D Project Description

D.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 hide 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 humans 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. 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, allowing for some incorporation of prior information, but retaining some vagueness in the form of the prior. The nonparametric component of this method produced an enhanced understanding of the agency environment that was not directly available by conventional means. Such information can be thought of as latent clustering in the data.

This proposal outlines plans to substantially improve the current state of clustering algorithms using *Generalized Linear Mixed Dirichlet Models* (GLMDM). Our key objectives are to:

- ▷ Improve understanding of latent clustering effects that are pervasive in social science datasets, notably with empirical studies of terrorism.
- ▷ Adapt GLMDM algorithms such that the subclustering assignments in the Gibbs sampler lead to substantive substantive clusters of interest using posterior probability.
- ▷ Develop an algorithmic approach that directly includes variable selection within clusters into a general clustering model.
- ▷ Speed up the simultaneous clustering and variable selection process by parallelization.
- ▷ Distribute this technology as an easy-to-use `R` package for general use by others.

We believe that this project has the following importance.

- ▶ The **Intellectual Merit** is in establishing a new paradigm for using Bayesian nonparametric approaches to produce clustering based on posterior probability. This approach promises to improve current designs since it inherently includes fit and ovefit criteria in the context of Bayesian posterior probability. This is not possible without new Bayesian stochastic simulation tools. Our application to the empirical study of terrorism is unique and will substantially improve our understanding of these groups.

- The **Broader Impact** comes from facilitating the development of more sophisticated modeling frameworks across social and behavioral sciences thus helping researchers understand complex phenomena in new and difficult datasets with measurement challenges. The algorithmic developments, which we will develop and disseminate widely, can be applied in any scientific field and will contribute to the statistical literature on Markov chain Monte Carlo. The development of nonparametric clustering algorithms promises to substantially improve the current state of data clustering. This project also aids in the intellectual development of both graduate *and* undergraduate students as well as post-doctorate researchers, all of which will benefit from the interdisciplinary focus. The work will be in both social/behavioral science journals and statistics journals.
- This is **Transformative Research** because accounting for data clusters is often omitted or done incompletely in statistical models, even though the problem is of great importance in the social sciences, computer science, and in many biological applications. Our proposed technology can substantially improve statistical practice when unobserved explanatory phenomena are present.

D.2 Background

We are concerned with how nonparametric priors can enhance the increasing use of Bayesian models in the social sciences. Consider modeling a dichotomous individual choices, Y_i, such as a vote choice, participation in a social event, or joining a political party. Typically logit or probit functions are specified to provide an smooth underlying utility curve that dictates such preferences, providing an estimated threshold, $\theta \in [0, 1]$, determining the choice. The individual's placement along this curve is given by the standard additive right-hand specification, $\mathbf{X}\boldsymbol{\beta}$. The θ threshold should actually be treated differently for each individual, but we can use the reasonable Bayesian approach of assuming that these vary but are still generated from a single distribution G which is itself conditional on a parameter α, thus $E[nG(\theta|\mathbf{X}, \boldsymbol{\beta}, \alpha)]$ is the expected number of positive outcomes.

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 G above as a single model. Here we are first concerned with latent clustering that adversely affects the quality of the model if not accounted for since unmeasured explanatory phenomena still affect the modeled relationship from the observed data. This is a very general problem since these "clusters" can be described in many ways. As an example, consider a vote choice in a congressional election where the Democratic candidate favors pulling US soldiers out of Afghanistan and the Republican candidate favors 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 respondent who has a relative in the military based in Afghanistan. This is not normally a question asked in such surveys, but may be an unmeasured strong causal reason for this vote. Or more generally, religious affiliation may determine sets of clusters that have a strong effect on support for continuing military operations.

Unknown clustering also has an effect on variable selection. Most literatures in the social sciences have a collection of explanatory phenomena that need to be included because the theories

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

One effective strategy for dealing with unmeasured 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 the choice is a normal distribution, even in the absence of evidence that it provides a good fit. A better alternative is a nonparametric Bayesian approach that draws ψ_i from more flexible class of distributions. We start with a general mixed effects model

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

where f and G are often taken to be normal. The addition of a link function, $g^{-1}()$, turns this into a generalized linear mixed model. As the random effects, unlike error terms, can not be checked (there are no corresponding residuals), the normal assumption is justified only as a convenience. a popular alternative being the Dirichlet Process with

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

where $\mathcal{DP}$ is the Dirichlet Process with base measure ϕ_0 and precision parameter m.

Dirichlet process mixture 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 m. 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 we developed in Gill and Casella (2009). We found that this approach can model difficult data and produce results that existing alternative methods fail to discover. In that work we were able to account for unobserved, important clustering structures that provided information about agency environment that was not explicitly available.

As a clarification, we are concerned here with accounting for an unknown number of unknown clusters *in the context of building a statistical model*. We are not working in the general area of spatial clustering with known coordinates using algorithms of the form: hierarchical, partitional, K-means, c-means, QT, graphing, spectral, etc. Thus all work pertains to Bayesian statistical model specification in the regression sense.

