathbf{A}'_\ell(\mathbf{y}_\ell-\mathbf{1}_{n_\ell}\mu_\ell-\mathbf{X}_\ell\boldsymbol{\beta}_\ell)}{\sigma^2_\ell}\right),\left(\sum_{\ell=1}^{m}\frac{\mathbf{A}'_\ell\mathbf{A}_\ell}{\sigma^2_\ell}+\frac{\mathbf{I}_k}{\tau^2}\right)^{-1}\right) \tag{32}$$

This draw for $\boldsymbol{\eta}$ is the final step in a single iteration of the Markov chain.

3.6 Detailed Balance and Ergodicity

At each stage of the sampler, detailed balance is ensured because the sampling is either Gibbs or a Metropolis correction is employed. For sampling $\mathcal{C}$ and $\mathbf{A}$, we employ a number of different proposal kernels. At each iteration, a kernel is chosen at random. As discussed in Tierney (1994), using a Metropolis correction on the selected kernel produces a chain which maintains detailed balance. One concern that arises is whether the composite chain is ergodic. As discussed in Jain and Neal (2004), the chain is ergodic since each proposal kernel is ergodic and has non-zero probability of leaving the state unchanged. Meaning we know that this is an aperiodic, Harris recurrent (there is σ-finite probability measure ψ for $(\Omega,\mathcal{F})$ such that at time n it has the property: $\psi(A)>0,\ \forall A\in\mathcal{F}$), and is therefore and ergodic Markov chain.

4 Application: Terrorism Data Analysis

The health and security of millions of people around the world depends on the understanding of the connections that exist in covert networks, especially terrorist networks. Terrorism is an important problem because it affects internal government policy, public perception, relations between states, and of course, personal safety. To protect citizens, governments and nongovernmental organizations invest enormous amounts of time and energy to understand and to thwart terrorist attacks. This remains a challenging social, political, and military problem because many of the variables that we would like to see are unobservable under even the most highly

visible situations. The study of terrorism has not made enough *empirical* progress due to inherent problems in the available data.

Data problems include: non-granular discrete measurement, insufficient explanatory variables, and the lack of access to classified collections. Another key problem is that there are unmeasured clusters in almost all terrorism data. These groups have a natural affinity for dispersion and segmentation, either for ideological or obvious tactical reasons. Some progress has been made: these groups are usually not unitary actors (Chai 1993, Crenshaw 1981), they tend to be cellular and distributed rather than hierarchical (Carley 2004, Krebs 2002, Rothenberg 2002), and they adapt over time (Carley 2003, 2006). We *know* there are clusters within terrorist organizations, even if they are not visible, since terrorists are known to be collaborative, imitative, and fluid. Kyung (2011) showed that Dirichlet Process Priors on random effects can improve regression models with this type of data by accounting for the heterogeneity from latent clustering. But, as pointed out before, the resulting *sub*clusters are not the substantive clusters of actual interest.

An additional problem with much of the empirical/statistical literature on terrorism is its focus on "events data," data analysis where the outcome variable is an attack and the explanatory variables describe the individuals involved, the location, the target, and the means used. Thus the datasets created are selecting on the observable outcome and ignore failed attacks, aborted attacks, and planned events. Obviously this is a necessary encumbrance since terrorist groups and governments have motivations to hide their activities. This has driven more work in terrorist networks (Tsvetovat and Carley 2006), and formal/mathematical modeling of psychology and motivation (Bueno de Mesquita 2005).

The model estimated here improves on the empirical study of terrorism since it addresses the main problems just listed: the challenge in producing real clustering, and not selecting on the observable outcome merely out of convenience. We use the product partition technology just developed in conjunction with Dirichlet process priors to see latent features of importance with real data on terrorist organizations worldwide.

