Volume 2 Number 3 2007

Bayesian Analysis

Bayesian Analysis (2007) **2**, Number 3, pp. 445–472

Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model

Sonia Jain[*] and Radford M. Neal[†]

Abstract. The inferential problem of associating data to mixture components is difficult when components are nearby or overlapping. We introduce a new split-merge Markov chain Monte Carlo technique that efficiently classifies observations by splitting and merging mixture components of a nonconjugate Dirichlet process mixture model. Our method, which is a Metropolis-Hastings procedure with split-merge proposals, samples clusters of observations simultaneously rather than incrementally assigning observations to mixture components. Split-merge moves are produced by exploiting properties of a restricted Gibbs sampling scan. A simulation study compares the new split-merge technique to a nonconjugate version of Gibbs sampling and an incremental Metropolis-Hastings technique. The results demonstrate the improved performance of the new sampler.

Keywords: Bayesian model, Markov chain Monte Carlo, split-merge moves, nonconjugate prior

1 Introduction

Bayesian mixture models have gained in popularity as an alternative to traditional density estimation and clustering techniques. In particular, Bayesian mixture models in which a Dirichlet process prior defines the mixing distribution are of interest due to their flexibility in fitting a countably infinite number of components (Ferguson (1983)). Much of the recent research related to the Dirichlet process mixture model has been devoted to developing computational techniques, usually Markov chain Monte Carlo methods, to sample from its posterior distribution (Neal (2000), MacEachern and Müller (1998)). Other techniques to estimate the Dirichlet process model include sequential importance sampling (MacEachern et al. (1999)) and variational methods (Blei and Jordan (2004)). The practical utility of these methods is illustrated by their recent use for complex biological and genetics problems, such as haplotype reconstruction (Xing et al. (2004)), estimation of rates of non-synonymous and synonymous nucleotide substitutions as evidence for natural selection in evolutionary biology problems (Huelsenbeck et al. (2006)), and determination of differential gene expression (Do et al. (2005)).

The focus of this article is on Markov chain sampling for nonconjugate Dirichlet process mixture models, building on our previous work for conjugate models (Jain and Neal (2004)). Conjugate models are appropriate for some problems, which is convenient due

[*] Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, CA, mailto:sojain@ucsd.edu

[†] Department of Statistics and Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, http://www.cs.toronto.edu/~radford/

to the analytical tractability of these priors. However, in many situations, conjugate priors can be too restrictive. Forcing conjugacy on the model can lead to undesirable or even nonsensical priors. A classic example is a simple model for normally distributed data, where conjugacy requires an assumption that the mean and variance are *a priori* dependent, which is often unrealistic in actual problems.

Computationally, Markov chain sampling procedures can operate differently depending on whether conjugacy is assumed. In the conjugate case, we can analytically integrate away the mixing proportions for the components and the parameters for each component. This leads to Markov chain Monte Carlo procedures that update only the latent indicator variable associating mixture components with data observations (MacEachern (1994), Neal (1992)). However, in the nonconjugate case, the parameters of the model cannot be integrated away and must be included in the Markov chain update. Further, since we lose the advantage of analytic tractability, computational difficulties arise, which makes it more difficult, but not impossible, to construct valid Markov chain Monte Carlo procedures.

Nonconjugate Markov chain sampling methods based on the Gibbs sampler have been proposed previously; see, for instance, MacEachern and Müller (1998) and Neal (2000). When the mixture components are nearby or overlapping, these incremental samplers (as well as those for conjugate models) suffer from computational difficulties, such as remaining stuck in isolated modes and poor mixing between components.

Alternative nonincremental Markov chain samplers for the Dirichlet process mixture model based on split-merge moves have been proposed by Green and Richardson (2001) and by ourselves (Jain and Neal (2004)). In a single iteration, these methods can split a mixture component moving all observations to an appropriate new component, or merge two distinct components together. The Green and Richardson (2001) method is based on the reversible-jump procedure, in which numerous ways to propose a split move are possible. Since specific moment conditions must be preserved, the split-merge proposals are model-dependent. Jain and Neal (2004) introduce a Metropolis-Hastings technique with split-merge proposals for conjugate Dirichlet process mixture models. The innovation in this work is exploiting properties of a Gibbs sampling scan to construct split-merge moves, such that their Metropolis-Hastings proposals are model-independent. In this article, we extend the conjugate split-merge technique to a class of nonconjugate Dirichlet process mixture models by developing a novel scheme to incorporate the model parameters into the sampling procedure.

This article is organized as follows. Section 2 defines the nonconjugate Dirichlet process mixture model. Section 3 briefly describes the Metropolis-Hastings split-merge technique based on Gibbs sampling proposals. The new split-merge technique for a class of nonconjugate models is proposed in Section 4. Next, in Section 5, we illustrate the utility of our method in by comparing it to an auxiliary Gibbs sampling method (Neal (2000), Algorithm 8). Section 6 is a general discussion and concluding remarks. Details of a simulation study are provided in the Appendix in Section 7.

2 The model

The Dirichlet process mixture model takes the following hierarchical model form for observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ that is considered exchangeable:

$$\begin{aligned} y_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim DP(G_0, \alpha) \end{aligned} \tag{1}$$

Here, $F(\theta_i)$ is a component parameterized by θ_i from a parametric distribution whose density will be written as $f(y;\theta)$. G is the mixing distribution. G_0 defines a base distribution for the Dirichlet process (DP) prior, while α is a concentration parameter that takes values greater than zero. The usual conditional independence assumptions for a hierarchical model apply, so that the only dependencies are those that are explicitly shown.

Realizations of the Dirichlet process are discrete with probability one. A consequence of this is that the mixture model in equation (1) can be viewed as a countably infinite mixture model (Ferguson (1983)). This is evident when we simplify the model in equation (1) by integrating G over its prior distribution. The θ_i follow a generalized Polya urn scheme (Blackwell and MacQueen (1973)) and the prior distribution for the θ_i may be represented by the following conditional distributions:

$$\begin{aligned} \theta_1 &\sim G_0 \\ \theta_i \mid \theta_1, \ldots, \theta_{i-1} &\sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta(\theta_j) \; + \; \frac{\alpha}{i-1+\alpha} G_0 \end{aligned} \tag{2}$$

where $\delta(\theta_j)$ is the distribution which is a point mass at θ_j.

We can represent the fact that (2) results in some of the θ_i being identical by setting $\theta_i = \phi_{c_i}$, where c_i represents the latent class associated with observation i, and all ϕ_c are independently drawn from G_0. The Polya urn scheme for sampling the θ_i is equivalent to the following scheme for sampling the latent variables, c_i, and associated ϕ_c:

$$\begin{aligned} P(c_i = c \mid c_1, \ldots, c_{i-1}) &= \frac{n_{i,c}}{i-1+\alpha}, \quad \text{for } c \in \{c_j\}_{j<i} \\ P(c_i \neq c_j \text{ for all } j<i \mid c_1, \ldots, c_{i-1}) &= \frac{\alpha}{i-1+\alpha} \end{aligned} \tag{3}$$

where $n_{i,c}$ is the number of c_k for $k < i$ that are equal to c. The probabilities shown in (3) define the Dirichlet process model. This notation will be employed in subsequent sections.

3 Jain and Neal's conjugate split-merge procedure

We have previously introduced a split-merge Metropolis-Hastings procedure for conjugate Dirichlet process mixture models (Jain and Neal (2004); Jain (2002)). In the conjugate version of the algorithm, we assume that F is conjugate to G_0 in equation (1), so the model parameters, ϕ_c, in addition to the mixing distribution, G, can be integrated away. The state of the Markov chain consists only of the mixture component indicators, c_i.

This sampler proposes nonincremental moves that can produce major changes to the configuration of observations to mixture components in a single iteration. The split-merge proposals are evaluated by a Metropolis-Hastings procedure, in which split proposals are constructed by exploiting properties of a *restricted* Gibbs sampling scan on the component indicators, c_i. The Gibbs sampling scan is restricted in that it is only performed on a subset of the data (the observations associated with the merged component that is proposed to be split) and will only allocate observations between two mixture components.

To achieve more reasonable split proposals, several intermediate restricted Gibbs sampling scans are conducted prior to the final restricted Gibbs sampling scan, which is used to calculate the Metropolis-Hastings acceptance probability. The result of the last intermediate Gibbs sampling scan is denoted as the random *launch* state, from which the restricted Gibbs sampling transition probability is explicitly calculated. The number of intermediate restricted Gibbs sampling scans is considered a tuning parameter of this algorithm.

Note that for a merge proposal, there is only one way to combine items in two components to one component. However, deciding whether to accept or reject a merge proposal requires hypothetical consideration of the reverse split, which requires computations similar to those done for an actual split. A description of the steps involved in this algorithm, details to compute the Metropolis-Hastings acceptance probability, and a discussion of the validity of the conjugate version of the split-merge Metropolis-Hastings algorithm are provided in Jain and Neal (2004).

4 The nonconjugate split-merge procedure

We adapt Jain and Neal's conjugate split-merge Markov chain procedure described in Section 3 to accommodate models with nonconjugate priors. As mentioned earlier, because conjugate priors are not appropriate for all modeling situations, much of the recent Bayesian mixture modeling literature has been dedicated to nonconjugate algorithms (for instance, MacEachern and Müller (1998), Green and Richardson (2001), and Neal (2000)). A major impediment in designing nonconjugate procedures is the computational difficulty that arises when the model is no longer analytically tractable.

We say the model is nonconjugate when G_0 is not conjugate to F in the mixture model (equation 1). Aside from being unable to simplify the state of the Markov chain by integrating away the model parameters, ϕ, the main obstacle occurs when trying to sample for a new mixture component. When a c_i is updated, it can be set either to one of the other components currently associated with some observation or to a new mixture component. The probability of setting c_i to a new component involves the integral, $\int F(y_i;\phi)\, dG_0(\phi)$, which is analytically intractable in most nonconjugate situations. Allowances that some previous nonconjugate methods have made when dealing with this integral include approximating the true posterior distribution by another stationary distribution (which can be extremely detrimental) or creating model-specific *ad hoc* algorithms (which fail to generalize well).

Neal (2000) proposed two incremental Markov chain sampling procedures: Gibbs sampling with auxiliary parameters (Algorithm 8), and an incremental Metropolis-Hastings technique (Algorithm 5). These are exact Markov chain Monte Carlo methods that sample the correct posterior distribution and are straightforward to implement. However, in situations where the mixture components are nearby or similar in structure, these incremental methods' performance is analogous to the incremental methods for conjugate models (see Jain and Neal (2004)). To overcome their problems, such as remaining stuck in isolated modes and poor mixing between mixture components, we have developed a nonincremental split-merge alternative. In the next section, we compare empirically the performance of the new sampler to Neal's two incremental algorithms.

In this article, we show how such a nonincremental split-merge procedure can be applied when the model uses a particular type of nonconjugate prior, the conditionally conjugate family of priors. In conditionally conjugate models, it is still impossible to efficiently compute the integral, $\int F(y_i;\phi)\,dG_0(\phi)$. However, the pair F and G_0 are conditionally conjugate in one model parameter if the remaining parameters are held fixed. A well-known instance of this is the following Normal model. Suppose the observations, $y_1,\ldots,y_n$, are distributed as $F(y_i;\mu,\sigma^2) = \text{Normal}(y_i;\mu,\sigma^2)$, and the prior is $G_0(\mu,\sigma^{-2}) = \text{Normal}(\mu;w,B^{-1}) \cdot \text{Gamma}(\sigma^{-2};r,R)$. The distributions, $F(y_i;\mu,\sigma^2)$ and $G_0(\mu,\sigma^{-2})$, are conjugate in μ when σ^2 is fixed, and conjugate in σ^2 if μ is fixed. But, the joint posterior distribution is not analytically tractable. For the sake of brevity, when this nonconjugate Normal-Gamma prior is applied to a Normal mixture model, we will refer to it as the Normal-Gamma mixture model. Note, however, that this model using a conjugate prior, in which the mean and variance are *a priori* dependent, is sometime referred to similarly.

4.1 Restricted Gibbs sampling split-merge proposals

The conjugate split-merge algorithm of Section 3 cannot be applied directly to the conditionally conjugate case, but the basic mechanism of creating restricted Gibbs sampling split-merge proposals can still be applied. Since the model parameters, ϕ_c, cannot be integrated away, the state of the Markov chain for the split-merge sampler consists of both the component indicators and model parameters, denoted by $\boldsymbol{\gamma} = (\boldsymbol{c},\boldsymbol{\phi})$, where $\boldsymbol{c} = (c_1,\ldots,c_n)$ and $\boldsymbol{\phi} = (\phi_c : c \in \{c_1,\ldots,c_n\})$.

Conditional conjugacy in the model is required so that restricted Gibbs sampling scans can be performed to allocate observations reasonably between two mixture components. During these scans, we do not need to compute the integral, $\int F(y_i;\phi)\,dG_0(\phi)$, since we are only allocating observations between two known components that have at least one observation already assigned to them. For a nonconjugate model, a restricted

Gibbs sampling scan also updates the parameters for the affected mixture components, while holding the parameters of the other components fixed. Note that use of a restricted Gibbs sampling scan (and consequently, conditional conjugacy) is only crucial for the final Gibbs sampling scan from the launch state, since it allows the Metropolis-Hastings proposal density can be calculated. The intermediate scans could be replaced by some other type of Markov chain update.

Due to the inclusion of the model parameters, when two separate components are being merged to a single component, there is no longer only one possible component to merge into. The merged component is now defined by component parameters, which must be accounted for in the Metropolis-Hastings acceptance probability (in Section 4.3). The algorithm addresses this problem by conducting intermediate restricted Gibbs sampling for the merged component's parameters to arrive at a *launch state* (in a similar fashion as the "split" intermediate Gibbs sampling). From this launch state, **one** final restricted Gibbs sampling scan is performed to obtain the model parameters of the proposed merged component. The number of intermediate Gibbs sampling scans for the merged component's parameters is an additional tuning parameter in this algorithm. In this generalized version of the split-merge algorithm, there are therefore two launch states, $\boldsymbol{\gamma}^{L_{split}}$ and $\boldsymbol{\gamma}^{L_{merge}}$, that are necessary in order to calculate Gibbs sampling transition kernels for the split and merge proposal distributions.

4.2 Restricted Gibbs sampling split-merge procedure for the nonconjugate case

Let the state of the Markov chain consist of $\boldsymbol{\gamma} = (\boldsymbol{c}, \boldsymbol{\phi})$ where $\boldsymbol{c} = (c_1, \ldots, c_n)$ and $\boldsymbol{\phi} = (\phi_c : c \in \{c_1, \ldots, c_n\})$.

1. Select two distinct observations, i and j, at random uniformly.
2. Let S denote the set of observations, $k \in \{1, \ldots, n\}$, for which $k \neq i$ and $k \neq j$, and $c_k = c_i$ or $c_k = c_j$.
3. Define **launch** states, $\boldsymbol{\gamma}^{L_{split}}$ and $\boldsymbol{\gamma}^{L_{merge}}$, that will be used to define Gibbs sampling distributions required for the split and merge proposals.
   - Obtain launch state $\boldsymbol{\gamma}^{L_{split}} = (\boldsymbol{c}^{L_{split}}, \boldsymbol{\phi}^{L_{split}})$ as follows:
     - If $c_i = c_j$, then let $c_i^{L_{split}}$ be set to a new component such that $c_i^{L_{split}} \notin \{c_1, \ldots, c_n\}$ and let $c_j^{L_{split}} = c_j$. Otherwise, when $c_i \neq c_j$, let $c_i^{L_{split}} = c_i$ and $c_j^{L_{split}} = c_j$. For every $k \in S$, randomly set $c_k^{L_{split}}$, independently with equal probability, to either of the distinct components, $c_i^{L_{split}}$ or $c_j^{L_{split}}$. Initialize model parameters, $\phi_{c_i^{L_{split}}}^{L_{split}}$ and $\phi_{c_j^{L_{split}}}^{L_{split}}$, associated with the two distinct components by drawing new values from their prior distribution.
     - Modify $\boldsymbol{\gamma}^{L_{split}}$ by performing t intermediate restricted Gibbs sampling scans to update $\boldsymbol{c}^{L_{split}}$, $\phi_{c_i^{L_{split}}}^{L_{split}}$, and $\phi_{c_j^{L_{split}}}^{L_{split}}$.
   - Obtain launch state $\boldsymbol{\gamma}^{L_{merge}} = (\boldsymbol{c}^{L_{merge}}, \boldsymbol{\phi}^{L_{merge}})$ as follows:

- If $c_i = c_j$, then let $c_i^{L_{merge}} = c_j^{L_{merge}} = c_j$ (which is the same as c_i). Similarly, if $c_i \neq c_j$, then set $c_i^{L_{merge}} = c_j^{L_{merge}} = c_j$. For every $k \in S$, set $c_k^{L_{merge}} = c_j$. Initialize model parameter, $\phi^{L_{merge}}_{c_j^{L_{merge}}}$, associated with the merged component by drawing a new value from its prior distribution.
- Modify $\boldsymbol{\gamma}^{L_{merge}}$ by performing r intermediate restricted Gibbs sampling scans to update $\phi^{L_{merge}}_{c_j^{L_{merge}}}$.

4. If items i and j are in the same mixture component, i.e. $c_i = c_j$, then:

   (a) Propose a new assignment of data items to mixture components, denoted as $\boldsymbol{c}^{split}$, in which component $c_i = c_j$ is split into two separate components, c_i^{split} and c_j^{split}, and propose new values for the corresponding components' parameters, $\phi^{split}_{c_i^{split}}$ and $\phi^{split}_{c_j^{split}}$. Define each element of the candidate state, $\boldsymbol{\gamma}^{split} = (\boldsymbol{c}^{split}, \boldsymbol{\phi}^{split})$, as follows:

   - Let $c_i^{split} = c_i^{L_{split}}$ (note that $c_i^{L_{split}} \notin \{c_1, \ldots, c_n\}$)
   - Let $c_j^{split} = c_j^{L_{split}}$ (which is the same as c_j)
   - By conducting **one** final Gibbs sampling scan from the launch state, $\boldsymbol{\gamma}^{L_{split}}$, for every observation $k \in S$, let c_k^{split} be set to either component c_i^{split} or c_j^{split} and draw values for the model parameters, $\phi^{split}_{c_i^{split}}$ and $\phi^{split}_{c_j^{split}}$.
   - For observations $k \notin S \cup \{i, j\}$, let $c_k^{split} = c_k$, and for $c \notin \{c_i^{split}, c_j^{split}\}$, let $\phi^{split}_{c^{split}} = \phi_c$.

   (b) Compute the proposal densities, $q(\boldsymbol{\gamma}^{split}|\boldsymbol{\gamma})$ and $q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{split})$, that will be used to calculate the Metropolis-Hastings acceptance probability.

   - Calculate the split proposal density, $q(\boldsymbol{\gamma}^{split}|\boldsymbol{\gamma})$, by computing the Gibbs sampling transition kernel from the split launch state, $\boldsymbol{\gamma}^{L_{split}}$, to the final proposed state, $\boldsymbol{\gamma}^{split}$. The Gibbs sampling transition kernel is the product of the individual probabilities of setting each element in the launch state to its final proposed value during the final Gibbs sampling scan.
   - Calculate the corresponding proposal density, $q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{split})$, by computing the Gibbs sampling transition kernel from the merge launch state, $\boldsymbol{\gamma}^{L_{merge}}$, to the original merged configuration, $\boldsymbol{\gamma}$. The Gibbs sampling transition kernel is the product of the probability of setting each element in the original merge state (in this case, elements of ϕ_{c_j}) to its original value in a (hypothetical) Gibbs sampling scan from the merge launch state.

   (c) Evaluate the proposal by the Metropolis-Hastings acceptance probability $a(\boldsymbol{\gamma}^{split}, \boldsymbol{\gamma})$. If the proposal is accepted, $\boldsymbol{\gamma}^{split}$ becomes the next state in the Markov chain. If the proposal is rejected, the original configuration and model parameter, $\boldsymbol{\gamma}$, remain as the next state.

5. Otherwise, if i and j are in different mixture components, i.e. $c_i \neq c_j$, then:

   (a) Propose a new assignment of data items to mixture components, denoted as $\boldsymbol{c}^{merge}$, in which distinct components, c_i and c_j, are combined into a single component, and propose a new value for the corresponding merged component's model parameter, $\phi^{merge}_{c_j^{merge}}$. Define each element of the candidate state, $\boldsymbol{\gamma}^{merge} = (\boldsymbol{c}^{merge}, \boldsymbol{\phi}^{merge})$, as follows:

- Let $c_i^{merge} = c_i^{L_{merge}}$ (which is the same as c_j)
- Let $c_j^{merge} = c_j^{L_{merge}}$ (which is the same as c_j)
- For every observation $k \in S$, let $c_k^{merge} = c_j^{L_{merge}}$ (which is the same as c_j)
- For observations $k \notin S \cup \{i, j\}$, let $c_k^{merge} = c_k$, and for $c \neq c^{merge}$, let $\phi_{c^{merge}}^{merge} = \phi_c$.
- Conduct **one** final restricted Gibbs sampling scan from the launch state, $\boldsymbol{\gamma}^{L_{merge}}$, in order to draw a new value for the model parameter, $\phi_{c_j^{merge}}^{merge}$.

(b) Compute the proposal densities, $q(\boldsymbol{\gamma}^{merge}|\boldsymbol{\gamma})$ and $q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{merge})$, that will be used to calculate the Metropolis-Hastings acceptance probability.

- Calculate the merge proposal density, $q(\boldsymbol{\gamma}^{merge}|\boldsymbol{\gamma})$, by computing the Gibbs sampling transition kernel from the merge launch state, $\boldsymbol{\gamma}^{L_{merge}}$, to the final proposed state, $\boldsymbol{\gamma}^{merge}$. The Gibbs sampling transition kernel is the probability of setting $\phi_{c_j^{L_{merge}}}^{L_{merge}}$ to its final proposed value, $\phi_{c_j^{merge}}^{merge}$, via one Gibbs sampling scan.
- Calculate the corresponding proposal density, $q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{merge})$, by computing the Gibbs sampling transition kernel from the split launch state, $\boldsymbol{\gamma}^{L_{split}}$, to the original split configuration, $\boldsymbol{\gamma}$. The Gibbs sampling transition kernel is the product of the probabilities of setting each element in the original split state to its original value in a (hypothetical) Gibbs sampling scan from the split launch state.

(c) Evaluate the proposal by the Metropolis-Hastings acceptance probability $a(\boldsymbol{\gamma}^{merge}, \boldsymbol{\gamma})$. If the proposal is accepted, $\boldsymbol{\gamma}^{merge}$ becomes the next state. If the merge proposal is rejected, the original configuration and model parameters, $\boldsymbol{\gamma}$, remain as the next state.

4.3 The Metropolis-Hastings acceptance probability

The Metropolis-Hastings acceptance probability (Metropolis et al. (1953), Hastings (1970)) takes the following form when updating $\boldsymbol{\gamma} = (\boldsymbol{c}, \boldsymbol{\phi})$:

$$a(\boldsymbol{\gamma}^*, \boldsymbol{\gamma}) \quad = \quad \min\left[1, \; \frac{q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^*)}{q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})} \, \frac{P(\boldsymbol{\gamma}^*)}{P(\boldsymbol{\gamma})} \, \frac{L(\boldsymbol{\gamma}^*|\boldsymbol{y})}{L(\boldsymbol{\gamma}|\boldsymbol{y})}\right] \qquad (4)$$

where $\boldsymbol{\gamma}^*$ is either $\boldsymbol{\gamma}^{split}$ or $\boldsymbol{\gamma}^{merge}$ depending on the type of proposal.

The prior distribution, $P(\boldsymbol{\gamma})$, will be a product of the individual prior distributions for $\boldsymbol{c}$ and $\boldsymbol{\phi}$, since they are *a priori* independent. As before, the prior distribution for $P(\boldsymbol{c})$ will be a product of factors in equation (3). The ϕ_c for different mixture components are independent. Therefore, the prior distribution for $P(\boldsymbol{\gamma})$ is:

$$\begin{aligned} P(\boldsymbol{\gamma}) \quad &= \quad P(\boldsymbol{c}) \prod_{c \,\in\, \boldsymbol{C}} P(\phi_c) &\quad (5) \\ &= \quad \alpha^D \, \frac{\prod_{c \,\in\, \boldsymbol{C}} (n_c - 1)!}{\prod_{k=1}^{n} (\alpha + k - 1)} \prod_{c \,\in\, \boldsymbol{C}} g(\phi_c) &\quad (6) \end{aligned}$$

where D is the number of distinct mixture components, n_c is the count of items belonging to mixture component $c \in \boldsymbol{c}$, and $g(\phi_c)$ is the prior probability density function for ϕ_c for mixture component $c \in \boldsymbol{c}$.

For the split proposal, the appropriate ratio of prior distributions is:

$$\frac{P(\boldsymbol{\gamma}^{split})}{P(\boldsymbol{\gamma})} \quad = \quad \alpha \, \frac{(n^{split}_{c^{split}_i} - 1)!\,(n^{split}_{c^{split}_j} - 1)!\,g(\phi^{split}_{c^{split}_i})\,g(\phi^{split}_{c^{split}_j})}{(n_{c_i} - 1)!\,g(\phi_{c_i})} \tag{7}$$

where $\boldsymbol{\gamma}$ is the original state in which i and j belong to the same mixture component, $n^{split}_{c^{split}_i}$ and $n^{split}_{c^{split}_j}$ are the number of observations associated with each split component. The ratio of the prior distributions simplifies because the denominator in equation (6) and factors not associated with components that are directly involved in the Metropolis-Hastings update cancel.

For the merge proposal, the prior ratio simplifies to:

$$\frac{P(\boldsymbol{\gamma}^{merge})}{P(\boldsymbol{\gamma})} \quad = \quad \frac{1}{\alpha} \, \frac{(n^{merge}_{c^{merge}_i} - 1)!\,g(\phi^{merge}_{c^{merge}_i})}{(n_{c_i} - 1)!\,(n_{c_j} - 1)!\,g(\phi_{c_i})\,g(\phi_{c_j})} \tag{8}$$

where $n^{merge}_{c^{merge}_i}$ denotes the number of observations associated with the single merged component. $\boldsymbol{\gamma}$ represents the original state in which items i and j belong to separate components.

The likelihood, $L(\boldsymbol{\gamma}|\boldsymbol{y})$, will be a product over n observations:

$$L(\boldsymbol{\gamma}|\boldsymbol{y}) \quad = \quad \prod_{k=1}^{n} f(y_k; \phi_{c_k}) \tag{9}$$

$L(\boldsymbol{\gamma}|\boldsymbol{y})$ can be expressed as a double product over components, c, and items, $k \in \{1, \ldots, n\}$, associated with each component:

$$L(\boldsymbol{\gamma}|\boldsymbol{y}) \quad = \quad \prod_{c=1}^{D} \prod_{k\,:\,c_k = c} f(y_k; \phi_c) \tag{10}$$

where D is the number of distinct components. This expression to calculate the likelihood is often easier to use in real examples.

Likelihood factors involving items associated with components not directly involved in the split proposal cancel. The ratio of likelihoods in equation (4) reduces to the following:

$$\frac{L(\boldsymbol{\gamma}^{split}|\boldsymbol{y})}{L(\boldsymbol{\gamma}|\boldsymbol{y})} \quad = \quad \frac{\displaystyle\prod_{k\,:\,c^{split}_k = c^{split}_i} f(y_k; \phi^{split}_{c^{split}_i}) \prod_{k\,:\,c^{split}_k = c^{split}_j} f(y_k; \phi^{split}_{c^{split}_j})}{\displaystyle\prod_{k\,:\,c_k = c_i} f(y_k; \phi_{c_i})} \tag{11}$$

Likewise, for the merge proposal, the ratio of likelihoods is:

$$\frac{L(\boldsymbol{\gamma}^{merge}|\boldsymbol{y})}{L(\boldsymbol{\gamma}|\boldsymbol{y})} = \frac{\prod\limits_{k\,:\,c_k^{merge}=c_i^{merge}} f(y_k;\phi^{merge}_{c_i^{merge}})}{\prod\limits_{k\,:\,c_k=c_i} f(y_k;\phi_{c_i}) \prod\limits_{k\,:\,c_k=c_j} f(y_k;\phi_{c_j})} \tag{12}$$

The Metropolis-Hastings proposal density, $q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma})$, is the restricted Gibbs sampling transition kernel from launch state $\boldsymbol{\gamma}^L$ to final state $\boldsymbol{\gamma}^*$. This is a product of the conditional probabilities of each individual update of the vector $\boldsymbol{c}^*$ from $\boldsymbol{c}^L$ and the conditional densities of assigning successive components of $\boldsymbol{\phi}^L$ to their final values, $\boldsymbol{\phi}^*$.

Typically, for each mixture component, ϕ is composed of more than one model parameter, i.e. each ϕ_c can be a vector of parameters. For example, in the normal model, there are two parameters per component, $\phi_c = (\mu_c, \sigma_c^2)$. In a Gibbs sampling scan, each element of parameter ϕ_c is updated individually, while holding the other elements of ϕ_c fixed. A single element of ϕ_c is updated in a restricted Gibbs sampling scan by drawing a new value from its full conditional distribution.

We will denote the product of conditional probabilities obtained from **one full scan** of restricted Gibbs sampling as P_{GS}. Since $\boldsymbol{\gamma}$ is comprised of both $\boldsymbol{c}$ and $\boldsymbol{\phi}$, for clarity, we can split the Gibbs sampling transition kernel into its factors. The order of updating the variables does not affect the validity of the method, but for presentation purposes, we assume that Gibbs sampling updates $\boldsymbol{\phi}$ first (as is done in the later examples):

$$q(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}) \quad = \quad P_{GS}(\boldsymbol{\phi}^* \,|\, \boldsymbol{\phi}^L,\, \boldsymbol{c}^L,\, \boldsymbol{y}) \cdot P_{GS}(\boldsymbol{c}^* \,|\, \boldsymbol{c}^L, \boldsymbol{\phi}^*,\, \boldsymbol{y}) \tag{13}$$

An individual update of a particular c_k is as follows:

$$P(c_k \,|\, c_{-k},\, \phi_{c_k},\, y_k) = \frac{n_{-k,c_k}\, f(y_k;\phi_{c_k})}{n_{-k,c_i}\, f(y_k;\phi_{c_i}) \,+\, n_{-k,c_j}\, f(y_k;\phi_{c_j})} \tag{14}$$

where c_{-k} represents the c_l for $l \neq k$ in $S \cup \{i,j\}$, $n_{-k,c}$ is the number of c_l for $l \neq k$ in $S \cup \{i,j\}$ that are equal to c, and $f(y_k;\phi_c)$ is the likelihood. Here, c_k is restricted to being either c_i or c_j. Each time a c_k or ϕ_{c_k} is incrementally modified during a restricted Gibbs sampling scan, it is immediately used in the subsequent Gibbs sampling computation.

The required ratios for the split and merge proposals are shown below in equations (15) and (16), respectively. For the merge proposal, there is still only one way to combine items in two components into one component, so $P_{GS}(\boldsymbol{c}|\boldsymbol{c}^{L_{merge}},\, \boldsymbol{\phi},\, \boldsymbol{y}) = 1$ in equation (15). The same is true for $P(\boldsymbol{c}^{merge}|\boldsymbol{c}^{L_{merge}},\, \boldsymbol{\phi}^{merge},\, \boldsymbol{y})$ in equation (16). However, since specific parameters now define the mixture components, there are numerous possibilities for choosing a particular mixture component. We address this, in a similar method as the split scenario, by conducting intermediate Gibbs sampling scans to decide the value of the merged component's parameters. One final Gibbs sampling scan is conducted from the launch state to calculate the Gibbs sampling transition kernel.

The ratio of transition densities for the split proposal is:

$$\frac{q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{split})}{q(\boldsymbol{\gamma}^{split}|\boldsymbol{\gamma})}$$

$$= \frac{P_{GS}(\phi_{c_i}|\phi_{c_i}^{L_{merge}}, \boldsymbol{c}^{L_{merge}}, \boldsymbol{y})\, P_{GS}(\boldsymbol{c}|\boldsymbol{c}^{L_{merge}}, \boldsymbol{\phi}, \boldsymbol{y})}{P_{GS}(\phi^{split}_{c_i^{split}}|\phi^{L_{split}}_{c_i^{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\phi^{split}_{c_j^{split}}|\,\phi^{L_{split}}_{c_j^{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\boldsymbol{c}^{split}|\boldsymbol{c}^{L_{split}}, \boldsymbol{\phi}^{split}, \boldsymbol{y})}$$

$$= \frac{P_{GS}(\phi_{c_i}|\phi_{c_i}^{L_{merge}}, \boldsymbol{c}^{L_{merge}}, \boldsymbol{y})}{P_{GS}(\phi^{split}_{c_i^{split}}|\phi^{L_{split}}_{c_i^{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\phi^{split}_{c_j^{split}}|\phi^{L_{split}}_{c_j^{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\boldsymbol{c}^{split}|\boldsymbol{c}^{L_{split}}, \boldsymbol{\phi}^{split}, \boldsymbol{y})} \quad (15)$$

To calculate $q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{split})$, the same intermediate Gibbs sampling operations that are performed when proposing a merge must be conducted here to arrive at a suitable merge launch state, even though no actual merge is performed. The Gibbs sampling transition probability is calculated from the launch state (which is the last intermediate Gibbs sampling state) to the original merged state. These operations are necessary to produce the correct proposal ratios.

For the merge proposal, the ratio of transition densities is:

$$\frac{q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{merge})}{q(\boldsymbol{\gamma}^{merge}|\boldsymbol{\gamma})} = \frac{P_{GS}(\phi_{c_i}|\phi_{c_i}^{L_{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\phi_{c_j}|\phi_{c_j}^{L_{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\boldsymbol{c}|\boldsymbol{c}^{L_{split}}, \boldsymbol{\phi}, \boldsymbol{y})}{P_{GS}(\phi^{merge}_{c_i^{merge}}|\phi^{L_{merge}}_{c_i^{merge}}, \boldsymbol{c}^{L_{merge}}, \boldsymbol{y})\, P_{GS}(\boldsymbol{c}^{merge}|\boldsymbol{c}^{L_{merge}}, \boldsymbol{\phi}^{merge}, \boldsymbol{y})}$$

$$= \frac{P_{GS}(\phi_{c_i}|\phi_{c_i}^{L_{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\phi_{c_j}|\phi_{c_j}^{L_{split}}, \boldsymbol{c}^{L_{split}}, \boldsymbol{y})\, P_{GS}(\boldsymbol{c}|\boldsymbol{c}^{L_{split}}, \boldsymbol{\phi}, \boldsymbol{y})}{P_{GS}(\phi^{merge}_{c_i^{merge}}|\phi^{L_{merge}}_{c_i^{merge}}, \boldsymbol{c}^{L_{merge}}, \boldsymbol{y})} \quad (16)$$

To obtain $q(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{merge})$, we similarly perform the same intermediate Gibbs sampling moves when proposing a split, even though no actual split is proposed (since it is already known). This time the Gibbs sampling transition probability is calculated from the launch state to the original split state. This ensures correct proposal ratios.

The number of intermediate Gibbs sampling scans used to arrive at suitable launch states for both split and merge proposals are tuning parameters of this algorithm. There is an additional tuning parameter for the nonconjugate split-merge procedure that is not present in the conjugate version, which did not require a merge launch state.

4.4 Validity of the algorithm

The nonconjugate split-merge procedure described here is justified as a valid two-stage random Metropolis-Hastings procedure. In the first stage, we randomly select of observations i and j to decide which subset of Metropolis-Hastings proposals will be considered. In the second stage, we randomly select a launch state from among all possible launch states (given the selection of observations i and j), by means of intermediate Gibbs sampling scans. We then perform a standard Metropolis-Hastings update with a proposal distribution that depends on the selection of i and j and on the launch state.

As discussed by Tierney (1994), a random selection among transitions (in this case, via random selection of a proposal distribution) is a valid way of constructing Markov chain Monte Carlo algorithms, as long as all the transitions that might be selected are valid on their own.