As a realistic example of the difference between the proposed approach and conventional Bayesian (or other) modeling, we look at terrorist activity in 22 Asian democracies over 8 years (1990-1997) with data subsetted from Koch and Cranmer (2007). Data problems (a recurrent problem in the empirical study of terrorism) reduce the number of cases to 162 and make fitting an standard model difficult. The outcome of interest is dichotomous, indicating whether not there was at least one violent terrorist act in a country/year pair. In order to control for the *level* of democracy, `DEM`, in these countries we use the Polity IV 21-point democracy scale ranging from -10, indicating a hereditary monarchy to +10, indicating a fully consolidated democracy (Gurr, Marshall, and Jaggers 2003). The variable `FED` is assigned zero if sub-national governments do not have substantial taxing, spending, and regulatory authority, and one otherwise. We look at three rough classes of government structure with the variable `SYS` coded as: (0) for direct presidential elections, (1) for strong president elected by assembly, and (2) dominant parliamentary government. Finally, `AUT` is a dichotomous variable indicating whether or not there are autonomous regions not directly controlled by the central government. The key substantive question evaluated here is whether specific structures of government and sub-governments lead to more or less terrorism.

Table 1: Probit Models for Asian Terrorism Incidents

	Standard Probit			GLMDM Probit		
Coefficient	COEF	SE	95% CI	COEF	SE	95% CI
`Intercept`	0.249	0.337	[-0.412 : 0.911]	0.127	0.188	[-0.241 : 0.495]
`DEM`	0.109	0.035	[0.041 : 0.177]	0.058	0.019	[0.020 : 0.095]
`FED`	0.649	0.469	[-0.269 : 1.566]	0.258	0.254	[-0.241 : 0.756]
`SYS`	-0.817	0.252	[-1.034 : -0.601]	-0.420	0.137	[-0.690 : -0.151]
`AUT`	1.619	0.871	[0.123 : 3.116]	0.450	0.371	[-0.277 : 1.176]

For purposes of contrast we specify a standard GLM model and a Generalized Linear Mixed Dirichlet Model (GLMDM) as we developed in Kyung, *et al.* (2009a), both with a probit link function. In the latter case, the Markov chain is run for 50,000 iterations disposing of the first half. There is no evidence of non-convergence in the remaining runs using standard diagnostic tools. Table 1 provides results from two approaches: a standard Bayesian probit model with flat

priors, and a Dirichlet random effects model. Notice first that while there are no changes in sign or statistical reliability for the estimated coefficients, the magnitude of effects uniformly smaller with the enhanced model: four of the estimates are roughly twice as large and one is about three times as large in the standard model. This indicates that there is extra variability in the data detected by the random effect that tends to dampen the size of the effect of these explanatory variables on explaining incidences of terrorist attacks. Specifically, running the standard probit model would find an *exaggerated* relationship between these explanatory variables and the outcome. Notice also that the credible intervals are smaller for the GLMDM model, providing more accurate estimates than the standard Bayesian GLM (in Kyung *et al.*(2009c) we prove that this is always the case).

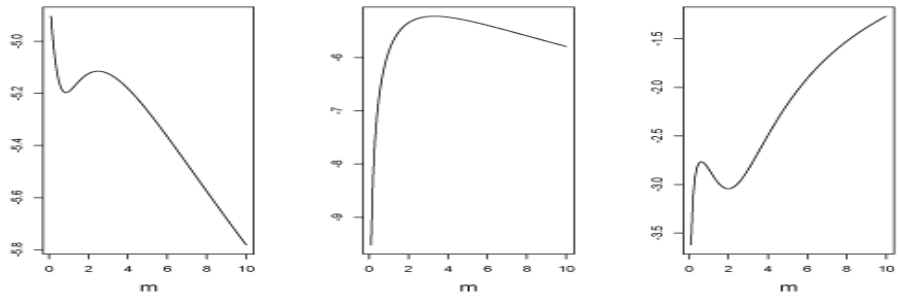
The results are also interesting substantively. The more democratic a country is, the more terrorist attacks they can expect. This is consistent with the literature in that autocratic nations tend to have more security resources per capita and fewer civil rights to worry about. Secondly, the more the legislature holds central power, the fewer expected terrorist attacks. This also makes sense, given what is known, disparate groups in society tend to have a greater voice in government when the legislature dominates the executive.

D.3 Previous Related Accomplishments

During the previous grant period we developed new models for social science data, and made substantial progress both on statistical estimation and computational implementation. In Gill and Casella (2009) we first explored the effectiveness of using a Dirichlet random effects model for social science data, implementing the model (1) with an ordered probit link, and standard priors on the model parameters in θ, to model the stress levels of presidential appointees. The conclusions drawn from his model were both interesting as well as suggestive, and were different from those that would have been obtained with classical modeling, including the aforementioned-mentioned observation that the Dirichlet model resulted in shorter credible intervals.

After seeing that these models had potential, we became concerned about two aspects. First, the estimation of the precision parameter m needed more attention. We found in Gill and Casella (2009) that the fit of the model was relatively insensitive to the value of m, and we wanted to find out if this was always the case and, if not, was there a preferred way to estimate it. Second, the MCMC algorithm that we used, which was a common one in the literature (see, for example, Neal 2000) was suspect in its mixing potential, and we wanted to look at other alternatives. Both of these goals were accomplished in Kyung *et al.* (2009a), where we first looked at maximum likelihood estimation of m, and found that the standard approach in finding the maximum likelihood estimate (MLE), given in Liu (1996), may have problems. In particular, by plotting the likelihood function for a selection of parameter values, we found that Liu's equation could yield a maximum, a minimum, or there might be a ridge. Figure 1, reproduced from Kyung *et al.* (2009a), shows possible shapes of the likelihood function. Thus, likelihood estimation is not reliable, and we were able to prove that by introducing a prior distribution on m, we could guarantee and interior mode, stabilizing the estimation procedure.