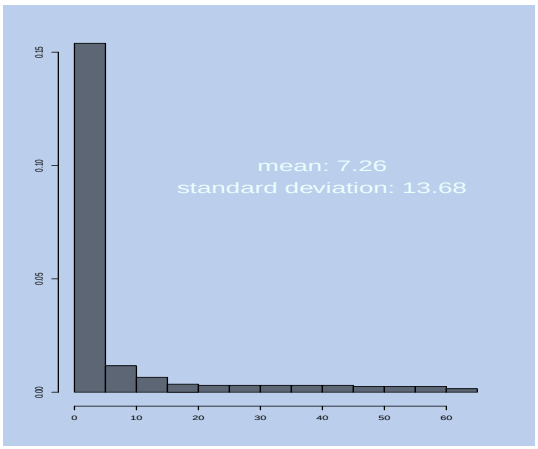
4.1 Data and Model

The approach taken here is to look at groups rather than individual events using the *Big Allied and Dangerous* (BAAD) Database 1 (Asal, Rethemeyer and Anderson 2008). This aggregates worldwide lethal attacks from 1998-2005 by terrorist organizations recording variables describing: geography, ideology, group size, organizational structure, funding sources, and network information on allies and state sponsors. These data were assembled from several established databases: Memorial Institute for the Prevention of Terrorism's (MIPT) Terrorism Knowledge Base (TKB), Correlates of War (COW), Polity, and Polity2. These are standard sources for terrorism events, international conflict between nations, and regime characteristics (respectively). Thus the BAAD dataset is not an original collection but an assembly of variables from reliable and routinely used resources, assembled with careful quality control.

All terrorism data comes with problems due to the nature of the actors being studied. The creators of the BAAD data are careful to identify the potential shortcomings. Many of the described events come from media reports, which may introduce measurement problems inherent in journalistic reporting. The

data miss about one-half of the recorded events worldwide since events for which no party claims responsibility, there is no data on the claimant, or the perpetrator is not an organized group. Strictly religious events are also excluded by Asal, Rethemeyer and Anderson (2008), since they deem these attacks to be fundamentally different in nature. Finally, we use the version of their dataset that excludes Al Qaeda since its scope, profile, and effectiveness place it in a unique category. This increases the generality of our findings but obviously excludes a pivotal event in the modern history of terrorism.

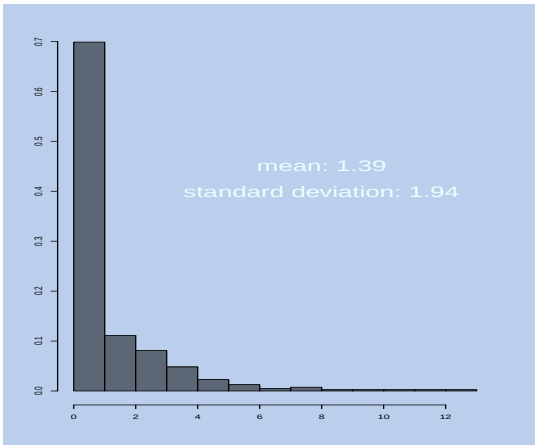
Figure 3: Histogram of Fatalities from BAAD Groups, 1998-2004



The data provide 395 cases (each a terrorist attack) and 16 possible explanatory variables. We use `fatalities` as the outcome variable to focus on the primary objective of these attacks. See the figure at the right. Even though these attacks are intended to cause personal harm through violence, it is important to note that the real underlying goal for terrorist organizations is not actually the death and mayhem that results from such attacks. This is only an intermediate objective and their actual goal is undermining citizens' confidence in their government's ability to protect them (Hoffman 1988).

Terrorist groups can be cleanly classified according to their sponsorship status. The variable `statespond` indicates whether the group is financially or logistically supported by one or more recognized governments (coded 1, $n_1 = 32$), or not (coded 0, $n_0 = 363$). Hezbollah is a archetypal example of a state-sponsored terrorist group since they receive support from Iran and Syria. The home-base of these groups can be important, either locally or regionally, so we include the variable `masterccode` which denotes the COW `CCODE` value. The size of the groups' membership is given by the ordinal measure `ordsize`, according to 0 for less than 100 members ($n_0 = 261$), 1 for 101-1,000 members ($n_1 = 77$)), 2 for 1,001-10,000 members ($n_2 = 45$), and 3 for more than 10,000 members ($n = 12$). Another important delineator for these groups is whether they control some land and therefore operate with impunity in some geographic area. The variable `terrStrong` is coded 1 ($n_1 = 43$) if they possess territory and 0 if they do not ($n_0 = 352$). There is evidence that interconnections between terrorist groups affects their endurance and effectiveness, and `degree` gives a count of alliance connections in the network sense, where the histogram

Figure 4: Histogram of Alliances from BAAD Groups, 1998-2004



of these is shown in the figure above left.