A subtle clarification should be pointed out regarding the construction of the Metropolis-Hastings acceptance probability for the nonconjugate procedure. When a split is proposed from a merged state, only one ϕ_c is included in the equations, since the merged component has only one set of parameters associated with it now. We happen to initially pick ϕ_{c_j} to be associated with the observations in the merged component, but this is equivalent to initially selecting ϕ_{c_i} since the labels are irrelevant. To avoid changing dimensions when we compute the Metropolis-Hastings acceptance probability, we could include the appropriate ϕ_{c_i} terms in the computations. Since ϕ_{c_i} is an extra parameter for the merged component that is no longer associated with the data, we choose to propose a new value for it during the restricted Gibbs sampling scan by drawing from its prior distribution. This choice conveniently allows the prior density for this term to implicitly cancel with the corresponding term in the proposal density of the acceptance probability, showing that the change in dimensionality is not a problem. Consider the following set-up for the prior and proposal ratios for a split proposal which include the ϕ_{c_i} terms. We intentionally omit the likelihoods and indicator terms for simplicity and space considerations:

$$\frac{P(\phi^{split}_{c_i^{split}})\,P(\phi^{split}_{c_j^{split}})}{P(\phi_{c_i})\,P(\phi_{c_j})}\;\frac{P_{GS}(\phi_{c_i}|\phi_{c_i}^{L_{merge}},\,\boldsymbol{c}^{L_{merge}})\,P_{GS}(\phi_{c_j}|\phi_{c_j}^{L_{merge}},\,\boldsymbol{c}^{L_{merge}},\,\boldsymbol{y})}{P_{GS}(\phi^{split}_{c_i^{split}}|\phi^{L_{split}}_{c_i^{split}},\,\boldsymbol{c}^{L_{split}},\,\boldsymbol{y})P_{GS}(\phi^{split}_{c_j^{split}}|\phi^{L_{split}}_{c_j^{split}},\,\boldsymbol{c}^{L_{split}},\,\boldsymbol{y})}$$

The proposal factor, $P_{GS}(\phi_{c_i}|\phi_{c_i}^{L_{merge}},\,\boldsymbol{c}^{L_{merge}})$ does not depend on the data, since the ϕ_{c_j} factor has been selected earlier to be the merged component's parameter. Therefore, a new draw from ϕ_{c_i}'s conditional distribution will be equivalent to drawing a new value from its prior distribution, and this will cancel with the prior term, $P(\phi_{c_i})$. As a result, the ratios described earlier do not need to include these terms. The identical situation occurs in the case when a merge is proposed from an original split state and is handled similarly.

Note that it is possible to propose any configuration of observations from any initial state via a sequence of split and then merge proposals. However, to ensure ϕ-irreducibility on a continuous state space, it must be possible to propose any set of parameter values for each component. This will be true if each individual restricted Gibbs sampling conditional distribution for parameters of components that are involved in a particular split or merge update has a positive probability density of proposing any value. To ensure that the split-merge algorithm is well-defined, the model should satisfy the condition that the distributions $F(y_i;\theta_i)$ be mutually absolutely continuous for all θ in the support of G_0.

5 Performance of the nonconjugate split-merge procedure

Suppose we consider a Normal mixture model, in which the data, $\boldsymbol{y} = (y_1, \ldots, y_n)$, are independent and identically distributed, such that each observation, y_i, given the class, c_i, has m Normally distributed attributes, $(y_{i1}, \ldots, y_{im})$. An observation's attributes are independent given the class, c_i. The Normal mixture model is commonly used in Bayesian mixture analysis because of its simplicity in constructing conditional distributions and flexibility in modeling a number of heterogeneous populations simultaneously.

5.1 The Normal mixture model with Normal-Gamma prior

We model data from a mixture of Normal distributions using a Dirichlet process mixture model with Normal-Gamma prior, as follows:

$$\begin{aligned}
y_i \mid \mu_i,\, \tau_i &\sim F(y_i;\, \mu_i, \tau_i) \;=\; N(y_i;\, \mu_i, \tau_i^{-1}\, \boldsymbol{I}_m) \\
(\mu_i,\, \tau_i) \mid G &\sim G \\
G &\sim DP(G_0, \alpha) \\
G_0(\mu, \tau) &= N(\mu;\, w, B^{-1}) \,\cdot\, \mathrm{Gamma}\,(\tau;\, r, R)
\end{aligned} \tag{17}$$

where τ, the *precision* parameter, is σ^{-2}. Hyperpriors could be placed on w, B, r, and R to add another stage to this hierarchy if desired. Here, we consider these parameters to be known.

The probability density function for the prior distribution of μ given in (17) is:

$$g(\mu \,|\, w, B) = \left(\frac{B}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-B}{2}(\mu - w)^2\right) \tag{18}$$

where B is a precision parameter.

The probability density function for the prior for τ is:

$$g(\tau \,|\, r, R) = \frac{1}{R^r\, \Gamma(r)}\, \tau^{r-1} \exp\left(\frac{-\tau}{R}\right) \tag{19}$$

This parameterization of the Gamma density is adopted throughout this section.

These priors, equations (18) and (19), are necessary to compute the priors for the parameters in the Metropolis-Hastings acceptance probability of equation (4).

It is straightforward to set up the conditional distributions required for the restricted Gibbs sampling in the split-merge procedure used in the Metropolis-Hastings proposal densities. For the model parameters, this amounts to sampling from the marginal posterior distributions for a particular parameter of component c. The conditional posterior distribution for μ_{ch} (when τ_{ch} is known) for a specific attribute h is:

$$\mu_{ch} \,|\, \boldsymbol{c}, \boldsymbol{y}, \tau_{ch}, w, B \quad \sim \quad N\left(\frac{w\, B + \bar{y}_{ch}\, n_c\, \tau_{ch}}{B + n_c\, \tau_{ch}},\; \frac{1}{B + n_c\, \tau_{ch}}\right) \tag{20}$$

where n_c is the number of observations belonging to component c and $\bar{y}_{ch}$ is the mean of these observations for attribute h.

Similarly, if μ_{ch} is fixed, the conditional posterior distribution for τ_{ch} for a particular attribute h is:

$$\tau_{ch} \,|\, \boldsymbol{c}, \boldsymbol{y}, \mu_{ch}, r, R \quad \sim \quad \text{Gamma}\left(r + \frac{n_c}{2}, \frac{1}{R^{-1} + \frac{1}{2} \sum\limits_{k:c_k=c} (y_{kh} - \mu_{ch})^2} \right) \tag{21}$$

The conditional posterior distribution for an indicator variable, c_i, is obtained by combining the probability of the data (given in equation 17) given a value for c_i with the prior for indicators, $P(\boldsymbol{c})$. This yields for $c \in \{c_j\}_{j \neq i}$:

$$\begin{aligned} P(c_i = c \mid c_{-i},\, \mu_c, \tau_c, y_i) \quad &\propto \quad P(c_i = c \mid c_{-i}) \cdot P(y_i \mid \mu_c, \tau_c, c_{-i}) \qquad (22) \\ &\propto \quad n_{-i,c} \prod_{h=1}^{m} \tau_{ch}^{\frac{1}{2}} \exp\left(\frac{-\tau_{ch}}{2} \, (y_{ih} - \mu_{ch})^2 \right) \end{aligned}$$

These conditional distributions are also employed in computations required for Gibbs sampling with auxiliary parameters and incremental Metropolis-Hastings updates that will be used as comparisons to the nonconjugate split-merge technique later in this article.

The likelihood used in computing acceptance probabilities for split-merge updates is much simpler to obtain than in the conjugate case, since the parameters are not integrated away. For the mixture of Normals, the likelihood (given component indicators) is

$$L(\boldsymbol{\gamma}|\boldsymbol{y}) \quad = \quad \prod_{c=1}^{D} \prod_{k\,:\,c_k=c} \prod_{h=1}^{m} \left(\frac{\tau_{ch}}{2\pi} \right)^{\frac{1}{2}} \exp\left(\frac{-\tau_{ch}}{2} \, (y_{kh} - \mu_{ch})^2 \right) \tag{23}$$

Interchanging the products over k and h of equation (23) yields the following:

$$L(\boldsymbol{\gamma}|\boldsymbol{y}) \quad = \quad \prod_{c=1}^{D} \prod_{h=1}^{m} \left(\frac{\tau_{ch}}{2\pi} \right)^{\frac{n_c}{2}} \exp\left(\frac{-\tau_{ch}}{2} \sum_{k:c_k=c} (y_{kh} - \mu_{ch})^2 \right) \tag{24}$$

5.2 Illustration: Beetle Data

The Dirichlet process mixture model is a useful tool in model-based, unsupervised cluster analysis. We illustrate the practical utility of our split-merge algorithm with a six-dimensional data set from Lubischew (1962) that has been previously used by West et al. (1994). The data consists of six measurements of physical characteristics of three species of male beetles for a total of $n = 74$ beetles. The three species are *chactocnema concina, chactocnema heikertinger,* and *chactocnema heptapotamica*, in which $n_{conc} = 21$, $n_{heik} = 31$, and $n_{hept} = 22$.

The measurements for the i^{th} beetle are denoted as: $y_{ij} = (y_{i1}, \ldots, y_{i6})$ for $i = (1, \ldots, 74)$. The six measurements are:

$y_{.1}$ = width of the first joint	$\hat{\mu}_1 = 177.3$	$\hat{\sigma}_1^2 = 865.1$
$y_{.2}$ = width of the second joint	$\hat{\mu}_2 = 124.0$	$\hat{\sigma}_2^2 = 71.9$
$y_{.3}$ = maximal width of the aedeagus	$\hat{\mu}_3 = 50.4$	$\hat{\sigma}_3^2 = 7.6$
$y_{.4}$ = front angle of the aedeagus	$\hat{\mu}_4 = 134.8$	$\hat{\sigma}_4^2 = 107.1$
$y_{.5}$ = maximal width of the head	$\hat{\mu}_5 = 13.0$	$\hat{\sigma}_5^2 = 4.6$
$y_{.6}$ = aedeagus side-width	$\hat{\mu}_6 = 95.4$	$\hat{\sigma}_6^2 = 204.6$

The objective of our analysis is to recover the three latent classes corresponding to the three different species of beetles **without** using the species information in the analysis. We apply the Normal-Gamma Dirichlet process mixture model to this data, identical to equation 17. The Dirichlet process parameter, α, is set to one. The values for the priors of the parameters have been set for each dimension as follows: $w_j = (w_1, \ldots, w_6) = (100, 100, 50, 100, 25, 100)$, $B_j^{-1} = (B_1^{-1}, \ldots, B_6^{-1}) = (500, 100, 25, 100, 25, 150)$ where B is a precision parameter, $r = 1$ across all six dimensions, and $R = 5$ across all six dimensions.

We applied the nonconjugate split-merge algorithm (5,1,1,5), in which five intermediate Gibbs sampling scans were each used to reach the launch states for the split and merge proposals. One split-merge update was used in a single iteration and one final incremental Gibbs sampling scan was conducted after the final split-merge update. For comparison purposes, we considered the Gibbs sampling technique of Neal (2000) with $v = 3$ auxiliary components to this data. Computation time per iteration is similar for both algorithms. For each algorithm, results are provided for the case in which all observations are initially assigned to the same mixture component, and each algorithm is run for 5000 iterations.

From the two top trace plots given in Figure 1, it is evident that Gibbs sampling is unable to separate the data and leaves all observations in the same mixture component. It is clear that Gibbs sampling will take longer to reach equilibrium. On the other hand, split-merge splits the data into three major clusters (corresponding to the correct proportion of observations to species, i.e. 42%, 30% and 28%.) within the first twenty iterations.

To generate the two bottom trace plots in Figure 1, we set the prior values of w_j and B^{-1} to be more reflective of the data. The values used are: $w_j = (w_1, \ldots, w_6) = (100, 100, 50, 100, 10, 100)$ and $B_j^{-1} = (B_1^{-1}, \ldots, B_6^{-1}) = (800, 100, 10, 100, 10, 200)$. While Gibbs sampling does recover the three different species groups almost immediately, it is important to note that it becomes stuck in a low probability two-component configuration and mixes poorly. However, split-merge continues to mix well in a three-component configuration.

As a final check, the simulations were repeated by starting the simulation from a typical state of the competing method's apparent equilibrium distribution. Gibbs sampling stayed in the three-component state that it was started from, confirming that the three-component state has high posterior probability, and that the difference seen is not the result of some bug in the split-merge procedure. When the simulations were repeated using an initial state in which each observation is in a different component, the Gibbs sampler is able to reach equilibrium sooner and performs better.

The results from the beetle data illustration show that Gibbs sampling experiences a long burn-in time compared to the nonconjugate split-merge technique and is not always suitable for high-dimensional analysis. While it is true that the values of the priors for the parameters may not be ideal and that more realistic values may yield better sampling, often in real data analysis, there is no *a priori* information to suggest reasonable priors. A Markov chain Monte Carlo technique that can overcome poor choices in priors is preferred, as illustrated here, since this leads to shorter burn-in times and full exploration of the posterior distribution.

6 Discussion

The nonincremental split-merge procedure for nonconjugate models introduced in this article avoids the problem of being trapped in local modes, allowing the posterior distribution to be fully explored. In general, the nonconjugate split-merge procedure can become computationally expensive, but when Gibbs sampling or some other incremental procedure fails to reach equilibrium in a sensible amount of time, this procedure becomes necessary. Another related issue is burn-in time. Even if an incremental procedure reaches stationarity within a desired time limit, one must often discard a large number of early iterations, which can lead to poor estimates. In split-merge type situations, the computational burden of using a nonincremental procedure is offset by its quick burn-in and dramatic improvement in performance. To further improve sampling performance in which both large changes to the clustering configuration and small refinements are required, we recommend combining split-merge and Gibbs sampling updates as a way to reap the benefits of both samplers.

In higher dimensions, split-merge procedures continue to work well as the components are moved closer together. Convergence to the equilibrium distribution is relatively quick. It is possible that the split-merge procedure may break down for very high dimensional problems, because appropriate splits will be rejected, since it will become unlikely that a merge operation from the split state would produce the same merged parameter values as the current state. However, we have not encountered an example of this. Perhaps this issue arises only in situations where the dimensionality is in the hundreds.

A possible extension of the split-merge technique is to employ the Dahl (2003) sequentially allocated split-merge sampler as a method to initialize the intermediate Gibbs sampling step. This method could potentially provide a better starting state than our method of performing a random split of items and selecting values for the parameters from the prior.

7 Appendix

The purpose of the following simulation study is to classify observations into appropriate latent classes using the Normal-Gamma Dirichlet process mixture model. We can make this problem computationally more difficult by increasing the dimensionality of the data and by moving the components closer together. Various combinations of these factors were tested on all procedures. We found that the split-merge procedures outperformed the incremental procedures even in very low-dimensional problems, in which distinct components were visible by eye, showing the difficulty that incremental samplers have in reaching equilibrium even in simple problems when the components are similar.

We will consider two simulated data sets with a finite number of components. We expect that the Dirichlet process mixture model will model the finite situation perfectly well without problems such as overfitting, even though the model allows an infinite number of components. For each of the two examples, the data are composed of five equally-probable mixture components, in which each component is a distribution over m dimensions. To maintain uniformity amongst the examples, we generated $n = 100$ observations, stratified so that 20 observations came from each of the five mixture components.

Data for the two examples were randomly generated from the mixture distributions shown in Tables 1 and 2. Scatterplots of the data are shown in Figures 2 and 3. A standard deviation of 0.2 was selected for all Normal distributions, so that only the means would vary. The first example holds the dimensionality at two. The second example differs from the first in that the dimensionality is increased to three, and the components are closer together. Intentional asymmetry is introduced so that three components are more similar than the other two. This is intended to test whether the nonconjugate split-merge techniques can split in three ways.

The Dirichlet process parameter, α, is set to one for all demonstrations. Recall that a small value of α places stronger belief that the number of mixture components in the data is likely to be small. The parameters of the priors for the parameters on the component distributions have been set to the same values over all dimensions as follows: $w = 5$, $B = 1/12$, $r = 1$, and $R = 5$. Here, B is a precision parameter. For consistency, these parameters are fixed at these values for all simulations. In actual problems, these parameters could be set either by prior knowledge or given higher-level priors.

7.1 Performance

For the two examples, two incremental procedures, Gibbs sampling with $v = 3$ auxiliary variables, and an incremental Metropolis-Hastings method, are compared to four versions of the nonconjugate split-merge procedure. We use four parameters to describe the various split-merge procedures:

1. Number of intermediate Gibbs sampling scans to reach the launch state for a split proposal

Table 1: True mixture distribution for Example 1.

c	$P(c_i = c)$	$P(y_{ih} \mid c_i = c), h = 1, 2$	
1	0.2	N(2.0, 0.04)	N(3.0, 0.04)
2	0.2	N(3.0, 0.04)	N(2.0, 0.04)
3	0.2	N(3.3, 0.04)	N(3.3, 0.04)
4	0.2	N(8.0, 0.04)	N(9.0, 0.04)
5	0.2	N(9.0, 0.04)	N(8.5, 0.04)

Table 2: True mixture distribution for Example 2.

c	$P(c_i = c)$	$P(y_{ih} \mid c_i = c), h = 1, 2, 3$		
1	0.2	N(2.0, 0.04)	N(2.0, 0.04)	N(3.0, 0.04)
2	0.2	N(2.0, 0.04)	N(3.0, 0.04)	N(2.0, 0.04)
3	0.2	N(2.0, 0.04)	N(2.5, 0.04)	N(2.5, 0.04)
4	0.2	N(8.0, 0.04)	N(8.0, 0.04)	N(8.0, 0.04)
5	0.2	N(8.0, 0.04)	N(9.0, 0.04)	N(9.0, 0.04)

2. Number of split-merge updates done in a single overall iteration
3. Number of complete incremental Gibbs sampling scans after the final split-merge update
4. Number of intermediate Gibbs sampling scans to reach the launch state for a merge proposal

The four split-merge procedures we tested are described using these numbers as Split-Merge (0,1,0,0), Split-Merge (5,1,0,5), Split-Merge (0,1,1,0), and Split-Merge (5,1,1,5).

We compared the split-merge procedures with both the auxiliary variable and Metropolis-Hastings incremental samplers because we did not know beforehand which incremental method would perform better in situations where splits and merges might be necessary. Performance of the auxiliary variable Gibbs sampling is expected to improve as we increase the number of auxiliary components, except that it also takes longer per iteration (Neal (2000)). We did vary this parameter, but will report findings for $v = 3$ for all examples, since this version is comparable to the best version of split-merge in terms of computation time per iteration. As the incremental final scan for the split-merge procedure, Gibbs sampling with one auxiliary variable is used for all examples.

Performance measures that were considered include trace plots over time (Figures 4 and 5) and computation time per iteration (Table 3). The trace plots show five values which represent the fractions of observations associated with the most common, two most common, three most common, four most common, and five most common mixture components. Since each of the five components appear equally in the samples, if the true situation were captured exactly, the five traces would occur at values of 0.2, 0.4, 0.6, 0.8, and 1.0.

For each algorithm, all observations were assigned to the same mixture component for the initial state, and each algorithm was run for 5000 iterations. All simulations were performed on Matlab, Version 6.1, on a Dell Precision 530 workstation (which has a 1.7 GHz Pentium 4 processor). Note that the computation times reported include the extra time spent due to Matlab's inefficiencies when copying and incrementally updating arrays, which are not inherent in the algorithm.

7.1.1 Example 1

The three types of procedures, incremental Metropolis-Hastings, incremental Gibbs sampling with auxiliary variables, and split-merge, correctly classify the data in Figure 2 into five distinct clusters. The main difference in performance is the number of burn-in iterations that must be discarded.

The trace plots in Figure 4 show that Gibbs sampling with three auxiliary parameters has fewer burn-in iterations than the incremental Metropolis-Hastings method (compare 1000 to 3200 burn-in iterations). However, since the incremental Metropolis-Hastings method is approximately 5.5 times faster per iteration than the auxiliary Gibbs

Table 3: Time per iteration (in seconds) for the algorithms tested.

Algorithm	*Example 1*	*Example 2*
Incremental M-H	0.08	0.09
Gibbs Sampling	0.45	0.60
Split-Merge (0,1,0,0)	0.05	0.10
Split-Merge (0,1,1,0)	0.27	0.35
Split-Merge (5,1,0,5)	0.16	0.24
Split-Merge (5,1,1,5)	0.40	0.53

sampling method, it actually converges sooner with respect to computation time. Split-Merge (5,1,0,5) almost immediately splits the data into five components, but notice that the proportions do not occur at exactly 0.2 intervals until after the first thousand iterations. It takes this procedure longer to move a few singleton observations between components, since there is no final incremental update to make these minor adjustments. In five thousand iterations, it is not clear if Split-Merge (5,1,0,5) has actually reached the equilibrium distribution. Split-Merge (0,1,0,0) does not reach the equilibrium distribution in the five thousand iterations shown. Because the split and merge proposals have no intermediate Gibbs sampling scans, the proposals are not expected to be realistic. Split-Merge (0,1,0,0) is essentially a simple random split procedure, except that one restricted Gibbs sampling scan is conducted to reach the final state, which of course will not lead to reasonable split and merge proposals.

However, either by adding intermediate Gibbs sampling scans (as in the case of Split-Merge (5,1,0,5)) or adding a final full incremental scan (as in Split-Merge (0,1,1,0)), the correct proportion of items in each cluster is established. Split-Merge (0,1,1,0) eventually reaches the five component configuration after 500 burn-in iterations. The final procedure of Figure 4, Split-Merge (5,1,1,5), finds the five components immediately, and it appears that there is negligible burn-in (four iterations). The computation time per iteration is higher for Split-Merge (5,1,1,5) versus Split-Merge (0,1,1,0) and (5,1,0,5), but the computation time to equilibrium is much lower.

7.1.2 Example 2

Example 2 has three dimensions and the mixture components are close together. A perspective scatterplot of the data is given in Figure 3, and it shows that the components are difficult to distinguish. Given the priors selected, there is significant posterior probability for both the four and five mixture component configurations. Only Split-Merge (5,1,0,5) and Split-Merge (5,1,1,5) mix between these configurations, as observed in Figure 5. The incremental samplers and the split-merge procedures with zero intermediate restricted Gibbs sampling scans do not find the five components over the 5000 iterations, but are stuck in either two or four components. If each item is initially assigned to a different mixture component (plots not included), these samplers do split the data into five components, but take a long time to move to four components, indicating poor mixing. Here, the problem is that the deletion of a component is rare under both incremental updates and poor split-merge proposals.

Comparing further the two procedures that appear to converge, the autocorrelation time for trace 1 is much lower for Split-Merge (5,1,1,5) than Split-Merge (5,1,0,5) (126 vs. 718). For the autocorrelation time of an indicator variable, $I_{26,57}$, coding if observations 26 and 57 are in the same component, the time is much lower for Split-Merge (5,1,1,5) (38 vs. 417). Even though both algorithms do mix between the two configurations and Split-Merge (5,1,0,5) is faster per iteration, the improvement in autocorrelation time for Split-Merge (5,1,1,5) cannot be ignored. The extra full scan of incremental sampling for minor adjustments is worth the computational effort.

7.1.3 Summary of findings

It appears that split-merge moves are necessary in nonconjugate problems of this sort. Incremental samplers perform adequately when the components are distinct clusters in low dimensions, but as the components become more difficult to distinguish, these samplers take much longer to reach equilibrium. It is important to note that the incremental samplers begin to break down even in low dimensions. The split-merge procedures are able to handle three-way splits without any problems, although this is done by two two-way splits.

The split-merge procedure with several intermediate Gibbs sampling scans followed by an incremental full scan is the best version of the split-merge procedure. The split-merge method relies on proposing appropriate new clusters, which is accomplished by conducting several intermediate scans to reach the split and merge launch states. The split-merge methods generally have a longer computation time per iteration. However, in the case of the Gibbs sampling procedure with $v = 3$ auxiliary parameters, the best version of the split-merge procedure, Split-Merge (5,1,1,5), is slightly faster in our implementation (see Table 3). Therefore, there does not appear to be any advantage in using only incremental procedures for these types of problems.

References

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *Annals of Statistics*, 1: 353–355. 447

Blei, D. M. and Jordan, M. I. (2004). "Variational methods for the Dirichlet process." *ACM International Conference Proceeding Series: Proceedings of the twenty-first international conference on machine learning*. Vol. 69, article no. 12. 445

Dahl, D. B. (2003). "An improved merge-split sampler for conjugate Dirichlet process mixture models." Technical Report 1086, Department of Statistics, University of Wisconsin. 460

Do, K.-A., Müller, P., and Tang, F. (2005). "A Bayesian mixture model for differential gene expression." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54: 627–644. 445

Ferguson, T. S. (1983). "Bayesian density estimation by mixtures of normal distributions." In Rizvi, H. and Rustagi, J. (eds.), *Recent Advances in Statistics*, 287–303. New York: Academic Press. 445, 447

Green, P. J. and Richardson, S. (2001). "Modelling heterogeneity with and without the Dirichlet process." *Scandinavian Journal of Statistics*, 28: 355–375. 446, 448

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57: 97–109. 452

Huelsenbeck, J. P., Jain, S., Frost, S. W. D., and Pond, S. L. K. (2006). "A Dirichlet process model for detecting positive selection in protein-coding DNA sequences." *PNAS*, 103: 6263–6268. 445

Jain, S. (2002). "*Split-Merge Techniques for Bayesian Mixture Models*." Unpublished Ph.D. dissertation, University of Toronto, Department of Statistics. 447

Jain, S. and Neal, R. 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. 445, 446, 447, 448, 449

Lubischew, A. (1962). "On the use of discriminant functions in taxonomy." *Biometrics*, 18: 455–477. 458

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics: Simulation and Computation*, 23: 727–741. 446

MacEachern, S. N., Clyde, M., and Liu, J. (1999). "Sequential importance sampling for nonparametric Bayes models: the next generation." *The Canadian Journal of Statistics*, 27: 251–267. 445

MacEachern, S. N. and Müller, P. (1998). "Estimating mixture of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7: 223–238. 445, 446, 448

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *Journal of Chemical Physics*, 21: 1087–1092. 452

Neal, R. M. (1992). "Bayesian mixture modeling." In Smith, C. R., Erickson, G. J., and Neudorfer, P. O. (eds.), *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle 1991*, 197–211. Dordrecht: Kluwer Academic Publishers. 446

— (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 445, 446, 448, 449, 459, 463

Tierney, L. (1994). "Markov chains for exploring posterior distributions (with discussion)." *Annals of Statistics*, 22: 1701–1762. 456

West, M., Müller, P., and Escobar, M. D. (1994). "Hierarchical priors and mixture models, with application in regression and density estimation." In Freeman, P. R. and Smith, A. F. M. (eds.), *Aspects of Uncertainty*, 363–386. New York: Wiley. 458

Xing, E., Sharan, R., and Jordan, M. I. (2004). "Bayesian Haplotype Inference via the Dirichlet Process." *ACM International Conference Proceeding Series: Proceedings of the twenty-first international conference on machine learning*. Vol. 69, article no. 111. 445

Acknowledgments

Sonia Jain's research is supported by a UCSD Academic Senate Research Grant (Grant Number RG012H). Radford Neal's research is supported by the Natural Sciences and Engineering Research Council of Canada. He holds a Canada Research Chair in statistics and machine learning. The authors would like to thank the Associate Editor and referees for suggestions that improved this manuscript.

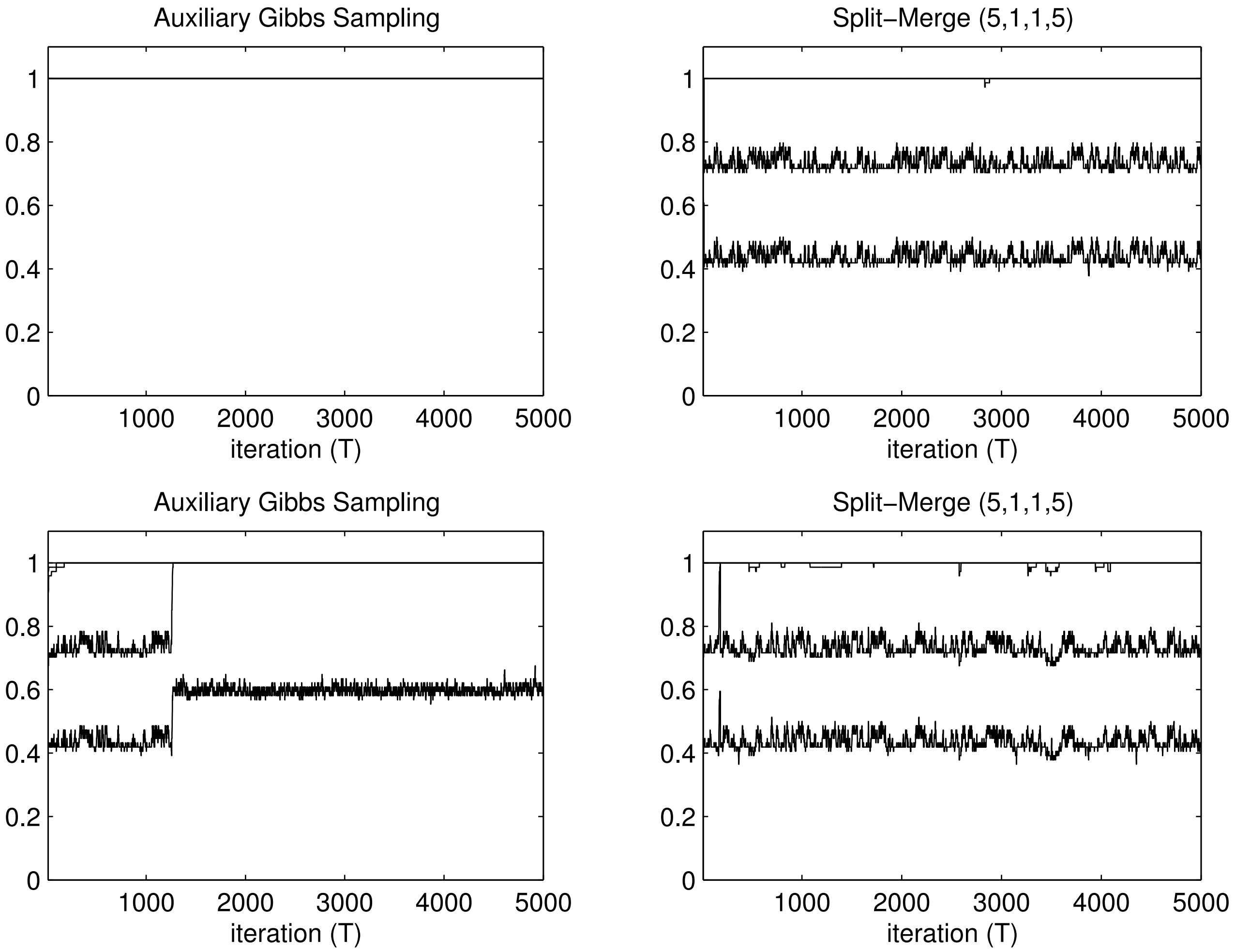


Figure 1: Trace plots comparing Auxiliary Gibbs Sampling to Split-Merge (5,1,1,5) for the beetle data using vague priors (top) and realistic priors (bottom). Trace plots show three traces which represent the fractions of observations associated with the most common, second most common, and third most common mixture components.

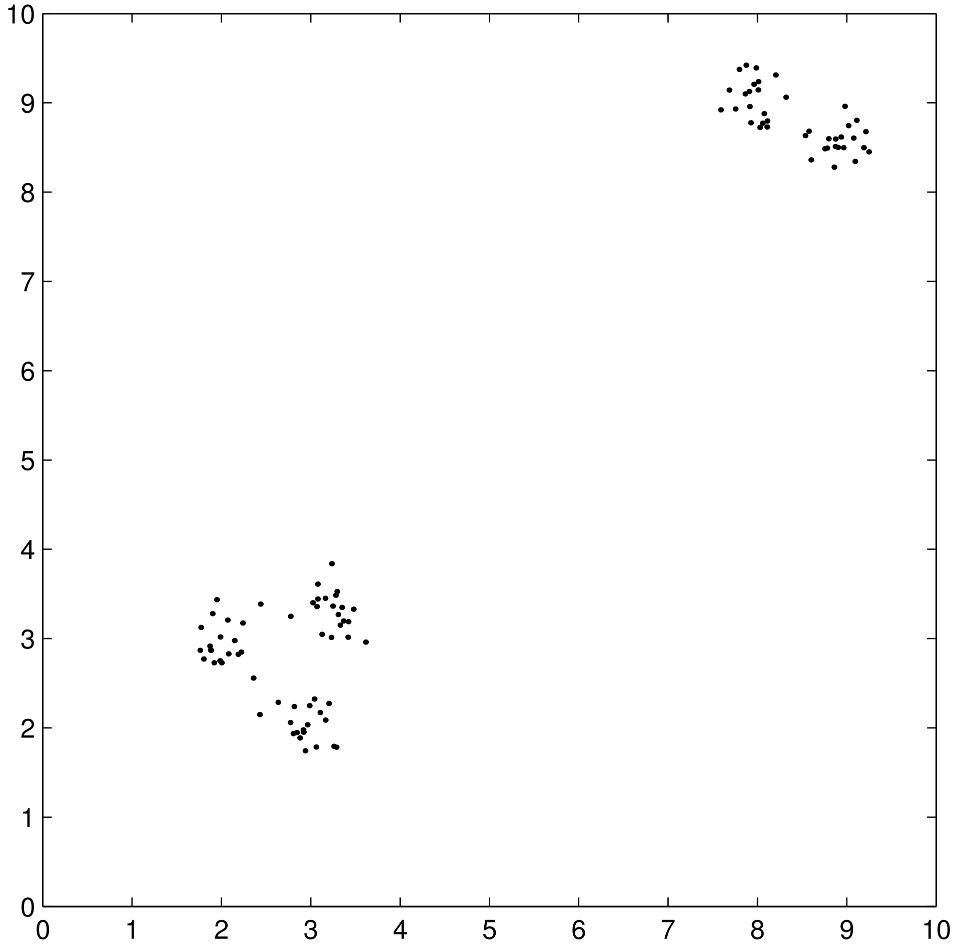


Figure 2: Scatterplot of the data in Example 1

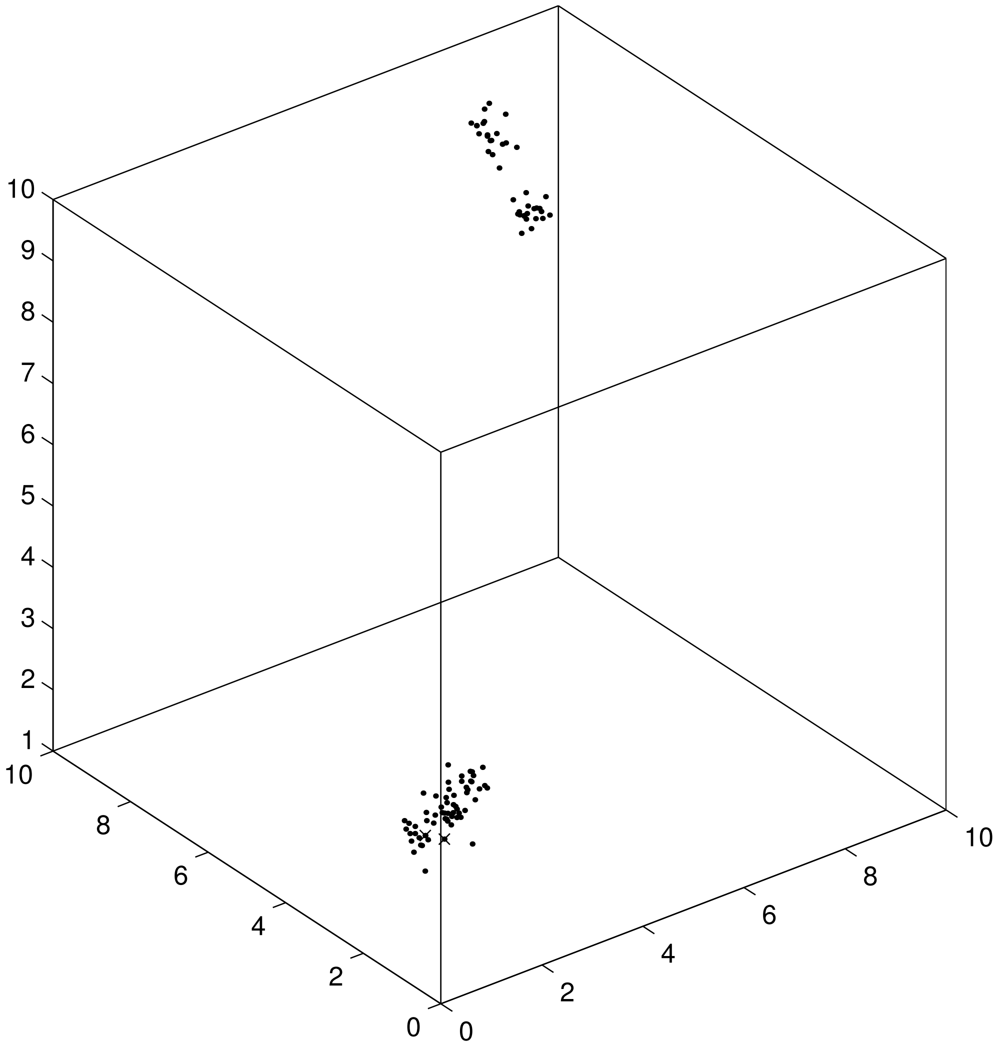


Figure 3: Scatterplot of the data in Example 2. The two x's represent observations 26 and 57 used in autocorrelation calculations for an indicator variable.