Figure 1: Log-likelihood functions for configurations of component likelihoods.



We next concentrated on a normal linear mixed model of the form:

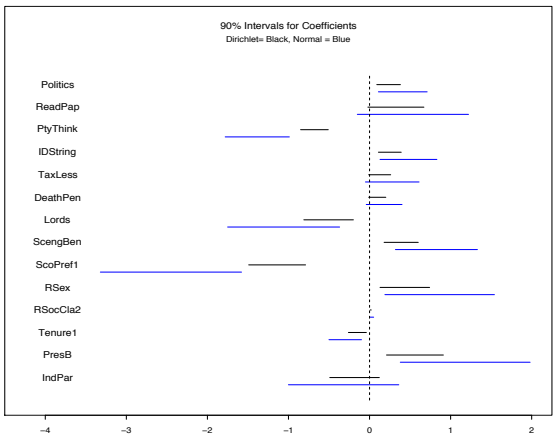
$$\underset{n\times 1}{\mathbf{Y}}|\boldsymbol{\psi} \sim N(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\psi}, \sigma^2 I), \tag{3}$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)'$, $\psi_i \sim \mathcal{DP}(m, N(0, \tau^2))$, $i = 1, \ldots, n$, independent. As the values of $\boldsymbol{\psi}$ are not distinct, we can represent this model as

$$\underset{n\times 1}{\mathbf{Y}}|\underset{k\times 1}{\boldsymbol{\eta}}, \mathbf{A} \sim N(X\beta + \mathbf{A}\boldsymbol{\eta}, \sigma^2 I), \quad \boldsymbol{\eta} \sim N_k(0, \tau^2 I), \tag{4}$$

where k is the number of distinct values in $\boldsymbol{\psi}$, $\mathbf{A}'_{k\times 1} = (a'_1, a'_2, \cdots, a'_n)$, and each a_i is a $1 \times k$ vector of all zeros except for a 1 in one position that indicates which group the observation is from. With this representation, MCMC implementation is straightforward, as the Dirichlet process is reduced to dealing with the binary matrix $\mathbf{A}$. Using this representation we presented a new Gibbs sampler that was proven to be more efficient that the current algorithms, in that the variances were proven to be uniformly smaller. Moreover, investigation into the size of the reduction showed that it could be more than 50%.

Figure 2: 90% credible intervals for the Dirichlet random effects model (black), and a normal random effects model (blue)



The normal linear mixed model is easily extended to a probit model by introducing a latent variable into the hierarchy. However, other models, such as logit or loglinear models are not available as straightforward extensions, and present new computational problems. These models, with Dirichlet random effects, are the subject of Kyung *et al.* (2009b).

After looking at a number of data sets, and comparing the fit with the Dirichlet random effects model to that of the more classical normal random effects, we noted that in every instance the Dirichlet model produces shorter intervals on the coefficient estimates as seen in the terrorism

example; recall Table 1. In another example in Kyung *et al.* (2009a) we analyzed data from the British General Election Study, Scottish Election Survey, 1997 (ICPSR Study Number 2617). There we saw that the widths of 90% credible intervals from the Dirichlet random effects model were all shorter than those from a normal random effects model, as shown in Figure 2. Such results were so persistent that it seemed unlikely to be a coincidence, and in Kyung *et al.* (2009c) we were able to prove that the Dirichlet model always results in smaller posterior variances than the normal model. Specifically, in that paper we proved the following theorem.

Theorem 1 *For all* $\mathbf{y}$ *not containing a within subcluster contrast, the mean of the posterior distribution of the variance from the Dirichlet random effects model*

$$\mathbf{Y}|\mu, \boldsymbol{\eta}, \sigma^2, \mathbf{A} \sim \mathcal{N}\left(\mu\mathbf{1} + \mathbf{A}\boldsymbol{\eta}, \sigma^2\mathbf{I}\right) \qquad \boldsymbol{\eta}|\sigma^2 \sim \mathcal{N}_n\left(\mathbf{0}, c\sigma^2\mathbf{I}_K\right)$$

$$\mu|\sigma^2 \sim \mathcal{N}\left(0, v\sigma^2\right) \qquad \sigma^2 \sim \mathcal{IG}\left(a, b\right),$$

is smaller than that from the normal random effects model.

D.4 Proposed Research

In this section we provide a brief background in order to put some perspective on our proposed work and then describe four main topics that we will address in the grant period: (*i*) a new approach for clustering, (*ii*) merging clustering and variable selection, (*iii*) implementing valid parallel processing of MCMC algorithms, and (iv) the application of this technology to the challenging data applications in the empirical study of terrorism.

D.4.1 Cluster Models

Model-based cluster analysis has a long and rich history, and recent developments using MCMC methods and hierarchical models are relevant to our work. There are two 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*). The mixture model begins with the assumption that $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are realizations of n independent and identically distributed (iid) random vectors from the K-component mixture density

$$\sum_{\alpha=1}^{K} \tau_\alpha \, f(\cdot|\boldsymbol{\beta}_\alpha, \Sigma_\alpha, G) \,, \tag{5}$$

where K is a fixed positive integer, $0 \leq \tau_\alpha \leq 1$, $\sum_{\alpha=1}^{K} \tau_k = 1$. A partition of the data is typically obtained as a byproduct of an EM algorithm. An alternative, the product partition model, starts with the assumption that there is some fixed, unknown partition of $\mathbb{N}_n := \{1, 2, \ldots, n\}$, $\mathcal{C}_m$, that has m clusters denoted by $C_1, \ldots, C_m$ and that the data are a realization from a density of the form

$$f(\mathbf{y}|\boldsymbol{\beta}, \Sigma, G, \mathcal{C}_m) = \prod_{\alpha=1}^{k} \prod_{i \in C_\alpha} f(\mathbf{y}_i|\boldsymbol{\beta}_\alpha, \Sigma_\alpha, G) \,. \tag{6}$$

Unlike the standard mixture model, (6) contains a parameter, $\mathcal{C}_m$, that is directly relevant 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.