Religion and ethnicity are obviously big drivers of modern terrorism, so we include four variables from a longer list that potentially highlight these effects. We use the variable `LeftNoReligEthno` where a 1 indicates that the group's ideology is leftist and it is not compounded with another ideological orientation ($n_1 = 94$), and a 0 indicates that group's ideology is either not leftist or is a mix of leftist and at other ideological dimensions ($n_0 = 301$). The variable `PureRelig` indicates with a 1 whether the group's ideology is purely religious and not associated with other political or social factors ($n_1 = 50$), and 0 otherwise ($n_0 = 345$). Similarly, the variable `PureEthno` indicates with a 1 whether the group is ethnonationalist (nationalist causes tied to ethnic identity) and not associated with other ideological factors ($n_1 = 26$), and 0 otherwise ($n_0 = 369$). Finally, since modern terrorism is strongly tied to history and politics in the Middle East, we also include a variable `Islam` where a 1 is assigned to groups inspired by some form of Islam ($n_1 = 287$) and 0 otherwise.

4.2 Model Results

We estimate the DPP/Product Partition model using the sampler developed in Section 3. The Markov chain is run for 22,500 iterations disposing of the first 2,500 as burn-in. Convergence was assessed with `superdiag`, a diagnostic suite provided by an `R` package (Tsai and Gill 2012) that calls all of the conventional convergence diagnostics typically used (Gelman & Rubin, Geweke, Heidelberger & Welch, Raftery & Louis). We also found no evidence of non-convergence with standard graphical tools (traceplots, cumsum diagrams, etc.).

Table 1: Selected Covariate Percentage for Modal Partition

	Cluster 1		Cluster 2		Cluster 3	
Covariate Value:	0	1	0	1	0	1
`statespond`	262 (96%)	10 (4%)	51 (78%)	14 (22%)	50 (86%)	8 (14%)
`terrStrong`	257 (94%)	15 (6%)	43 (66%)	22 (34%)	52 (90%)	6 (10%)
`LeftNoReligEthno`	196 (72%)	76 (28%)	52 (80%)	13 (20%)	53 (91%)	5 (9%)
`PureRelig`	249 (92%)	23 (8%)	54 (83%)	11 (17%)	42 (72%)	16 (28%)
`PureEthno`	256 (94%)	16 (6%)	60 (92%)	5 (8%)	53 (91%)	5 (9%)
`Islam`	223 (82%)	49 (18%)	37 (57%)	28 (43%)	27 (47%)	31 (53%)

During the stochastic search, 5,681 unique partitions were visited. In total, the sampler spent 581 iterations in partitions with 3 clusters, 18837 iterations with 4 clusters, and 3082 iterations with 5 clusters. There were no observed partitions with less than three clusters. The highest posterior probability ($p(C|\mathbf{y}) = 0.8543$) partition had three clusters with the counts: $[272, 65, 58]$. In contrast, the next four partition probabilities were: 0.0989, 0.0129, 0.0055, 0.0031. The top 15 partitions all have 3 clusters and the prior on the space of clusters behaves as expected. The hierarchical uniform prior appropriately penalizes these partitions for their complexity, which would not have occurred under the uniform prior on the space of clusters.

Table 1 uses the highest probability partition to evaluate covariate differences between the three clusters. For this partition, the individual clusters have fundamentally different, and substantively interesting, covariate mixes indicating non-independence of explanatory effects between clusters. Cluster 1 is almost strictly free of government sponsorship where the other two are noticeably more in receipt of such support. Cluster 2 seems

to be identified strongly with holding actual territory. Notice also that Cluster 3 has many fewer cases where the group's ideology is leftist and it is not compounded with another ideological orientation (9% versus 28% for Cluster 1 and 20% for Cluster 2). Finally, Cluster 1 is significantly less made of cases that are inspired by some form of Islam. As a whole, Table 1 provides evidence that this partition reveals substantively important cluster differences, which provide potentially useful information for policy-makers engaged in anti-terrorist planning.