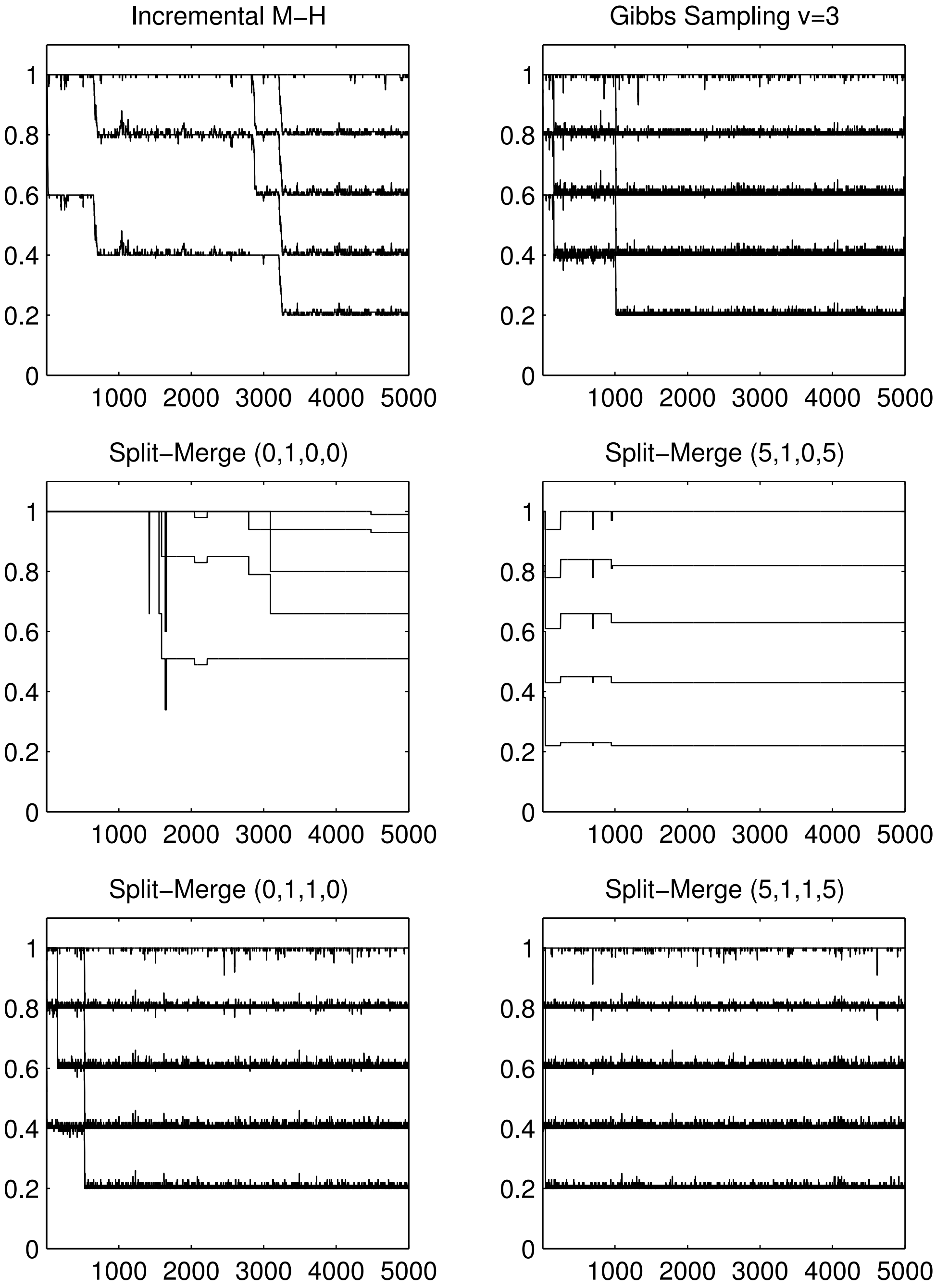


Figure 4: Trace plots of the six algorithms in Example 1.

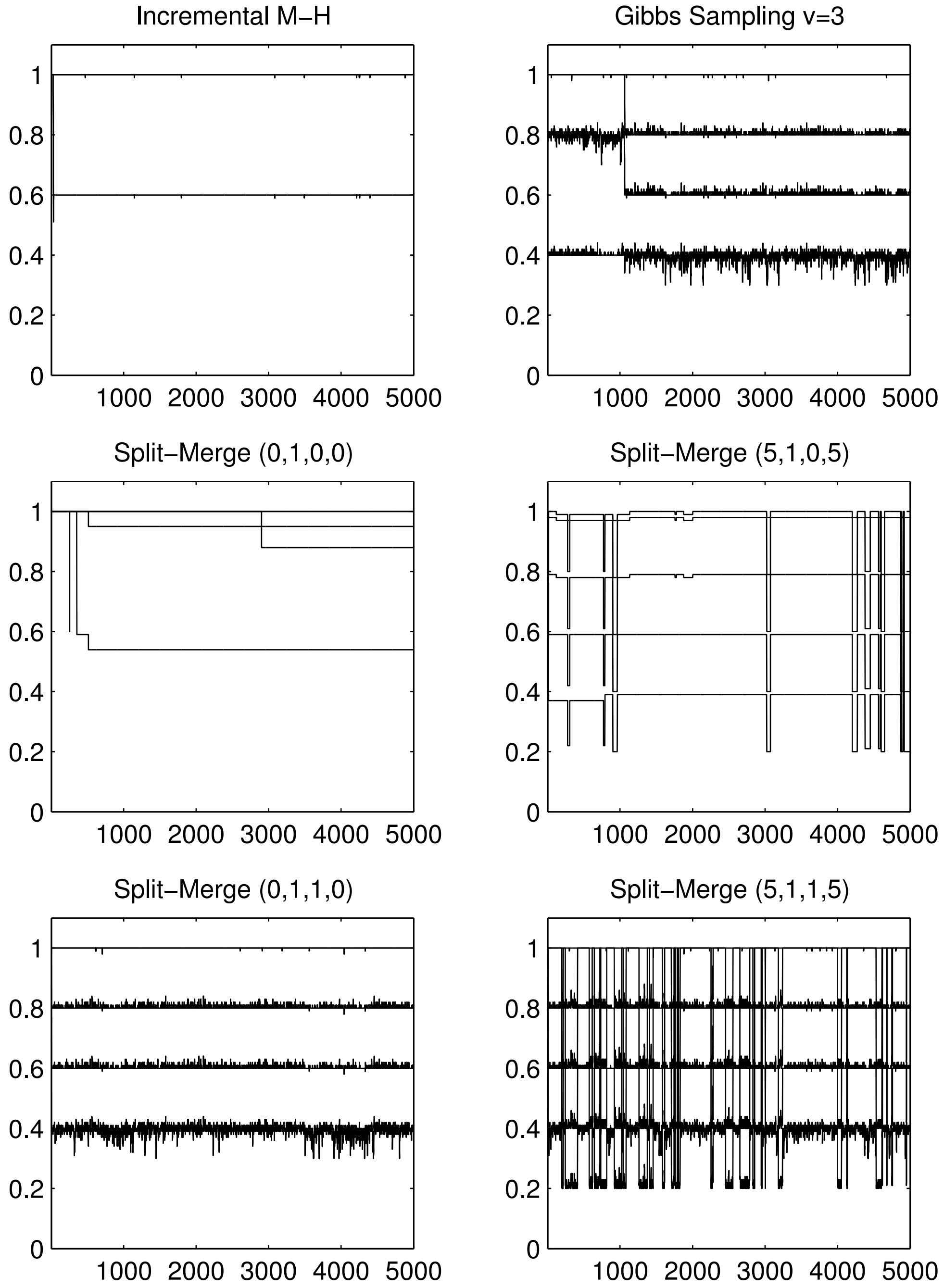


Figure 5: Trace plots of the six algorithms in Example 2.

Bayesian Analysis (2007) **2**, Number 3, pp. 473–478

Comment on Article by Jain and Neal

David B. Dahl*

1 Introduction

Sonia Jain and Radford Neal (JN) make a significant contribution to the literature on Markov chain Monte Carlo (MCMC) sampling techniques for Dirichlet process mixture (DPM) models. The paper presents some very nice ideas and will be on my required reading list for students working with me. DPM models are widely used for Bayesian nonparametric analyses and efficient sampling techniques are essential for their routine application. Incremental samplers for nonconjugate DPM models, such as the Auxiliary Gibbs sampler in Neal (2000), are easily implemented and potentially very efficient. Unfortunately, these samplers can also have difficulty mixing over the entire sample space and standard MCMC diagnostics may fail to indicate the problem. JN's paper represents a significant advance by providing a non-incremental sampler for conditionally conjugate DPM models.

The authors have a history of influential papers in this area, including Neal (2000) and Jain and Neal (2004). Their 2004 paper provided a split-merge sampler for conjugate DPM models, where the base distribution G_0 in the Dirichlet process prior is conjugate to the likelihood F. By exploiting this conjugacy, the model parameters of a cluster can be integrated away. The state of the Markov chain is merely the clustering of observations. Thus, sampling algorithms for conjugate DPM model attempt to sample from the posterior clustering distribution.

In nonconjugate DPM models, the model parameters of a cluster cannot be integrated away. Sampling algorithms must simultaneously address the clustering and the model parameters associated with each cluster. Green and Richardson (2001) were the first to propose a split-merge sampler for nonconjugate DPM models. Their procedure is based on reversible jump MCMC (Green 1995; Richardson and Green 1997) where the Metropolis-Hastings proposals are model-specific. In this paper, JN provide an MCMC sampler that can be generically applied to any conditionally conjugate DPM model.

2 Conditional Conjugate vs. Nonconjugate

It is important to note that conditional conjugacy is a necessary prerequisite for the application of JN's sampler. Suppose the model parameters for the cluster containing observation i are $\phi_1, \ldots, \phi_H$ with likelihood $F(y_i|\phi_1, \ldots, \phi_H)$ and prior $G_0(\phi_1, \ldots, \phi_H)$. A DPM model is conditionally conjugate if, for each $\phi_h \in \{\phi_1, \ldots, \phi_H\}$, $G_0(\phi_1, \ldots, \phi_H)$ is conjugate to $F(y_i|\phi_1, \ldots, \phi_H)$ in ϕ_h. JN's procedure relies on conditional conjugacy

[*] Department of Statistics, Texas A&M University, College Station, TX, http://www.stat.tamu.edu/~dahl

and hence their procedure is not applicable to all nonconjugate DPM model. Whether the conditional conjugacy constraint imposes a practical limitation is perhaps problem-specific.

3 Cluster Labels, Set Partition, and Implementation

JN describe their algorithm using notation involving cluster labels $c_1, \ldots, c_n$. An alternative way of describing sampling algorithms for DPM models uses set partition notation. In my experience, the set partition notation provides a straightforward presentation with simple notation. A set partition $\pi = \{S_1, \ldots, S_q\}$ of $S_0 = \{1, \ldots, n\}$ divides the n integers into mutually-exclusive, non-empty, and exhaustive clusters $S_1, \ldots, S_q$. I especially find the set partition notation helpful when translating sampling algorithms for DPM models to computer code. I use an array of length n whose elements point to C++ classes representing clusters containing the model parameters and a set of integers (for the cluster membership).

Regardless of notational preference, readers should be assured that the actual implementation of JN's sampler need not be complex. The core of my implementation of JN's split-merge procedure is 158 lines of C++ code, whereas the core of my implementation of the Auxiliary Gibbs sampler is 53 lines of C++ code. The extra time and mental effort needed to implement their split-merge sampler can pay large dividends for models and datasets where the Auxiliary Gibbs sampler is likely to have problems.

4 Initial States & Benefits of Split/Merge Samplers

I applied JN's split-merge algorithm to their Normal-Gamma mixture model and the beetle data used in their example. In the top two plots of JN's Figure 1, they use their vague priors with hyperparameters $w_j = (100, 100, 50, 100, 25, 100)$,
$B_j^{-1} = (500, 100, 25, 100, 25, 150)$, and $r = R = 1$ across all six dimensions. The top left plot corresponds to Auxiliary Gibbs sampling and shows that this sampler never moves away from the configuration with all observations in one cluster. JN contrast that with their Split-Merge (5,1,1,5) sampler (shown in the top right plot of JN's Figure 1) which is able to readily find the true three-component structure in the data.

In my implementation with 100 different random number seeds, the Auxiliary Gibbs sampler was able to find the three-component structure in 98 instances and a two-component structure (hinted at in the bottom left plot of JN's Figure 1) in the remaining two instances. Why could my implementation of the algorithm find the true structure, but their implementation of the same algorithm could not? The issue was the initial values of the model parameters. I sampled the initial value of the model parameters from the prior G_0. If, instead, I set the initial values of the model parameters to the sample means and precisions, I am able to replicate the results of JN in 100 of 100 instances. Also, if each observation is initially placed in its own cluster, the Auxiliary Gibbs sampler performed well (regardless of the method used to set the model parameters). My Figure 1 summarizes the results, showing that the problem with the Auxiliary Gibbs sampler

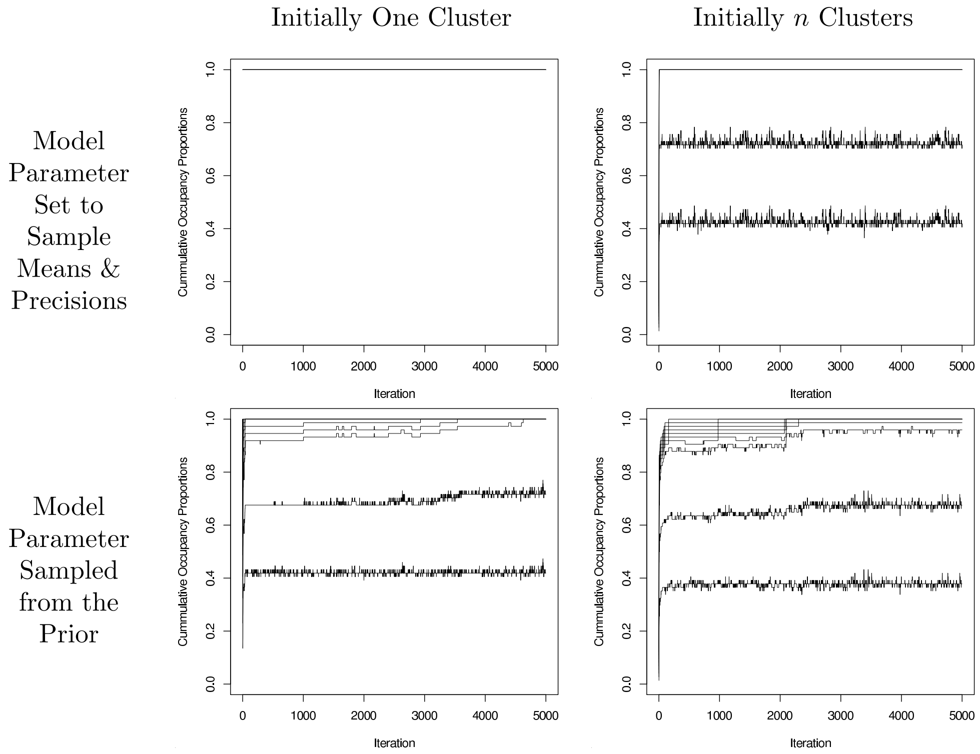


Figure 1: Trace plots from the Auxiliary Gibbs Sampler for the beetle data using JN's vague priors and four typical initial states for the Markov chain. The poor performance of the Auxiliary Gibbs sampler (shown in top left plot of JN's Figure 1) is only present when every observation is initially clustered together and the model parameters are initially set to the sample means and precisions.

is only present when every observation is initially clustered together and the model parameters are initially set to the sample means and precisions.

In my experience replicating the JN's Figure 1, their split-merge sampler is not sensitive to the initial values. For the beetle data and their Normal-Gamma mixture model, their sampler immediately finds the true three-component structure as shown in the plots on the right in JN's Figure 1.

It is interesting to observe that the posterior distribution apparently has virtually no support for anything other than three components. Notice that the Split-Merge (5,1,1,5) sampler never moves away from three clusters (to a configuration with two or four clusters, for example). The jitter present in the right hand size of JN's Figure 1 is due purely to the fact that their split-merge sampler embeds the Auxiliary Gibbs sampler (whose number of scans per split-merge attempt is given as the the third argument in

	Jain & Neal		*Dahl*	
Algorithm	*Example 1*	*Example 2*	*Example 1*	*Example 2*
Gibbs Sampling $v = 3$	1.00	1.00	1.00 (0.80, 1.13)	1.00 (0.84, 1.12)
Split-Merge (0,1,0,0)	0.11	0.17	0.17 (0.17, 0.18)	0.18 (0.17, 0.19)
Split-Merge (0,1,1,0)	0.60	0.58	0.61 (0.60, 0.62)	0.61 (0.60, 0.63)
Split-Merge (5,1,0,5)	0.36	0.40	0.84 (0.79, 0.88)	0.86 (0.82, 0.89)
Split-Merge (5,1,1,5)	0.89	0.88	1.32 (1.30, 1.35)	1.34 (1.32, 1.37)
Seconds per Iteration	0.45	0.60	5.50×10^{-4}	6.75×10^{-4}

Table 1: Comparison of relative CPU time of the various samplers depending on the dataset and the implementation. Jain & Neal columns are derived from Table 3. The Dahl columns show averages from 100 replications and the 2.5^{th} and 97.5^{th} quantiles. The data have been standardized by the "Seconds per Iteration" row to make them comparable across computers and programming languages.

the quad specifying the details of their sampler). Thus, the CPU time spent on trying to merge and split is wasted and time would be better spent on just the Auxiliary Gibbs sampler. The same can be said concerning the first simulated dataset in JN's Figure 4. In contrast, Example 2 (shown in JN's Figure 5) does provide a compelling case for the split-merge sampler. It freely moves between four and five components, whereas the Auxiliary Gibbs sampler is unlikely to easily switch between four and five components.

5 Timing

My final point concerns inherent variability in the implementation of algorithms due to the chosen programming language and data structures. JN have two simulated datasets (labeled Example 1 and Example 2) which they use to compare the various samplers. My Table 1 compares the CPU time of my C++ implementation of the various algorithms with that of JN's Matlab implementation. The first two columns are taken from JN's Table 3. The Dahl columns show averages from 100 replications and the 2.5^{th} and 97.5^{th} quantiles. The important point is the relative performance of the various sampling algorithms (not the speeds of different computers or programming languages), so the data has been scaled by the "Seconds per Iteration" row. Specifically, the Auxiliary Gibbs sampler with three auxiliary parameters (labeled as "Gibbs sampling $v = 3$") is set at 1.0 within each column.

Notice that relative CPU time taken by each of the samplers, within an implementation, is relatively constant across the two example datasets. There are, however, very different relative CPU times across implementations within a dataset. Recall that the Split-Merge(5,1,1,5) sampler embeds one Auxiliary Gibbs update with one auxiliary parameter per split-merge attempt. The Split-Merge(5,1,0,5) does not have any embedded Auxiliary Gibbs updates, leading to a $1 - 0.36/0.89 = 60\%$ reduction in the CPU time per iteration for JN's implementation of Example 1. In contrast, my implementation of Split-Merge(5,1,0,5) provides only $1 - 0.84/1.32 = 36\%$ reduction from my Split-Merge(5,1,1,5).

JN (2007) compare an Auxiliary Gibbs sampler with three auxiliary parameters with their Split-Merge(5,1,1,5) sampler which embeds an Auxiliary Gibbs sampler with one auxiliary parameter. They chose three and one auxiliary parameters respectively to make the CPU times comparable per iteration and then run each sampler for a fixed number of iterations. In my experience, additional auxiliary parameters are often not worth the extra CPU effort. For the sake of comparison, it might be more useful to have the number of auxiliary parameters be the same for both samplers. Comparisons would then be based on a fixed CPU time rather than a fixed number of iterations.

6 Conclusion

JN have made a significant contribution to the literature on sampling algorithms for DPM models. In implementing their algorithm and model and in using their example datasets, I found their method can have substantial benefits over the Auxiliary Gibbs sampler when used to sample from the posterior distribution of conditionally conjugate DPM models. Their algorithm is certainly more complicated than the Auxiliary Gibbs sampler, but perhaps not as difficult as one might initially expect. My experience with JN's Figure 1 reinforced the importance of using a variety of starting states, particularly when using the Auxiliary Gibbs sampler. It was nice to see that initial starting values were not an issue for JN's split-merge sampler. Although the relative CPU timings of JN's implementation and mine can be quite different, the salient point is that split-merge samplers can finding high-probability regions in posterior distributions that may be missed by incremental samplers.

References

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732. 473

Green, P. J. and Richardson, S. (2001). "Modelling heterogeneity with and without the Dirichlet process." *Scandinavian Journal of Statistics*, 28: 355–375. 473

Jain, S. and Neal, R. M. (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182. 473

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 473

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures With An Unknown Number of Components (Disc: P758-792) (Corr: 1998V60 P661)." *Journal of the Royal Statistical Society, Series B, Methodological*, 59: 731–758. 473

Bayesian Analysis (2007) **2**, Number 3, pp. 479–482

Comment on Article by Jain and Neal

C.P. Robert*

From a stylistic point of view, I think this paper reads very much like a sequel to the important paper Jain and Neal (2004) and therefore it is not exactly self-contained since the main bulk of the paper is a commentary of the program provided in Section 4.2. Instead of the current version, I would thus have preferred a truly self-contained version with a more user-friendly introduction, for instance when reading and re-reading Sections 3 and 4.1...[1]

The central point of the paper is to extend Jain and Neal (2004) so that the lack of complete conjugacy of the prior does not prevent the algorithm from being run. Indeed, in Jain and Neal (2004), the model parameters are completely hidden in that the likelihood and the prior only depend on the cluster index vector $\mathbf{c}$, which means working in a finite set. The difficulty with priors G_0 that do not lead to closed form marginals is that the parameters must take part in the simulation process. The idea at the core of the current paper is to take advantage of the conditional conjugacy, i.e. the fact that the prior on a given parameter is still conjugate and thus manageable, conditional on all the other parameters, so that a Gibbs sampling version can be implemented.

At this stage, I understand the rationale of the partial conjugacy for the Metropolis-Hastings ratio to be computed (Section 4.1) but I wonder how difficult it would be to extend the idea to any type of prior distribution. I also note that at both split and merge stages the algorithm simulates new values of the parameter from the *prior* distribution, rather than from a more adapted distribution. This is as generic as it can be, but simulating from vague priors usually slows down algorithms and it is of course impossible for improper priors. It thus seems to me that the factor t directing the number of intermediate Gibbs (or Metropolis-Hastings) iterations in Step 3 must be influential in the overall behaviour of the algorithm and that large values of t may be necessary to overcome the dependence on the starting value.

More generally, I also wonder why a more global tempering strategy would not fare better than the local split-merge proposals used in the paper. For illustration purposes, I implemented below the regular Gibbs sampler in the [BetaBinomial] Example 1 of Jain and Neal (2004) and compared it with a naïve tempered version where the tempered likelihood L_τ is made of a product of $\tau \geq 1$ (sub)likelihoods based on a partition of the observations in τ random clusters, τ being itself uniform on $\{1, \ldots, n/2\}$. (The advantages of using this form of tempering are (a) that the same Gibbs sampler can be used for the sublikelihoods and (b) that the normalising constant of the tempered version is still available, as opposed to the choice of a power of the likelihood. The acceptance probability at the end of the tempered moves is then function of the likelihood ratio $L(\theta|x)/L_\tau(\theta|x)$ and can be directly computed.) As shown on Figure 1 *(bottom)*,

*CREST and CEREMADE, Uni. Paris Dauphine, France, mailto:xian@ceremade.dauphine.fr

[1] This may explain why the following reads more like an eloped referee's report than like a true discussion!

explained below, the mixing and the exploration of various likelihood values is quite improved with this tempered scheme, since no column sticks to a single colour theme.

Since Dirichlet mixtures are closely related to mixtures, I would have liked to read some discussion on the label switching phenomenon (see, e.g., Stephens 2000; Marin et al. 2005; Jasra et al. 2005). Indeed, while the original model of Jain and Neal (2004) is somehow impervious to the issue of label switching, since the clustering parameterisation only focus on class allocations, the introduction of the parameter in the game means that a proper exploration of the posterior requires the reproduction of the symmetry in the various components of the mixture. Using a split-merge basis for this exploration may then prove to be insufficiently powerful for this task.

In fact, it is close to impossible to judge of the overall convergence performances from the simulation output, which solely concentrates on the cluster sizes. Additional graphical summaries would be welcome, like the "allocation map" advertised in Robert and Casella (2004) and represented on both Figures 1 and 2. The pixelised lines on the pictures represent the cluster index via different colours for all observations, the index on the first axis being the index of the observation. The second axis corresponds to the iteration index. Long vertical stripes of similar colours indicate poor mixing of the algorithm.

In this illustration, we see clearly that the 5 equal groups of Example 1 of Jain and Neal (2004) are identified by the Gibbs sampler–as signalled by the homogeneous columns $1-20$, $21-40$, $41-60$, $61-80$ and $81-100$—and, furthermore, that label switching does occur, even if at a very slow pace—as shown by columns $61-80$ for instance.

A point of detail (?) is that the algorithm must be (is) validated as a Gibbs procedure rather than as a Metropolis-Hastings algorithm, given that at any stage only a subset of the parameters and of the clustering indicators is updated. In addition, this is quite an interesting example of algorithmic bypassing the varying dimension pitfalls, since it avoids dealing with the measure theoretic subtleties encountered by reversible jump for instance (Green 1995) while being in a continuous varying dimension state space, contrary to the setup of Jain and Neal (2004).

References

Green, P. (1995). "Reversible jump MCMC computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. 480

Jain, S. and Neal, R. (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *J. Computat. Graphical Statist.*, 13(1): 158–182. 479, 480, 481

Jasra, A., Holmes, C., and Stephens, D. (2005). "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." *Statistical Science*, 20(1): 50–67. 480

Marin, J., Mengersen, K., and Robert, C. (2005). "Bayesian Modelling and Inference

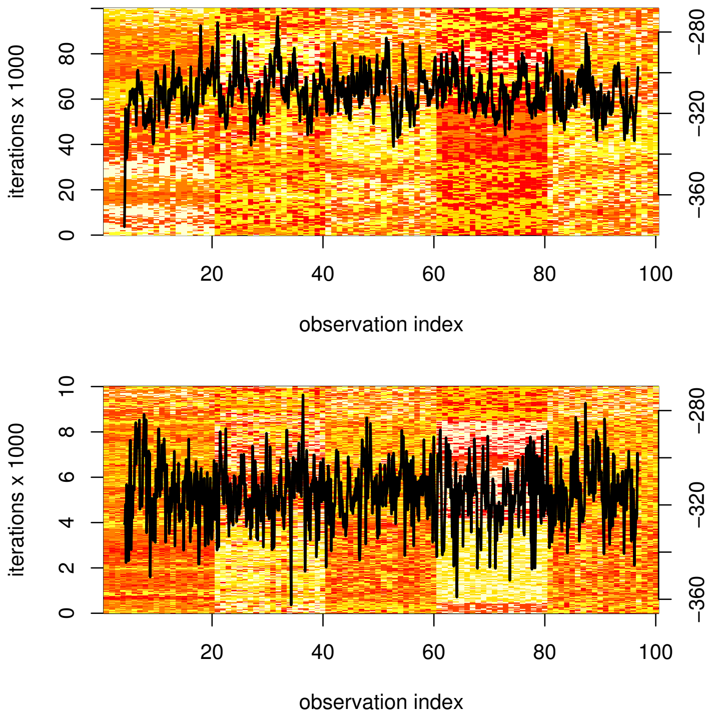


Figure 1: *(top)* Allocation map of the simulated cluster index vector $\mathbf{c}^{(t)}$ for $m = 6$, $n = 100$ observations and $T = 10^5$ Gibbs iterations (subsampled every 1000 iteration), in the setup of Example 1 of Jain and Neal (2004). The colours used in the graphs range from red (1) to white (6) and identify the labels of the cluster indicators c_i along the iterations. The superimposed graph is the corresponding sequence of likelihood values over the $T = 10^5$ Gibbs iterations, associated with the scale on the right hand side. *(bottom)* Same representation for a tempered version with $T = 10^3$ iterations made of $T_o = 10^2$ tempered moves.

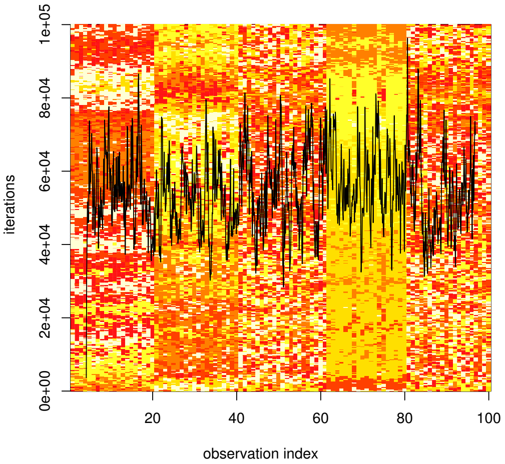


Figure 2: Same representation as Figure 1 for another run of the Gibbs sampler,

on Mixtures of Distributions." In Rao, C. and Dey, D. (eds.), *Handbook of Statistics*, volume 25. Springer-Verlag, New York. 480

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition. 480

Stephens, M. (2000). "Dealing with label switching in mixture models." *J. Royal Statist. Soc. Series B*, 62(4): 795–809. 480

Bayesian Analysis (2007) **2**, Number 3, pp. 483–494

Comment on Article by Jain and Neal

Steven N. MacEachern*

1 Introduction

It was with great interest that I read Jain and Neal's paper. In the paper, they address a tough problem, namely how to improve the mixing/convergence of Markov chain Monte Carlo (MCMC) algorithms for an important class of models. The models are those involving mixtures of Dirichlet processes, ranging from a fairly straightforward mixture of Dirichlet processes model to the more complex models that are springing up in a wide variety of applications. The algorithms are in the split-merge vein, allowing a different kind of step than incremental Gibbs samplers. The extension of the split-merge technology with targeted proposals to conditionally conjugate models is a welcome addition to the collection of transitions available for fitting models that include the Dirichlet process as a component.

Jain and Neal's algorithms (see also Dahl, 2005) have refined the technology of split-merge samplers so that proposals are no longer "blind", but, through intermediate Gibbs scans, move toward a region of higher posterior probability. The ability to target better proposals results in algorithms that naturally make better proposals, and this improves mixing of the Markov chain. An important element of these intermediate Gibbs scans is their ability to move toward a more appropriate launch state.

This discussion focuses on two features that are hidden in the innards of the algorithm. The first is the notion of identifiability and the second is that of a random scan. Jain and Neal's algorithms make nice use of a non-identifiable model for the intermediate Gibbs scans (section 4.2, step 3 and following) to produce what are presumably better proposals. They also implicitly use a random scan for split and merge proposals in the sense that cases i and j are selected at random (section 4.2, step 1). The remainder of this discussion looks at these issues in the context of a simple, artificial example where one can explicitly calculate rates of convergence for a variety of incremental Gibbs algorithms. The hope is that the example, in spite of its simplicity, provides insight into the effectiveness of the algorithms and suggests potential directions for their further refinement.

2 Identifiability

While details of various algorithms are left for the next section, one recurring issue in proposals for novel algorithms for Dirichlet based models is identifiability. This issue is not limited to mixture models, but arises in many other contexts. There is

[*] Department of Statistics, The Ohio State University, Columbus, OH, mailto:snm@stat.ohio-state.edu

often a connection between identifiability and the convergence rate of a Markov chain: Identifiable models may show quicker convergence to the limiting distribution than do non-identifiable models. This has led some to suggest a general principle that non-identifiable models be avoided when MCMC methods are to be used to fit the model. This section reviews the arguments raised against non-identifiable models, and the following section develops the arguments in more detail through consideration of a simple example.

Consider a model where there is a parameter space, say Θ. The distribution of the data depends on the value of the parameter, so that $X \sim F_\theta$ for some $\theta \sim \Theta$. A model is non-identifiable if there exist $\theta_1, \theta_2 \in \Theta$, with $\theta_1 \neq \theta_2$, for which $F_{\theta_1} = F_{\theta_2}$. Models that are not non-identifiable are called identifiable models. Typically, when the model is non-identifiable, it will be the case that for every $\theta_1 \in \Theta$ there exists a $\theta_2 \in \Theta$, with $\theta_2 \neq \theta_1$ for which $F_{\theta_1} = F_{\theta_2}$.

Several reasons have been given for avoiding the use of non-identifiable models. First, while a Bayesian approach places a prior over the parameter space, and so, in principle, there is no difficulty in creating estimates with this methodology, there is the question of consistency. Identifiability is closely connected with parameter estimation. Methods such as maximum likelihood cannot distinguish between parameter values that imply the same distribution for the data, and so may not produce unique estimates. Bayes estimates, heavily based on the likelihood, are typically also inconsistent for non-identifiable models. However, if consideration is restricted to identifiable functionals, the Bayes estimates will typically be consistent, as they are under identifiable models. A desire to interpret parameter values directly is closely related to a desire for consistency. Restricting interpretation to identifiable quantities $g(\theta)$, such that if $g(\theta_1) \neq g(\theta_2)$ then $F_{\theta_1} \neq F_{\theta_2}$, the worry about non-identifiability disappears. A complete, identical Bayes analysis could be done on an identifiable model. This first objection has no connection to the use of MCMC methods.

Second, there are examples where the convergence rate of a Markov chain is improved by the choice of an identifiable model. The convergence here is convergence of π_n, the distribution of $\theta^n|\theta^0$, to the limiting distribution of the Markov chain. The limiting distribution is, by construction, also the posterior distribution of θ. The main purpose of the simulation is to provide estimates of posterior summaries, and, although there is a difference between the accuracy of these estimates and the convergence rate, in most circumstances the two produce qualitative agreement: A better convergence rate means more accurate estimators. This issue is examined in the next section.

Third, there is the practical issue of how well the MCMC algorithm works when actually implemented. The main concerns are the numerical accuracy and stability of the computations. In some instances, particularly with very diffuse posterior distributions, some of the parameter values generated during the course of the simulation may be enormous. This can lead to unstable computations and hence to inaccurate estimates.

Fourth, there is the issue of prior elicitation. The choice of a model has an impact on the particular prior that is chosen. This choice is not directly tied to the use of MCMC methods, but is an issue of increasing importance now that more complex models are being fit. Examples include collinearity and the variable selection problem where priors are chosen according to prescription, problems based on the hierarchical model, and nonparametric Bayes problems.

Consider fitting models with MCMC methods. The Markov chain upon which the simulation is based is realized through successive generations of a parameter vector, θ. The chain, assumed to be irreducible and aperiodic, is also assumed to have a fixed transition matrix, say P. Consequently, it has a limiting distribution, π. The transition matrix is chosen in such fashion that π is the posterior distribution for $\theta|X$. A realization of the chain consists of a sequence $\theta^1, \theta^2, \ldots, \theta^N$. Convergence is often described in terms of the total variation norm: We wish $||\pi_n - \pi||$ to approach 0 quickly. For finite state chains, the rate of convergence is governed by the second largest eigenvalue of the transition matrix. The convergence rate of more complex chains is determined by a similar quantity.

The MCMC method constructs P by creating a set of transition kernels. For a fixed scan algorithm, the overall transition kernel is the product of, say, p transition kernels, $P = P_1 \ldots P_p$. A random scan sampler selects one of the P_i at random. A popular choice is to select the P_i with equal probabilities, so that $P = p^{-1} \sum P_i$.

Two useful techniques for improving convergence of a sampler are (i) to generate a block of parameters at a time (say $\theta_1, \ldots, \theta_c$ is generated from $[\theta_1, \ldots, \theta_c | \theta - \{\theta_1, \ldots, \theta_c\}, X]$), and (ii) to collapse or coarsen the state space of the Markov chain by reducing the dimension of θ. The dimension of θ is reduced through integration. For example, θ_p may be marginalized, leaving only $\theta_1, \ldots, \theta_{p-1}$. For discussion, theory and examples of (i) and (ii) see Liu (1995); of (ii) see also MacEachern (1994). The impact of non-identifiability on MCMC algorithms is closely connected to blocking and coarsening.

3 Illustration

Nonparametric Bayesian models have been considered for several decades. Early models, such as those of Kraft and van Eeden (1964) and Ramsey (1972) for the bioassay problem, provided a start in the area. These models were based on the notion of a Dirichlet distribution being the conjugate prior for multinomial data. The models were nonparametric in the sense that the prior had full support on the set of multinomial probability vectors. This work was followed by the well-known work of Ferguson (1973) and Antoniak (1974). Early work exploiting mixtures of Dirichlet processes includes Berry and Christensen (1979) and Lo (1984).

The mixture of Dirichlet process model has many applications beyond bioassay. The basic mixture of Dirichlet processes model may be written as follows:

$$\begin{aligned} F &\sim Dir(\alpha) \\ \theta_1, \ldots, \theta_p | F &\sim F \\ X_i | \theta_i &\sim G_{\theta_i}, \text{ for } i = 1, \ldots, p. \end{aligned}$$

Here, following Ferguson's notation, α, the positive, finite measure that parameterizes the Dirichlet process, is often split into its total mass, M, and its shape, say F_0. Thus if α is a measure on the real line, F_0 is a distribution function, $M > 0$, and $\alpha((-\infty, x]) = MF_0(x)$.

G_θ is a distribution indexed by the parameter θ. The models are easily generalized to include hyperparameters that index α, groups of observations associated with each θ_i, observation specific covariates, and additional parameters common to some or all observations. The bioassay problem is one which fits into this framework.

There are three main types of MCMC methods that have been widely used for the mixture of Dirichlet process models. The first is based directly on the hierarchical model written above. It makes use of the sequence of conditional generations $[F|\theta], [\theta|F]$. See Kuo and Smith (1992), Gelfand and Kuo (1991) and, in a general setting, Ishwaran and Zarepour (2000) for details. See also Diebolt and Robert (1994) in the context of a related finite mixture model.