Booth *et al.* (2008) argue strongly in favor of the product partition model, noting that not only does the mixture model lack a parameter that defines the clusters, the inference is confounded with the application of the EM algorithm. That is, even if the parameters of the model are known, there needs to be one final run of the EM algorithm to obtain the clusters. McCullaugh and Yang (2008) agree, arguing that the mixture model is not appropriate for determining clusters. A further deficiency is that the model needs to be run with a fixed K; the typical strategy is to run a selection of K 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 k, 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.

We can classify related previous work according to the model used for the clustering, and the model used for the random effects, summarized in the following table:

	Mixture Model	*Product Partition Model*
Dirichlet	Dahl (2006) Kim *et al.* (2006) Rodriguez *et al.* (2008) Green and Richardson (2001)	Lau and Green (2007) Quintana and Iglesias (2003)
Normal	Fraley and Raftery (2002) Richardson and Green (2002) Tadesse *et al.* (2005)	Booth *et al.* (2008) Heard *et al.* (2006)

Clustering based on mixture models was used by Fraley and Rafterty (2002), estimating the allocation probabilities and the model parameters with the EM algorithm, and using the Bayesian information criteria (BIC) to 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, the 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.

It is important to understand the clustering strategy that has previously been used in all applications of the Dirichlet random effects model. The cluster search has been carried out only using the random clustering that appears in realizations of the Dirichlet process. In particular, consider model (3), 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, then choose the best of these by a post-hoc scheme–processing the MCMC output through some objective function to find the best cluster. All previous applications of the Dirichlet random effects model only cluster the subjects according to the generated values of the random effects parameter, ψ_i, and then fit the $\boldsymbol{\beta}$ vector, that is, the cluster search then ignores the covariates. Thus, the clusters are produced in a realization of the Dirichlet process and are not substantive in any way, and cannot reflect any cluster structure driven by the covariates. Importantly, *Dirichlet process subcluster assignments are not disaggregated pieces of real clusters, they are temporary model assignments of random effects assignments that make the model fit better.* Furthermore, since there is no overfitting penalty in the Dirichlet process, there will always be more subclusters than actual latent clusters in the data. This problem is our starting point and the next section describes a solution for recovering estimates of substantive clusters in social science data.

D.4.2 Substantive Clustering

Here, we start with the model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}\psi_i + \boldsymbol{\epsilon}_i, \tag{7}$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ir})'$ are the responses of subject i, $i = 1, \ldots, n$, $\mathbf{X}_i$ is an $r \times p$ design matrix for subject i, $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector, ψ_i is a subject specific random effect (multiplied by the vector of ones $\mathbf{1}_{r\times 1}$ modeled with the Dirichlet process (2) with $\phi_0 = N(0, \tau^2)$, and $\boldsymbol{\epsilon}_i \sim N(0, \Sigma_{r\times r})$ are error terms. We assume that $\boldsymbol{\epsilon}$ and $\boldsymbol{\psi}$ are mutually independent. Starting from (7), we will search for a partition of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ into clusters $C_\alpha, \alpha = 1, \ldots, m$ where for Y_i in cluster C_α we have

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_\alpha + \mathbf{1}\psi_i + \boldsymbol{\epsilon}_{\alpha_i}, \tag{8}$$

where $\boldsymbol{\beta}_\alpha$ and $\epsilon_{\alpha_i} \sim N(0, \Sigma_\alpha)$ are specific to cluster C_α. For ease of notation we take $r = 1$; the more general case merely increases the algebraic overhead. Using the $\mathbf{A}$ matrix notation of (4), and letting the base measure of ψ be $N(0, \tau^2)$, we can integrate out the ψ_i to obtain the marginal distribution of Y_i in cluster C_α is

$$\mathbf{Y}_\alpha|\mathbf{A}, \mathcal{C}_m, \boldsymbol{\theta} \sim N_{n_\alpha}\left(\mathbf{X}^{(\alpha)}\boldsymbol{\beta}_\alpha, \sigma_\alpha^2\boldsymbol{\Sigma}_\alpha + \tau^2\mathbf{A}_\alpha\mathbf{A}_\alpha'\right), \tag{9}$$

where $\mathbf{Y}_\alpha$ is a vector of length n_α containing the Y_i in cluster C_α, $\mathbf{A}_\alpha$ is made of the rows of $\mathbf{A}$ corresponding to observations in $\mathbf{Y}_\alpha$, $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \sigma_1^2, \ldots, \sigma_m^2)$ are unknown, and $\boldsymbol{\Sigma}_\alpha$ is a known matrix describing the within-cluster structure.