4.3 Analyzing A Single Result

A key question is what modeling strategy should be pursued now that there is reliable posterior clustering information. Recall that the simultaneous *product partition clustering/Dirichlet process prior* sub*clustering* model used the explanatory variable specification to produce a sample of partitions from the MCMC draws. Now we would like to use that information to produce an analogous regression model to make substantive inferences with the same mix of right-hand side covariates. Given the selection of one of the partitions, three canonical models are available to us: a "fully unpooled" form with separate specifications for the clusters in a selected partition, a "fully pooled" model that ignores clustering information, and a hierarchical specification where the estimated clusters in the chosen partition form the group level definitions.[2] Since choosing a preferred partition from the product partition/Dirichlet process sampler fully defines how these models are identified with regard to group (partition) identification, each of them can easily run with existing samplers. The fully unpooled model is run with the generalized linear mixed Dirichlet model (GLMDM) on the partition with the highest posterior probability noted above. These results give separate GLMDM models corresponding to the three separate clusters in this partition and displayed in Table 2. The fully pooled model and the multilevel model are estimated with the `jags` MCMC estimation software (`http://www-ice.iarc.fr/~martyn/software/jags/`). We apply diffuse conjugate priors distributions for the regression parameters and standard methods for assessing convergence with `superdiag`.

Several of the results from the two models differ in important ways. For the majority of terrorist acts (Cluster 1), the only reliable predictor is the ordinal size of the organization. However, there are important differences in the effect size of `ordsize` between the clusters. The 95% credible intervals for the effects size do not overlap for the different clusters, with the second cluster (which happens to have the largest percentage of Islamic groups performing the acts) having a very large effect due to organization size. This second cluster also contains the acts for which whether the organization has a strong hold (`terrStrong`) is strongly positively associated with the number of fatalities. It is also interesting to note that the degree of association (`degree`) is positively associated with fatalities for the second and third clusters, with the largest positive effect in the second cluster, as is the case with whether the organization is Islamic (`Islam`). We can now give a substantive interpretation to our clusters. The first represents terrorist acts where simply the size of the organization weakly influences the number of fatalities. The second represents terrorist acts where organizational size, associations, territory, and Islamic fundamentalism provide large increases in the number of fatalities. The

[2] We borrow this pooling language from Gelman and Hill (2007), especially Chapter 12.

Table 2: Cluster Model Results: BAAD Data

		95% CI Lower	0.25 Quantile	Median	0.75 Quantile	95% CI Upper
Cluster 1, $n = 272$	(Intercept)	-0.402	0.408	0.911	1.330	2.047
	statespond	-0.596	-0.301	-0.156	-0.013	0.265
	masterccode	-0.007	-0.003	-0.001	0.001	0.005
	ordsize	0.084	0.174	0.220	0.269	0.361
	terrStrong	-0.378	-0.130	-0.001	0.126	0.370
	degree	-0.003	0.035	0.055	0.075	0.114
	LeftNoReligEthno	-0.253	-0.129	-0.065	-0.001	0.126
	PureRelig	-0.204	0.021	0.141	0.258	0.486
	PureEthno	-0.444	-0.223	-0.108	0.010	0.235
	Islam	-0.270	-0.067	0.030	0.132	0.328
Cluster 2, $n = 65$	(Intercept)	16.219	23.906	27.617	31.418	38.901
	statespond	-8.182	-3.758	-1.643	0.559	4.734
	masterccode	-0.412	-0.311	-0.260	-0.210	-0.104
	ordsize	4.767	6.921	8.094	9.247	11.556
	terrStrong	0.408	4.547	6.738	8.954	13.351
	degree	0.146	0.794	1.138	1.459	2.097
	LeftNoReligEthno	-12.387	-7.238	-4.701	-2.134	2.490
	PureRelig	-2.737	1.620	3.947	6.260	10.908
	PureEthno	-14.041	-7.412	-4.186	-1.163	4.979
	Islam	4.551	8.868	11.102	13.352	17.678
Cluster 3, $n = 58$	(Intercept)	0.827	2.675	3.613	4.497	6.311
	statespond	-2.040	-0.900	-0.333	0.214	1.326
	masterccode	-0.060	-0.031	-0.016	-0.001	0.028
	ordsize	0.396	0.943	1.221	1.495	2.041
	terrStrong	-1.244	0.044	0.694	1.398	2.817
	degree	0.039	0.305	0.438	0.576	0.845
	LeftNoReligEthno	-2.269	-0.807	-0.137	0.539	1.982
	PureRelig	-0.851	0.179	0.699	1.231	2.316
	PureEthno	-1.466	-0.064	0.622	1.384	2.924
	Islam	0.395	1.521	2.096	2.682	3.788