The second type of Markov chain method makes use of an alternative representation of the Dirichlet process known as the Polya urn scheme (Blackwell and MacQueen, 1973). Under the Polya urn scheme, the random distribution function F is marginalized, resulting in the model

$$\begin{aligned} \theta_1, \ldots, \theta_p &\sim F_{\theta_1, \ldots, \theta_p} \\ X_i|\theta_i &\sim G_{\theta_i}, \text{ for } i = 1, \ldots, p. \end{aligned}$$

To simplify description, take F_0 to be continuous. With this model, the components θ_i are no longer conditionally independent. Instead, they have a distribution that is built up sequentially: $\theta_1 \sim F_0$. For $i > 1$, θ_i is set equal to θ_j with probability $1/(M + i - 1)$ and is drawn from F_0, independent of previous draws from F_0, with probability $M/(M + i - 1)$. The induced distribution on the vector θ is often thought of in two parts. The first is the partition of θ into distinct values, and the second is the location of the, say k, elements of the partition. Each partition receives positive probability under the prior. Given a partition, the k locations of the elements, denoted $\theta_1^*, \ldots, \theta_k^*$, are i.i.d. draws from F_0. A Markov chain based on this representation of the model involves sequential generation of $[\theta_i|\theta_{-i}]$, for $i = 1, \ldots, p$, with the updating performed immediately in each case. See Escobar (1994) and Escobar and West (1995) for algorithms of this sort. These algorithms may be refined by discarding the locations of the clusters and running a Markov chain on only the space of partitions of θ (Neal, 1992; MacEachern, 1994). Such chains tend to produce quicker convergence to the posterior and naturally suggest better estimators. The calculations below refer to this last refinement of the algorithm, though they can be replicated when the locations are present.

The third type of algorithm is the split-merge algorithm with its ability to make large moves in directions not easily traveled in with algorithms of the first two types. The simple example can be fit with the simple split-merge algorithm of Jain and Neal (2000). In this case, the improvements in the algorithm do not change its performance.

The Markov chain runs on a state space which consists of all partitions of θ into clusters. This is a finite state space, which is denoted by S. An element in the state space is a p-dimensional vector, $s = (s_1, \ldots, s_p)$, with component s_i indicating to which cluster θ_i belongs. If there are k clusters of θ_i, there will be k distinct integers in the partition vector. If θ_i and θ_j are in the same cluster, $s_i = s_j$; if in different clusters, $s_i \neq s_j$. The fact that the state space is finite allows us to perform exact calculations on the transition matrix of the Markov chain in small examples. Several chains are compared for the case of $p = 3$. A major issue is the labelling of the state space. Two identifiable labellings and one non-identifiable labelling are considered. The labelling/identifiability issue is cleanest for Type II algorithms. The labellings are presented in Table 1.

The first Type II scheme numbers the clusters consecutively from 1 to k as they are built up from the Polya urn scheme. Thus $s_1 = 1$, and for all i for which $\theta_i = \theta_1$, $s_i = 1$. The second cluster is begun by the first $\theta_i \neq \theta_1$, and so $s_i = 2$ for $i = inf[j|\theta_j \neq \theta_1]$. All other θ_j equal to this θ_i are in this cluster and so are assigned $s_j = 2$. The numbering of the later clusters proceeds in a similar fashion, so that for a legitimate partition vector (i.e., one which receives positive probability under the prior) representing k clusters, the numbers 1 through k will appear and their first appearances will occur in increasing order. The final legitimate values of s for the case $p = 3$ appear in Table 1 under the heading scheme 1. With this parameter space, the model is identifiable. Each legitimate configuration vector produces a distinct partition of the θ and hence (under the mild regularity condition that there is a set of θ_i with positive F_0 probability such that $G_{\theta_1} = G_{\theta_2}$ iff $\theta_1 = \theta_2$) produces a distinct distribution for X.

The second Type II scheme is similar to the first in that there is a 1-1 mapping between partitions and legitimate configuration vectors. The difference is in how the clusters are labelled. Again, all θ_i in a cluster will have the same index in the configuration vector. Those θ_i in the cluster with θ_1 have $s_i = 1$. Further clusters have an index equal to $inf[j|\theta_j$ in cluster$]$. For example, define $i = inf[j|\theta_j \neq \theta_1]$. Then $s_j = i$ for all j such that $\theta_j = \theta_i$. The legitimate values for s under this labelling scheme when $p = 3$ appear in Table 1 under the heading scheme 2. Since there is a $1 - 1$ mapping between this labelling and the previous one, identifiability for this model follows from identifiability of scheme 1.

The third Type II scheme produces a non-identifiable model. With this scheme, the clusters will each receive a distinct integer from 1 to n, and each θ_i in a particular cluster will receive the same index. There is, however, no other restriction on the index values assigned to the clusters. To create this scheme formally, begin with the first labelling scheme. Probabilities of the legitimate states are determined by the Polya urn scheme. Then the probability for a particular configuration is distributed among the possible labellings for the configuration. For a configuration with k clusters, there are $n!/(n-k)!$ distinct labellings. The probability for this configuration is distributed uniformly among these labellings. This model is clearly non-identifiable, since there are several parameter values (here several different configuration vectors) which produce the same distribution for the data. Interestingly, $[F|\theta, X, s]$ depends on s only through the configuration. Hence, any inference depends only on the equivalence class on s defined by the configuration itself.

State	Configuration	scheme 1	scheme 2
a	$\theta_1, \theta_2, \theta_3$	1,2,3	1,2,3
b	$\theta_1 = \theta_2; \theta_3$	1,1,2	1,1,3
c	$\theta_1 = \theta_3; \theta_2$	1,2,1	1,2,1
d	$\theta_1; \theta_2 = \theta_3$	1,2,2	1,2,2
e	$\theta_1 = \theta_2 = \theta_3$	1,1,1	1,1,1

Table 1: Labellings of configurations under schemes 1 and 2.

Gibbs samplers were developed for each of the labelling schemes above for the no-data problem. The transition matrix for a fixed scan, in the order $[s_1|s_2, s_3]$, $[s_2|s_1, s_3]$, and then $[s_3|s_1, s_2]$ was calculated analytically. For the first two schemes, the second largest eigenvalue of the transition matrix was determined. To compare the third scheme to the first two, identifiable functions are considered. In order to determine an effective rate of convergence for these functions, the transition matrix for the sampler is rewritten in terms of an identifiable model. Happily, all of the transition vectors from each non-identifiable state corresponding to a particular configuration to the distinct configurations are identical (e.g., the transition probability for moving from the state $s = (1, 1, 3)$ to the configuration $\theta_1 = \theta_2 = \theta_3$ is the same as the transition probability for moving from the state $s = (2, 2, 1)$ to the configuration $\theta_1 = \theta_2 = \theta_3$). The chain, in terms of this identifiable state space, retains the Markov property. The implication is that the second largest eigenvalue of the rewritten transition matrix governs the rate of convergence in the identifiable space.

The three Gibbs samplers corresponding to the three labelling schemes were compared by means of the second largest eigenvalue of their transition matrices, presented in Table 2. The comparison of the three schemes shows that scheme 3, based on the non-identifiable model, produces the best performance. The non-identifiable model results in better mixing.

Simulations were carried out to compare the Type I algorithm to the Type II algorithms. The simulation made use of a non-identifiable version of the Type I algorithm. The estimated second largest eigenvalue of the Type I algorithm appears in Table 2 along the row labelled Type I. Scheme 3 appears to dominate this type of algorithm. This conclusion agrees with results that suggest a collapse of the state space improves the convergence rate of a Markov chain, since the scheme 3 algorithm may be constructed by adding generations to a Type I algorithm and then collapsing the state space. Interestingly, this is in opposition to the sometimes expressed intuition that a two-stage Gibbs sampler, as the Type I method, should show quicker convergence than a three-stage Gibbs sampler, as the scheme 3 algorithm is. These results in this simple context are in agreement with the careful simulations for more realistic settings in Papasiliopoulos and Roberts (2008).

The random scan Gibbs sampler was investigated in a similar fashion. Table 2 contains a summary of the results for 3 transitions (so chosen to match the three transitions

M	1	5	10	100
scheme 1	.222	.327	.389	.485
scheme 2	.222	.0408	.0139	.000192
scheme 3	.0370	.00292	.000579	9.42e-7
Type I	.301	.0837	.0332	.000559

M	1	5	10	100
scheme 1	.559	.630	.669	.726
scheme 2	.559	.395	.352	.303
scheme 3	.171	.0787	.0588	.0393
Split-merge	1.00	.152	.216	.287

Table 2: Second largest eigenvalues for MCMC algorithms. The top table is for fixed scan samplers; the bottom table is for random scan samplers. M is the mass of the base measure of the Dirichlet process.

of the fixed scan sampler). Notice that the second largest eigenvalues are considerably larger for random scan samplers, corresponding to the potentially long lags between successive sampling of a component. Again, scheme 3, corresponding to the non-identifiable model, is preferable to the Type II schemes. The Type III (split-merge) sampler is, for the larger values of M, preferable to the Type II samplers that impose identifiability. In this example, it does not mix as well as the non-identifiable algorithm. Interestingly, when $M = 1$, the sampler yields a periodic Markov chain, and so mixing is poor although estimation (barring an even subsampling rate) is fine. It should be noted that this periodicity is very special to this example.

4 Heuristics

The simplicity of the example allows us to focus on features of the algorithms that impact mixing: Comparisons among the Type II algorithms suggest that non-identifiability (of a certain sort) improves mixing; the comparison between fixed and random scans suggests that fixed scans lead to better mixing; a good Type II algorithm leads to better mixing than a Type I algorithm; for *small* clusters, the Type II algorithm mixes better than the Type III algorithm.

Within Type II algorithms, the example shows a remarkable advantage for the non-identifiable model. This appears to follow from the conditioning sets used to create the Gibbs sampler. The non-identifiable model leads to conditioning sets that contain the conditioning sets arising from the identifiable model. To illustrate this point, a schematic of the transition matrices is provided in Table 3. Comparing the two P_1's, for instance, under scheme 1 the transition matrix is the identity while under scheme 3 it is a block diagonal matrix with only two blocks. Both chains are based on conditional generations. For each current state, the set conditioned upon for the generation under

P_1											
From	To	a	b	c	d	e	a	b	c	d	e
a		x	-	-	-	-	x	x	x	-	-
b		-	x	-	-	-	x	x	x	-	-
c		-	-	x	-	-	x	x	x	-	-
d		-	-	-	x	-	-	-	-	x	x
e		-	-	-	-	x	-	-	-	x	x
P_2											
From	To	a	b	c	d	e	a	b	c	d	e
a		x	-	-	-	-	x	x	-	x	-
b		-	x	-	x	-	x	x	-	x	-
c		-	-	x	-	x	-	-	x	-	x
d		-	x	-	x	-	x	x	-	x	-
e		-	-	x	-	x	-	-	x	-	x
P_3											
From	To	a	b	c	d	e	a	b	c	d	e
a		x	-	x	x	-	x	-	x	x	-
b		-	x	-	-	x	-	x	-	-	x
c		x	-	x	x	-	x	-	x	x	-
d		x	-	x	x	-	x	-	x	x	-
e		-	x	-	-	x	-	x	-	-	x

Table 3: Scheme 1 transition matrices on the left, scheme 3 transition matrices on the right. The states are described in Table 1. A dash indicates that a transition cannot take place, an x that it can. Note the enlargement of the sets over which conditional generations take place with scheme 3.

the scheme 3 chain contains the set conditioned upon for the generation under the scheme 1 chain. Thus the conditioning sets for the scheme 1 chain are nested in those for the scheme 3 chain. The following result connects the nesting of conditioning sets to total variation distance.

Proposition 1. Suppose that we have a countable state space, and a distribution π which assigns positive probability to each state. Further suppose that this state space is partitioned into conditioning sets C_i. Define row i of the transition matrix P to consist of the distribution π, restricted to the conditioning set in which state i lies. Consider two partitions, A and B, where $\{C_{A,i}\}$ is a refinement of $\{C_{B,i}\}$ and the corresponding transition matrices P_A and P_B. Then, for any initial distribution, π_I, $||\pi_I^{'} P_A - \pi|| \geq ||\pi_I^{'} P_B - \pi||$.

Proof. The total variation distance between the distributions F and G is defined by $||F - G|| = sup_A(|F(A) - G(A)| + |F(A^C) - G(A^C)|)$ where A ranges over all subsets of the state space. When the initial distribution π_I is modified through a transition governed by a conditional distribution over a partition, the supremum is attained by a set A for which each element of the partition is either entirely contained in A or entirely contained in A^C. Since the conditioning sets used to create P_A are a refinement of those used to create P_B, we may view the supremum in the former case as being taken over a larger set. Hence, $||\pi_I^{'} P_A - \pi||$ is at least as large as $||\pi_I^{'} P_B - \pi||$.

Proposition 1 shows that one step of the chain based on larger conditioning sets (i.e., the sampler based on the non-identifiable model) is preferable to one step of the chain based on the smaller conditioning sets. However, the proof given here does not extend to more steps. Presumably, the quicker one-step movement toward the posterior will often carry over into a quicker rate of convergence for the chain, as it does in the example of Section 3. Consideration of the impact of identifiability underlay, in part, the development of nonconjugate algorithms in MacEachern and Muller (1998).

As Jain and Neal comment, the Type III algorithms are most beneficial when there are large clusters of observations. With only a few large clusters, all observations will frequently have a chance to switch clusters. However, my experience with models involving the Dirichlet process is that the posterior distribution typically includes a number of small clusters (in addition to the large clusters). The simple example suggests that including Type II steps is important to facilitate mixing for these small clusters.

5 Conclusions

The example presented herein, as well as others that I have examined, lead to the following viewpoint on the four reasons presented earlier for avoiding non-identifiable models. The first, interpretation of the model, has no connection to whether MCMC methods are used to fit the model, and so in no way suggests that one restrict themself to use of identifiable models. The second reason seems to be largely irrelevant. The important convergence rate (if an identifiable model is to be considered at all) is convergence for estimates of identifiable functionals. This may be quicker than the convergence rate of the chain in the non-identifiable space. In any event, if an effective chain can be created based on the identifiable form of the model, the same chain can be created based on the non-identifiable form of the model. The third concern, for numerical stability of the computations, remains a concern. The fourth issue is one of prior elicitation. Since models and prior distributions are subjective and situation specific, any recommendation for one form of model over another is open to criticism. Nevertheless, some classes of models seem much more natural than do others. Often, as in the case of the hierarchical model, these classes contain non-identifiable models. A decision to replace a natural, non-identifiable model with an identifiable model that seems to be less natural seems unwise without a demonstrated improvement in the ease or effectiveness with which the model is fit.

My own view on problems necessitating MCMC methods is this. One should first write down the most natural model, whether it be identifiable or non-identifiable. Next, lay out several MCMC methods for this version of the model. Further consider expanding the parameter space to create non-identifiable models. Particular consideration should be given to inducing non-identifiability by adding symmetries such as the relabelling of the clusters in the simple Dirichlet process example. Again, examine a batch of MCMC algorithms, with attention to generating blocks of parameters and to marginalizing parameters. Finally, select an algorithm based on the heuristics of preferring those derived from larger conditioning sets, those that have collapsed the state space, and those that generate blocks of parameters at a time. To this algorithm, add steps that target particularly difficult transitions–such as splitting and merging large clusters.

The hints in Jain and Neal's paper and the simple example suggest a natural direction for extension of the split-merge moves: a move away from a random scan (i.e., random selection of observations i and j that determine the attempted split/merge) and toward a scan with reduced randomness. The randomness of the scan can be lessened, for example, by permuting the indices from 1 through n, and using successive pairs for i and j. This type of permutation bounds the time between successive attempts at updating each observation's cluster membership. In turn, this ensures that the number of iterates until every observation-specific parameter has had a chance to be updated is controlled. I suspect that the benefits that Jain and Neal have demonstrated of combining both incremental and split-merge moves in an algorithm are partly due to the implicit reduction in randomness–a complete incremental Gibbs scan ensures that all cases have had the opportunity to move.

A second possible extension is to reserve the split/merge moves for clusters of substantial size. To do so, one could partition the parameter space into two parts–one part where the combined number of cases in clusters identified by observations i and j exceeds some threshold and the second part where the combined number of cases is small. If the current state were in the first part, a split-merge move would be attempted, and the state after transition would also remain in the first part. If the current state were in the second part, slightly modified incremental steps would be attempted, with the modification ensuring that the state after transition would also remain in the second part. Alternatively, for this second part, one could make no transition at all. With the posterior distribution invariant for each potential step, the posterior distribution would remain invariant for the chain as a whole. Supplementing this type of move with incremental Gibbs scans would yield irreducibility of the chain.

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2: 1152–1174.

Berry, D. A. and Christensen, R. (1979). "Empirical Bayes Estimation of a Binomial Parameter Via Mixtures of Dirichlet Processes." *The Annals of Statistics*, 7: 558–568.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions Via Pólya Urn Schemes." *The Annals of Statistics*, 1: 353–355.

Dahl, D. (2005). “Sequentially-allocated merge-split sampler for conjugate and nonconjugate Dirichlet process mixture models.” Technical report, Department of Statistics, Texas A& M University.

Diebolt, J. and Robert, C. P. (1994). “Estimation of Finite Mixture Distributions through Bayesian Sampling.” *Journal of the Royal Statistical Society, Series B: Methodological*, 56: 363–375.

Escobar, M. D. (1994). “Estimating Normal Means with a Dirichlet Process Prior.” *Journal of the American Statistical Association*, 89: 268–277.

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.” *The Annals of Statistics*, 1: 209–230.

Gelfand, A. E. and Kuo, L. (1991). “Nonparametric Bayesian Bioassay Including Ordered Polytomous Response.” *Biometrika*, 78: 657–666.

Ishwaran, H. and Zarepour, M. (2000). “Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-parameter Process Hierarchical Models.” *Biometrika*, 87(2): 371–390.

Jain, S. and Neal, R. M. (2000). “A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.” Technical report, Department of Statistics, University of Toronto.

Kraft, C. H. and van Eeden, C. (1964). “Bayesian Bio-assay.” *The Annals of Mathematical Statistics*, 35: 886–890.

Kuo, L. and Smith, A. F. M. (1992). “Bayesian Computations in Survival Models Via the Gibbs Sampler (Disc: P22-24).” In Klein, J. P. and Goel, P. K. (eds.), *Survival Analysis: State of the Art*, 11–22. Kluwer Academic Publishers Group.

Liu, J. S. (1994). “The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem.” *Journal of the American Statistical Association*, 89: 958–966.

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

MacEachern, S. N. (1994). “Estimating Normal Means with a Conjugate Style Dirichlet Process Prior.” *Communications in Statistics: Simulation and Computation*, 23: 727–741.

MacEachern, S. N. and Müller, P. (1998). “Estimating Mixture of Dirichlet Process Models.” *Journal of Computational and Graphical Statistics*, 7: 223–238.

Neal, R. (1991). "Bayesian mixture modelling." In C.R. Smith, G. E. and Neudorfer, P. (eds.), *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, 197–211. Kluwer Academic Publishers.

Papaspiliopoulos, O. and Roberts, G. (2008). "Retrospective MCMC for Dirichlet process hierarchical models." *Biometrika(to appear)*.

Ramsey, F. L. (1972). "A Bayesian Approach to Bioassay (Com: V29 P225-226, V29 P830)." *Biometrics*, 28: 841–858.

Acknowledgments

This research was funded by the NSF under grant numbers SES-0437251 and DMS-0605041 and by the NSA under award number H98230-05-1-0065.

Bayesian Analysis (2007) **2**, Number 3, pp. 495–500

Rejoinder

Sonia Jain* and Radford M. Neal†

We thank discussants Drs. MacEachern, Robert, and Dahl for their thoughtful comments. Since many of their comments are related, we will address them by topic below.

1 Creation, Deletion, Identifiability, and Tempering

Our conditionally conjugate split-merge technique belongs to the family of trans-dimensional MCMC algorithms, which includes, for example, reversible-jump MCMC (Green 1995), birth-death MCMC (Stephens 2000a), and split-merge MCMC (Jain and Neal 2004), (Dahl 2003). Trans-dimensional MCMC algorithms construct Markov chain transitions between states that vary in dimension. For Dirichlet process mixture models, this involves the creation or deletion of mixture components.

Of course, even plain Gibbs Sampling updates for this model must be able to create and delete mixture components, but they do so only in an incremental fashion, in which a new component must start off explaining only a single observation — which may be a rather unlikely state. A key strength of trans-dimensional MCMC procedures is the ability to traverse the parameter space efficiently without having to pass through such low-probability states. For simple problems, these techniques can save computation time by reducing the required burn-in, and improving sampling thereafter. For more complex and difficult problems, such as are encountered in areas such as genetics and image analysis, these techniques may be essential if the problem is to be solved in any reasonable amount of time.

The mixture components created are given arbitrary labels, which could be permuted without affecting fit to the data, or prior probability. This "non-identifiability" has been seen by some as raising issues with regard to proper interpretation of the results, as discussed, for example by Stephens (2000b). These issues are of no relevance to our paper, which is concerned only with efficiently sampling from the posterior distribution. We agree with MacEachern that forcing the Dirichlet process mixture model to be identifiable is a hindrance to efficient MCMC sampling.

In this regard, one should note that sampling of all equivalent labellings can easily be obtained by simply introducing an additional MCMC update (applied at any desired interval) that permutes the labels — though this would be pointless for most purposes, since the labelling doesn't matter. Robert demonstrates that Gibbs sampling alone may fail to move easily between modes with different labellings. In itself, this failure is of no practical significance. Lack of movement between these equivalent modes should be

*Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, University of California at San Diego, La Jolla, CA, mailto:sojain@ucsd.edu

†Department of Statistics and Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, http://www.cs.toronto.edu/~radford/

ba0007

worrying only to the extent that one thinks it is a sign that the MCMC method would also fail to move between non-equivalent modes (if any) that correspond to substantively different interpretations of the data. It is unclear to us that failure to move amongst equivalent modes is actually indicative of a real problem of this sort. Conversely, there is no guarantee that a method that moves amongst equivalent modes can also move easily between non-equivalent modes.

Robert suggests that perhaps global tempering would perform better than a split-merge procedure, with regard to movement between isolated modes. (We are not sure which tempering method Robert used for his example, as it is not specified.) However, his example considers the benefits of tempering only when transitions are done using Gibbs sampling, without any split-merge updates. Moreover, his example concerns a fully conjugate model of the type treated in our earlier work (Jain and Neal 2004), rather than the nonconjugate models discussed in this article.

Also, the comparison looks only at mixing amongst equivalent modes, which as mentioned above is of no importance in itself. For these reasons, this demonstration does not convince us that tempering would work better than split-merge methods. However, we do expect that for some complex problems, such as very high-dimensional clustering, split-merge may not be sufficient. We hypothesize that tempering methods may also have difficulty with such problems, but that applying tempering in conjunction with split-merge updates might allow for their solution.

2 The Role of Conditional Conjugacy in Our Algorithm

As the discussants highlight, our split-merge method applies only to models in which the prior for parameters of component distributions exhibits conditional conjugacy. Though this limits the the usefulness of our algorithm, its domain is perhaps wider than one might expect. For instance, consider MacEachern (1998), in which he describes how a nonconjugate model can be treated as a conditionally conjugate problem by using piecewise log-concavity.

Robert wonders whether it might be possible to extend the algorithm beyond conditionally conjugate models. Conditional conjugacy is needed so that we can do a Gibbs sampling scan from the launch state, and also compute the probability density for choosing the value chosen at each stage of this scan. The underlying requirement is that we have a way of proposing a new parameter vector based on the launch state such that (a) the distribution of the proposed state is similar to the posterior distribution of parameter values, and (b) we can for this proposed state compute the probability density of its having been proposed. This allows us to implement efficient Metropolis-Hastings updates, with a particular update being chosen randomly by the procedure for selecting a launch state. As an aside, note that from its very origins (Metropolis et al. 1953) the Metropolis algorithm has commonly been used with proposals that change only a subset of the variables. It is not necessary to justify such partial Metropolis-Hastings updates in a special way (as Robert suggests), or to refer to such updates as anything other than Metropolis-Hastings updates.

An MCMC transition that leaves the posterior distribution invariant is a natural way of trying to get a proposal for parameters of split/merged components that comes from close to the posterior distribution. More than one such transition would be better, but would make computing the density for a proposal impossible, as that would require integrating over intermediate states. (Such an integral is avoided in our algorithm by treating the intermediate Gibbs sampling updates not as part of the proposal distribution but rather as a procedure for choosing a launch state.) One could certainly imagine using MCMC transitions other than Gibbs sampling for this purpose. One could, for example, use a series of Metropolis-Hastings updates applied to each parameter in turn. The probability density for proposing a state that differs in all components from the launch state would then be easily computed, as the product of all the proposal densities and all the acceptance probabilities. Unfortunately, the probability density for a state in which any of these Metropolis-Hastings updates was rejected (so that at least component is the same in the launch state and in the proposal) will be infinite, which will result in a zero acceptance probability for the split-merge update.

So, although one can imagine such variations, they may not be useful in practice. One possibility that would be worth investigating is using some approximation to the posterior distribution (for the model restricted to two components), such as a Gaussian. The conditional distributions from this approximation could be used as proposals (resulting in Gibbs sampling in the limit as the approximation becomes perfect). If the rejection rate is small enough, this might work well. Alternatively, the approximation could be used directly — the validity of our algorithm does not depend on the transition from the launch state (or the intermediate transitions) leaving the actual posterior distribution invariant, though use of a bad approximation will of course lead to a low acceptance rate for the split-merge updates.

3 The Usefulness of Incremental Markov Chain Updates Together with Split-Merge

Our split-merge algorithm has four tuning parameters, controlling the number of intermediate restricted Gibbs sampling scans for splits proposals and merge proposals, and the number of split-merge updates and incremental Markov chain updates (e.g. Gibbs sampling scans) done as part of a full iteration. Both MacEachern and Dahl remark on the importance of a final incremental Gibbs sampling scan. We agree with MacEachern that the inclusion of such a step is important to facilitate mixing, as we have demonstrated in the article. However, though MacEachern emphasizes the role of such updates in mixing for small clusters, we believe that they are at least as important for moving observations back and forth between large clusters, as this cannot be done efficiently with split-merge updates.

Indeed, as we have described in our article and as Dahl observes, the “jitter” that is observed in the trace plots of the beetle example can be attributed to the final auxiliary Gibbs sampling scan. However, we disagree with Dahl’s conclusion from this that the CPU time spent on split-merge updates is “wasted” when these moves are not accepted. How can we know *a priori* that these moves will not be accepted unless they are proposed? Since split-merge updates are required to obtain a correct solution (in a reasonable amount of time) for some problems, it is necessary to perform them for all problems in order to determine if they are actually needed, and hence ensure that the answer obtained is correct.

Further, Dahl's demonstration with only auxiliary Gibbs sampling (no split-merge updates) is not entirely convincing. On close inspection, the lower plots of his Figure 1 show that auxiliary Gibbs sampling is not actually performing that well! For several thousand iterations, a number of observations seem to have been incorrectly allocated to small clusters, with the Gibbs sampler making only slow progress in correcting this. It is possible that just a few split-merge iterations could take care of these orphan clusters. By performing both incremental and non-incremental split-merge updates, one can take advantage of both large-scale changes to the cluster configuration via split-merge moves and small-scale adjustments that move a few observations between clusters, as is necessary depending on the problem.

4 MCMC Initialization

Robert suggests that sampling from the prior to initialize the intermediate restricted Gibbs sampling could lead to wasted computational effort. In higher-dimensional problems, we agree that overcoming bad initial values could be a problem — i.e. many restricted Gibbs sampling scan might be required. In the Discussion section of the paper, we had suggested alternatives to sampling from the prior to initialize the restricted Gibbs sampling, such as adapting a method used by Dahl (2003), or some other posterior estimation method.

A feature of the split-merge technique that Dahl discusses is its insensitivity to the initial value that the Markov chain is started with, whereas the Gibbs sampler is susceptible to poor choices (as illustrated in the Beetle example). We agree with this, but are puzzled by the discrepancies in the simulations. We also initialized the chain by sampling the model parameters from the prior and not by setting the initial values to the sample mean and precision. One possible explanation is differing orders of updates — we sampled the indicators first and then the model parameters (means before precisions).

5 Random versus Fixed Scan Sampling

MacEachern investigates in detail how MCMC performance differs for fixed versus random scans, in the context of Gibbs sampling. He proposes a systematic scan as an alternative to the random scan that we utilize to initiate the split-merge process (i.e. select two observations, denoted as i and j, uniformly at random). MacEachern suggests permuting the indices from 1 to n and then using successive pairs as i and j, thereby reducing randomness. This gives a feasible scan length (unlike systematically using all possible pairs of observations). We agree that this is likely to improve performance, but perhaps not by much. Partly, this is because there will still be considerable randomness in which *clusters* are chosen for split/merge operations — in particular, the same clusters might well be chosen several times in a row.

More generally, however, MacEachern may be overestimating the difference between fixed and random scans. The interpretation of his Table 2 is perhaps not obvious. The number of iterations required to reach some small total variation distance is proportional to $-1/\log(v)$, where v is the second-largest eigenvalue. So, for example, using scheme 3, with $\alpha = 1$ (which is $M = 1$ in MacEachern's notation), the fixed scan method is not better by a factor of $0.171/0.037 = 4.6$, as one might naively think, but rather by a factor of $\log(0.037)/\log(0.171) = 1.9$. As α approaches infinity (approximated by $\alpha = 100$, i.e. corresponding to $M = 100$ in the table), the second largest eigenvalue for the fixed scan approaches zero — all the variables are independent, so a single fixed Gibbs sampling scan immediately reaches equilibrium. The random scan has a non-zero second eigenvalue in the $\alpha \to \infty$ limit, reflecting the fact that after any number of iterations there is a non-zero probability that some variable could still be left unchanged. Technically, the asymptotic convergence rate of the fixed scan is infinitely better than that of the random scan, but in practice a modest number of iterations is sufficient to give the correct result with very high probability.

In this small example, Markov chain sampling is based on the prior distribution of clusterings for three data points, but the likelihood factors deriving from the data are omitted. However, in practical problems, where many split-merge proposals are likely to be made for each that is accepted, the randomness in choice of clusters to split/merge may be negligible compared to the randomness in proposing how to split or merge them, and in whether or not to accept the result. Finding ways of further improving the split/merge proposals may be a better focus for future research.

References

Dahl, D. B. (2003). "An improved merge-split sampler for conjugate Dirichlet process mixture models." Technical Report 1086, Department of Statistics, University of Wisconsin. 495, 498

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732. 495

Jain, S. and Neal, R. 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. 495, 496

MacEachern, S. N. (1998). "Computational methods for mixture of Dirichlet process models." In Dey, D., Müller, P., and Sinha, D. (eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, 23–43. New York: Springer-Verlag. 496

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). "Equation of state calculations by fast computing machines." *Journal of Chemical Physics*, 21: 1087–1092. 496

Stephens, M. (2000a). “Bayesian analysis of mixtures with an unknown number of components – an alternative to reversible jump methods.” *Annals of Statistics*, 28: 40–74. 495

— (2000b). “Dealing with label-switching in mixture models.” *Journal of the Royal Statistical Society, Series B*, 62: 795–809. 495

Bayesian Analysis (2007) **2**, Number 3, pp. 501–528

Hidden Markov Dirichlet Process: Modeling Genetic Inference in Open Ancestral Space

Eric P. Xing[*] and Kyung-Ah Sohn[†]

Abstract. The problem of inferring the population structure, linkage disequilibrium pattern, and chromosomal recombination hotspots from genetic polymorphism data is essential for understanding the origin and characteristics of genome variations, with important applications to the genetic analysis of disease propensities and other complex traits. Statistical genetic methodologies developed so far mostly address these problems separately using specialized models ranging from coalescence and admixture models for population structures, to hidden Markov models and renewal processes for recombination; but most of these approaches ignore the inherent uncertainty in the genetic complexity (e.g., the number of genetic founders of a population) of the data and the close statistical and biological relationships among objects studied in these problems. We present a new statistical framework called hidden Markov Dirichlet process (HMDP) to jointly model the genetic recombinations among a possibly infinite number of founders and the coalescence-with-mutation events in the resulting genealogies. The HMDP posits that a haplotype of genetic markers is generated by a sequence of recombination events that select an ancestor for each locus from an unbounded set of founders according to a 1st-order Markov transition process. Conjoining this process with a mutation model, our method accommodates both between-lineage recombination and within-lineage sequence variations, and leads to a compact and natural interpretation of the population structure and inheritance process underlying haplotype data. We have developed an efficient sampling algorithm for HMDP based on a two-level nested Pólya urn scheme, and we present experimental results on joint inference of population structure, linkage disequilibrium, and recombination hotspots based on HMDP. On both simulated and real SNP haplotype data, our method performs competitively or significantly better than extant methods in uncovering the recombination hotspots along chromosomal loci; and in addition it also infers the ancestral genetic patterns and offers a highly accurate map of ancestral compositions of modern populations.

Keywords: Dirichlet Process, Hierarchical DP, hidden Markov model, MCMC, statistical genetics, recombination, population structure, SNP.

1 Introduction

The availability of nearly complete genome sequences for organisms such as humans makes it possible to begin to explore individual differences between DNA sequences, known as *genetic polymorphisms*, on a genome-wide scale, and to search for associations

[*] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, http://www.cs.cmu.edu/~epxing/

[†] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, http://www.cs.cmu.edu/~ksohn/

of such genotypic variations with diseases and other phenotypes. Most human variation that is influenced by genes can be related to a particular kind of genetic polymorphism known as the *single nucleotide polymorphisms*, or SNPs. A SNP refers to the existence of two possible kinds of nucleotides from $\{A, C, G, T\}$ at a single chromosomal *locus* (i.e., a position on the chromosome) in a population; each variant is called an *allele* [*]. A *haplotype* is a list of alleles at contiguous sites in a local region of a single chromosome. Assuming no recombination in this local region, a haplotype is inherited as a unit. But under many realistic biological or genetic scenarios, repeated recombinations between ancestral haplotypes during generations of inheritance may confound the genetic origin of modern haplotypes (Figure 1).

Recombinations between ancestral chromosomes during meiosis play a key role in shaping the patterns of linkage disequilibrium (LD)—the non-random association of alleles at different *loci*—in a population. When a recombination occurs between two loci, it tends to decouple the alleles carried at those loci in its decedents and thus reduce LD; uneven occurrence of recombination events along chromosomal regions during genetic history can lead to "block structures" in molecular genetic polymorphisms such that within each block only low level of diversities are present in a population.

Statistically, for a pair of loci with genetic polymorphic markers, say, X and Y, the LD between these two loci can be characterized by a number of so-called *LD measures*. For example, for bi-allelic markers (i.e., markers that have only two possible states, say "0" and "1"), LD can be measured by the *gametic disequilibrium*, $D = p_{00}p_{11} - p_{01}p_{10}$, where $p_{00} := \text{Prob}(X = 0, Y = 0)$, $p_{11} := \text{Prob}(X = 1, Y = 1)$, $p_{01} := \text{Prob}(X = 0, Y = 1)$, and $p_{10} := \text{Prob}(X = 1, Y = 0)$, are the empirical frequencies of joint allele-state configurations. Another popular LD measure is the p-value for Fisher's exact test over samples of X and Y. When $D = 0$, which means that the two loci of interest are not arranged randomly during inheritance (due to recombination of their host chromosomes at a position between the two loci), they often emerge (e.g., from all possible pairs in a large number of loci being surveyed) as candidates of marker pairs on the chromosome whose locations are physically close so that there is a low probability of having recombination events between them. However, to the best of our knowledge, extant LD-measures remain primarily focused on offering population-level descriptive statistics of the sample, rather than on modeling and inferring the underlying genetic mechanisms and processes that may have generated the data. For example, the pairwise LD measure ignores the global context and overall pattern of the genetic polymorphisms, and thus can not distinguish linkages due to spurious statistical association (e.g., due to problems in sample procedures) from those resulting from true physical proximity, or from genetic coupling due to co-evolution [†]. Such an approach also

[*] In general, an *allele* represents a variant of a SNP, a gene, or some other entity associated with a locus on DNA. In our case (SNPs), the locus harbors a single nucleotide, and therefore the alleles can generally be assumed to be binary, reflecting the fact that "lightning doesn't tend to strike twice in the same place". That is, nucleotide substitutions (i.e., mutations) do not occur to the same locus twice during the inheritance course from a common ancestor. More generally, e.g., in case of *microsatellite* polymorphism, the allele-state can be k-nary, a scenario to which our proposed model also applies.

[†] Co-evolution can occur for DNA sequences that are far apart in the genome if they encode genes or regulatory elements that jointly or corporately perform an indispensable biology function. For example,

provides no information regarding the demographical history and ancestral composites of each individual in the study population. In this paper, we propose a new model-based approach to address these issues.