Our goal is to find the best partition $\mathcal{C}_m$, but the $\mathbf{A}$ matrix cannot be ignored. From (9), we want to find the posterior probability of the pair $(\mathcal{C}_m, \mathbf{A})$, marginalized over the covariates, and use this function to drive a stochastic search. Using the priors

$$\boldsymbol{\beta}_\alpha \sim N_p\left(\mathbf{0}, d\sigma_\alpha^2\mathbf{I}\right) \qquad \sigma_\alpha^2 \sim IG(a_1, b_1) \qquad \text{and} \qquad \tau^2 \sim IG(a_2, b_2).$$

we can integrate out $\boldsymbol{\theta}$ and τ^2 to obtain the posterior probability of $\mathbf{A}$ and $\mathbf{C}$

$$\pi\left(\mathbf{C},\mathbf{A},\boldsymbol{\eta}|\mathbf{y}\right)\propto\frac{\pi(\mathbf{A})\pi(\mathbf{C})\Gamma(\frac{k}{2})}{[\frac{1}{2}|\boldsymbol{\eta}|^2+b_2]^{\left(\frac{k}{2}+a_2\right)}}\prod_{\alpha=1}^{m}\frac{|\boldsymbol{\Sigma}_{\mathbf{y}_\alpha}|^{-1/2}\Gamma\left(\frac{n_\alpha}{2}+a_1\right)}{\left[b+\frac{1}{2}\left(\mathbf{y}_\alpha-\mathbf{A}_\alpha\boldsymbol{\eta}\right)'\boldsymbol{\Sigma}_{\mathbf{y}_\alpha}^{-1}\left(\mathbf{y}_\alpha-\mathbf{A}_\alpha\boldsymbol{\eta}\right)\right]^{\left(\frac{n_\alpha}{2}+a_1\right)}},\tag{10}$$

where $\boldsymbol{\Sigma}_{\mathbf{y}_\alpha}=\boldsymbol{\Sigma}_\alpha+d\mathbf{X}^{(\alpha)}\mathbf{X}^{(\alpha)\prime}$. We will use (10) to drive a stochastic search. Specifically, we create a Markov chain $(\mathbf{A},\boldsymbol{\eta},\mathcal{C}_m,\boldsymbol{\theta})$ with stationary distribution (10), which will therefore explore the modes of the posterior probability surface. To do this we will adapt our Gibbs sampler in Kyung *et al.* (2009a) to go from $(\mathbf{A},\boldsymbol{\eta},\mathcal{C}_m,\boldsymbol{\theta})$ to $(\mathbf{A}',\boldsymbol{\eta}',\mathcal{C}'_{m'},\boldsymbol{\theta}')$ using the conditionals:

$$(\mathcal{C}'_{m'},\boldsymbol{\theta}')|\mathbf{A},\boldsymbol{\eta},\mathcal{C}_m,\boldsymbol{\theta},\qquad(\mathbf{A}',\boldsymbol{\eta}')|\mathbf{A},\boldsymbol{\eta},\mathcal{C}'_{m'},\boldsymbol{\theta}'.\tag{11}$$

We then take this candidate and apply a Metropolis-Hasting correction (see Robert and Casella 2004, Chapter 7) using (10), which insures that the posterior probability will drive the search.

We give a few details of the stochastic search, as they show the possible difficulties we will encounter. Given we are at $(\mathbf{A},\boldsymbol{\eta},\mathcal{C}_m,\boldsymbol{\theta})$, we first update to $(\mathcal{C}'_{m'},\boldsymbol{\theta}')$. To do this we first obtain a new candidate $\mathcal{C}^*_{m^*}$ using our algorithm in Kyung *et al.* (2009a). We now need a candidate $\boldsymbol{\theta}^*$, but the previous $\boldsymbol{\theta}$ is useless, as it is based on different clusters. This would not be a problem if we made the somewhat unrealistic assumption that the cluster variances are all equal, as done in Tadesse *et al.* 2005 and Booth *et al.* 2008.

Given $\mathcal{C}^*_{m^*}$ we can calculate the mean and variance of the conditional distribution of $\boldsymbol{\beta}'_\alpha$ up to a constant that depends on the incalculable cluster variance. A way to do this is with an Accept-Reject algorithm (see Robert and Casella 2004, Section 2.3). Thus we generate a candidate $\boldsymbol{\theta}^*=(\boldsymbol{\beta}^*_\alpha,\sigma^{2*}_\alpha$ according to

$$g\left(\sigma^2,\boldsymbol{\beta}\right)=IG\left(\sigma^2:a_1^*,b_1^*\right)\times N_p\left(\boldsymbol{\beta}:\boldsymbol{\beta}^*,\sigma^2\Sigma^*_{\boldsymbol{\beta}}\right)$$

where the parameters of the candidate can all be calculated:

$$\begin{aligned}\Sigma^*_{\boldsymbol{\beta}} &= \left(\frac{1}{d}\mathbf{I}_p+\mathbf{X}^{(\alpha)\prime}\boldsymbol{\Sigma}_\alpha^{-1}\mathbf{X}^{(\alpha)}\right)^{-1}\qquad \boldsymbol{\beta}^*=\boldsymbol{\Sigma}^*_{\boldsymbol{\beta}}\mathbf{X}^{(\alpha)\prime}\boldsymbol{\Sigma}_\alpha^{-1}\left(\mathbf{y}_\alpha-\mathbf{A}_\alpha\boldsymbol{\eta}\right)\\ a_1^* &= \frac{n^*_\alpha}{2}+a_1\qquad b_1^*=b_1+\frac{1}{2}\left(\mathbf{y}_\alpha-\mathbf{A}_\alpha\boldsymbol{\eta}\right)'\boldsymbol{\Sigma}_{\mathbf{y}_\alpha}^{-1}\left(\mathbf{y}_\alpha-\mathbf{A}_\alpha\boldsymbol{\eta}\right).\end{aligned}$$