third represents acts where size, associations, and Islamic fundamentalism have a positive but less pronounced effect. Finally, four marginal posteriors are disappointingly centered near zero with large variance: those for `statespond`, `LeftNoReligEthno`, `PureRelig`, and `PureEthno`. This suggests that the most reliable effect in terms of religion and (associated) ethnicity is associated with Islamic groups rather than others.

This analysis so far falls under the completely unpooled approach in the multilevel modeling sense where the three clusters are considered as completely separate collections, each deserving its own model. As a contrast we run a standard Bayesian multilevel linear model (diffuse proper priors) with the three estimated clusters as group definitions. Table 3 gives these results (using `jags`) and a comparison with a standard non-hierarchical Bayesian linear model. Here τ is the precision of the residuals (standard linear model), or the precision of the between group categories (multilevel linear model).

The completely pooled standard (Bayesian) linear model in the left-hand side of Table 3 serves only as a benchmark analysis since the clusters are completely ignored. We see reliable explanatory effects for the coefficient posteriors: `ordsize`, `terrStrong`, `degree`, and `Islam`. This naïve model suggests that: larger groups are more deadly, those holding territory are also more deadly, as are those that are more networked and also those with an Islamic identity. However, the more principled multilevel model, that does not ignore the

Table 3: Flat and Hierarchical Model Results: BAAD Data

	Standard Linear Model						Multilevel Linear Model				
Quantiles:	0.025	0.25	0.5	0.75	0.975		0.025	0.25	0.5	0.75	0.975
α	-2.842	-1.171	-0.303	0.566	2.255	α_1	-5.842	-4.765	-4.191	-3.617	-2.512
						α_2	16.756	18.228	19.001	19.771	21.237
						α_3	-4.290	-2.953	-2.246	-1.545	-0.200
`statespond`	-1.810	-0.313	0.492	1.302	2.848		2.140	3.207	3.760	4.323	5.403
`masterccode`	-0.056	-0.015	0.007	0.029	0.070		-0.075	-0.051	-0.038	-0.025	0.000
`ordsize`	3.328	4.256	4.744	5.223	6.152		1.723	2.300	2.606	2.911	3.489
`terrStrong`	1.194	2.942	3.850	4.755	6.492		0.299	1.534	2.188	2.848	4.095
`degree`	1.727	2.108	2.309	2.511	2.898		0.870	1.095	1.214	1.333	1.562
`LeftNoReligEthno`	-1.809	-0.437	0.287	1.013	2.389		-0.590	0.306	0.782	1.258	2.175
`PureRelig`	-1.451	0.239	1.117	2.003	3.692		-0.563	0.671	1.309	1.950	3.166
`PureEthno`	-3.726	-1.913	-0.958	-0.005	1.809		-3.501	-2.169	-1.461	-0.758	0.579
`Islam`	0.473	2.022	2.839	3.646	5.182		1.188	2.284	2.853	3.425	4.527
τ	0.007	0.008	0.009	0.009	0.010		0.023	0.026	0.027	0.029	0.031
	Summed Deviance: 3001						Summed Deviance: 2547				
							Variance		Std.Dev.		
						σ_α:	37.037		6.086		
						σ_y:	1.31		1.15		

high posterior probability clusters identified above, tells a more informed story. In this second specification in the right-hand side of Table 3, each of these four explanatory effects is also statistically reliable, but in all cases the posterior median is noticeably lower, suggesting that the model ignoring clusters exaggerates the impact of the associated variables. In addition, the posterior distribution for `statespond` was not 95% bounded away from zero in the first model, but it a large and reliable effect in the second model. This is consistent with a large body of descriptive literature on terrorist organizations: having a government benefactor is instrumental in securing resources for attacks, recruitment, and defensive operations.