The problem of inferring chromosomal recombination hotspots is essential for understanding the origin and characteristics of genome variations; several combinatorial and statistical approaches have been developed for uncovering optimum block boundaries from single nucleotide polymorphism haplotypes
(Daly et al. 2001; Anderson and Novembre 2003; Patil et al. 2001; Zhang et al. 2002). For example, Zhang et al. (2002) proposed a dynamic programming algorithm for partitioning single nucleotide polymorphism (SNP) haplotypes (explained in the sequel) into low-diversity blocks; Daly et al. (2001) and Greenspan and Geiger (2004a) have developed hidden Markov models for locating recombination hotspots in haplotypes; and Anderson and Novembre (2003) proposed a minimum description length (MDL) method for optimal haplotype block finding. Some recent studies resorted to more sophisticated population genetics arguments that more explicitly capture the mechanistic and population genetic foundations underlying recombination and LD pattern formation. For example, Li and Stephens (2003) used a tractable approximation to the recombinational coalescence, via a (latent) genealogy of the population, to capture the conditional dependencies between haplotypes. Rannala and Reeve (2001) also use a coalescence-based model and an MCMC method to integrate over the unknown gene genealogy and coalescence times. These advances have important applications in genetic analysis of disease propensities and other complex traits.

The deluge of SNP data also fuels the long-standing interest of analyzing patterns of genetic variations to reconstruct the evolutionary history and ancestral structures of human populations, using, for example, variants of admixture models on genetic polymorphisms (Pritchard et al. 2000; Rosenberg et al. 2002; Falush et al. 2003). These models are instances of a more general class of hierarchical Bayesian models known as *mixed membership models* (Erosheva et al. 2004), which postulate that genetic markers of each individual are iid (Pritchard et al. 2000) or spatially coupled (Falush et al. 2003) samples from multiple population-specific fixed-dimensional multinomial-distributions of marker alleles. However, the admixture models developed so far do model genetic drift due to mutations from the ancestor allele and therefore do not enable inference of the founding genetic patterns and the age of the founding alleles (Excoffier and Hamilton 2003).

This progress notwithstanding, the statistical methodologies developed so far mostly deal with LD analysis and ancestral inference separately, using specialized models that do not capture the close statistical and genetic relationships of these two problems. Moreover, most of these approaches ignore the inherent uncertainty in the genetic complexity (e.g., the number of genetic founders of a population) of the data and rely on inflexible models built on a pre-fixed, closed genetic space. Recently, we have developed a nonparametric Bayesian framework for modeling genetic polymorphisms based on the Dirichlet process (DP) mixtures and extensions, which attempts to allow more

proteins that form a complex to carry out enzymatic activities usually co-evolve.

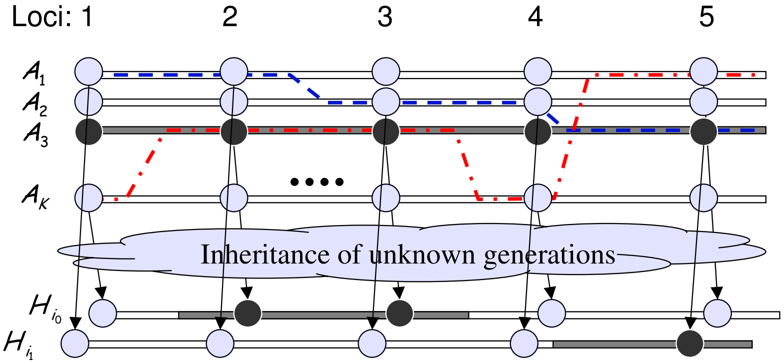


Figure 1: An illustration of a hidden Markov Dirichlet process for haplotype recombination and inheritance. Note that the total number of ancestors is unknown.

flexible control over the number of genetic founders than has been provided by the statistical methods proposed thus far (Xing et al. 2004) . In this paper, we leverage on this approach and present a unified framework to model complex genetic inheritance processes that allows recombinations among possibly infinite founding alleles and coalescence-with-mutation events in the resulting genealogies.

We assume that individual chromosomes in a modern population are originated from an unknown number of ancestral haplotypes via biased random recombinations and mutations (Figure 1). The recombinations between the ancestors follow a state-transition process we refer to as hidden Markov Dirichlet process (originated from the infinite HMM by Beal et al. (2002)), which travels in an open ancestor space, with nonstationary recombination rates depending on the genetic distances between SNP loci. Our model draws inspiration from the HMM proposed in Greenspan and Geiger (2004b), but we employ a two-level Pólya urn scheme akin to the hierarchical DP (Teh et al. 2006) to accommodate an open ancestor space, and allow full posterior inference of the recombination sites, mutation rates, haplotype origin, ancestor patterns, etc., conditioning on phased SNP data, rather than estimating them using information theoretic or maximum likelihood principles. On both simulated and real genetic data, our model and algorithm show competitive or superior performance on a number of genetic inference tasks over the state-of-the-art parametric methods.

The remainder of this paper is presented as follows. In section 2, we formulate the problem, and present details of the proposed model. In section 3, we describe a block Gibbs sampling algorithm for posterior inference of the latent variables. In section 4, we present experimental results on a simulated data haplotype data set, and on two published real data sets, one from a single population, and the other from two populations. We conclude with a brief discussion in section 6. A short version of this manuscript was presented earlier in Sohn and Xing (2006), but the current version offers more details on the biological background, the model specifications, and the experimental results.

2 The Statistical Model

Sequentially choosing recombination targets from a set of ancestral chromosomes can be modeled as a hidden Markov process (Niu et al. 2002; Greenspan and Geiger 2004b), in which the hidden states correspond to the index of the candidate chromosomes, the transition probabilities correspond to the recombination rates between the recombining chromosome pairs, and the emission model corresponds to a mutation process that passes the chosen chromosome region in the ancestors to the descents. When the number of ancestral chromosomes is not known, it is natural to consider an HMM whose state space is countably infinite (Beal et al. 2002; Teh et al. 2006). In this section, we describe such an infinite HMM formalism, which we would like to call *hidden Markov Dirichlet process*, for modeling recombination in an open ancestral space.

2.1 Dirichlet Process mixtures

For self-containedness, we begin with a quick overview of the fundamentals of the Dirichlet process and its connection to the coalescent process in population genetics, followed by a brief recapitulation of the basic Dirichlet process mixture model we proposed in Xing et al. (2004) for haplytope inheritance without recombination.

As mentioned earlier, a *haplotype* refers to the joint allele configuration of a contiguous list of SNPs located on a chromosome. Under a well-known genetic model known as *coalescence with infinite-many-alleles (IMA) mutations* (but without recombination), one can treat a haplotype from a modern individual as a descendent of a most recent common ancestor (MRCA) of unknown haplotype via random mutations that alter the allelic states of some SNPs (Kingman 1982). Hoppe (1984) observed that a coalescent process in an infinite population leads to a partition of the population at every generation that can be succinctly captured by the following Pólya urn scheme.

Consider an urn that at the outset contains a ball of a single color. At each step we either draw a ball from the urn and replace it with two balls of the same color, or we are given a ball of a new color which we place in the urn. One can see that such a scheme leads to a partition of the balls according to their color. Mapping each ball to a haploid individual [‡] and each color to a possible haplotype, this partition is equivalent to the one resulting from the *coalescence-with-mutation* process (Hoppe 1984), and the probability distribution of the resulting *allele spectrum*—the numbers of colors (resp. haplotypes) with every possible number of representative balls (resp. decedents)—is captured by the well-known Ewens' sampling formula (Tavare and Ewens 1998).

Letting parameter α define the probabilities of the two types of draws in the aforementioned Pólya urn scheme, and viewing each (distinct) color as a sample from Q_0, and each ball as a sample from Q [§], Blackwell and MacQueen (1973) showed that this

[‡] A haploid individual refers to an individual with only one haplotype — a simplifying assumption often used on population genetics when the paternal and maternal haplotypes of a diploid individual are inherited independently.

[§] Here we deviate from the conventional notations in the statistics literature (e.g., Neal (2000); Escobar and West (2002); Ishwaran and James (2001)) and use Q and Q_0, instead of G and G_0 (or

Pólya urn model yields samples whose distributions are those of Q_0 the marginal probabilities under the *Dirichlet process* (Ferguson 1973). Formally, a random probability measure Q is generated by a DP if for any measurable partition $B_1, \ldots, B_k$ of the sample space, the vector of random probabilities $Q(B_i)$ follows a Dirichlet distribution: $(Q(B_1), \ldots, Q(B_k)) \sim \mathrm{Dir}(\alpha Q_0(B_1), \ldots, \alpha Q_0(B_k))$, where α denotes a *scaling parameter* and Q_0 denotes a *base measure*. The Pólya urn model makes explicit that the association of data points to colors defines a "clustering" of the data. Specifically, having observed n values $(\phi_1, \ldots, \phi_n)$ sampled from a Dirichlet process $DP(\alpha, Q_0)$, the probability of the $(n+1)$th value is given by:

$$\phi_{n+1}|\phi_1, \ldots, \phi_n, \alpha, Q \;\sim\; \sum_{i=1}^{n} \frac{1}{n+\alpha}\delta_{\phi_i}(\cdot) + \frac{\alpha}{n+\alpha}Q_0(\cdot), \tag{1}$$

where $\delta_{\phi_i}(\cdot)$ denotes a point mass at value ϕ_i. Another very useful representation of DP is the stick-breaking construction by Sethuraman (1994). This construction is based on independent sequences of independent random samples $\{\pi'_{k,i}\}_{i=1}^{\infty}$ and $\{\phi_i\}_{i=1}^{\infty}$ generated in the following way: $\pi'_i|\alpha, Q_0 \sim \mathrm{Beta}(1, \alpha)$ and $\phi_i|\alpha, Q_0 \sim Q_0$, where $\mathrm{Beta}(a, b)$ is the Beta distribution with parameter a and b. Let $\pi_i = \pi'_i \prod_{l=1}^{k-1}(1 - \pi'_l)$ (analogous to a process of repetitively breaking a stick at fraction π'_l), Sethuraman showed that the random measure arising from $DP(\alpha, Q_0)$ admits the representation $Q = \sum_{i=1}^{\infty} \pi_i \delta_{\phi_i}$. The ϕ_i's can be understood as the *locations* of samples in their space, and the π_i's are the *weights* of these samples.

The discrete nature of the DP, as obviated from the stick-breaking construction, is well suited for the problem of placing priors on mixture components in mixture modeling. In the context of mixture models, one can associate mixture component centroids (e.g., haplotype founders, as explained in the sequel) with colors in the Pólya urn model and thereby define a "clustering" of the (possibly noisy) data (e.g., modern haplotypes that are "recognizable" variants of their corresponding founders). This mixture model is known as a DP mixture (Antoniak 1973; Escobar and West 2002) (also known as "infinite" mixture model in machine learning community). Note that a DP mixture requires no prior specification of the number of components, which is typically unknown in genetic demography and general data clustering problems. It is important to emphasize that here DP is used as a *prior distribution* of mixture components. Multiplying this prior by a likelihood that relates the mixture components to the actual data yields a *posterior distribution* of the mixture components, and the design of the likelihood function is completely up to the modeler based on specific problems. MCMC algorithms have been developed to sample from the posterior associated with DP priors (Escobar and West 2002; Neal 2000; Ishwaran and James 2001). This nonparametric Bayesian formalism forms the technical foundation of the haplotype modeling and inference algorithms to be developed in this paper.

Back to haplotype modeling, a straightforward statistical genetics argument shows that the distribution of haplotypes can be formulated as a mixture model, where the set

H), to denote the random probability measure under DP and the base measure of DP, because in the genetic context, G and H have been used to denote the genotype and haplotype of polymorphic markers (Pritchard et al. 2000; Stephens et al. 2001; Li and Stephens 2003; Xing et al. 2004).

of mixture components corresponds to the pool of ancestor haplotypes, or *founders*, of the population (Excoffier and Slatkin 1995; Niu et al. 2002; Kimmel and Shamir 2004). Crucially, however, the size of this pool is unknown; indeed, knowing the size of the pool would correspond to knowing something significant about the genome and its history. On the other hand, despite its elegance, with a purely coalescence-based model for genetic patterns, it is hard to perform statistical inference of ancestral features and many other interesting genetic variables (for a large population, the number of hidden variables in a coalescence tree is prohibitively large) (Stephens et al. 2001). In most practical population genetic problems, usually the detailed genealogical structure of a population (as provided by the coalescent trees) is of less importance than the population-level features such as the pattern of major common ancestor alleles (i.e., founders) in a population bottleneck ¶, the age of such alleles, etc. In this case, the DP mixture offers a principled approach to generalize the finite mixture model for haplotypes to an infinite mixture model that models uncertainty regarding the size of the ancestor haplotype pool; at the same time, it provides a reasonable approximation to the coalescence model by utilizing the partition structure resulting therefrom (but allows further mutations within each partite to introduce further diversity among descents of the same founder, which correspond to the balls with the same color in the Pólya urn metaphor). Without further digression, below we summarize the Dirichlet process mixture model we proposed in Xing et al. (2004) for haplytope inheritance without recombination.

Write $H_i = [H_{i,1}, \ldots, H_{i,T}]$ for a haplotype over T SNPs from chromosome i ‖; let $A_k = [A_{k,1}, \ldots, A_{k,T}]$ denote an ancestor haplotype (indexed by k) and θ_k denote the *mutation rate* of ancestor k; and let C_i denote an *inheritance variable* that specifies the ancestor of haplotype H_i. Under a DP mixture, we have the following Pólya urn scheme for sampling modern haplotypes:

- Draw first haplotype:

  $a_1 \mid \mathrm{DP}(\tau, Q_0) \sim Q_0(\cdot)$, sample the 1st founder;

  $h_1 \sim P_h(\cdot|a_1, \theta_1)$, sample the 1st haplotype from an inheritance model defined on the 1st founder;

- for subsequent haplotypes:

  – sample the founder indicator for the ith haplotype:

  $$c_i|\mathrm{DP}(\tau, Q_0) \sim \begin{cases} p(c_i = c_j \text{ for some } j < i|c_1, \ldots, c_{i-1}) = \frac{n_{c_j}}{i-1+\alpha_0} \\ p(c_i \neq c_j \text{ for all } j < i|c_1, \ldots, c_{i-1}) = \frac{\alpha_0}{i-1+\alpha_0} \end{cases}$$

  where n_{c_i} is the *occupancy number* of class c_i—the number of previous samples belonging to class c_i.

¶ A stage in coalescence when there are only a very small number of founding haplotype patterns surviving and giving rise to all the haplotypes in the modern population.

‖ We ignore the parental origin index of haplotypes as used in Xing et al. (2004), and assume that the paternal and maternal haplotypes of each individual are given unambiguously (i.e., *phased*, as known in genetics), as is the case in many LD and haplotype-block analyses (Daly et al. 2001; Anderson and Novembre 2003). But it is noteworthy that our model can generalize straightforwardly to unphased genotype data by incorporating a simple genotype model as in Xing et al. (2004).

– sample the founder of haplotype i (indexed by c_i):

$$\phi_{c_i}|\mathrm{DP}(\tau,Q_0)\begin{cases} =\{a_{c_j},\theta_{c_j}\} & \text{if } \ c_i=\{a_{c_j},\theta_{c_j}\} \ \text{ for some } j<i \text{ (i.e., } c_i \text{ refers to an inherited founder)} \\ \sim Q_0(a,\theta) & \text{if } \ c_i\neq c_j \ \text{ for all } \ j<i \text{ (i.e., } c_i \text{ refers to a new founder)} \end{cases}$$

– sample the haplotype according to its founder:

$h_i \mid c_i \sim P_h(\cdot|a_{c_i},\theta_{c_i})$.

The usefulness of the DP mixture framework for the haplotype problem should be clear—using a Dirichlet process prior we in essence maintain a pool of haplotype founders that grows as observed individual haplotypes are processed. But notice that the above generative process assumes each modern haplotype originates from a single ancestor, which is only true for haplotypes spanning a short region on a chromosomal. Now we consider long haplotypes possibly bearing multiple ancestors due to recombinations between an unknown number of founders.

2.2 Hidden Markov Dirichlet Process (HMDP)

In a standard HMM, state-transitions across a discrete time- or space-interval take place in a fixed-dimensional state space, thus it can be fully parameterized by, say, a K-dimensional initial-state probability vector π_0 and a $K \times K$ state-transition probability matrix $\Pi_{K\times K}$. As first proposed in Beal et al. (2002), and later discussed in Teh et al. (2006), one can "open" the state space of an HMM by treating the now infinite number of discrete states of the HMM as the support of a DP, and the transition probabilities to these states from some source as the masses associated with these states. In particular, for each source state (say, state j), the possible transitions to the target states need to be modeled by a unique DP Q_j. Since all possible source states and target states are taken from the same infinite state space, overall we need an open set of DPs with different mass distributions on the SAME support (to capture the fact that different source states can have different transition probabilities to any target state). In the sequel, we describe such a nonparametric Bayesian HMM using an intuitive hierarchical Pólya urn construction. We call this model a **hidden Markov Dirichlet process**.

In an HMDP, both the columns and rows of the transition matrix Π are infinite dimensional. To construct such an stochastic matrix, we will exploit the fact that in practice only a finite number of states (although we don't know what they are) will be visited by each source state, and we only need to keep track of these states. The following sampling scheme based on a hierarchical Pólya urn scheme captures this spirit and yields a constructive definition of HMDP.

We set up a single "stock" urn at the top level, which contains balls of colors that are represented by at least one ball in one or multiple urns at the bottom level. At the bottom level, we have a set of *distinct* urns which are used to define the initial and transition probabilities of the HMDP model (and are therefore referred as HMM-urns).

Specifically, one of the HMM urns, Q_0, is set aside to hold colored balls to be drawn at the onset of the HMM state-transition sequence [**]. Each of the remaining HMM urns is painted with a color represented by at least one ball in the stock urn, and is used to hold balls to be drawn during the execution of a Markov chain of state-transitions. Now let's suppose that at time t the stock urn contains n balls of K distinct colors indexed by an integer set $\mathcal{C} = \{1, 2, \ldots, K\}$; the number of balls of color k in this urn is denoted by $n_k, k \in \mathcal{C}$. For urn Q_0 and urns $Q_1, \ldots, Q_K$, let $m_{j,k}$ denote the number of balls of color k in urn Q_j, and $m_j = \sum_{k \in \mathcal{C}} m_{j,k}$ denote the total number of balls in urn Q_j. Suppose that at time $t-1$, we had drawn a ball with color k'. Then at time t, we either draw a ball randomly from urn $Q_{k'}$, and place back two balls both of that color; or with probability $\frac{\tau}{m_j+\tau}$ we turn to the top level. From the stock urn, we can either draw a ball randomly and put back two balls of that color to the stock urn and one to $Q_{k'}$, or obtain a ball of a new color $K+1$ with probability $\frac{\gamma}{n+\gamma}$ and put back a ball of this color to both the stock urn and urn $Q_{k'}$ of the lower level. Essentially, we have a master DP Q_0 (the stock urn) that serves as a source of atoms for infinite number of child DPs $\{Q_j\}$ (the HMM-urns). As pointed out in Teh et al. (2006), this model can be viewed as an instance of the hierarchical Dirichlet process mixture model, with an infinite number of DP mixtures as components. Specifically, we have:

$$\begin{aligned} Q_0|\alpha, F &\sim \mathrm{DP}(\alpha, F), \quad \text{The master DP over target states common for all sources;} \\ Q_j|\tau, Q_0 &\sim \mathrm{DP}(\tau, Q_0), \quad \text{The HMM DP over target states of source } j. \end{aligned}$$

From the above equation we see that the base measure of the DP mixture associated each of the source states in the HMM is itself drawn from a Dirichlet process DP(α, F). Since a draw from a DP is a discrete measure with probability 1, atoms drawn from this measure—atoms which are used as targets for each of the (unbounded number of) source states—are not generally distinct. Indeed, the transition probabilities from each of the source states have the same support—the atoms in Q_0.

The Pólya urn scheme described above is similar in spirit to the "Chinese restaurant franchise" scheme discussed in Teh et al. (2006), but it differs in that it avoids having separate occupancy counters in each lower-level DP for repeated draws of the same atom from a top-level DP, and it also motivates a simpler sampling scheme for inference as discussed in Section 3.

Associating each color k with an ancestor configuration $\phi_k = \{a_k, \theta_k\}$ whose values are drawn from the base measure F, and recalling our discussion in the previous section, we know that draws from the stock urn can be viewed as marginals from a random measure distributed as a Dirichlet Process Q_0 with parameter (α, F). Specifically, for n random draws $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_n\}$ from Q_0, the conditional prior for $(\phi_n|\boldsymbol{\phi}_{-n})$, where the subscript "$-n$" denotes the index set of all but the n-th ball, is

$$\phi_n|\boldsymbol{\phi}_{-n} \sim \sum_{k=1}^{K} \frac{n_k}{n-1+\alpha} \delta_{\phi_k^*}(\phi_n) + \frac{\alpha}{n-1+\alpha} F(\phi_n), \tag{2}$$

[**] Purposely, we overload the symbol Q_j to let it denote both the urns in the hierarchical Pólya urn scheme, and the Dirichlet processes distributions represented by each of these urns.

where $\phi^*_k, k = 1, \ldots, K$ denote the K distinct values (i.e., colors) of $\boldsymbol{\phi}$ (i.e., all the balls in the stock urn), n_k denote the number of balls of color k in the top urn, and $\delta_a(\phi_i)$ denotes a unit point mass at $\phi_i = a$.

Conditioning on the Dirichlet process underlying the stock urn, the samples in the jth bottom-level urn are also distributed as marginals under a Dirichlet measure:

$$\begin{aligned}\phi_{m_j}|\boldsymbol{\phi}_{-m_j} &\sim \sum_{k=1}^{K} \frac{m_{j,k} + \tau \frac{n_k}{n-1+\alpha}}{m_j - 1 + \tau} \delta_{\phi^*_k}(\phi_{m_j}) + \frac{\tau}{m_j - 1 + \tau} \frac{\alpha}{n - 1 + \alpha} F(\phi_{m_j}) \\ &= \sum_{k=1}^{K} \pi_{j,k} \delta_{\phi^*_k}(\phi_{m_j}) + \pi_{j,K+1} Q_0(\phi_{m_j}), \qquad (3)\end{aligned}$$

where $\pi_{j,k} \equiv \frac{m_{j,k}+\tau\frac{n_k}{n-1+\alpha}}{m_j-1+\tau}$, $\pi_{j,K+1} \equiv \frac{\tau}{m_j-1+\tau}\frac{\alpha}{n-1+\alpha}$. Let $\boldsymbol{\pi}_j \equiv [\pi_{j,1}, \pi_{j,2}, \ldots]$. Now we have an infinite-dimensional Bayesian HMM that, given F, α, τ, and all initial states and transitions sampled so far, follows an initial states distribution parameterized by $\boldsymbol{\pi}_0$, and transition matrix Π whose rows are defined by $\{\boldsymbol{\pi}_j : j > 0\}$.

Finally, as in, e.g., Escobar and West (2002) and Rasmussen (2000), we can also introduce vague priors such as a Gamma or an inverse Gamma for the scaling parameters α and τ.

2.3 HMDP Model for Recombination and Inheritance

Now we describe a stochastic model, based on an HMDP, for generating individual haplotypes in a modern population from a hypothetical pool of ancestral haplotypes via recombination and mutations (i.e., random mating with neutral selection). See Figure 1 for an illustration.

First recall that a base measure F at the top of our hierarchical Pólya urn scheme is defined as a distribution from which ancestor haplotype templates ϕ_k are drawn. We define the base measure F as a joint measure on both ancestor A and mutation rate θ, and let $F(A, \theta) = p(A)p(\theta)$, where $p(A)$ is uniform over all possible haplotypes and $p(\theta)$ is a beta distribution, $Beta(\alpha_h, \beta_h)$, with a small value for $\beta_h/(\alpha_h + \beta_h)$ corresponding to a prior expectation of a low mutation rate. For simplicity, we assume each $A_{k,t}$ (and also each $H_{i,t}$) takes its value from an allele set B.

Now for each modern chromosome i, let $C_i = [C_{i,1}, \ldots, C_{i,T}]$ denote the sequence of inheritance variables specifying the index of the ancestral chromosome at each SNP locus. When no recombination takes place during the inheritance process that produces haplotype H_i (say, from ancestor k), then $C_{i,t} = k, \forall t$. When a recombination occurs, say, between loci t and $t+1$, we have $C_{i,t} \neq C_{i,t+1}$. We can introduce a Poisson point process to control the duration of non-recombinant inheritance. That is, given that $C_{i,t} = k$, then with probability $e^{-dr} + (1 - e^{-dr})\pi_{kk}$, where d is the physical distance between two loci, r reflects the rate of recombination per unit distance, and π_{kk} is the self-transition probability of ancestor k defined by HMDP, we have $C_{i,t+1} = C_{i,t}$; otherwise, the source state (i.e., ancestor chromosome k) pairs with a target state (e.g., ancestor chromosome k') between loci t and $t+1$, with probability $(1-e^{-dr})\pi_{kk'}$. Hence, each haplotype H_i is a mosaic of segments of multiple ancestral chromosomes from the ancestral pool $\{A_k\}_{k=1}^{\infty}$. Essentially, the model we described so far is a time-inhomogeneous infinite HMM. When the physical distance information between loci is not available, we can simply set r to be infinity (hence $e^{-dr} \approx 0$) so that we are back to a standard stationary HMDP model with infinite dimensional transition probability matrix $\Pi_{\infty\times\infty}$ described earlier.

The emission process of the HMDP corresponds to an inheritance model from an ancestor to the matching descendent. For simplicity, we adopt the *single-locus mutation model* in Xing et al. (2004):

$$p(h_t|a_t,\theta) = \theta^{\mathbb{I}(h_t=a_t)}\Big(\frac{1-\theta}{|B|-1}\Big)^{\mathbb{I}(h_t\neq a_t)}, \tag{4}$$

where h_t and a_t denote the alleles at locus t of an individual haplotype and its corresponding ancestor, respectively; θ indicates the ancestor-specific mutation rate; and $|B|$ denotes the number of possible alleles. As discussed in Liu et al. (2001), this model corresponds to a star genealogy resulting from infrequent mutations over a shared ancestor, and is widely used in statistical genetics as an approximation to a full coalescent genealogy starting from the shared ancestor.

Assume that the mutation rate θ admits a Beta prior with hyperparameter (α_h,β_h) [††], the marginal conditional likelihood of all the haplotype instances $\boldsymbol{h} = \{h_{i,t} \ : \ i \in \{1,2,\ldots,I\},\ t\in\{1,2,\ldots,T\}\}$ given the set of ancestors $\boldsymbol{a} = \{a_1,\ldots,a_K\}$ and the ancestor indicators $\mathbf{c} = \{c_{i,t} \ : \ i\in\{1,2,\ldots,I\},\ t\in\{1,2,\ldots,T\}\}$ can be obtained by integrating out θ from the joint conditional probability starting from Equation (4) as follows:

$$\begin{aligned} p(\boldsymbol{h}|\mathbf{c},\boldsymbol{a}) &= \prod_k \Big(\int \prod_{i,t|c_{i,t}=k} p(h_{i,t},\theta_k|a_{k,t})R(\alpha_h,\beta_h)\theta_k^{\alpha_h-1}(1-\theta_k)^{\beta_h-1}d\theta_k\Big) \\ &= \prod_k R(\alpha_h,\beta_h)\frac{\Gamma(\alpha_h+l_k)\Gamma(\beta_h+l'_k)}{\Gamma(\alpha_h+\beta_h+l_k+l'_k)}\Big(\frac{1}{|B|-1}\Big)^{l'_k} \end{aligned} \tag{5}$$

where $\Gamma(\cdot)$ is the gamma function, $R(\alpha_h,\beta_h) = \frac{\Gamma(\alpha_h+\beta_h)}{\Gamma(\alpha_h)\Gamma(\beta_h)}$ is the normalization constant associated with $\text{Beta}(\alpha_h,\beta_h)$ (which is a prior distribution for θ), $l_k = \sum_t\sum_i \mathbb{I}(h_{i,t} = a_{k,t})\mathbb{I}(c_{i,t} = k)$ is the number of alleles that were not mutated with respect to the ancestral allele, and $l'_k = \sum_t\sum_i \mathbb{I}(h_{i,j} \neq a_{k,j})\mathbb{I}(c_{i,t} = k)$ is the number of mutated alleles. The counting record $\mathbf{l}_k = \{l_k, l'_k\}$ is a sufficient statistic for the parameter θ_k.

The generative process and likelihood functions described above point naturally to an algorithm for population genetic inference. Unlike the classical coalescence models for recombination (Hudson 1983), which have been primarily used for theoretical analysis and simulation, but are hardly feasible for reverse ancestral inference based on

[††] For simplicity, we assume that the mutation rates pertaining to different ancestors follow the same prior $\text{Beta}(\alpha_h,\beta_h)$.

observed genetic data, the HMDP model described above for recombination and inheritance provides a semi-parametric Bayesian formalism that is well suited for data-driven posterior inference on the latent variables that can yield rich information on the population ancestry and genetic structure of the study population. For example, under a HMDP, given the haplotype data, one can infer the ancestral pattern, LD structure and recombination hotspot of a population using the posterior distribution of inheritance variable $\mathbf{c}$ and ancestral state $\boldsymbol{a}$, as we will elaborate in the sequel. It is also possible to infer the age of the haplotype alleles and/or the time of recombination events by exploring the posterior estimates of the mutation and recombination rates under HMDP.

3 Posterior Inference

In this section, we describe a Gibbs sampling algorithm for posterior inference under HMDP. Recall that a Gibbs sampler draws samples of each random variable (or subset of random variables) in the model from the conditional distribution of the variable(s) given (previously sampled) values of all the remaining variables. The variables of interest in our model include $\{C_{i,t}\}$, the inheritance variables specifying the origins of SNP alleles of all loci on each haplotype, and $\{A_{k,t}\}$, the founding alleles at all loci of each ancestral haplotype. All other variables in the model, e.g., the mutation rate θ, are integrated out.

We assume that the individual haplotypes $\{H_{i_e,t}\}$ are given unambiguously for the study population, as is the case in many LD and haplotype-block analyses (Daly et al. 2001; Anderson and Novembre 2003); but it is noteworthy that our model can generalize straightforwardly to unphased genotype data by incorporating a simple genotype model as in Xing et al. (2004). Given that haplotypes are unambiguous, we can now treat the paternal and maternal haplotypes of N individual as $2N$ *iid* samples from the HMDP process and omit the parental index e.

The Gibbs sampler alternates between two sampling stages. First it samples the inheritance variables $\{C_{i,t}\}$, conditioning on all given individual haplotypes $\boldsymbol{h} = \{h_1, \ldots, h_{2N}\}$, and the most recently sampled configuration of the ancestor pool $\boldsymbol{a} = \{a_1, \ldots, a_K\}$; then given $\boldsymbol{h}$ and current values of the $C_{i,t}$'s, it samples every ancestor a_k.

To improve the mixing rate, we sample the inheritance variables one block at a time. That is, every time we sample δ consecutive states $c_{t+1}, \ldots, c_{t+\delta}$ starting at a randomly chosen locus $t+1$ along a haplotype. (For simplicity we omit the haplotype index i here and in the forthcoming expositions when it is clear from context that the statements or formulas apply to all individual haplotypes). Let $\mathbf{c}^-$ denote the set of previously sampled inheritance variables. Let $\mathbf{n}$ denote the totality of occupancy records of the top-level DP (i.e. the "stock urn") — $\{n\} \cup \{n_k \ : \ \forall k\}$, and $\boldsymbol{m}$ denote the totality of the occupancy records of each lower-level DP (i.e., the urns corresponding to the recombination choices by each ancestor) — $\{m_k \ : \ \forall k\} \cup \{m_{k,k'} \ : \ \forall k, k'\}$. Let $\mathbf{l}_k$ denote the sufficient statistics associated with all haplotype instances originating from ancestor k. The predictive distribution of a δ-block of inheritance variables can be written as:

$$\begin{aligned} p(c_{t+1:t+\delta}\ |\mathbf{c}^{-}, \boldsymbol{h}, \boldsymbol{a}) &\propto p(c_{t+1:t+\delta}\ |c_t, c_{t+\delta+1}, \boldsymbol{m}, \mathbf{n}) p(h_{t+1:t+\delta}|a_{c_{t+1},t+1}, \ldots, a_{c_{t+\delta},t+\delta}) \\ &\propto \prod_{j=t}^{t+\delta} p(c_{j+1}|c_j, \boldsymbol{m}, \mathbf{n}) \prod_{j=t+1}^{t+\delta} p(h_j|a_{c_j,j}, \mathbf{l}_{c_j}). \end{aligned} \tag{6}$$

This expression is simply Bayes' theorem with $p(h_{t+1:t+\delta}|a_{c_{t+1},t+1}, \ldots, a_{c_{t+\delta},t+\delta})$ playing the role of the likelihood and $p(c_{t+1:t+\delta}\ |\mathbf{c}^{-}, \boldsymbol{h}, \boldsymbol{a})$ playing the role of the posterior. One should be careful that the sufficient statistics $\mathbf{n}$, $\boldsymbol{m}$ and $\mathbf{l}$ employed here should exclude the contributions by samples associated with the δ-block to be sampled. Note that naively, the sampling space of an inheritance block of length δ is $|A|^{\delta}$ where $|A|$ represents the cardinality of the ancestor pool. However, if we assume that the recombination rate is low and block length is not too big, then the probability of having two or more recombination events within a δ-block is very small and thus can be ignored. This approximation reduces the sampling space of the δ-block to $O(|A|\delta)$, i.e., $|A|$ possible recombination targets times δ possible recombination locations. Accordingly, Eq. (6) reduces to:

$$p(c_{t+1:t+\delta}\ |\mathbf{c}^{-}, \boldsymbol{h}, \boldsymbol{a}) \propto p(c_{t'}\ |c_{t'-1} = c_t, \boldsymbol{m}, \mathbf{n}) p(c_{t+\delta+1}\ |c_{t+\delta} = c_{t'}, \boldsymbol{m}, \mathbf{n}) \prod_{j=t'}^{t+\delta} p(h_j|a_{c_{t'},j}, \mathbf{l}_{c_{t'}}), \tag{7}$$

for some $t' \in [t+1, t+\delta]$. Recall that in an HMDP model for recombination, given that the total recombination probability between two loci d-units apart is $\lambda \equiv 1 - e^{-dr} \approx dr$ (assuming d and r are both very small), the transition probability from state k to state k' is:

$$\begin{aligned} & p(c_{t'} = k'\ |c_{t'-1} = k, \boldsymbol{m}, \mathbf{n}, r, d) \\ = & \begin{cases} \lambda\pi_{k,k'} + (1-\lambda)\delta(k, k') & \text{for } k' \in \{1, ..., K\}, \text{ i.e., transition to an existing ancestor,} \\ \lambda\pi_{k,K+1} & \text{for } k' = K+1, \text{ i.e., transition to a new ancestor,} \end{cases} \end{aligned} \tag{8}$$

where π_k represents the transition probability vector for ancestor k under HMDP, as defined in Eq. (3). Note that when a new ancestor a_{K+1} is instantiated, we need to immediately instantiate a new DP under F to model the transition probabilities from this ancestor to all instantiated ancestors (including itself). Since the occupancy record of this DP, $\boldsymbol{m}_{K+1} := \{m_{K+1}\} \cup \{m_{K+1,k}\ :\ k = 1, \ldots, K+1\}$, is not yet defined at the onset, with probability 1 we turn to the top-level DP when departing from state $K+1$ for the first time. Specifically, we define $p(\cdot|c_{t'} = K+1)$ according to the occupancy record of ancestors in the stock urn. For example, at the distal border of the δ-block, since $c_{t+\delta+1}$ always indexes a previously inherited ancestor (and therefore must be present in the stock-urn), we have:

$$p(c_{t+\delta+1}\ |c_{t+\delta} = K+1, \boldsymbol{m}, \mathbf{n}) \quad = \quad \lambda \times \frac{n_{c_{t+\delta+1}}}{n - 1 + \alpha}. \tag{9}$$

Now we can substitute the relevant terms in Eq. (6) with Eqs. (8) and (9). The marginal likelihood term in Eq. (6) can be readily computed based on Eq. (4), by integrating out the mutation rate θ under a Beta prior (and also the ancestor a under a uniform prior if $c_{t'}$ refers to an ancestor to be newly instantiated) (Xing et al. 2004). Putting everything together, we have the proposal distribution for a block of inheritance variables. Upon sampling every c_t, we update the sufficient statistics $\mathbf{n}$, $\boldsymbol{m}$ and $\{\mathbf{l}_k\}$ as follows. First, before drawing the sample, we erase the contribution of c_t to these sufficient statistics. In particular, if an ancestor gets no occupancy in either the stock or the HMM urns afterwards, we remove it from our repository. Then, after drawing a new c_t, we increment the relevant counts accordingly. In particular, if $c_t = K+1$ (i.e., a new ancestor is to be drawn), we update $n = n+1$, set $n_{K+1} = 1$, $m_{c_t} = m_{c_t}+1$, $m_{c_t,K+1} = 1$, and set up a new (empty) HMM urn with color $K+1$ (i.e. instantiating $\boldsymbol{m}_{K+1}$ with all elements equal to zero).