The candidate $(\mathcal{C}^*_{m^*},\boldsymbol{\theta}^*)$ is then subjected to the Accept-Reject test using the target density

$$f\left(\sigma^2,\boldsymbol{\beta}\right)=IG\left(\sigma^2:a_1^*,b_1^*+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\Sigma^{*-1}_{\boldsymbol{\beta}}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right)\times N_p\left(\boldsymbol{\beta}:\boldsymbol{\beta}^*,\sigma^2\Sigma^*_{\boldsymbol{\beta}}\right).$$

This process is repeated until a candidate is accepted. Note that a Metropolis-Hastings step cannot be used, as there is no direct way to compare $(\mathcal{C}'_{m^*},\boldsymbol{\theta}^*)$ to $(\mathcal{C}_m,\boldsymbol{\theta})$ We now are at $(\mathbf{A},\boldsymbol{\eta},\mathcal{C}'_{m'},\boldsymbol{\theta}')$, and to proceed to $(\mathbf{A}\prime,\boldsymbol{\eta}\prime,\mathcal{C}'_{m'},\boldsymbol{\theta}')$ by adapting our Gibbs sampler in Kyung *et al.* (2009a)

D.4.3 Variable Selection

The idea of combining variable selection and clustering has a long history in statistics, dating back to at least Fowlkes *et al.* (1988). More recent work includes Raftery and Dean (2006), who use a mixture model for the clustering and a greedy search to select variables; Wang and Zhu (2008), who also use mixture models but use a lasso-type penalty to select variables; and Xie *et al.* (2008), who extend the model of Wang and Zhu (2008), looking at computational properties. Tadesse *et al.* (2005) also use mixture models, but partition variables into those that discriminate the mixture components, and those that do not, and only cluster the former. None of the previous papers used the Dirichlet process, but Kim *et al.* (2006) adapted the Tadesse *et al.* (2005) approach to Dirichlet random effects. However, all of these papers start from the mixture model which, as previously argued, is not a good clustering model. In addition, the applied practice of variable selection has an inglorious history in the social sciences (Leamer 1978, Gill 1999), and remains an ongoing source of controversy. We seek to provide a means whereby empirical researchers can avoid: arbitrary decisions, difficult covariate tradeoffs, and the fallacy that only one specification was tried.

Variable selection, independent of clustering, got a huge boost with the work of George and McCulloch (1993), who showed how hierarchical models and stochastic search can be used for variable selection. Since then there has been a lot of development of this topic, with the recent papers by Casella and Moreno (2006) and Casella *et al.* (2009) being most relevant to our application. They showed that by introducing a vector γ of 0s and 1s, indicating which variables are to be included, and selecting on Bayes factors (posterior probabilities), one can do consistent variable selection.

In our application we define, for cluster C_α, a vector γ_α by: $\gamma_{\alpha_i} = \begin{cases} 0 & \text{if } \beta_{\alpha_i} = 0 \\ 1 & \text{otherwise,} \end{cases}$

and our model (8) becomes

$$\mathbf{Y}_i = \mathbf{X}_i(\boldsymbol{\beta}_\alpha * \gamma_\alpha) + \mathbf{1}\psi_i + \boldsymbol{\epsilon}_{\alpha_i}, \quad \text{for } Y_i \text{ in cluster } C_\alpha, \tag{12}$$

where "*" is elementwise multiplication. We now want to include the vector γ_α in the Markov chain (11) using a random walk/independent jump algorithm that moves $\gamma_\alpha \to \gamma'_\alpha$ by

$$\begin{aligned} \text{with probability } p &: \text{ select one element of } \gamma_\alpha \text{ at random and switch } 0 \to 1 \text{ or } 1 \to 0 \\ \text{with probability } 1-p &: \text{ select } \gamma'_\alpha \text{ at random from all possible vectors.} \end{aligned}$$

However, we encounter the same problem as in Section D.4.2. Once we iterate to a new cluster structure, the γ of the previous structure is irrelevant, so implementing a random walk is not straightforward. There are two possible ways to overcome this problem:

(*i*) Take $\gamma_\alpha = \gamma$, the same for each cluster, and update γ one time for each step of the $(\mathbf{A}, \boldsymbol{\eta}, \mathcal{C}_m, \boldsymbol{\theta})$ chain.

(*ii*) Given $(\mathbf{A}, \boldsymbol{\eta}, \mathcal{C}_m, \boldsymbol{\theta})$, start with each $\gamma_\alpha = \mathbf{1}$, and update γ_α a total of ℓ times, using a hybrid Metropolis-Hastings step based on (10) and including random jumps.

Clearly (*i*) is computationally faster, but restricts each cluster to select the same variables, which

will often be too limiting. The more flexible (*ii*) requires restarting the γ_α chain at each iteration, again increasing the computational overhead. However, this is the more realistic approach, and to better deal with it we will also investigate ways of substantially increasing computation speed.

D.4.4 Parallel Processing

Our last objective is purely computational, but has the potential of increasing the speed of related MCMC algorithms greatly. The chains that we have described in Sections D.4.2 and D.4.3 will require chains within chains, greatly increasing the computational demand.