We believe that these results can be very helpful to policy-makers concerned with controlling and responding to international terrorism. The largest cluster ($n = 272$) is majority non-Islam: 223 to 49, and majority not of a single (pure) religious base: 249 to 23, and majority not state sponsored: 260 to 10, whereas the other clusters contain a much higher percentage of Islamic groups and are more likely to be state sponsored. Since the results point strongly to Islamic terrorist groups being more deadly, perhaps these cluster is a lower priority to policy-makers, perhaps based on cost to defeat. However, if policy makers can effectively isolate these groups, the deadly nature of their attacks can be mitigated. Forcing terrorist groups in the second cluster out of their controlled territory would also help to reduce the number of fatalities due to these groups.

5 Conclusion

In this paper we have proposed an innovative model-based Bayesian clustering approach which not only provides substantive clustering through the product partition model, but also incorporates Dirichlet process random effects to further account for individual level variation. The model thus provides a means of clustering observations based on response to covariates while relaxing assumptions about the residual structure. Since the

partition posterior probabilities cannot be analytically calculated, we have proposed an MCMC algorithm that alternates between sampling substantive clusters and Dirichlet process *sub*clusters. The partition posterior probabilities are calculated using a Rao-Blackwell style estimator using the entire MCMC output. This estimator allows for reasonable estimation of the partition posterior probabilities. Beyond the important substantive analysis of terrorism data, this paper also provides an interesting insight into the hierarchical uniform prior on the space of partitions. In contrast to the uniform prior, the hierarchical uniform prior (HUP) provides an appropriate complexity penalization to allow the identification of a small number of substantive clusters.

6 References

Victor Asal, R. Karl Rethemeyer, Ian Anderson. (2009). Big Allied and Dangerous (BAAD) Database 1 - Lethality Data, 1998-2005, website.

Barry, D. and Hartigan, J. A. (1992). Product Partition Models for Change Point Problems. *Annals of Statistics* **20**, 260-279.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics* **1**, 353-355.

Booth, J. G., Casella, G. and Hobert, J. P. (2008). Clustering Using Objective Functions and Stochastic Search. *Journal of the Royal Statistical Society, Series B* **70**, 119-140

Bueno de Mesquita, Ethan. (2005). Conciliation, Counterterrorism, and Patterns of Terrorist Violence. *International Organization* **59**, 145-176

Carley, Kathleen M. (2003). Dynamic Network Analysis. In *Dynamic Social Network Modeling and Analysis*, R. Breiger, K. Carley, and P. Pattison (Eds.), pp. 1-13. Committee on Human Factors. National Research Council. Washington D.c. The National Academies Press.

Carley, Kathleen M. (2004). Estimating Vulnerabilities in Large Covert Networks Using Multi-Level Data. *Proceedings of the NAACSOS Conference*. Pittsburgh, PA.

Carley, Kathleen M. (2006). A Dynamic Network Approach to the Assessment of Terrorist Groups and the Impact of Alternative Courses of Action. *Pre-Proceedings, Visualizing Network Information* IST-063/RWS-010. Royal Danish Defence College, Copenhagen.

Casella, G. and Moreno, E. (2006). Objective Bayes Variable Selection. *Journal of the American Statistical Association* **101**, 157-167.

Casella, G., Giron, F. J. and Moreno, E. (2009). Consistent Model Selection in Regression. *Annals of Statistics* **37**, 1207-1228.

Casella, G., Moreno, E. and Giron, F. J. (2011). Cluster Analysis, Model Selection, and Prior Distributions on Models. University of Florida Technical Report.

Celeux, G. (1998). Bayesian Inference For Mixtures: The Label-Switching Problem. In *Compstat 1998-Proceedings in Computational Statistics*, R. Payne, P. Green (Eds.), pp. 227-232. Physica Verlag, Heidelberg.