Now we move on to sample the founders $\{a_{k,t}\}$. From the mutation model in Equation (4), we can derive the following posterior distribution to sample the founder a_k [‡‡]:

$$
\begin{aligned}
p(a_{k,t}|\mathbf{c},\boldsymbol{h}) &\propto \int \Big(\prod_{i|c_{i,t}=k} p(h_{i,t}|a_{k,t},\theta)\Big) Beta(\theta|\alpha_h,\beta_h)d\theta \\
&= \frac{\Gamma(\alpha_h + l_{k,t})\Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h+\beta_h+l_{k,t}+l'_{k,t})(|B|-1)^{l'_{k,t}}} R(\alpha_h,\beta_h), \qquad (10)
\end{aligned}
$$

where $l_{k,t}$ is the number of allelic instances originating from ancestor k at locus t that are identical to the ancestor, when the ancestor has the pattern $a_{k,t}$; and $l'_{k,t} = \sum_i \mathbb{I}(c_{i,t} = k|a_{k,t}) - l_{k,t}$ represents the complement. The normalization constant of this proposal distribution can be computed by summing the R.H.S. of Eq. (10) over all possible allele states of an ancestor at the locus being sampled. If k is not represented previously, we can just set $l_{k,t}$ and $l'_{k,t}$ both to zero. Note that when sampling a new ancestor, we can only condition on a small segment of an individual haplotype. To instantiate a complete ancestor, after sampling the alleles in the ancestor corresponding to the segment according to Eq. (10), we first fill in the rest of the loci with random alleles. When another segment of an individual haplotype needs a new ancestor, we do not naively create a new full-length ancestor; rather, we use the *empty* slots (those with random alleles) of one of the previously instantiated ancestors, if any, so that the number of ancestors does not grow unnecessarily.

4 Experiments

We applied the HMDP model to both simulated and real haplotype data. Our analyses focus on the following three popular problems in statistical genetics: 1. Ancestral Inference: estimating the number of founders in a population and reconstructing the ancestor haplotypes; 2) LD-block Analysis: inferring the recombination sites in each individual haplotype and uncover population-level recombination hotspots on the chromosome region; 3) Population Structural Analysis: mapping the genetic origins of all

[‡‡] In deriving Equation (10), instead of assuming a common mutation rate θ_k for all loci of ancestor a_k, we endow each locus with its own mutation parameter $\theta_{k,t}$, with all parameters admitting the same prior Beta(α_h, β_h). This is arguably a more accurate reflection of reality.

loci of each individual haplotype in a population.

4.1 Analyzing simulated haplotype population

To simulate a population of individual haplotypes, we started with a fixed number, K_s (unknown to the HMDP model), of randomly generated ancestor haplotypes, on each of which a set of recombination hotspots were pre-specified. Then we applied a hand-specified recombination process, which is defined by a K_s-dimensional HMM, to the ancestor haplotypes to generate N_s individual haplotypes, via sequentially recombining segments of different ancestors according to the simulated HMM states at each locus, and mutating certain ancestor SNP alleles according to the emission model. All the ancestor haplotypes were set to be 100 SNPs long. At the hotspots (pre-specified at every 10-th loci in the ancestor haplotypes), we defined the recombination rate to be 0.05, otherwise it is 0.00001. We simulated the recombination process for each progeny haplotype; but to force every progeny haplotype to have at least one recombination, in the rare cases where no recombination event was simulated for an progeny haplotype, we sampled one of the hotspots randomly and forced it to recombine with another ancestor chosen at random at that loci. (Thus our simulated samples were not exactly distributed according to the generative model we used, but such samples were arguably more close to the real data.) Overall, 30 datasets each containing 100 individuals (i.e., 200 haplotypes) with 100 SNPs were generated from $K_s = 5$ ancestor haplotypes.

As baseline models, we also implemented 3 standard fixed-dimensional HMM, with K' equal to 3, 5 (the true number of ancestors for the simulated) and 10 hidden states, respectively, which correspond to the number of ancestors available for recombination. For these baseline HMMs, we follow the same mutation model for emission as that of the HMDP (i.e., Eq. (4)), and we also subject the mutation rate to a Beta prior. In these HMMs, the SNP-types of the ancestors at every locus, e.g., $a_{t,k}$, are treated as the mean parameters of the observed SNPs samples at the corresponding locus; the inheritance variables $\{C_{i,t}\}$ correspond to the latent states following a 1-st order Markov process; and the transition models governing recombinations amongst the ancestors as indicated by the values $c_{i,t}$'s are parameterized by a K'-dimensional stochastic matrix. We estimate these parameters via a maximal likelihood principle using the Balm-Welch algorithm. Note that since K' is chosen *a priori*, we cannot estimate the number of ancestors using these HMMs.

Following a *collapsed* Gibbs sampling scheme (Liu 1994), we integrated out the mutation rate θ, and sample variables $\{A_{k,t}\}$ and $\{C_{i,t}\}$ iteratively. We monitor convergence based on the occupancy counts of the top factors in the master DP. Typically, convergence was achieved after around 3000 samples (Figure 2), and the samples obtained after convergence (with proper de-autocorrelation, i.e., by using samples from every 10 iterations over 5000 $\sim$ 10000 samples) are used for computing relevant sufficient statistics. To increase the chance of proper mixing, 10 independent runs of sampling, with different random seeds, are simultaneously performed. Convergence is monitored at runtime using an on-line minimal pairwise Gelman-Rubin (GR) statistic (Gelman 1998) of scalar summaries of the model parameters (e.g., average occupancy of top fac-

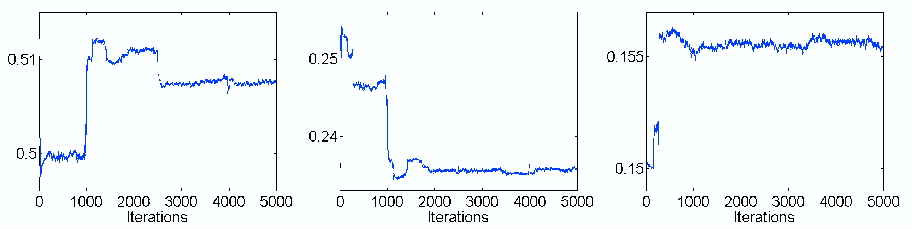


Figure 2: Sampling trace of the top three most occupied factors (ancestor chromosomes). The x-axis represents the sampling iteration, and the y-axis represent the fraction of the occupancy (i.e., be chosen as recombination target) of each factor over total occupancy.

tors) obtained in each Markov chain. The total running time for posterior inference on a simulated data set described below was around 3.5 hours using a matlab implementation on a Dell PowerEdge 1850 workstation with an Intel Xeon 3.6 GHz processor. (This computation includes a huge disk-writing overhead for recording the running trace. The actual CPU time for computing is less than 10% of that. We intend to soon release a C++ implementation which is expected to further reduce computation cost.)

Ancestral Inference Using HMDP, we successfully recovered the correct number (i.e., $K = 5$) of ancestors in 21 out of 30 simulated populations; for the remaining 9 populations, we inferred 6 ancestors. From samples of ancestor states $\{a_{k,t}\}$, we reconstructed the ancestral haplotypes under the HMDP model. For comparison, we also inferred the ancestors under the 3 standard HMM using an EM algorithm. We define the *ancestor reconstruction error* ϵ_a for each ancestor to be the ratio of incorrectly recovered loci over all the chromosomal sites. The average ϵ_a over 30 simulated populations under 4 different models are shown in Figure 3a. In particular, the average reconstruction errors of HMDP for each of the five ancestors are 0.026, 0.078, 0.116, 0.168, and 0.335, respectively. There is a good correlation between the reconstruction quality and the population frequency of each ancestor. Specifically, the average (over all simulated populations) fraction of SNP loci originated from each ancestor among all loci in the population is 0.472, 0.258, 0.167, 0.068 and 0.034, respectively. As one would expect, the higher the population frequency of an ancestor is, the better its reconstruction accuracy. Interestingly, under the fixed-dimensional HMM, even when we use the correct number of ancestor states, i.e., $K = 5$, the reconstruction error is still very high (Figure 3), typically 2.5 times or higher than the error of HMDP. We conjecture that this is because the non-parametric Bayesian treatment of the transition rates and ancestor configurations under the HMDP model leads to a desirable adaptive smoothing effect and also less constraints on the model parameters, which allow them to be more accurately estimated. Whereas under a parametric setting, parameter estimation can easily be sub-optimal due to lack of appropriate smoothing or prior constraints, or deficiency of the learning algorithm (e.g., local-optimality of EM).

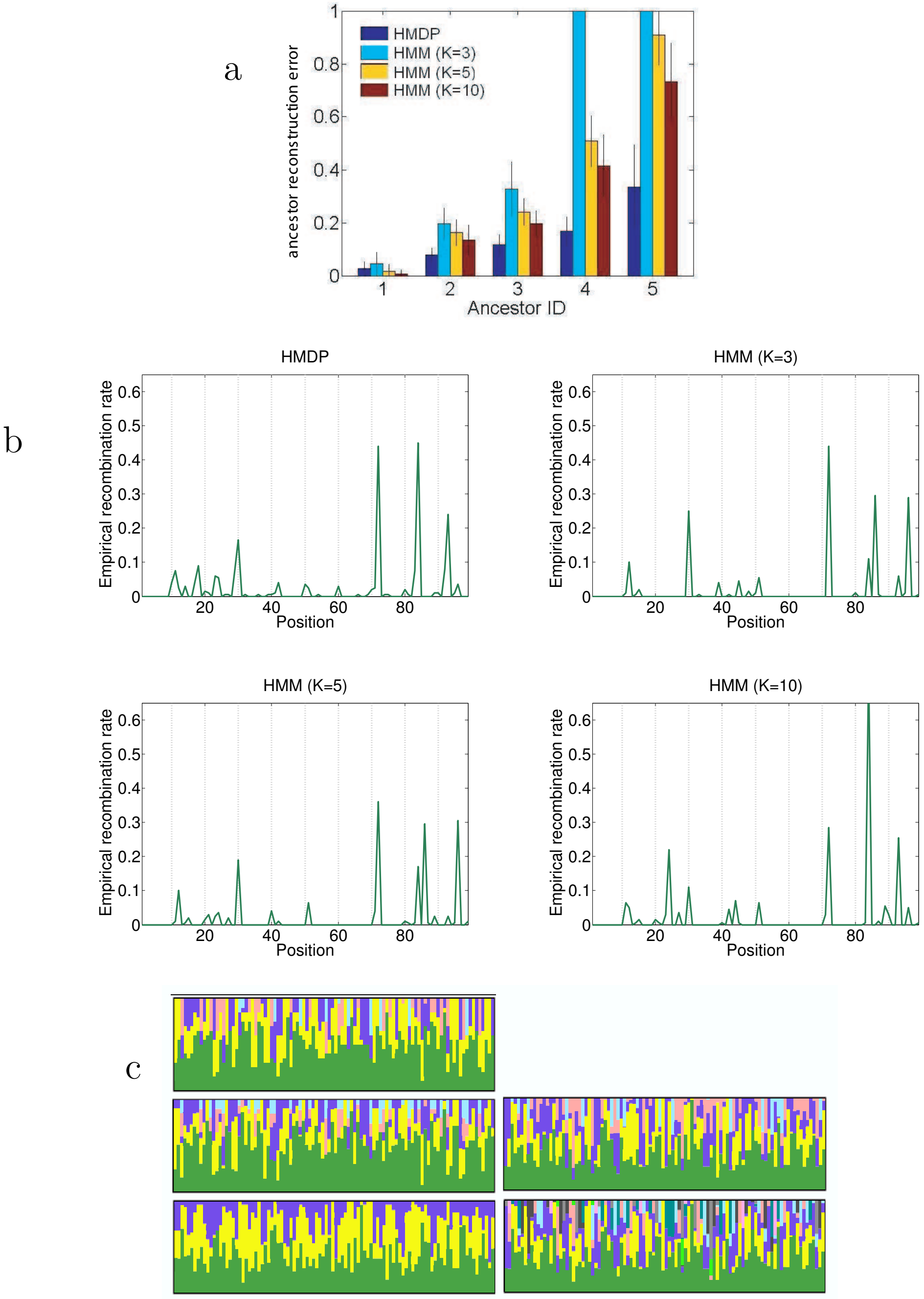


Figure 3: Analysis of simulated haplotype populations. (a) A comparison of ancestor reconstruction errors for the five ancestors (indexed along x-axis). The vertical lines show ±1 standard deviation over 30 populations. (b) Plots of the empirical recombination rates along 100 SNP loci in one of the 30 populations for HMDP and 3 HMMs. The dotted lines show the pre-specified recombination hotspots. (c) The true (panel 1) and estimated (panel 2 for HMDP, and panel 3-5 for 3 HMMs) population maps of ancestral compositions in a simulated population. Figures were generated using the software *distruct* from Rosenberg *et al* [2002].

threshold	0.01			0.03		
tolerance window	0	± 1	± 2	0	± 1	± 2
False positive rate	0.16	0.12	0.067	0.08	0.04	0.03
False negative rate	0	0	0	0.77	0.55	0.55

Table 1: False positive and false negative rates for recombination hotspot detection using medians of the empirical recombination rates over 30 population samples as shown in Figure 4.

LD-block Analysis From samples of the inheritance variables $\{c_{i,t}\}$ under HMDP, we can infer the recombination status of each locus of each haplotype. We define the empirical recombination rates λ_e at each locus to be the ratio of individuals who had recombinations at that locus over the total number of haploids in the population. Figure 3b shows plots of the λ_e from HMDP and the 3 HMMs in one of the 30 simulated populations. We can identify the recombination *hotspots* directly from such a plot based on an empirical threshold λ_t (i.e., $\lambda_t = 0.05$). For comparison, we also give the true recombination hotspots (depicted as dotted vertical lines) chosen in the ancestors for simulating the recombinant population. The inferred hotspots (i.e., the λ_e peaks) show reasonable agreement with the reference in both HMDP and HMMs, but it appears that in the HMMs the hotspots around position 20 and 60 are less obvious. Figure 4 shows a boxplot of the empirical recombination rates at the 100 SNP loci estimated from the the 30 different population samples simulated from these ancestors. The gray vertical lines along the x-axis correspond to the locations of pre-specified recombination hotspots. A simple thresholding at 0.01 would identify 24 hotspots which include all the 9 true hotspots and 15 false positive sites. This leads to the false negative rate to be 0 and the false positive rate to be 0.16. To give credit to the false positive sites which are close to the true hotspots, we may allow small discrepancy between the true hotspots and the detected ones. By allowing ±2 sites discrepancy and eliminating possibly redundant ones in the detection, (e.g., the two detected sites 70 and 71 would be just counted as 1 site of 70), the number of false positive sites decreased to 6, which resulted in the false positive rate of 0.067 and the false negative rate unchanged. Using a threshold of 0.03, 10 hotspots would be detected, among which two sites agree with the true ones. After allowing ±2 sites discrepancy 4 true hotspots could be identified with 3 remaining false positive sites. The false positive and negative rates using these two thresholds are summarized in Table 1.

Population Structural Analysis Finally, from samples of the inheritance variables $\{c_{i,t}\}$, we can also uncover the genetic origins of all loci of each individual haplotype in a population. For each individual, we define an empirical *ancestor composition vector* η_e, which records the fractions of every ancestor in all the $c_{i,t}$'s of that individuals. Figure 3c displays a *population map* constructed from the η_e's of all individual. In the population map, each individual is represented by a thin vertical line which is partitioned into colored segments in proportion to the ancestral fraction recorded by η_e. Five population maps, corresponding to (1) true ancestor compositions, (2) ancestor compositions inferred by HMDP, and (3-5) ancestor compositions inferred by HMMs with 3, 5, 10 states, respectively, are shown in Figure 3c. To assess the accuracy of

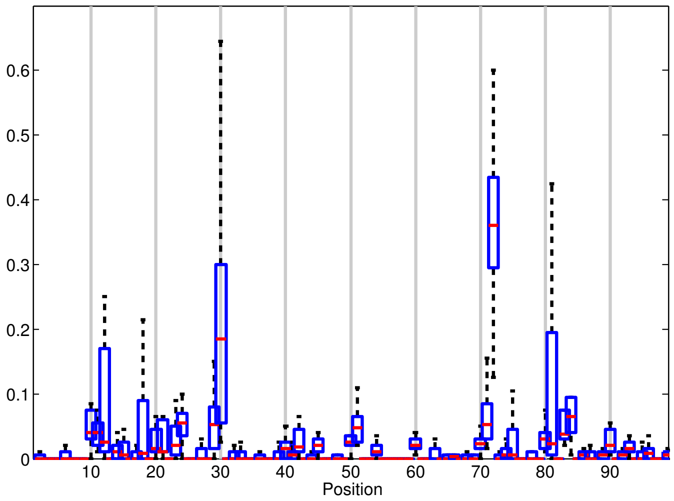


Figure 4: Boxplot of the empirical recombination rates at the 100 SNP loci over 30 different simulated population samples. The gray vertical lines show the pre-specified recombination hotspots used for simulating the data.

our estimation, we calculated the distance between the true ancestor compositions and the estimated ones as the mean squared distance between true and the estimated η_e over all individuals in a population, and then over all 30 simulated populations. We found that the distance between the HMDP-derived population map and the true map is 0.190 ± 0.0748, whereas the distance between HMM-map and true map is 0.319 ± 0.0676, significantly worse than that of HMDP even though the HMM is set to have the true number of ancestral states (i.e., $K = 5$). Because of dimensionality incompatibility and apparent dissimilarity to the true map for other HMMs (i.e., $K = 3$ and 10), we forgo the above quantitative comparison for these two cases.

To summarize our analyses on the simulated data, although the fixed dimensional HMMs are fast and easy to implement, they appear to offer much less accurate results than that of the HMDP model on ancestor reconstruction, and population-map estimation, even when the number of HMM states is set to the true number of haplotype ancestors (which is in practice unknown). When the number of HMM states is chosen incorrectly, the inference results degrade significantly. For hotspot prediction, qualitatively we have not seen significant differences in the accuracy, although the HMDP model appeared to be slightly better. We will look into this issue via a more quantitative analysis in our later study.

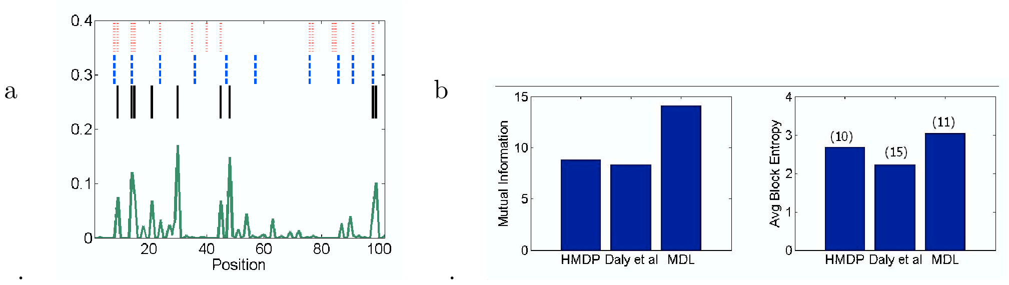


Figure 5: Analysis of the Daly data. (a) A plot of λ_e estimated via HMDP; and the haplotype block boundaries according to HMDP (black solid line), HMM (Daly et al. 2001) (red dotted line), and MDL (Anderson and Novembre 2003) (blue dashed line). (b) IT scores for haplotype blocks from each method. The left panel shows cross-block MI and the right shows the average within-block entropy. The total number of blocks inferred by each method are given on top of the bars.

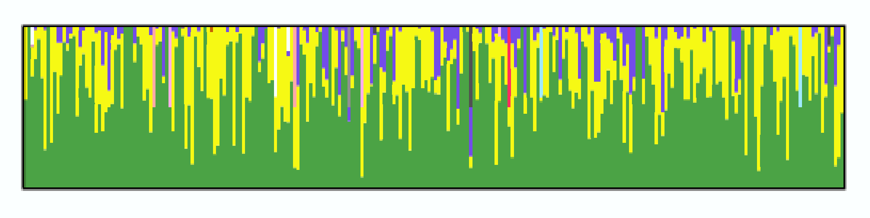


Figure 6: The estimated population map of the Daly dataset.

4.2 Analyzing two real haplotype datasets

We applied HMDP to two real haplotype datasets, the single-population Daly data (Daly et al. 2001), and the two-population (CEPH: Utah residents with northern/western European ancestry; and YRI: Yoruba in Ibadan and Nigeria) HapMap data (Consortium" 2005; Thorisson et al. 2005). These data consist of trios of genotypes, so most of the true haplotypes can be directly inferred from the genotype data. Note that for these real biological data, there is no ground truth regarding the ancestral history, hotspot location, and population composition, based on which we can validate our results, or compare to other methods. We present our analysis as a demonstration of the utilities of our model, which, to our knowledge, are not offered jointly under a unified model by extant methods in statistical genetics. (As we discuss in the sequel, some extant methods can perform some of the inference tasks that HMDP does, and in these cases we show a comparison.)

The single-population Daly dataset We first analyzed the 256 individuals from Daly data. This data set consists of the haplotypes 103 SNPs across a 616.7-kb region on chromosome 5q31 of 129 trios from a European-derived population. Earlier studies indicate that this region contains a genetic risk factor for Crohn disease. Earlier analysis of this data set using a hidden Markov model revealed the existence of discrete haplotype blocks, each with low diversity, in this region (Daly et al. 2001).

We compared the recovered recombination hotspots with those reported in Daly et al. (2001) (which is based on an HMM employing different number of states at different chromosome segments) and in Anderson and Novembre (2003) (which is based on a minimal description length (MDL) principle applied to Daly's HMM). Note that the HMM used by Daly et al. (2001) and Anderson and Novembre (2003) is different from the ones we used in our simulation study in section 4.1. Their HMM models a stochastic process that selects haplotype-segments from pools of "ancestors" without mutation for a concatenating list of haplotype-block regions constituting the study SNP sequences. Each region has their own ancestor pool of possibly unequal sizes; thus between each pair of adjacent blocks, the HMM needs a unique (possibly rectangular) stochastic matrix for ancestor transitions. The block boundaries are fixed under this HMM (and the only stochasticity lies in the choice of local "ancestors" for each block), and determining the block boundaries is treated as a model-selection problem based on a maximal-likehood (Daly et al. 2001) or MDL (Anderson and Novembre 2003) principle. Strictly speaking, Daly's HMM model itself offers little means to infer recombination events and the ancestor association map, because the "ancestors" thereof are defined independently for each block rather than as whole founding chromosomes; different blocks have different number of ancestors; and the determination of these "local ancestors" employs an initial heuristic scan for regions of low haplotype diversity, whose formal connection to the HMM model is not clear.

Figure 5a shows the plot of empirical recombination rates estimated under HMDP, side-by-side with the reported recombination hotspots. There is no ground truth to judge which one is correct; hence we computed information-theoretic (IT) scores based on the estimated within-block haplotype frequencies and the between-block transition probabilities under each model for a comparison. Figure 5b shows a comparison of these scores for haplotype blocks obtained from HMDP and the other two sources. The left panel of Figure 5b shows the total pairwise mutual information between adjacent haplotype blocks segmented by the recombination hotspots uncovered by the three methods. The right panel shows the average entropies of haplotypes within each block. The number above each bar denotes the total number of blocks. The pairwise mutual information score of the HMDP block structure is similar to that of the Daly structure, but smaller than that of MDL. Similar tendencies are observed for average entropies. Note that the Daly and the MDL methods allow the number of haplotype founders to vary across blocks to get the most compact local ancestor constructions. Thus their reported scores are an underestimate of the true global score because certain segments of an ancestor haplotype that are not or rarely inherited are not counted in the score. Thus the low IT scores achieved by HMDP suggest that HMDP can effectively avoid inferring spurious global and local ancestor patterns. This is confirmed by the population map shown in Figure 6, which shows that HMDP recovered 6 ancestors and among them the 3 dominant ancestors account for 98% of all the modern haplotypes in the population.

We did not compare our results with that of Daly et al. (2001) and Anderson and Novembre (2003) exhaustively, e.g., on ancestor reconstruction and population map estimation, because their methods cannot perform these inferential tasks. Indeed, to our knowledge there is no single model that does all the inferential tasks HMDP is capable of. Thus we can only compare HMDP with specialized models on certain tasks, as described above. Since implementations of the methods in Daly et al. (2001) and Anderson and Novembre (2003) are not available, we can only compare with their results reported on the original papers, which are obtained on the Daly data. But we cannot apply their methods to our simulated data or the HapMap data for more informative comparisons. The total running time of our algorithm on the Daly data set (with the 3000 burn-in steps, 3000 samples, and 1 per 5 sample deceleration sampling interval) is about 14hr, which includes the disk-writing overhead for trace-recording.

The two-population HapMap dataset The HapMap data was generated by the International HapMap Project that attempts to identify and catalog genetic similarities and differences in human beings of different ethnic origins (Consortium" 2005; Thorisson et al. 2005). The current release of the whole HapMap data contains over 1 million SNPs, from 269 individuals belonging to four populations. In this study, we only focus on a small subset of SNPs common to all populations; we use data from two of the four populations, YRI and CEPH. Specifically, we have 30 trios of YRI and 30 trios of CEPH (i.e., 180 individuals in total), of which the 120 unrelated phase-known individuals corresponding to the parents in the trios were used in the experiment (the children's haplotypes are inherited from the parents and are redundant in the population). We concern ourselves with 254 SNPs, which are located in the region of $ENm010.7p15.2$ spanning 497.5 kilo-basepair (kb). The computation time for analyzing this data set is comparable to that of the Daly data set.

We applied HMDP to the union of the populations, with a random individual order. Delightfully, the two-population structure is clearly retrieved from the population map constructed from the population composition vectors η_e for every individual. As seen in Figure 7a, the left half of the map clearly represents the CEPH population and the right half the YRI population. We found that the two dominant haplotypes covered over 85% of the CEPH population (and the overall breakup among all four ancestors is 0.5618, 0.3036, 0.0827, 0.0518). On the other hand, the frequencies of each ancestor in YRI population are 0.2141, 0.1784, 0.3209, 0.1622, 0.1215 and 0.0029, showing that the YRI population is much more diverse than CEPH. This might explain an earlier observation that genetic inference on the YRI population appeared to be more difficult than for CEPH (Marchini et al. 2006). The recombination maps of the two different populations also show noticeably different spatial patterns of recombination hotspots (Figure 7b), which may reflect different recombination histories of the founders of the two populations.

Note that the population partition result reported in Figure 7b is trivial because it is inferred purely based on SNPs haplotypes without knowledge of ethnic labels of the samples. In most genetic samples, ethnic labels are either not available or ambiguous (e.g., the Daly data has no subpopulation details). By discovering the right population separation, one can perform hotspot estimation for each population and capture population-specific LD (as in Figure 7a); whereas in a mixed population, one may not be able to correctly estimate such patterns.

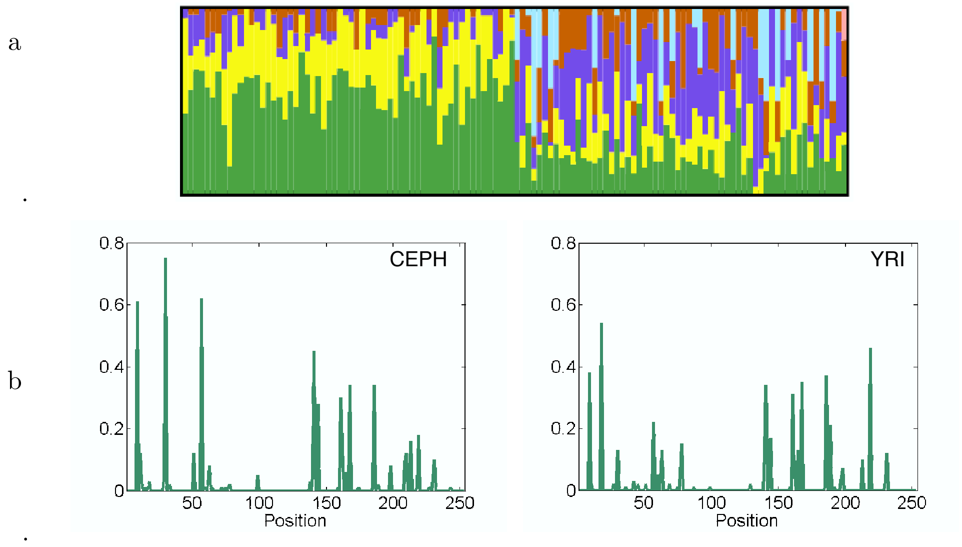


Figure 7: Result on the two-population (CEPH and YRI) HapMap data. (a) The estimated population map of the whole dateset with two populations. (b) The estimated recombination rates along the chromosomal position in the two populations.

5 Conclusion

We have proposed a new Bayesian approach for joint modeling of genetic recombinations among possibly infinite founding alleles and coalescence-with-mutation events in the resulting genealogies. By incorporating a hierarchical DP prior to the stochastic matrix underlying an HMM, which facilitates a well-defined transition process between infinitely many ancestors, our proposed method can efficiently infer a number of important genetic variables, such as recombination hotspot, mutation rates, haplotype origin, and ancestor patterns, jointly underly a unified statistical framework.

Empirically, on both simulated and real data, our approach compares favorably to its parametric counterpart—a fixed-dimensional HMM (even when the number of its hidden states, i.e., the ancestors, is correctly specified) and a few other specialized methods, on ancestral inference, haplotype-block uncovering and population structural analysis. We are interested in further investigating the behavior of an alternative scheme based on reverse-jump MCMC over Bayesian HMMs with different latent states in comparison with HMDP; we also intend to apply our methods to genome-scale LD and demographic analysis using the full HapMap data. While our current model employs only phased haplotype data, it is straightforward to generalize it to unphased genotype data as provided by the HapMap project. HMDP can also be easily adapted to many engineering and information retrieval contexts such as object and theme tracking in open space.

References

Anderson, E. C. and Novembre, J. (2003). "Finding haplotype block boundaries by using the minimum-description-length principle." *Am J Hum Genet*, 73: 336–354. 503, 507, 512, 520, 521, 522

Antoniak, C. E. (1973). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *Annals of Statistics*, 2: 1152–1174. 506

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). "The Infinite Hidden Markov Model." In Dietterich, T., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*, 577–584. MIT Press. 504, 505, 508

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions Via Pólya Urn Schemes." *Annals of Statistics*, 1: 353–355. 505

Consortium", I. H. (2005). "A haplotype map of the human genome." *Nature*, 437: 1299–1320. 520, 522

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). "High-resolution haplotype structure in the human genome." *Nature Genetics*, 29(2): 229–232. 503, 507, 512, 520, 521, 522

Erosheva, E., Fienberg, S., and Lafferty, J. (2004). "Mixed-membership models of scientific publications." *Proc Natl Acad Sci U S A*, 101 (Suppl 1): 5220–5227. 503

Escobar, M. D. and West, M. (2002). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. 505, 506, 510

Excoffier, L. and Hamilton, G. (2003). "Comment on Genetic Structure of Human Populations." *Science*, 300(5627): 1877b–. 503

Excoffier, L. and Slatkin, M. (1995). "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." *Molecular Biology and Evolution*, 12(5): 921–7. 507

Falush, D., Stephens, M., and Pritchard, J. K. (2003). "Inference of population structure: Extensions to linked loci and correlated allele frequencies." *Genetics*, 164(4): 1567–1587. 503

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230. 506

Gelman, A. (1998). "Inference and monitoring convergence." In Gilks, W. E., Richardson, S., and Spiegelhalter, D. J. (eds.), *Markov Chain Monte Carlo in Practice*. Boca Raton, Florida: Chapman & Hall/CRC. 515

Greenspan, G. and Geiger, D. (2004a). "High density linkage disequilibrium mapping using models of haplotype block variation." *Bioinformatics*, 20 (Suppl.1): 137–144. 503

— (2004b). "Model-Based Inference of Haplotype Block Variation." *Journal of Computational Biology*, 11(2/3): 493–504. 504, 505

Hoppe, F. M. (1984). "Pólya-like urns and the Ewens' sampling formula." *Journal of Math. Biol.*, 20(1): 91–94. 505

Hudson, R. R. (1983). "Properties of a neutral allele model with intragenic recombination." *Theor Popul Biol.*, 23(2): 183–201. 511

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 90: 161–173. 505, 506

Kimmel, G. and Shamir, R. (2004). "Maximum likelihood resolution of multi-block genotypes." In *In proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)*, 2–9. The Association for Computing Machinery. 507

Kingman, J. (1982). "On the genealogy of large populations." *J. Appl. Prob.*, 19A: 27–43. 505

Li, N. and Stephens, M. (2003). "Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data Genetics." *Genetics*, 165: 2213–2233. 503, 506

Liu, J. S. (1994). "The collapsed Gibbs sampler with applications to a gene regulation problem." *J. Amer. Statist. Assoc*, 89: 958–966. 515

Liu, J. S., Sabatti, C., Teng, J., Keats, B., and Risch, N. (2001). "Bayesian analysis of Haplotypes for Linkage Disequilibrium Mapping." *Genome Res.*, 11: 1716–1724. 511

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z., Munro, H., Abecasis, G., Donnelly, P., and Consortium, I. H. (2006). "A Comparison of Phasing Algorithms for Trios and Unrelated Individuals." *The American Journal of Human Genetics*, 78: 437–450. 522

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *J. Computational and Graphical Statistics*, 9(2): 249–256. 505, 506

Niu, T., Qin, S., Xu, X., and Liu, J. (2002). "Bayesian haplotype inference for multiple linked single nucleotide polymorphisms." *American Journal of Human Genetics*, 70: 157–169. 505, 507

Patil, N., Berno, A. J., et al. (2001). "Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21." *Science*, 294: 1719–1723. 503

Pritchard, J. K., Stephens, M., Rosenberg, N., and Donnelly, P. (2000). "Association mapping in structured populations." *Am. J. Hum. Genet.*, 67: 170–181. 503, 506

Rannala, B. and Reeve, J. P. (2001). "High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence." *Am J Hum Genet.*, 69(1): 159–78. 503

Rasmussen, C. E. (2000). "The Infinite Gaussian Mixture Model." In *Advances in Neural Information Processing Systems 12*, 554–560. Cambridge, MA: MIT Press. 510

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). "Genetic Structure of Human Populations." *Science*, 298: 2381–2385. 503

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 1(4): 639–50. 506

Sohn, K.-A. and Xing, E. (2006). "Hidden Markov Dirichlet Process: Modeling Genetic Recombination in Open Ancestral Space." In *Advances in Neural Information Processing Systems 19*. Cambridge, MA: MIT Press. 504

Stephens, M., Smith, N., and Donnelly, P. (2001). "A new statistical method for haplotype reconstruction from population data." *American Journal of Human Genetics*, 68: 978–989. 506, 507

Tavare, S. and Ewens, W. (1998). "The Ewens Sampling Formula." *Encyclopedia of Statistical Sciences*, Update Volume 2.: 230–234. 505

Teh, Y., Jordan, M. I., Beal, M., and Blei, D. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association* (to appear). 504, 505, 508, 509

Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005). "The International HapMap Project Web site." *Genome Research*, 15: 1591–1593. 520, 522

Xing, E., Sharan, R., and Jordan, M. (2004). "Bayesian Haplotype Inference via the Dirichlet Process." In *Proceedings of the 21st International Conference on Machine Learning*, 879–886. New York: ACM Press. 504, 505, 506, 507, 511, 512, 514

Zhang, K., Deng, M., Chen, T., Waterman, M., and Sun, F. (2002). "A dynamic programming algorithm for haplotype block partitioning." *Proc. Natl. Acad. Sci. USA*, 99(11): 7335–39. 503

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0523757, and and by the Pennsylvania Department of Health's Health Research Program under Grant No. 2001NF-Cancer Health Research Grant ME-01-739. E.P.X. is also supported by a NSF CAREER Award under Grant No. DBI-0546594.

Bayesian Analysis (2007) **2**, Number 3, pp. 529–556

Re-considering the variance parameterization in multiple precision models

Yi He*, James S. Hodges†, and Bradley P. Carlin‡

Abstract. Recent developments in Bayesian computing allow accurate estimation of integrals, making advanced Bayesian analysis feasible. However, some problems remain difficult, such as estimating posterior distributions for variance parameters. For models with three or more variances, this paper proposes a simplex parameterization for the variance structure, which has appealing properties and eases the related burden of specifying a reference prior. This parameterization can be profitably used in several multiple-precision models, including crossed random-effect models, many linear mixed models, smoothed ANOVA, and the conditionally autoregressive (CAR) model with two classes of neighbor relations, often useful for spatial data. The simplex parameterization has at least two attractive features. First, it typically leads to simple MCMC algorithms with good mixing properties regardless of the parameterization used to specify the model's reference prior. Thus, a Bayesian analysis can take computational advantage of the simplex parameterization even if its prior was specified using another parameterization. Second, the simplex parameterization suggests a natural reference prior that is proper, invariant under multiplication of the data by a constant, and which appears to reduce the posterior correlation of smoothing parameters with the error precision. We use simulations to compare the simplex parameterization, with its reference prior, to other parameterizations with their reference priors, according to bias and mean-squared error of point estimates and coverage of posterior 95% credible intervals. The results suggest advantages for the simplex approach, particularly when the error precision is small. We offer results in the context of two real data sets from the fields of periodontics and prosthodontics.