A drawback of MCMC algorithms for large social science datasets is the need to run long chains, not only to reach the stationary distribution, but also to insure that the space has been adequately searched. These requirements seem to preclude the use of parallel computing, because if we start two Markov chains on two different nodes, there is no obvious way to put them together into one longer chain. In fact, if the same MCMC algorithm is run for M iterations on two separate nodes, the chains cannot be spliced together in a valid way. However, we can circumvent this problem by creating a *split chain* though the process of *regeneration*.

If we denote the current state of our Markov chain by $\Delta^{(t)} = (\mathbf{A}^{(t)}, \boldsymbol{\eta}^{(t)}, \mathcal{C}_m^{(t)}, \boldsymbol{\theta}^{(t)})$, then the transition kernel is the function $k(\Delta^{(t)}, \Delta^{(t+1)})$ that produces the next state in the Markov chain. For the chains described in this proposal, k is composed of the steps of the Gibbs sampler and Metropolis-Hastings algorithms. If there are functions s and w that satisfy

$$k(\Delta^{(t)}, \Delta^{(t+1)}) \geq s(\Delta^{(t)}) w(\Delta^{(t+1)}) \text{ for all } \Delta^{(t)} \in \mathcal{D}, \tag{13}$$

then $\mathcal{D}$ is a *small set* and (13) is a *minorization condition*. The existence of $\mathcal{D}$ allows us to construct the *split chain*, a Markov chain that is identical to $k(\Delta^{(t)}, \Delta^{(t+1)})$ with the exception that whenever the chain enters $\mathcal{D}$, there is a positive probability of emerging independently. This allows use to join chains together and thus take advantage of parallel processing.

In particular, suppose that we have R nodes and, in each node we start our Markov chain in the set $\mathcal{D}$. We then have

$$\begin{aligned} \text{Node } 1 &: \Delta_1^{(1)}, \ldots, \Delta_1^{(n_{\kappa_1})} \\ &\vdots \\ \text{Node } R &: \Delta_R^{(1)}, \ldots, \Delta_R^{(n_{\kappa_R})} \end{aligned}$$

where κ_i is the random time that chain i returns to $\mathcal{D}$. Since the Markov chain has been regenerated independently at each start in $\mathcal{D}$, we can validly splice the chain into

$$\Delta_1^{(1)}, \ldots, \Delta_1^{(n_{\kappa_1})}, \Delta_2^{(1)}, \ldots, \Delta_2^{(n_{\kappa_2})}, \ldots, \Delta_R^{(1)}, \ldots, \Delta_R^{(n_{\kappa_R})},$$

one chain of length $R \sum_i n_{\kappa_i}$. Moreover, the independent pieces allow for valid calculation of the Monte Carlo variance (Hobert *et al.* 2002).

The use of regeneration was first explored by Mykland *et al.* (1995) and Rosenthal (1995), but implementation was not apparent as the identification of small sets was very difficult. Moreover, such sets need to be entered with a large enough probability to be practical. Progress has been

made by Roy and Hobert (2007) and Tan (2009), and we intend to use those results to identify useful small sets. In particular, Tan (2009) identifies small sets for a two-stage block Gibbs sampler. Our Markov chain (11) is in this form, as we can write it as the conditionals:

$$\mathcal{C}'_{m'}|\mathbf{A}, \boldsymbol{\eta}, \boldsymbol{\theta}, \qquad (\boldsymbol{\theta}', \mathbf{A}', \boldsymbol{\eta}')|\mathcal{C}'_{m'}. \tag{14}$$

To implement this strategy we need to first need to verify that our chain is *geometrically ergodic*, that is, having a convergence rate that is at least geometric. As our chain is a block Gibbs sampler, geometric ergodicity is a strong possibility (Hobert and Geyer 1998). We have already adapted these results in other work to prove geometric ergodicity of another Gibbs sampler used in Bayesian lasso calculations (Kyung *et al.* 2009d), and we expect that a similar calculation will hold in this case. We then will identify small sets to investigate the practicality of parallelization.

D.4.5 Application: Reanalysis of Terrorism Data

The study of terrorism has not made substantial *empirical* progress due to inherent problems in the available data. However, terrorism is an important problem because it affects internal government policy, public perception, relations between states, and of course, personal safety. Data problems include: selection on the outcome variable (a visible event), non-granular discrete measurement, insufficient explanatory variables, and the lack of access to classified collections. Another key problem, and the one addressed here, 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 (becoming less transparent). Some progress has been made: we know not to treat these groups as unitary actors (Chai 1993, Crenshaw 1981), they tend to be cellular and distributed rather than hierarchically organized (Carley 2004, Krebs 2002, Rothenberg 2002), they evolve over time (Carley 2003, 2006), and semi-automatic parsing of signal traffic is informative (Tsvetovat and Carley 2006). Yet much more data-analytic work is needed.