Chai, S-K. (1993). An Organizational Economics Theory of Antigovernment Violence. *Comparative Politics* **26**, 99-110.

Chung, H., Loken, E., and Schafer, J. L. (2004). Difficulties In Drawing Inferences With Finite-Mixture Models: A Simple Example With a Simple Solution. *The American Statistician* **58**, 152-158.

Crenshaw, M. (1981). The Causes of Terrorism. *Comparative Politics* **13**, 379-399.

Crowley, E. M. (1997). Product Partition Models for Normal Means. *Journal of the American Statistical Association* **92**, 192-198.

Dahl, D. B. (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In *Bayesian Inference for Gene Expression and Proteomics*, Kim-Anh Do, Peter Müller, Marina Vannucci (Eds.), pp. 201-218. Cambridge: Cambridge University Press.

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90**, 577-588.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics* **1**, 209-230.

Fraley, Chris and Raftery, Adrian E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* **97**, 611-631.

Frühwirth-Schnatter, Sylvia. (2006). Finite Mixture and Markov Switching Models New York: Springer-Verlag.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

George, E. I. and McCulloch, R.E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* **88**, 881-889.

Gill, J. and Casella, G. (2009). Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation. *Journal of the American Statistical Association* **104**, 453-464.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **82**, 711-732.

Green, P. J. and Richardson, S. (2001). Modelling Heterogeneity With and Without the Dirichlet Process. *Scandinavian Journal of Statistics* **28**, 355-375.

Grün, Bettina, and Leisch, Friedrich. (2009). Dealing With Label Switching In Mixture Models Under Genuine Multimodality. *Journal of Multivariate Analysis* **100**, 851-861.

Hartigan, J. A. (1990). Partition Models. *Communications in Statistics* **19**, 2745-2756.

Heard, N. A., Holmes, C. C. and Stephens, D. A. (2006). A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Application of Bayesian Hierarchical Clustering of Curves. *Journal of the American Statistical Association* **101**, 18-29.

Hobert, J. P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association* **91**, 1461-73.

Hoffman, Bruce. (1988). *Inside Terrorism*. New York: Columbia University Press.

Jain, Sonia and Neal, Radford M. (2004). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model *Journal of Computational and Graphical Statistics* **13**, 158-182.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science* **20**, 50-67

Kim, S., Tadesse, M. G. and Vannucci, M. (2006). Variable Selection in Clustering via Dirichlet Process Mixture Models. *Biometrika* **93**, 877-893.

Korwar, R. M. and Hollander, M. (1973). Contributions to the Theory of Dirichlet Processes. *Annals of Probability* **1**, 705-711.

Krebs, Valis E. (2002) Mapping Terrorist Cells. *Connections* **24**, 43-52.

Kyung, M., Gill, J. and Casella, G. (2009). Characterizing the Variance Improvement in Linear Dirichlet Random Effects Models. *Statistics and Probability Letters* **79**, 2343-2350.

Kyung, M., Gill, J. and Casella, G. (2010). Estimation in Dirichlet Random Effects Models. *Annals of Statistics* **38**, 979-1009.

Kyung, M., Gill, J. and Casella, G. (2011). New Findings from Terrorism Data: Dirichlet Process Random Effects *Journal of the Royal Statistical Society, Series C* **60**, 701-721. Models for Latent Groups. *Journal of the Royal Statistical Society, Series C* **60**, 701-721.

Lau, J. W. and Green, P. J. (2007). Bayesian Model-Based Clustering Procedures. *Journal of Computational and Graphical Statistics* **16**, 526-558.

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**, 911-930.

Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *Annals of Statistics* **12**, 351-357.

MacEachern, S. N. and Müller, P. (1998). Estimating Mixtures of Dirichlet Process Model. *Journal of Computational and Graphical Statistics* **7**, 223-238.

Marin, J.-M. Mengersen, K., Robert, C. P. (2005). Bayesian Modelling and Inference On Mixtures of Distributions. In *Bayesian Thinking, Modeling and Computing, Handbook of Statistics* **25**, D. Dey, C. Rao (Eds.), pp. 459-507. New York: Elsevier.