1 Introduction

Recent developments in Bayesian computing have made it possible to analyze many previously intractable models, but some problems remain difficult, such as estimating posterior distributions for variance parameters. This paper considers the class of multiple-precision linear models, having linear mean structure, normal errors, and at least three precision parameters. This class includes the conditionally autoregressive (CAR) model with two types of neighbor relations (2NRCAR; Besag & Higdon 1999, Reich et al 2007), crossed random-effects models (Box & Tiao 1992, Chapter 5), some dynamic linear models (West & Harrison 1999, Chapter 4), smoothed analysis of variance (Gelman 2005a, Hodges et al 2007), some spatio-temporal models with 1 or 2 spa-

*Sanofi-Aventis Corp,Bridgewater, NJ, mailto:Yi.He@sanofi-aventis.com

†Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN,mailto:hodges@ccbr.umn.edu

‡Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN,http://www.biostat.umn.edu/~brad

tial neighbor relations and 1 temporal relation (2NRCAR or 3NRCAR), several linear mixed models (Zhao et al 2006), e.g., additive mixed models and bivariate smoothing, and, finally, many problem-specific models (e.g., Gelman & Huang 2007). To make this discussion concrete, we use the 2NRCAR model applied to periodontal data, as follows.

In periodontics, attachment loss is used to assess cumulative damage to a patient's periodontium and to monitor disease progression (Darby & Walsh 1995). Attachment loss is measured at six sites on each tooth; Figure 1 shows one patient's data. Each measurement site is indicated by a small circle whose shade of grey indicates measured attachment loss, with darker shade indicating larger (worse) attachment loss. Excluding the four "wisdom teeth" (third molars), a full mouth of 28 teeth gives 168 measurements. If the two jaws are treated as isolated from each other, this spatial structure has at least 2 "islands", i.e., disconnected groups of measurement sites.

Attachment loss measurements are spatially correlated, but the correlation may not simply be a function of distance. Instead of using point-data (geostatistical) methods, it is practical and intuitive to model attachment loss as measurements on a lattice, which suggests conditionally autoregressive (CAR) models. However, the 168 measurement sites have a complex topography, so more than one smoothing parameter may be needed for adequate fidelity. We consider CAR models with two classes of neighbor relations. Pairs of neighboring sites come in four types (Figure 2): direct neighbor (Type a), same-side neighbors crossing the gap between teeth (Type b), opposite-side neighbors on the same tooth (Type c), and opposite-side neighbors crossing the gap between teeth (Type d). These four types of neighbor pairs can be grouped into two classes in various ways (Reich et al 2007). This paper considers the classes shown in Figure 2, with solid and dashed lines for class 1 and 2 pairs respectively (Grid A in Reich et al., 2007).

Figure 1 summarizes one patient's data, to which we fit the 2NRCAR model, as follows. Let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ denote the attachment loss measurements, where the subscript indexes measurement sites, and specify this 2NRCAR model:

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\theta}, \tau_0 &\sim N(\boldsymbol{\theta}, \tau_0 I_n) \\ \boldsymbol{\theta}|\tau_1, \tau_2 &\propto c(\tau_1, \tau_2)^{1/2} exp\left(-\frac{1}{2}\boldsymbol{\theta}'\{\tau_1 Q_1 + \tau_2 Q_2\}\boldsymbol{\theta}\right), \end{aligned} \tag{1}$$

where τ_0, τ_1, and τ_2 are precisions and Q_1 and Q_2 specify the spatial neighbor relations smoothed by τ_1 and τ_2 respectively. Q_k, $k = 1, 2$, is $n \times n$ with off-diagonal entries $q_{k,ij} = -1$ if sites i and j are class-k neighbors and 0 otherwise, and diagonal entries $q_{k,ii}$ the number of site i's class-k neighbors.

Models are often reparameterized to improve computing or interpretation, e.g., a density with long, narrow contours can be transformed to have more circular contours. This paper proposes an alternative parameterization for variance-structure parameters, the simplex parameterization (Besag & Higdon 1999), and a slice sampler for MCMC draws in this parameterization. The simplex parameterization and its associated reference prior are then compared to other parameterizations and their reference priors. Often, the posterior for variance-structure parameters is sensitive to the prior because the data give little information about them, e.g., because of the spatial structure

(Reich et al 2007). Thus, reference priors for variance-structure parameters are an active research area (e.g., Browne & Draper 2006; Gelman 2005b).

Section 2 illustrates some problems that can arise in the posterior distributions of variance-structure parameters, motivating the simplex parameterization. Section 3 develops the new parameterization and a slice sampler for it. Section 4 uses effective sample size to compare the computing performance of MCMC algorithms arising from the simplex parameterizations and three competing parameterizations: precisions with gamma priors; standard deviations with flat priors (Gelman 2005b); and log precision ratios (defined below; Reich et al 2007) with flat priors. Our MCMC routine on the simplex parameterization generally outperforms MCMC routines on other parameterizations, even for reference priors specified on those other parameterizations. Section 5 uses simulation studies to explore statistical properties of the reference priors associated with each parameterization. Section 6 summarizes our findings. The computer code (in R) used for the simplex parameterization is available at http://www.biostat.umn.edu/~brad/software.html.

2 Problems with commonly-used parameterizations

For Bayesian analysis of multiple-precision models, several parameterizations have been proposed for the variance structure, including precisions τ_k, standard deviations $\sigma_k = \tau_k^{-1/2}$ (Gelman 2004), precision ratios $r_k = \tau_k/\tau_0$, $k = 1, 2, \ldots$, and log precision ratios $z_k = \log r_k$ (Reich et al 2004). These parameterizations are often associated with specific reference priors. For the precision parameterization, the standard "vague" prior is $\tau_k \sim Gamma(\epsilon, \epsilon)$ for $\epsilon = 0.01$ or 0.001. For the standard deviation parameterization, Gelman (2005b) proposed $\sigma_k \sim Unif(0, L)$ for a suitable upper bound L. The precision ratios, r_k, are positive and somewhat like precisions, which suggests $r_k \sim Gamma(\epsilon, \epsilon)$ as a "vague" prior. Finally, the log precision ratios take values anywhere in the real line, which suggests $z_k \sim Unif(-L, L)$ for a suitable L.

These parameterizations are all subject to problems that we illustrate using the 2NRCAR model (1) and Figure 1's data. Figure 3 suggests how the problems arise. Specifically, for each panel in Figure 3, we re-parameterized model (1) in terms of that panel's parameterization, applied the reference prior described above, derived the exact marginal posterior distribution of the smoothing parameters, and plotted its contours. For the precision ratios (r_1, r_2) and log precision ratios (z_1, z_2), Figure 3's panels c and d respectively show contours of the log marginal posterior after integrating all other parameters out of the posterior. Panels a and b show the log conditional posterior for the precisions (τ_1, τ_2) and standard deviations (σ_1, σ_2) after integrating $\boldsymbol{\theta}$ out of the posterior and fixing $\tau_0 = 1$ and $\sigma_0 = 1$, respectively. (These values of τ_0 and σ_0 are typical of those estimated in calibration studies.)

The contours for (τ_1, τ_2), (σ_1, σ_2), and (r_1, r_2) (panels a, b, and c, respectively) are L-shaped with two long arms and modes pressed tightly against one or both coordinate axes. While each plot assumes a particular reference prior, the same qualitative problems are present for other reference priors. The contours of (z_1, z_2)'s posterior are long and narrow here (panel d) but are distinctly L-shaped for other periodontal datasets (Reich et al 2007). Bimodal posteriors have been observed in the (r_1, r_2) and (z_1, z_2) parameterizations (Reich et al 2007), and indeed bimodality occurs readily even in the simplest hierarchical models (Liu & Hodges 2003).

Posterior distributions like these create predictable difficulties. First, standard MCMC approaches tend to give chains with high lagged autocorrelations and small effective sample sizes. For example, for the parameterizations in Figure 3 a, b, c, the autocorrelations at lag 10 are 0.2 to 0.4. Second, the parameters can be poorly identified, that is, either they are highly correlated *a posteriori*, or the posterior has a large flat mode indicating poor ability to distinguish between possible parameter values. Reich et al (2007) showed that for a variety of 2NRCAR spatial structures, posteriors for the precision parameters are either very flat or have pronounced ridges, inducing bad MCMC convergence and mixing (Gelfand et al 1995).

Different problems affect other aspects of Bayesian analysis. The posterior correlation between the error precision and the smoothing precisions is often high because the error precision in effect specifies the data's scale, and the data generally give much more information about this precision than about higher-level precisions. The variance, precision and standard deviation parameterizations are scale-dependent, so for example if the measurement unit is changed from centimeters to millimeters, these parameters are multiplied by 100, 0.01, and 10 respectively. This affects interpretation of hyperparameters and makes it difficult to specify a reference prior. The precision ratio and log precision ratio parameters r_k and z_k are scale-invariant, i.e., invariant if the data are multiplied by a constant, but as mentioned are prone to bimodality and highly autocorrelated MCMC draws. Sections 4.2 and 4.3 illustrate the latter point in detail. The simplex parameterization (Besag & Higdon 1999), which we now introduce, appears to avoid or mitigate these difficulties.

3 The simplex parameterization and associated methods

3.1 Definition of the simplex parameterization

For a multiple-precision model with precisions $(\tau_0, \tau_1, \cdots, \tau_m)$, define the total relative precision

$$\lambda = \sum_{k=1}^{m} r_k = \frac{1}{\tau_0} \sum_{k=1}^{m} \tau_k,$$

where $r_k = \tau_k/\tau_0$. Define the allocation of total relative precision as $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$, where

$$\beta_k = \frac{r_k}{\lambda} = \frac{r_k}{\sum_{j=1}^{m} r_j} = \frac{\tau_k}{\sum_{j=1}^{m} \tau_j};$$

$\sum_{k=1}^{m} \beta_k = 1$, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)$ takes values in the m-dimensional simplex. The 2, 3, and 4-dimensional simplices are a line segment, equilateral triangle, and tetrahedron, respectively.

This parameterization has two *a priori* attractive features. First, it is scale-invariant, that is, it does not change when the data are multiplied by a constant. Also, the simplex parameter $\boldsymbol{\beta}$ lies in a bounded space, so a natural reference prior, the flat prior, is proper and exchangeable. The rest of this paper uses a flat prior on $\boldsymbol{\beta}$ and gamma priors on λ and τ_0.

3.2 Computing strategy for the simplex parameterization

For a multiple-precision model like (1), the vector of unknown parameters is $(\boldsymbol{\theta}, \tau_0, \lambda, \boldsymbol{\beta})$, where $\boldsymbol{\theta}$ is the mean-structure parameters, τ_0 the error precision, λ the total relative precision, and $\boldsymbol{\beta}$ the allocation of total relative precision. To avoid MCMC sampling variation, we analytically integrate $\boldsymbol{\theta}$ and τ_0 out of the joint posterior and run a slice sampler on the marginal posterior of $(\lambda, \boldsymbol{\beta})$. Posterior summaries for $\boldsymbol{\theta}$ and τ_0 are then obtained by Rao-Blackwellizing.

Suppose the precision parameters in the 2NRCAR model (1) have prior $p(\tau_0, \tau_1, \tau_2)$. Then the joint posterior of all the unknowns is

$$\begin{aligned} p(\boldsymbol{\theta}, \tau_0, \tau_1, \tau_2 | \boldsymbol{y}) &\propto p(\tau_0, \tau_1, \tau_2) p(\boldsymbol{y}|\boldsymbol{\theta}, \tau_0) p(\boldsymbol{\theta}|\tau_1, \tau_2) \\ &\propto p(\tau_0, \tau_1, \tau_2) \tau_0^{n/2} \exp\left(-\frac{\tau_0}{2} \sum (y_i - \theta_i)^2\right) \\ &\quad \times \prod_{j=1}^{n-G} (\tau_1 d_{1j} + \tau_2 d_{2j})^{1/2} \exp\left(-\frac{1}{2} \boldsymbol{\theta}' (\tau_1 Q_1 + \tau_2 Q_2) \boldsymbol{\theta}\right), \end{aligned} \quad (2)$$

where G is the number of islands in the spatial map and d_{kj} is defined as follows. Simultaneously diagonalize the two positive semi-definite matrices Q_k as $B' D_k B$, where B is nonsingular (Newcomb 1961), and let D_k have j^{th} diagonal element d_{kj}. It is easy to see $\boldsymbol{\theta}|\boldsymbol{y}, \tau_0, r_1, r_2 \sim N((Q_r + I_n)^{-1} X' \boldsymbol{y}, \tau_0 (Q_r + I_n))$, where $Q_r = r_1 Q_1 + r_2 Q_2$ and $r_k = \tau_k / \tau_0$. After integrating out $\boldsymbol{\theta}$,

$$\begin{aligned} p(\tau_0, r_1, r_2 | \boldsymbol{y}) &\propto p(\tau_0, r_1, r_2) \tau_0^{\frac{n-G}{2}} |Q_r + I_n|^{-\frac{1}{2}} \prod_{j=1}^{n-G} (r_1 d_{1j} + r_2 d_{2j})^{1/2} \\ &\quad \times \exp\left(-\frac{\tau_0}{2} [\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}' (Q_r + I_n)^{-1} \boldsymbol{y}]\right). \end{aligned}$$

Then if τ_0's prior is $Gamma(a_0, b_0)$, with mean $\frac{a_0}{b_0}$, integrate out τ_0 to give

$$p(r_1, r_2 | \boldsymbol{y}) \propto p(r_1, r_2) \prod_{j=1}^{n-G} (r_1 d_{1j} + r_2 d_{2j})^{1/2} |Q_r + I_n|^{-\frac{1}{2}} R^{-b},$$

where $R = b_0 + \frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'(Q_r + I_n)^{-1}\boldsymbol{y}\right]$, and $b = a_0 + \frac{n-G}{2}$. Now change to the simplex parameterization $\lambda = r_1 + r_2$ and $\beta = r_1 / \lambda$, giving

$$p(\lambda, \beta | \boldsymbol{y}) \propto p(\lambda, \beta) \lambda^{\frac{n-G}{2}} |I + \lambda Q_\beta|^{-\frac{1}{2}} \prod_{j=1}^{n-G} (\beta(d_{1j} - d_{2j}) + d_{2j})^{\frac{1}{2}} R^{-b},$$

where $R = b_0 + \frac{1}{2}(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'(I + \lambda Q_\beta)^{-1}\boldsymbol{y})$, $Q_\beta = \beta Q_1 + (1-\beta)Q_2$, and the Jacobian is implicit in the change of variables in the prior, from $p(r_1, r_2)$ to $p(\lambda, \beta)$. B is orthogonal if and only if Q_1Q_2 is symmetric, in which case

$$p(\lambda, \beta|\boldsymbol{y}) \quad \propto \quad p(\lambda, \beta) \prod_{j=1}^{n-G} \left(\frac{\lambda\gamma_j}{1+\lambda\gamma_j} \right)^{\frac{1}{2}} \left[b_0 + \frac{1}{2} \left(\sum_j \frac{\lambda\gamma_j}{1+\lambda\gamma_j} \boldsymbol{y}_j^{*2} \right) \right]^{-b}, \tag{3}$$

where $\boldsymbol{y}^* = B\boldsymbol{y}$ and $\gamma_j = \beta(d_{1j} - d_{2j}) + d_{2j}$, so (3) depends on λ and γ_j only through $\frac{\lambda\gamma_j}{1+\lambda\gamma_j}$.

For this problem, we propose a slice sampler with one auxiliary variable. A slice sampler can be more efficient than an ordinary Metropolis-Hastings algorithm, e.g., Neal (1997, 2003), Tierney & Mira (1999). Generally, the slice sampler can be described as follows (Damien et al 1999). Suppose an MCMC has stationary distribution $\pi(\lambda, \boldsymbol{\beta}) \propto p(\lambda, \boldsymbol{\beta})l(\lambda, \boldsymbol{\beta})$. Introduce an auxiliary random variable U with a conditional uniform distribution $U|\beta, \lambda \sim \text{Unif}(0, l(\lambda, \boldsymbol{\beta}))$. Then $(\lambda, \boldsymbol{\beta}, U)$ has joint distribution

$$f(\lambda, \boldsymbol{\beta}, u) \propto p(\lambda, \boldsymbol{\beta}) I_{\{u < l(\lambda, \boldsymbol{\beta})\}}(\lambda, \boldsymbol{\beta}, u).$$

The slice sampler is then a special case of the Gibbs sampler:

1. Initialize $\boldsymbol{\beta}^{(0)}, \lambda^{(0)}$;

2. Generate $U|(\lambda, \boldsymbol{\beta})$ from a uniform distribution:
   $U^t|(\lambda^{t-1}, \boldsymbol{\beta}^{t-1}) \propto \text{Unif}(0, l(\lambda^{t-1}, \boldsymbol{\beta}^{t-1}))$.

3. Generate $\boldsymbol{\beta}|(u, \lambda)$ from $p(\lambda, \boldsymbol{\beta})$ restricted to $l(\lambda, \boldsymbol{\beta}) > u$:
   $\boldsymbol{\beta}^t|(\lambda^{t-1}, U^t) \propto p(\lambda, \boldsymbol{\beta}) I(l(\lambda^{t-1}, \boldsymbol{\beta}) > U^t)$.

4. Generate $\lambda|(u, \boldsymbol{\beta})$ from $p(\lambda, \beta)$ restricted to $l(\lambda, \boldsymbol{\beta}) > u$:
   $\lambda^t|(\boldsymbol{\beta}^t, U^t) \propto p(\lambda, \boldsymbol{\beta}) I(l(\lambda, \boldsymbol{\beta}^t) > U^t)$.

Repeat steps 2-4; after convergence, (β^t, λ^t) are samples from the stationary distribution $\pi(\lambda, \boldsymbol{\beta})$.

A natural $p(\lambda, \boldsymbol{\beta})$ is $p(\lambda, \boldsymbol{\beta}) = p_1(\lambda)p_2(\boldsymbol{\beta})$, where p_1 is a gamma density and p_2 is uniform on the simplex, a special case of the Dirichlet distribution. With this choice, candidate β_j can be generated as $X_j / \sum_{j=1}^m X_j$, where $X_1, \cdots, X_m$ are independent exponential variates. An informative prior for $\boldsymbol{\beta}$ can be $Dirichlet(\alpha_1, \cdots, \alpha_m)$, from which samples can also be generated using draws from gamma distributions. For the 2NRCAR model, $l(\lambda, \boldsymbol{\beta}) = \lambda^{\frac{n-G}{2}} \prod_{j=1}^{n-G} (\beta d_{1j} + (1-\beta)d_{2j})^{\frac{1}{2}} |I + \lambda Q_\beta|^{-\frac{1}{2}} R^{-b}$ and $p(\lambda, \beta) = \frac{1}{\Gamma(a_\lambda)} \lambda^{a_\lambda - 1} e^{-b_\lambda \lambda} I(\beta \in [0, 1])$.

The posterior distributions of $\boldsymbol{\theta}$ and τ_0 can be estimated by Rao-Blackwellizing (Casella & Robert 1996). For posterior samples $(\lambda^t, \boldsymbol{\beta}^t)$, $t = 1, 2, \cdots, M$, $\boldsymbol{\theta}$'s posterior density can be estimated as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\lambda, \boldsymbol{\beta}, \boldsymbol{y}) p(\lambda, \boldsymbol{\beta}|\boldsymbol{y}) d\lambda d\boldsymbol{\beta} \approx \frac{1}{M} \sum_{t=1}^{M} p(\boldsymbol{\theta}|\lambda^t, \boldsymbol{\beta}^t, \boldsymbol{y}), \tag{4}$$

where $p(\boldsymbol{\theta}|\lambda^t, \boldsymbol{\beta}^t, \boldsymbol{y})$ is $\boldsymbol{\theta}$'s conditional posterior given $(\lambda^t, \boldsymbol{\beta}^t)$. For the normal-error model (1), $\boldsymbol{\theta}|\lambda, \boldsymbol{\beta}, \boldsymbol{y}$ is multivariate-t with center $(P^t)^{-1}\boldsymbol{y}$, scale $(P^t)^{-1}R^t/b$ and $2b$ degrees of freedom, where $R^t = b_0 + \frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'(P^t)^{-1}\boldsymbol{y}\right]$, and $P^t = \lambda^t B'(\beta^t D_1 + (1-\beta^t)D_2)B + I_n$. Thus $\boldsymbol{\theta}$'s posterior mean and variance are estimated by

$$\begin{aligned}
E(\boldsymbol{\theta}|\boldsymbol{y}) &= E(E(\boldsymbol{\theta}|\lambda, \boldsymbol{\beta}, \boldsymbol{y})) \approx \frac{1}{M}\sum_{t=1}^{M} E(\boldsymbol{\theta}|\lambda^t, \boldsymbol{\beta}^t, \boldsymbol{y}) = \frac{1}{M}\sum_{t=1}^{M}\mu^t_{\boldsymbol{\theta}} = \bar{\mu}_{\boldsymbol{\theta}} \\
Var(\boldsymbol{\theta}|\boldsymbol{y}) &= E(Var(\boldsymbol{\theta}|\boldsymbol{y}, \lambda, \boldsymbol{\beta})) + Var(E(\boldsymbol{\theta}|\boldsymbol{y}, \lambda, \boldsymbol{\beta})) \\
&\approx \frac{1}{M}\left[\sum_{t=1}^{M}\Sigma^t_{\boldsymbol{\theta}} + \sum_{t=1}^{M}(\mu^t_{\boldsymbol{\theta}} - \bar{\mu}_{\boldsymbol{\theta}})(\mu^t_{\boldsymbol{\theta}} - \bar{\mu}_{\boldsymbol{\theta}})'\right], \qquad (5)
\end{aligned}$$

where $\mu^t_{\boldsymbol{\theta}}$ and $\Sigma^t_{\boldsymbol{\theta}}$ are the posterior mean and variance of $p(\boldsymbol{\theta}|\lambda^t, \boldsymbol{\beta}^t, \boldsymbol{y})$, respectively. Similarly, $\tau_0|\lambda^t, \beta^t, \boldsymbol{y}$ is gamma distributed with shape b and rate R^t, so posterior summaries for τ_0 can be obtained analogously.

4 MCMC algorithm performance in the different parameterizations

4.1 Effective sample size (ESS)

Effective sample size (ESS) is commonly used to assess MCMC mixing (e.g., Carlin & Louis 2000, Chapter 5; Sargent et al 2000; Chen et al 2000; Ridgeway & Madigan 2003). The ESS of a sampled quantity is defined (Kass et al 1998) as

$$ESS = \frac{M}{1 + 2\sum_{l=1}^{\infty}\rho_l}, \qquad (6)$$

where M is the number of MCMC samples for that quantity and ρ_l is the estimated lag l autocorrelation of the samples. ESS can be interpreted as the size of an independent, identically distributed sample giving information equivalent to the autocorrelated MCMC sample. In practice ρ_l is estimated with error, and past a certain l the $\hat{\rho}_l$ are dominated by noise (Gilks et al 1996; Chapter 3). To avoid summing noise, Geyer (1992) proposed the initial convex sequence estimator, which requires a sequence of empirical Γ_m estimates that are positive, monotone, and convex, where Γ_m is the sum of two lagged autocovariances γ_{2m} and γ_{2m+1}. The natural estimator of the lagged autocovariance is the empirical autocovariance $\hat{\gamma}_l = \frac{1}{M}\sum_{t=1}^{M-l}(X_t - \bar{X})(X_{t+l} - \bar{X})$, where $\{X_t\}$ is the sequence of MCMC samples. Priestley (1981, p. 323) suggests using this "biased" estimate with divisor M rather than the "unbiased" estimate with divisor $M - l$. Define m^* as the largest integer such that $\hat{\Gamma}_m$ is a positive, monotonely decreasing, and convex sequence in m. Then the ESS in (6) sums only estimated autocorrelations $\hat{\rho}_l$ for $l \leq 2m^*$.

4.2 Periodontal data analyzed using 2NRCAR

This section compares MCMC algorithms specified in each of four parameterizations, for the 2NRCAR model applied to Figure 1's data. For each parameterization, the data were analyzed three times, using three different prior distributions, each a reference prior for one of the parameterizations. This is an unusual simulation study design; the point is that one may prefer inferences using a reference prior specified on one parameterization, while it is advantageous to specify the MCMC algorithm on a different parameterization.

The four parameterizations are simplex, log precision ratios (z_1, z_2), precisions (τ_0, τ_1, τ_2), and standard deviations $(\sigma_0, \sigma_1, \sigma_2)$. The three reference priors are as follows: for the simplex parameterization, we put a Gamma$(0.01, 0.01)$ prior on λ, and on β, a uniform distribution on the unit interval; for the parameterization with three precisions, we gave each precision a Gamma$(0.01, 0.01)$ prior; and for the parameterization with three standard deviations, we gave each standard deviation a uniform prior on the interval $(0, 10)$. For each parameterization, for each prior, 10000 MCMC draws were made with 5000 discarded for burn-in. Table 1 describes the MCMC algorithm for each parameterization. Except for the simplex parameterization, the algorithms were Metropolis-Hastings with normal candidate draws for the working parameters (Table 1), centered on the current draw. For each working parameter, the sample standard deviation of the 5000 burn-in draws was used as the standard deviation of the candidate draws in the subsequent 5000 retained iterations. A dynamic search procedure (see the Appendix) was used to accelerate the slice sampler.

Table 2 shows effective sample size (ESS) for the four parameterizations and three priors. The simplex parameterization has the largest ESS for two priors, and roughly the same ESS as (z_1, z_2) for the flat prior on $(\sigma_0, \sigma_1, \sigma_2)$. The simplex parameterization's sample autocorrelations decrease quickly as lag increases and generally vanish by lag 10, while the alternatives have much larger autocorrelations at all lags (data not shown). As currently programmed, the simplex parameterization's slice sampler usually runs more slowly than the other algorithms, so it has a smaller advantage in ESS per second of run time (Table 3), and is roughly tied with the log precision ratio parameterization (z_1, z_2).

Section 2 suggested that the simplex parameters (λ, β) might have smaller posterior correlations with the error precision τ_0, compared to other parameterizations' smoothing parameters. This was true for the present dataset, with the prior distribution having little effect. For each parameterization, we report the posterior correlation only for the parameter having the largest absolute correlation. In the simplex parameterization, β had the largest absolute posterior correlation with τ_0, about 0.33 for all three priors. The analogous results for the other three parameterizations were: (z_1, z_2), 0.53 for z_1; precisions, 0.53 for τ_1; and standard deviations, 0.76 for σ_1. Contrary to our expectation, (z_1, z_2) — which, like the simplex parameterization, is invariant when the data are multiplied by a constant — gave the same maximum absolute posterior correlations as did the precision parameterization.

Figure 4 shows a contour plot of the log marginal posterior arising from the simplex parameterization and its reference prior. While this is not especially like a bivariate normal density, it does seem rather less irregular than the analogous contour plots for the other parameterizations (Figure 3).

4.3 Smoothed ANOVA (SANOVA) model

The smoothed ANOVA model used here was introduced by Sargent & Hodges (1997) and fully developed in Hodges et al (2007; see also Smith 1973, Gelman 2005a). Suppose the experimental design has one error term, c design cells, and n replications per cell. Parameterize each effect so the design matrix has orthogonal columns. Group the L columns for main effects, including the intercept, into a matrix A_1, and the N columns for interactions into a matrix A_2, and scale A_1 and A_2 so $A_1'A_1 = cnI_L$ and $A_2'A_2 = cnI_N$; $A_1'A_2 = 0$. The SANOVA model is

$$\boldsymbol{y} \quad = \quad A_1\boldsymbol{\Theta}_1 + A_2\boldsymbol{\Theta}_2 + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{y}$ is the cn-vector of observed outcomes, $\boldsymbol{\epsilon} \sim N(0, \Gamma_1)$, the grand mean and main effects in $\boldsymbol{\Theta}_1$ have an improper flat prior, the interactions in $\boldsymbol{\Theta}_2$ have a $N(0, \Gamma_2)$ prior, $\boldsymbol{\epsilon}$ and $[\boldsymbol{\Theta}_1|\boldsymbol{\Theta}_2]$ are independent *a priori*, and the two covariance matrices Γ_1 and Γ_2 are specified as $\Gamma_1 = \frac{1}{\tau_0}I_{cn}$ and $\Gamma_2^{-1} = diag(\phi_1, \cdots, \phi_N)$. For a set of distinct smoothing precisions $(\tau_1, \cdots, \tau_s)$, $s \leq N$, define a deterministic assignment function $j(k)$ that specifies groups of ϕ_k within which $\phi_k = \tau_{j(k)}$, and let n_j be the number of ϕ_k mapping to τ_j. The joint posterior after integrating out $\boldsymbol{\Theta}$ is

$$f(\tau_0, \boldsymbol{r}|\boldsymbol{Y}) \quad \propto \quad \pi(\tau_0, \boldsymbol{r})\tau_0^{\frac{cn-L}{2}} \exp(-\frac{1}{2}\tau_0 W(\boldsymbol{r})) \prod_{j=1}^{s} \left(\frac{r_j}{r_j + cn}\right)^{n_j/2}, \tag{8}$$

where $r_j = \frac{\tau_j}{\tau_0}$ and $W(\boldsymbol{r}) = \boldsymbol{y}'\boldsymbol{y} - \frac{1}{cn}\boldsymbol{y}'A_1A_1'\boldsymbol{y} - \boldsymbol{y}'A_2 diag((cn + r_{j(k)})^{-1})A_2'\boldsymbol{y}$.

This model has s smoothing precisions $\tau_1, \cdots, \tau_s$, so the simplex parameter $\boldsymbol{\beta}$ is s-dimensional. If τ_0 has a gamma prior $G(a_0, b_0)$, with mean $\frac{a_0}{b_0}$, then τ_0's full conditional posterior is also gamma. After integrating out τ_0, $(\lambda, \boldsymbol{\beta})$ has marginal posterior

$$f(\lambda, \boldsymbol{\beta}|\boldsymbol{Y}) \quad \propto \quad \pi(\lambda, \boldsymbol{\beta}) \prod_{j=1}^{s} \left[1 + \frac{cn}{\lambda\beta_j}\right]^{-n_j/2} R^{-b}, \tag{9}$$

where $R = b_0 + \frac{1}{2}\boldsymbol{y}'\boldsymbol{y} - \frac{1}{2cn}\boldsymbol{y}'A_1A_1'\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}'A_2 diag((cn + \lambda\beta_{j(k)})^{-1})A_2'\boldsymbol{y}$ and $b = a_0 + \frac{cn-L}{2}$.

Hodges & Sargent (2001, Section 6) applied smoothed ANOVA to a 2^3 factorial experiment testing a material's tensile strength (Lai & Hodges 1999). The three design factors were the type of mold, presence of pigment, and type of cure, with $n = 6$ replications per cell. The dataset is in Hodges & Sargent (2001). We used this dataset to compare MCMC routines for different parameterizations and priors, as in Section 4.2's comparison for the 2NRCAR model, and using the same priors as in Section 4.2. For all three priors, the MCMC on the simplex parameterization has by far the largest ESS (Table 4) and the smallest autocorrelations (data not shown). The MCMC on the simplex parameterization also has the largest ESS/sec for two of the three priors (Table 5). Overall, the smoothed ANOVA results are consistent with the 2NRCAR results.

5 Statistical performance of each parameterization's reference prior

5.1 2NRCAR model

To reduce computing time, we simulated periodontal measurements on upper and lower jaws with 5 teeth each, for 60 total measurements in one "patient". The two neighbor classes are as in Figure 2. This simulation experiment's design considered three factors: (1) true error precision τ_0; (2) the true degree of smoothness in the two classes of neighbor pairs, (τ_1, τ_2); and (3) the 4 parameterizations, each with its associated reference prior (Table 6). Table 7 gives the specific true values of (τ_0, τ_1, τ_2).

For each design cell, the 1000 simulated datasets were drawn as follows. By the spectral decomposition, $\tau_1 Q_1 + \tau_2 Q_2 = \Gamma' \Lambda \Gamma$, where Γ is an orthogonal matrix and Λ is diagonal. Then $\boldsymbol{\theta}^* = \Gamma \boldsymbol{\theta}$ has density

$$p(\boldsymbol{\theta}^*) \propto \exp(-\frac{1}{2}\boldsymbol{\theta}^{*\prime}\Lambda\boldsymbol{\theta}^*) = \exp(-\frac{1}{2}\boldsymbol{\theta}^{*\prime}_{n-G}\Lambda_{n-G}\boldsymbol{\theta}^*_{n-G})$$

where the subscript $n - G$ indicates the first $n - G$ rows and/or columns. Thus, the first $n - G$ elements of $\boldsymbol{\theta}^*$ were drawn from independent normal distributions, for $G = 2$ islands in the "mouth". The last 2 elements of $\boldsymbol{\theta}^*$ have flat priors under $p(\boldsymbol{\theta}^*)$ and were drawn from a uniform on $(-10, 10)$. Then the sample of true $\boldsymbol{\theta}$ were obtained as $\boldsymbol{\theta} = \Gamma' \boldsymbol{\theta}^*$.

For the simplex and log precision ratio (Z) parameterizations, MCMC samples were drawn from the marginal posterior after integrating out $\boldsymbol{\theta}$ and τ_0, and the posterior mean and interval coverage were estimated by Rao-Blackwellizing. For the precision and SD parameterizations, MCMC samples were drawn from the marginal posterior after integrating out only $\boldsymbol{\theta}$. For the simplex parameterization, we used the slice sampler (Section 3.2) with starting values $\beta_k = \frac{1}{s}$, where s is the number of smoothing precisions, and for the other parameterizations we used adaptive Metropolis algorithms as described in Section 4. Trace plots were checked for a sample of artificial datasets and in all cases indicated sampler convergence.

The parameterization/reference prior combinations (henceforth, "methods") were compared according to their results on the standard deviation scale, i.e., $\sigma_k = 1/\sqrt{\tau_k}$, the same scale as the data, using bias and MSE of posterior means as point estimates, and coverage of equal-tailed 95% credible intervals. (The Appendix gives equations for Rao-Blackwellizing the Z and simplex parameters in the standard deviation scale.) To remove effects that obscure comparisons, we report bias as a percent of the true value and we scale MSE according to the true error variance.

Figure 5 displays scaled bias, scaled MSE, and 95% interval coverage for the four methods. All methods have small biases for the error standard deviation σ_0 except the Z method in case 3, where the posterior mean overestimates σ_0 by about 30%. By contrast, the Z method consistently underestimates σ_1, while the other methods have small biases. For σ_2, all methods have larger bias and the SD method performs worst, overestimating substantially in all cases. For all methods and cases, the MSEs for σ_0 and σ_1 are small. The Z method has the largest MSE for σ_1. For σ_2, all methods' MSEs vary a lot, but the simplex method consistently gives the smallest MSE and the SD method the largest. Finally, all methods give coverage close to 95% for σ_0 and σ_1 except for Z, which gives low coverage. For σ_2, the precision and simplex methods give coverage 95% or higher for all cases, while the Z and SD methods had quite low coverage for some cases.

5.2 SANOVA model

This simulation experiment used artificial data from a 2^3 design with $n = 6$ replications per cell, as in Hodges et al's (2007, section 3) simulation study. The three design factors were: (1) the true error precision τ_0 (note that increasing n and τ_0 have the same effect); (2) the number of truly present interactions (1 or 3); and (3) the four parameterizations with associated reference priors, described in Table 6. Two further cases were simulated to examine the effect of multiplying the data by a constant. Table 7 gives the design values for the 8 cases considered.

We again generated 1000 simulated datasets for each "case". The design matrix for the 2^3 mean structure was orthogonal, so without loss of generality the true grand mean and main effects $\theta_1, \theta_2, \theta_3, \theta_4$ were set to zero. If an interaction term was present, its θ_k was set to 1, otherwise to zero. The interaction terms were *a priori* exchangeable and each was smoothed by its own smoothing precision, so as in Hodges et al. (2007, Section 3), we need only consider how many interactions are truly present, not which ones.

The four methods were compared according to their performance for three groups of parameters: the four interaction $\theta_k,\ k = 5, \cdots, 8$; the error precision τ_0; and the eight cell means $c_j,\ j = 1, \cdots, 8$. For each group of parameters, the methods were compared according to bias and MSE of posterior means as point estimates, and coverage probability of the 95% equal-tail credible interval, with one exception: cell-mean bias is a simple linear function of bias of the interaction θ_k and is thus omitted. By design, all methods give identical bias and MSE for the grand mean and main effects, so they are not considered further. We follow Hodges et al (2007) in calling truly present interactions "target interactions" and truly absent interactions "null interactions". By the simulation design's exchangeability, all target interactions have the same true bias, MSE, and coverage for a given method, as do all null interactions, so we present average bias and MSE for the targets and for the nulls. For the interactions θ_k and cell means c_j, we scaled bias and MSE as percents of the true error standard deviation $\frac{1}{\sqrt{\tau_0}}$ and the true error variance $\frac{1}{\tau_0}$, respectively. Similarly, for the estimates of the error precision τ_0, we report bias and square root of MSE as percents of τ_0.

Figure 6 displays the bias and MSE of posterior mean estimates of the interaction θ_k, and coverage of their 95% posterior intervals. For the target interactions, the number of truly present interactions has little effect on bias or MSE. Compared to the simplex method, the SD method has smaller bias (Figure 6a). In general, the SD method performs better than the precision method, which in turn performs better than the Z method. For the null interactions, all methods are essentially unbiased and the Z method has the smallest MSE (Figure 6b). As for 95% posterior intervals (Figure 6c,d), for the target interactions, the simplex and SD methods give coverage much closer to the nominal 95% than the Z and precision methods, which are too low for cases with small error precision. For the null interactions, the simplex and SD methods have about 95% coverage while coverage for the other two methods is too high. Broadly speaking, for the interaction θ_k, the simplex method gives good performance that improves relative to the other methods as the error precision decreases.