We intend to apply the GLMDM technology to this problem to reveal clusters in the data. Furthermore, we will use *every* publicly available organized dataset on terrorist events, including (but not limited to): START (University of Maryland), text-processed DOTS (Directory of Terrorists; Vinyardsoftware, Inc), Harrison-Israel, Koch and Cranmer, Global Terrorism Database II (GTD2), Multiple Regions MIPT Terrorism Knowledge Database, Varshney-Wilkinson Dataset on Hindu-Muslim Violence in India 1950-1995, Terrorism in Western Europe: Events Data (TWEED), Worldwide Incidents Tracking System, International Terrorism–Attributes of Terrorist Events (ITERATE), Pinkerton Global Intelligence Service (PGIS), and several on the Irish "Troubles" (Index of Deaths in Northern Ireland and others). Where possible we will also merge information across datasets since some have a limited number of possible covariates. Because these datasets are of modest size, we expect the nonparametric effect to be important, providing estimates of the unmeasured clusters from various sub-groups among these actors. Unfortunately, it is not always clear which group or sub-group committed the violence since the attackers may not reveal themselves, multiple groups attempt to take credit, or governments conceal information for strategic reasons. The GLMDM approach will also provide guidance on variable selection in a literature where there is no strong set of conventions for explanatory variables. The study of terrorism data is an ideal

application for our technology: progress is stymied by unmeasured effects in the data, strong qualitative and journalistic evidence points toward clustering, and nonparametric approaches to variable selection have the potential to improve our understanding of this pervasive political and military challenge.

D.5 Expected Outcomes

This work is expected to make research progress in both statistics and the social sciences. If supported, this project will lead to the following expected outcomes.

- ▷ Research papers in methodological statistics producing usable posterior forms from general nonparametric priors. At least one paper will be directed at statistical inference and properties of estimators, and another will address computational issues.
- ▷ Applied research papers demonstrating the value of specifying Dirichlet process priors to complex data problems in terrorism datasets.
- ▷ Software tools for MCMC written in `R` and `C++`, which will be made freely available to researchers on a dedicated webpage.
- ▷ Training of post-docs and graduate students in the development of statistical methodology and algorithms, with an emphasis on data problems encountered by social scientists. We will also include undergraduate students in the research by having them process datasets, test software, and participate in research meetings. See the attached Dean's letter as confirmation of our intent.

D.6 Results from Prior NSF Support

D.6.1 PI Gill

- ▷ NSF-MSBS. "Adaptive Nonparametric Markov Chain Monte Carlo Algorithms for Social Data Models with Nonparametric Priors," DMS-0631632 and SES-0631588 with George Casella. Award period: January 2007 to January 2010.
- ▷ During the grant period one post-doc was supervised (Minjung Kyung) and three PhD students were supervised: Skyler Cranmer (UC-Davis Political Science Ph.D, now Assistant Professor, UNC Chapel Hill Political Science), Xun Pang (Washington University Political Science), Andrew Womack (Washington University Mathematics).
- ▷ Relevant Publications:

1. Kyung, M., Gill, J. and Casella, G. (2009). Estimation in Dirichlet Random Effects Models. To appear in *Annals of Statistics*.
2. 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.
3. Garabed, R.B., Johnson, W.O., Gill, J., Perez, A.M. and Thurmond, M.C. (2008). Effects of Politics and Economics on Country-Level Foot-and-Mouth-Disease Status. *Journal of the Royal Statistical Society, Series A* **171** 699-722.

4. Gill, J. (2008). Is Partial-Dimension Convergence a Problem for Inferences From MCMC Algorithms? *Political Analysis* **16** 153-178.
5. Altman, M., Gill, J. and McDonald, M. P. (2007). Accuracy: Tools for Accurate and Reliable Statistical Computing. *Journal of Statistical Software* **21**. (Abstract published in *Journal of Computational Graphics and Statistics*.)

D.6.2 PI Casella

▷ NSF-MSBS. "Adaptive Nonparametric Markov Chain Monte Carlo Algorithms for Social Data Models with Nonparametric Priors," DMS-0631632 and SES-0631588 with Jeff Gill (Washington University). Award period: January 2007 to January 2010, and NSF-DMS "Cluster Analysis, Predictive Distributions, and Stochastic Search Algorithms," DMS-0405543. Award period: July 2004 to June 2009.

▷ During the grant period the main focus of the research was in development of algorithms for valid inferential procedures, with particular attention to cluster algorithms and stochastic search. One post-doc was supervised (Minjung Kyung) and four graduate students: Jessica Zhen Li. (PhD completed December 2008), Nabanita Mukherjee (PhD expected December 2009), Claudio Fuentes (PhD expected May 2010), and Viknesh Gopal (PhD expected May 2010).

▷ Relevant Publications (in addition to 1. and 2. above):

1. 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
2. Brumback, B. A, Winner, L. H., Casella, G., Ghosh, M., Hall, A., Zhang, J., Chorba, Lorna, and Duncan, P. (2008). Estimating a Weighted Average of Stratum-Specific Parameters. *Statistics in Medicine* **27** 4972-4991.
3. Casella, G., Giron, F. J. and Moreno, E. (2009). Consistent Model Selection in Regression. *Annals of Statistics* **37** 1207-1228
4. Casella, G. and Moreno, E. (2009). Assessing Robustness of Intrinsic Tests of Independence in Two-way Contingency Tables. To appear in the *Journal of the American Statistical Association*.
5. Fuentes, C. and Casella, G. (2009). Testing for the existence of Clusters (with discussion). To appear in *Statistics and Operations Research Transactions*
6. Joo, Y. , Booth, J. G., Namkoong, Y. and Casella, G. (2008). Model-Based Bayesian Clustering (MBBC). *Bioinformatics* **24** 874-875
7. Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103** 681-686

▷ Other Specific Products

Two `R` packages were created. The package `bayesclust` implements the cluster test of Fuentes, C. and Casella, G. (2009), and `BAMD` implements a missing data Gibbs sampler to do Bayesian association mapping, the subject of the PhD thesis of Jessica Li.