Maugis, C., Celeux, G., and Martin-Magniette. (2009). Variable Selection In Model-Based Clustering: A General Variable Role Modeling. *Computational Statistics & Data Analysis* **53**, 3872-3882.

McCullagh, P. and Yang, J. (2006). Stochastic Classification Models. *International Congress of Mathematicians* **III** 669-686.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **9**, 249-265.

Pan, Wei, and Shen Xiaotong. (2007). Penalized Model-Based Clustering with Application to Variable Selection. *The Journal of Machine Learning Research* **8**, 1145-1164.

Park, J. H. and Dunson, D. B. (2010). Bayesian Generalized Product Partition Model. *Statistica Sinica* **20**, 1203-1226.

Pitman, J. W. (1976). On Coupling of Markov Chains. *Probability Theory and Related Fields* **4** 315-322.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society, Series B* **65**, 557-574.

Raftery. A. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association* **101**, 168-178.

Richardson, S., Green, P. J. (1997). On Bayesian Analysis of Mixtures With an Unknown Number of Components. *Journal of the Royal Statistical Society, Series B* **59**, 731-792.

Richardson, S. and Green, P. J. (2002). On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B* **59**, 731-792.

Robert, C. P. and Casella, G. (2004). Monte Carlo Statistical Methods. Second Edition. New York: Springer-Verlag

Robert, C. P. and Casella, G. (2009). Introducing Monte Carlo Methods with `R`. New York: Springer-Verlag

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008) The Nested Dirichlet Process (with discussion). *Journal of the American Statistical Association* **103**, 1131-1154.

Rothenberg, R. (2002). From Whole Cloth: Making up the Terrorist Network. *Connections* **23**, 36-42.

Roy, V. and Hobert, J.P. (2007). Convergence Rates and Asymptotic Standard Errors for Markov Chain Monte Carlo Algorithms for Bayesian Probit Regression. *Journal of the Royal Statistical Society, Series B* **69**, 607-623.

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica* **4**, 639-650.

Sperrin, M, Jaki, T., and Wit, E. (2010). Probabilistic Relabelling Strategies For the Label Switching Problem in Bayesian Mixture Models. *Statistics and Computing* **20**, 357-366.

Steele, Russell J. and Raftery Adrian E. (2009). Bayesian Model Selection and Hypothesis Tests. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, Ming-Hui Chen, Dipak K. Dey, Peter Müller, Dongchu Sun and Keying Ye (Eds.), pp. 113-155. New York: Springer-Verlag.

Stephens, M. (2000). Dealing With Label Switching in Mixture Models. *Journal of the Royal Statistical Society, Series B* **62**, 795-809.

Tadesse, M. G., Sha, N. and Vannucci, M. (2005). Bayesian Variable Selection in Clustering High-Dimension Data. *Journal of the American Statistical Association* **100**, 602-617.

Tsai, Tsung-han. and Gill, Jeff. (2012). `superdiag`: A Comprehensive Test Suite for Markov Chain Non- Convergence. *The Political Methodologist* **19** (Spring), 12-18.

Tsvetovat, M. and Carley, K. M. (2006). Improving Effectiveness of Communications Sampling of Covert Networks. *E-Social Science Conference* Manchester, UK.

Viallefont, Valérie, Richardson, Sylvia, and Green, Peter J. (2002). Bayesian Analysis of Poisson Mixtures. *Journal of Nonparametric Statistic* **14**, 181-202.

Wang, S. and Zhu, J. (2008). Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. *Biometrics* **64**, 440-448.

Xie, Benhuai, Pan, Wei, and Shen, Xiaotong. (2008). Penalized Model-Based Clustering with Cluster-Specific Diagonal Covariance Matrices and Grouped Variables. *Electronic Journal of Statistics* **2**, 168-212.

Xie, B., Pan, W., and Shen, X. (2008). Variable Selection in Penalized Model-Based Clustering Via Regularization on Grouped Parameters. *Biometrics* **64**, 921-930.

Zhong, Shi and Ghosh, Joydeep. (2003). A Unified Framework For Model-Based Clustering. *The Journal of Machine Learning Research* **4**, 1001-1037.