Figure 7 shows scaled bias and MSE for the error precision τ_0 (panels a,b), and MSE and coverage probability for the cell means (panels c,d). For τ_0, the SD method outperforms the others in both bias and MSE (Figure 7a,b). The 95% CI coverage is close to the nominal 95% for all methods and cases (data not shown). For the cell means, Figure 7c,d show the scaled MSE (as a percent of $\frac{1}{\tau_0}$) and 95% interval coverage averaged over the 8 cells. The simplex and SD methods perform similarly. When 1 target interaction is present, these methods have higher bias than the other two, but when 3 target interactions are present, they have smaller bias. Coverage of 95% credible intervals is close to the nominal 95% for all methods, except for the Z method for small error precisions when 3 target interactions are present.

5.3 Crossed random effect model

The crossed random effect model (10) has error precision τ_0 and two smoothing precisions τ_1 and τ_2 for rows and columns respectively in the two-way layout, as follows:

$$y_{ijk} \quad = \quad \mu + \alpha_i + \gamma_j + \epsilon_{ijk} \;\; i = 1, \cdots, I; \; j = 1, \cdots, J; \; k = 1, \cdots, K, \qquad (10)$$

where $\alpha_i \sim N(0, \tau_1)$, $\gamma_j \sim N(0, \tau_2)$, and $\epsilon_{ijk} \sim N(0, \tau_0)$ for unknown τ_0, τ_1, τ_2. This simulation experiment's design had three factors: (1) the true error precision τ_0; (2) the true τ_1 and τ_2, considering equal and unequal smoothness in rows and columns; and (3) the four parameterizations with their reference priors, described in Table 6.

Each of the 1000 artificial datasets per simulation design cell had 5 row levels (α_i, $i = 1, \cdots, 5$), 5 column levels (γ_j, $j = 1, \cdots, 5$), and 5 replicates (ϵ_{ijk}, $k = 1, \cdots, 5$). Without loss of generality, the grand mean μ was set to zero. We generated artificial datasets as follows: Generate row effects $\alpha_1, \cdots, \alpha_5$, column effects $\gamma_1, \cdots, \gamma_5$, then in each of the 25 cells, add 5 random normal errors to give 125 total observations. The algorithms and outcome measures in this simulation study are the same as for the 2NRCAR simulation study (Section 5.1).

Figure 8 shows bias and MSE of posterior means as point estimates and 95% credible interval coverage, for the three standard deviations σ_0, σ_1, and σ_2. For the error standard deviation σ_0, all methods are essentially unbiased and have small MSE. However, bias is complex for the two smoothing standard deviations σ_1 and σ_2. The simplex method has much smaller bias than the SD method for most cases (Figure 8a), but otherwise it is difficult to generalize. For MSE (Figure 8b), the simplex method is lower than the alternatives except for cases 3 and 6 for σ_2. For coverage of 95% intervals (Figure 8c), all methods are consistently close to the nominal 95% for σ_0. For σ_1 and σ_2, the simplex, precision, and SD methods perform similarly and fairly well, while the Z method performs worse, particularly for σ_2.

6 Discussion

We have developed a parameterization for multiple-precision models, first mentioned for 2NRCAR by Besag & Higdon (1999). Based on Sections 4 & 5, the simplex parameterization appears to have two advantages. First, it gives simple MCMC algorithms with good mixing properties for various reference priors. Thus Bayesian analyses may benefit from this parameterization even for priors specified in another parameterization. Second, $\boldsymbol{\beta}$ has a proper natural reference prior that is invariant when the data are multiplied by a constant; λ has the same invariance property. Section 5 showed that compared to other proposed reference priors, this prior yields posterior means with generally good bias and mean squared error, and 95% credible intervals with close to nominal coverage, for the range of cases considered. Its worst performance was for smoothed ANOVA in Section 5.2. If one were designing a software package solely to do smoothed ANOVA, these results suggest that the simplex parameterization — with the reference prior used here — might not be the best choice for a prior distribution. However, if one were seeking an all- purpose off-the-shelf prior, these results are not so discouraging: while the simplex parameterization was not the best prior for smoothed ANOVA, it did not lose badly to the other priors, while each of the other priors did perform poorly for at least one example.

The obvious question is: can we improve the statistical performance of the simplex parameterization? The first consideration in this vein is the reference prior. The allocation parameter $\boldsymbol{\beta}$ has a natural reference prior, but the total relative precision λ does not. Sections 4 & 5 used the conventional “vague” Gamma(0.01,0.01) prior, which, with 50th and 90th percentiles 4×20^{-29} and 0.0015 respectively, is in fact quite informative. Other priors for λ may improve statistical or computing performance, though we do not yet have a firm basis for proposing an alternative. One simple alternative would be a log-normal prior. In preliminary results from a simulation study of smoothed ANOVA, giving λ a lognormal prior with a large variance seems to improve coverage of posterior 95% intervals compared to the gamma prior considered here, but otherwise the operating characteristics are similar.

It seems pertinent that λ is unitless or, put another way, that λ has the same scale for all problems. Thus, for the smoothed ANOVA and crossed random-effects models, it should be possible to determine universally-applicable large and small values of λ, and perhaps use that information to specify, say, a uniform prior for λ. The 2NRCAR example is more complicated in a manner that is beyond the present paper’s scope, but it might be possible to extend this general idea.

Some literature on priors for hierarchical models (e.g., Daniels 1999; Gustafson et. al. 2007) suggests that a prior may be judged by the relative weight it gives to information arising from the data (governed by the error precision τ_0) and information arising from the model (governed by the smoothing parameters τ_k). One way to implement this idea is to consider, in our notation, $\tau_k / \sum_{j=0}^{s} \tau_j$ for $k = 0, \ldots, s$. The simplex parameterization lends itself readily to this suggestion. The error precision's fraction of total precision is easily shown to be $1/(1+\lambda)$, which is readily computed in the context of MCMC. As for the smoothing precisions τ_k, $k = 1, \ldots, s$, their aggregate fraction of total precision is $\lambda/(1+\lambda)$, and τ_k's fraction of total precision is $\beta_k \lambda/(1+\lambda)$, also easily computed using MCMC. A flat prior on $\boldsymbol{\beta}$ treats $\tau_k, k = 1, \ldots, s$, exchangeably; priors on λ might be compared according to how they weigh τ_0 against individual τ_k or the ensemble of τ_ks.

The simplex parameterization extends straightforwardly in two ways. First, it extends immediately if any of the models presented here is extended by adding one or more random effects parameterized by variances or precisions. For example, the 2NR-CAR model (1) can be extended to a spatio-temporal model for multiple dental visits by adding a third class of neighbor pairs representing two consecutive observations at a given measurement site. This adds a third smoothing precision, which can be handled in the obvious manner. A second extension is for models with many smoothing precisions that naturally fall into, say, two groups. In such a model, a separate simplex parameter pair (λ, β) can be used for each of the groups of smoothing precisions.

Although the simplex parameterization is applicable to a broad class of models (Section 1), extension to models with covariance matrices would be desirable. The approach of Barnard et al (2000), in which the covariance matrix is decomposed into standard deviations and correlations, is one possible extension, where the simplex parameterization would be applied to the vector of standard deviations, after standardizing the regressors to put them all on the same scale.

Appendix

6.1 Rao-Blackwellizing on the standard deviation scale

In Section 5, the four parameterizations with their associated priors were compared according to point-estimate and interval-coverage performance on the standard deviation scale, with Rao-Blackwellizing done as follows. Suppose $\tau_0 | \lambda, \beta, \boldsymbol{y} \sim Gamma(b, R)$, then $p(\tau_0|\boldsymbol{y}) \approx \frac{1}{M} \sum_{t=1}^{M} Gamma(\tau_0|b^t, R^t)$. Changing variables to $\sigma_0 = \tau_0^{-1/2}$ and including the Jacobian, $p(\sigma_0|\boldsymbol{y}) \approx \frac{1}{M}\sum_{t=1}^{M} 2\sigma_0^{-3} Gamma(\sigma_0^{-2}|b^t, R^t)$, so

$$\begin{aligned} E(\sigma_0|\boldsymbol{y}) &\approx \frac{1}{M}\sum_{t=1}^{M}\int 2\sigma_0^{-2} Gamma(\sigma_0^{-2}|b^t, R^t)d\sigma_0 \\ &= \frac{1}{M}\sum_{t=1}^{M}\int \tau_0^{-\frac{1}{2}} Gamma(\tau_0|b^t, R^t)d\tau_0 \\ &= \frac{1}{M}\sum_{t=1}^{M} E(\tau_0^{-\frac{1}{2}}|b^t, R^t) = \frac{1}{M}\sum_{t=1}^{M}\frac{\Gamma(b^t-\frac{1}{2})}{\Gamma(b^t)}(R^t)^{\frac{1}{2}} \end{aligned}$$

Similarly, noting that $\sigma_1 = r_1^{-\frac{1}{2}}\tau_0^{-\frac{1}{2}}$ and $\sigma_2 = r_2^{-\frac{1}{2}}\tau_0^{-\frac{1}{2}}$,

$$\begin{aligned} E(\sigma_1|\lambda,\beta,\boldsymbol{y}) &= r_1^{-\frac{1}{2}} E(\tau_0^{-\frac{1}{2}}|\lambda,\beta,\boldsymbol{y}) = r_1^{-\frac{1}{2}}\frac{\Gamma(b-\frac{1}{2})}{\Gamma(b)}(R)^{\frac{1}{2}} \\ E(\sigma_1|\boldsymbol{y}) &\approx \frac{1}{M}\sum_{t=1}^{M}(r_1^t)^{-\frac{1}{2}}\frac{\Gamma(b^t-\frac{1}{2})}{\Gamma(b^t)}(R^t)^{\frac{1}{2}} \\ E(\sigma_2|\boldsymbol{y}) &\approx \frac{1}{M}\sum_{t=1}^{M}(r_2^t)^{-\frac{1}{2}}\frac{\Gamma(b^t-\frac{1}{2})}{\Gamma(b^t)}(R^t)^{\frac{1}{2}} \qquad (11) \end{aligned}$$

6.2 Dynamic search for the slice sampler

In the simplex parameterization's slice sampler (Section 3.2), to accept one sample, generally a large number of samples need to be drawn from $p(\lambda,\beta)$. The slice sampler can be accelerated by improving this acceptance rate. The following dynamic search is one approach for a low-dimensional parameter space; we show it for a scalar β.

1. Choose grid points for λ,β by a preliminary analysis, say, $\lambda^1 < \cdots < \lambda^\Omega$ and $\beta^1 < \cdots < \beta^\Pi$.

2. Calculate $l_{ij} = l(\lambda^i,\beta^j|\boldsymbol{y})$ at these grid points (λ^i,β^j).

3. At the t^{th} MCMC cycle, given λ^t and U^t, β is conditionally uniform on $\{l(\lambda^t,\beta) > U^t\}$. Thus, β can be generated from a uniform distribution on $(a_\beta,b_\beta) \supset \{l(\lambda^t,\beta) > U^t\}$, chosen as follows.

   (a) From the pre-selected grid for λ, find the two λ^i that bracket λ^t. Call them L_λ and U_λ.

   (b) Find the bounds of β, (a_β^*,b_β^*) among (L_λ,β^j) and (U_λ,β^j) such that $l(L_\lambda,\beta|\boldsymbol{y}) > U^t$ and $l(U_\lambda,\beta|\boldsymbol{y}) > U^t$.

   (c) Extend both ends of the interval (a_β^*,b_β^*) until $l(\lambda^t,a_\beta^*|\boldsymbol{y}) \le U^t$ and $l(\lambda^0,b_\beta^*|\boldsymbol{y}) \le U^t$, giving (a_β,b_β).

4. Draw β from $Unif(a_\beta, b_\beta)$, until $l(\lambda^t, \beta|\boldsymbol{y}) > U^t$.

The pre-processing steps 1 and 2 are done before the MCMC draws. The interval (a_β, b_β) is in general much narrower than the original $(0, 1)$, so the acceptance rate is improved.

We present this accelerator as part of a proof of principle and do not claim it can be used generally. Obviously the efficiency of our slice sampler can and should be improved.

References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage." *Statistica Sinica*, 10: 1281–1311.

Besag, J. and Higdon, D. (1999). "Bayesian Analysis of Agricultural Field Experiments (Disc: P717-746)." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61: 691–717.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis, Classic Edition*. John Wiley & Sons.

Browne, W. J. and Draper, D. (2006). "A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models (with discussion)." *Bayesian Analysis*, 1: 473–550.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis, 2nd edition*. Chapman & Hall Ltd.

Casella, G. and Robert, C. P. (1996). "Rao-Blackwellisation of Sampling Schemes." *Biometrika*, 83: 81–94.

Chen, L., Qin, Z., and Liu, J. (2000). "Exploring Hybrid Monte Carlo in Bayesian Computation." In George, E. I. (ed.), *Bayesian Methods with Applications to Science, Policy, and Official Statistics. Selected Papers from ISBA 2000*, 71–80. International Society for Bayesian Analysis.

Damien, P., Wakefield, J., and Walker, S. (1999). "Gibbs Sampling for Bayesian Non-conjugate and Hierarchical Models by Using Auxiliary Variables." *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61: 331–344.

Daniels, M. J. (1999). "A Prior for the Variance in Hierarchical Models." *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 27: 567–578.

Darby, M. and Walsh, M. (1995). *Periodontal and Oral Hygiene Assessment. Dental Hygiene Theory and Practice*. W.B. Saunders.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). "Efficient Parametrisations for Normal Linear Mixed Models." *Biometrika*, 82: 479–488.

Gelman, A. (2004). “Parameterization and Bayesian Modeling.” *Journal of the American Statistical Association*, 99: 537–545.

— (2005a). “Analysis of Variance – Why It Is More Important Than Ever.” *The Annals of Statistics*, 33: 1–53.

— (2005b). “Prior Distributions for Variance Parameters in Hierarchical Models.” *Bayesian Analysis*, 1: 1–19.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, 2nd edition*. Chapman & Hall/CRC.

Gelman, A. and Huang, Z. (2007). “Estimating Incumbency Advantage and Its Variation, As An Example of a Before/After Study (with discussion).” *Journal of the American Statistical Association*, to appear.

Geyer, C. J. (1992). “Practical Markov Chain Monte Carlo (Disc: P483-503).” *Statistical Science*, 7: 473–483.

Gilks, W. R. e., Richardson, S. e., and Spiegelhalter, D. J. e. (1998). *Markov Chain Monte Carlo in Practice*. Chapman & Hall Ltd.

Gustafson, P., Hossain, S., and MacNab, Y. (2007). “Conservative Prior Distributions for Covariance Parameters in Hierarchical Models.” *Canadian Journal of Statistics*, to appear.

Hodges, J., Cui, Y., Sargent, D., and Carlin, B. (2007). “Smoothing Balanced Single-Error-Term Analysis of Variance.” *Technometrics*, 49: 12–25.

Hodges, J. and Sargent, D. (2001). “Counting Degrees of Freedom in Hierarchical and Other Richly-Parameterised Models.” *Biometrika*, 88: 367–379.

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003). “On the Precision of the Conditionally Autoregressive Prior in Spatial Models.” *Biometrics*, 59: 317–322.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). “Markov Chain Monte Carlo in Practice: A Roundtable Discussion.” *The American Statistician*, 52: 93–100.

Lai, J. and Hodges, J. S. (1999). “Effects of Processing Parameters on Physical Properties of the Silicon Maxillofacial Prosthetic Materials.” *Dental Materials*, 15: 450–455.

Liu, J. and Hodges, J. S. (2003). “Posterior Bimodality in the Balanced One-way Random-effects Model.” *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 65: 247–255.

Neal, R. (1997). “Markov Chain Monte Carlo Methods Based On “Slicing” The Density Function.” *Technical Report 9722, Department of Statistics, University of Toronto*.

Neal, R. M. (2003). “Slice Sampling.” *The Annals of Statistics*, 31: 705–767.

Newcomb, R. (1961). "On The Simultaneous Diagonalization of Two Semi-Definite Matrices." *Quarterly of Applied Mathematics*, 19: 144–146.

Priestley, M. B. (1981). *Spectral Analysis and Time Series. (Vol. 1): Univariate Series*. Academic Press.

Reich, B., Hodges, J., and Carlin, B. (2007). "Spatial Analysis of Periodontal Data Using Conditional Autoregressive Priors Having Two Types of Neighbor Relations." *Journal of the American Statistical Association*, 102: 44–55.

Ridgeway, G. and Madigan, D. (2003). "A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets." *Data Mining and Knowledge Discovery*, 7: 301–319.

Sargent, D. and Hodges, J. (1997). "Smoothed ANOVA With Application To Subgroup Analysis." *Research Report rr97-002, Division of Biostatistics, University of Minnesota, ftp://ftp.biostat.umn.edu/pub/1997/rr97-002.ps.Z*.

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000). "Structured Markov Chain Monte Carlo." *Journal of Computational and Graphical Statistics*, 9: 217–234.

Smith, A. F. M. (1973). "Bayes Estimates in One-way and Two-way Models." *Biometrika*, 60: 319–329.

Tierney, L. and Mira, A. (1999). "Some Adaptive Monte Carlo Methods for Bayesian Inference." *Statistics in Medicine*, 18: 2507–2515.

West, M. and Harrison, J. (1999). *Bayesian Forecasting and Dynamic Models, 2nd edition*. Springer-Verlag Inc.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006). "General Design Bayesian Generalized Linear Mixed Models." *Statistical Science*, 21: 35–51.

	Simplex	(z_1, z_2)
Algorithm	Slice sampler with 1 uniform auxiliary variable	adaptive Metropolis w/ Normal candidate
Working par.	$(\lambda, \boldsymbol{\beta})$	(z_1, z_2)
Initial values	$(4, 0.2)$	$(1, 1)$
Initial tuning constants	—	0.5
	(τ_0, τ_1, τ_2)	$(\sigma_0, \sigma_1, \sigma_2)$
Algorithm	adaptive Metropolis w/ Normal candidate	adaptive Metropolis w/ Normal candidate
Working par.	$(\log \tau_0, \log \tau_1, \log \tau_2)$	$(\log \sigma_0, \log \sigma_1, \log \sigma_2)$
Initial values	$(1, 2, 1)$	$(1, 2, 1)$
Initial tuning constants	—	0.2

Table 1: Description of algorithms for the 2NRCAR model

	Parameterization used in MCMC algorithm			
Prior	$(\lambda, \boldsymbol{\beta})$	(r_1, r_2)	(τ_1, τ_2, τ_0)	$(\sigma_1, \sigma_2, \sigma_0)$
$\lambda \sim Gamma(0.01, 0.01)$	λ: 573	$\log(r_1)$: 577	$\log(\tau_1)$: 275	$\log(\sigma_1)$: 379
$\boldsymbol{\beta} \sim$ uniform on simplex	$\boldsymbol{\beta}$: 1009	$\log(r_2)$: 531	$\log(\tau_2)$: 591	$\log(\sigma_2)$: 672
			$\log(\tau_0)$: 301	$\log(\sigma_0)$: 359
$Gamma(0.01, 0.01)$ for	λ: 1037	$\log(r_1)$: 640	$\log(\tau_1)$: 261	$\log(\sigma_1)$: 261
τ_0, τ_1 and τ_2	$\boldsymbol{\beta}$: 1035	$\log(r_2)$: 713	$\log(\tau_2)$: 697	$\log(\sigma_2)$: 253
			$\log(\tau_0)$: 248	$\log(\sigma_0)$: 418
flat for SDs	λ: 1389	$\log(r_1)$: 648	$\log(\tau_1)$: 175	$\log(\sigma_1)$: 265
$\sigma_0, \sigma_1, \sigma_2$	$\boldsymbol{\beta}$: 860	$\log(r_2)$: 651	$\log(\tau_2)$: 406	$\log(\sigma_2)$: 156
			$\log(\tau_0)$: 215	$\log(\sigma_0)$: 234

Table 2: Effective sample size (ESS) comparison of various parameterizations for the CAR model with two classes of neighbor relations.

	Parameterization used in MCMC algorithm			
Prior	$(\lambda, \boldsymbol{\beta})$	(r_1, r_2)	(τ_0, τ_1, τ_2)	$(\sigma_0, \sigma_1, \sigma_2)$
$\lambda \sim Gamma(0.01, 0.01)$	λ: 0.43	$\log(r_1)$: 0.94	$\log(\tau_1)$: 0.17	$\log(\sigma_1)$: 0.24
$\boldsymbol{\beta} \sim$ uniform on simplex	$\boldsymbol{\beta}$: 0.75	$\log(r_2)$: 0.87	$\log(\tau_2)$: 0.36	$\log(\sigma_2)$: 0.43
			$\log(\tau_0)$: 0.18	$\log(\sigma_0)$: 0.23
$Gamma(0.01, 0.01)$ for	λ: 0.59	$\log(r_1)$: 1.02	$\log(\tau_1)$: 0.17	$\log(\sigma_1)$: 0.16
τ_0, τ_1 and τ_2	$\boldsymbol{\beta}$: 0.59	$\log(r_2)$: 1.14	$\log(\tau_2)$: 0.44	$\log(\sigma_2)$: 0.26
			$\log(\tau_0)$: 0.16	$\log(\sigma_0)$: 0.16
flat for SDs	λ: 0.78	$\log(r_1)$: 1.04	$\log(\tau_1)$: 0.11	$\log(\sigma_1)$: 0.29
$\sigma_0, \sigma_1, \sigma_2$	$\boldsymbol{\beta}$: 0.48	$\log(r_2)$: 1.05	$\log(\tau_2)$: 0.26	$\log(\sigma_2)$: 0.17
			$\log(\tau_0)$: 0.14	$\log(\sigma_0)$: 0.26

Table 3: Effective sample size per second (ESS/sec) comparison of various parameterizations for the CAR model with two classes of neighbor relations.

	Parameterization used in MCMC algorithm			
Prior	$(\lambda, \boldsymbol{\beta})$	$\boldsymbol{r}$	$\boldsymbol{\tau}$	$\boldsymbol{\sigma}$
$\lambda \sim Gamma(0.01, 0.01)$	λ: 1615	$\log(r_1)$: 336	$\log(\tau_0)$: 280	$\log(\sigma_0)$: 313
$\boldsymbol{\beta} \sim$ uniform on simplex	$\boldsymbol{\beta}$: 3965	$\log(r_2)$: 231	$\log(\tau_1)$: 244	$\log(\sigma_1)$: 194
	4617	$\log(r_3)$: 294	$\log(\tau_2)$: 227	$\log(\sigma_2)$: 169
	4971	$\log(r_4)$: 331	$\log(\tau_3)$: 219	$\log(\sigma_3)$: 343
			$\log(\tau_4)$: 230	$\log(\sigma_4)$: 329
$Gamma(0.01, 0.01)$ for	λ: 2498	$\log(r_1)$: 210	$\log(\tau_0)$: 436	$\log(\sigma_0)$: 287
$\tau_0, \tau_1, \tau_2, \tau_3, \tau_4$	$\boldsymbol{\beta}$: 4110	$\log(r_2)$: 365	$\log(\tau_1)$: 336	$\log(\sigma_1)$: 172
	5000	$\log(r_3)$: 244	$\log(\tau_2)$: 198	$\log(\sigma_2)$: 240
	5000	$\log(r_4)$: 370	$\log(\tau_3)$: 321	$\log(\sigma_3)$: 204
			$\log(\tau_4)$: 149	$\log(\sigma_4)$: 137
flat for SDs	λ: 4614	$\log(r_1)$: 638	$\log(\tau_0)$: 484	$\log(\sigma_0)$: 453
$\sigma_0, \sigma_1, \sigma_2, \sigma_3, \sigma_4$	$\boldsymbol{\beta}$: 4657	$\log(r_2)$: 752	$\log(\tau_1)$: 506	$\log(\sigma_1)$: 591
	4576	$\log(r_3)$: 626	$\log(\tau_2)$: 503	$\log(\sigma_2)$: 470
	5000	$\log(r_4)$: 655	$\log(\tau_3)$: 629	$\log(\sigma_3)$: 516
			$\log(\tau_4)$: 500	$\log(\sigma_4)$: 199

Table 4: Comparison of effective sample size (ESS) in SANOVA model

	Parameterization used in MCMC algorithm			
Prior	$(\lambda, \boldsymbol{\beta})$	$\boldsymbol{r}$	$\boldsymbol{\tau}$	$\boldsymbol{\sigma}$
$\lambda \sim Gamma(0.01, 0.01)$	λ: 152.5	$\log(r_1)$: 60.1	$\log(\tau_0)$: 46.0	$\log(\sigma_0)$: 50.7
$\boldsymbol{\beta} \sim$ uniform on simplex	$\boldsymbol{\beta}$: 374.4	$\log(r_2)$: 41.3	$\log(\tau_1)$: 40.1	$\log(\sigma_1)$: 31.4
	436.0	$\log(r_3)$: 52.6	$\log(\tau_2)$: 37.3	$\log(\sigma_2)$: 27.4
	469.4	$\log(r_4)$: 59.2	$\log(\tau_3)$: 36.0	$\log(\sigma_3)$: 55.6
			$\log(\tau_4)$: 37.8	$\log(\sigma_4)$: 53.3
$Gamma(0.01, 0.01)$ for	λ: 215.3	$\log(r_1)$: 36.6	$\log(\tau_0)$: 75.2	$\log(\sigma_0)$: 50.8
$\tau_0, \tau_1, \tau_2, \tau_3, \tau_4$	$\boldsymbol{\beta}$: 354.3	$\log(r_2)$: 63.6	$\log(\tau_1)$: 57.9	$\log(\sigma_1)$: 30.4
	431.0	$\log(r_3)$: 42.5	$\log(\tau_2)$: 34.1	$\log(\sigma_2)$: 42.5
	431.0	$\log(r_4)$: 64.5	$\log(\tau_3)$: 55.3	$\log(\sigma_3)$: 36.1
			$\log(\tau_4)$: 25.7	$\log(\sigma_4)$: 24.2
flat for SDs	λ: 36.5	$\log(r_1)$: 80.1	$\log(\tau_0)$: 61.2	$\log(\sigma_0)$: 57.9
$\sigma_0, \sigma_1, \sigma_2, \sigma_3, \sigma_4$	$\boldsymbol{\beta}$: 36.8	$\log(r_2)$: 94.4	$\log(\tau_1)$: 64.0	$\log(\sigma_1)$: 75.5
	36.2	$\log(r_3)$: 78.5	$\log(\tau_2)$: 63.6	$\log(\sigma_2)$: 60.0
	39.5	$\log(r_4)$: 82.2	$\log(\tau_3)$: 79.5	$\log(\sigma_3)$: 65.9
			$\log(\tau_4)$: 63.2	$\log(\sigma_4)$: 25.4

Table 5: Comparison of effective sample size per second (ESS/sec) in SANOVA model

Table 6: Parameterization and associated reference priors

Method	Parameter	Prior	Integrate out τ_0?
Simplex	$\beta_k = \frac{\tau_k}{\sum \tau_j}$;	$\lambda \sim Gamma(0.01, 0.01)$,	Yes
	$\lambda = \frac{\sum \tau_j}{\tau_0}$	$\boldsymbol{\beta} \sim Unif$ on the simplex	
Precision	$\tau_0, \tau_1, \cdots, \tau_s$	$\tau_k \sim Gamma(0.01, 0.01),\ k = 0, \cdots, s$	No
SD	$\sigma_0 = \frac{1}{\sqrt{\tau_0}}$,	$\sigma_k \sim Unif(0, 100),\ k = 0, \cdots, s$	No
	$\sigma_k = \frac{1}{\sqrt{\tau_k}}$	except SANOVA $\sigma_k \sim Unif(0, 10)$	
Z	$z_k = \log(\frac{\tau_k}{\tau_0})$	$z_k \sim Unif(-15, 15),\ k = 1, \cdots, s$	Yes

Table 7: Design values in the simulation studies

	2NRCAR			SANOVA									Crossed RE		
Case	τ_0	τ_1	τ_2	τ_0	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	τ_0	τ_1	τ_2
1	1	1	1	1	0	0	0	0	1	0	0	0	$\frac{1}{4}$	1	1
2	1	1	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0	0	1	0	0	0	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{16}$
3	1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{16}$	0	0	0	0	1	0	0	0	$\frac{1}{4}$	$\frac{1}{16}$	1
4	1	$\frac{1}{4}$	1	1	0	0	0	0	1	1	1	0	$\frac{1}{16}$	1	1
5	$\frac{1}{4}$	1	1	$\frac{1}{4}$	0	0	0	0	1	1	1	0	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
6	$\frac{1}{4}$	1	$\frac{1}{4}$	$\frac{1}{16}$	0	0	0	0	1	1	1	0	$\frac{1}{16}$	$\frac{1}{16}$	1
7	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{100}$	0	0	0	0	10	0	0	0	$\frac{1}{100}$	$\frac{1}{25}$	$\frac{1}{25}$
8	$\frac{1}{4}$	$\frac{1}{4}$	1	1	0	0	0	0	10	0	0	0	–	–	–

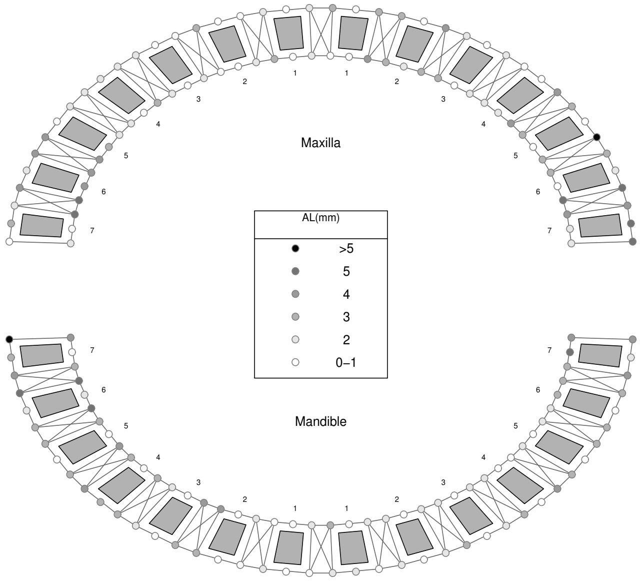


Figure 1: Attachment loss measurements for one patient. The maxilla is the upper jaw, the mandible is the lower jaw, the gray boxes are teeth, the small number counting from the center of each jaw is the tooth number. Small circles indicate the six measurement sites per tooth.

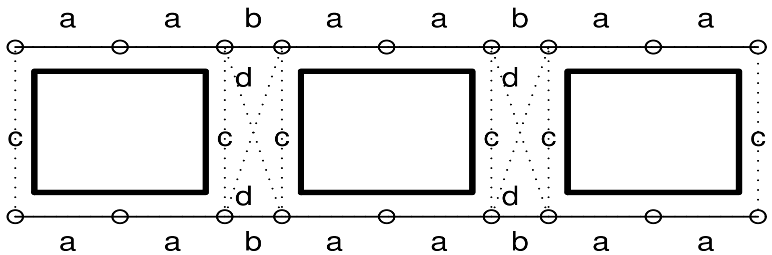


Figure 2: Neighbor types in periodontal measurements. Letters a-d specify neighbor types. Solid and dotted lines indicate the two classes of neighbors considered in this paper.

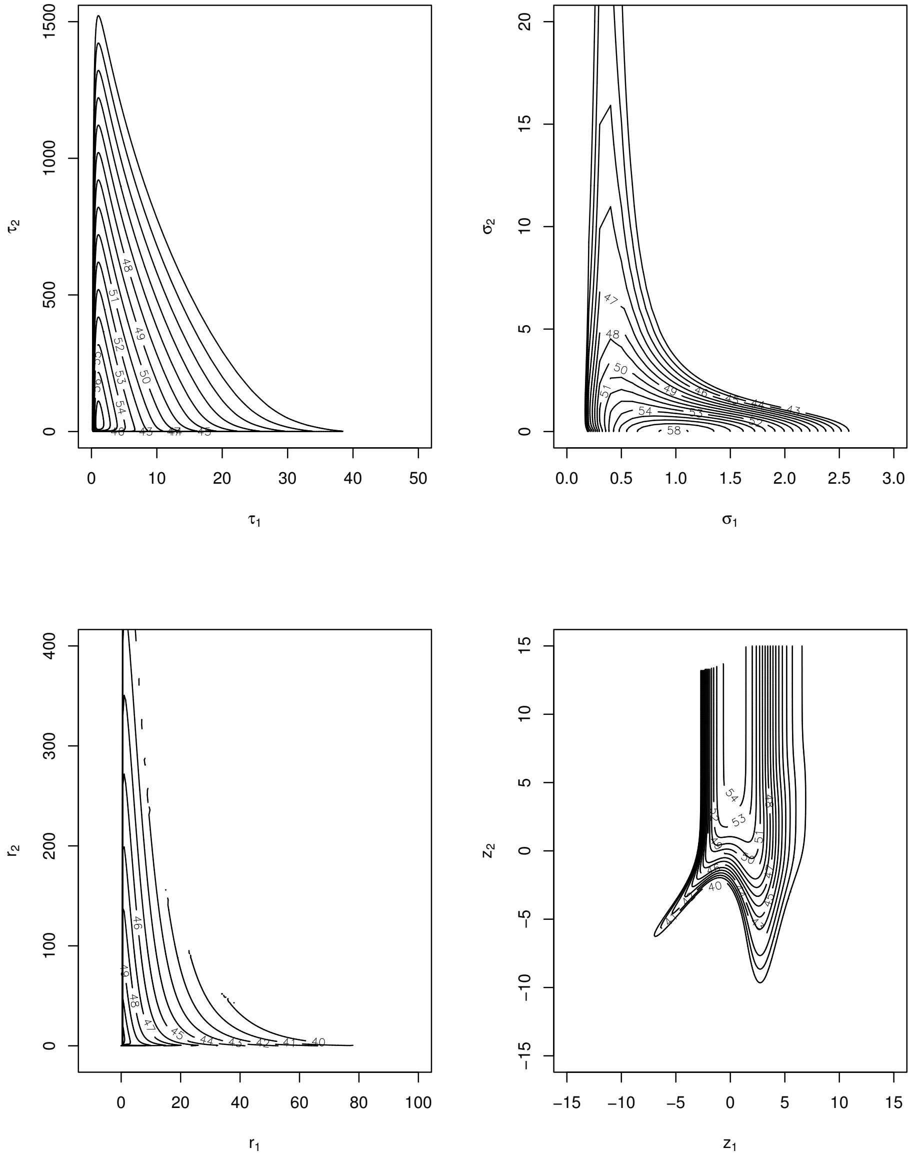


Figure 3: 2NRCAR model: Logarithm posterior contour plots with contours at 1 log intervals for four parameterizations with their own reference priors: $\tau_0, \tau_1, \tau_2 \sim Gamma(0.01, 0.01)$, $\sigma_0, \sigma_1, \sigma_2 \sim Unif(0, L)$, $r_1, r_2 \sim Gamma(0.01, 0.01)$, $z_1, z_2 \sim Unif(-15, 15)$. The contours for (τ_0, τ_1, τ_2) and $(\sigma_0, \sigma_1, \sigma_2)$ are drawn for the slice $\tau_0 = 1$ and $\sigma_0 = 1$, respectively.

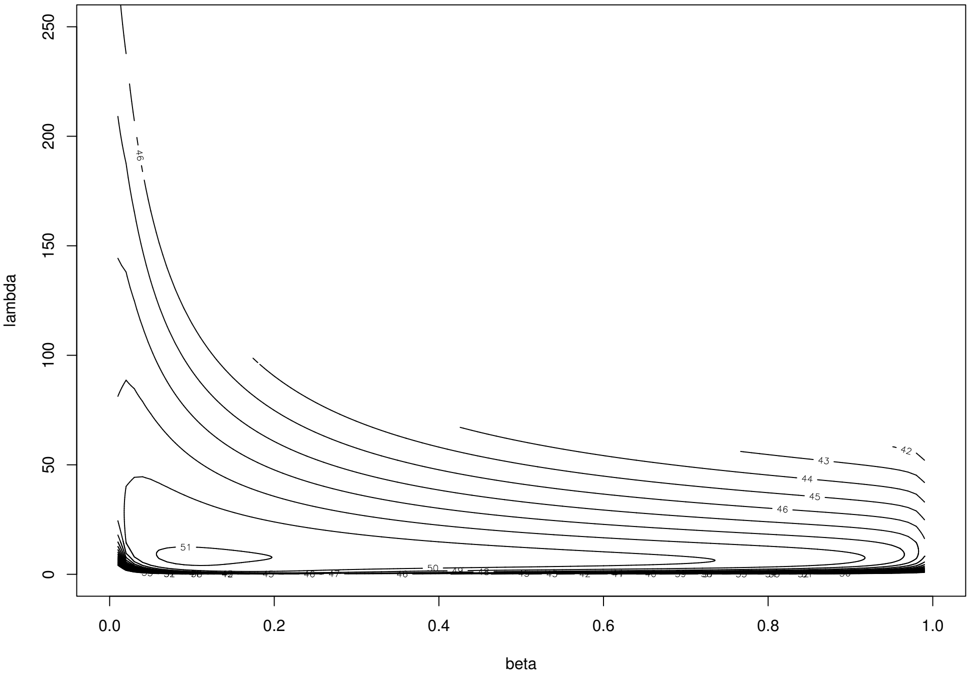


Figure 4: 2NRCAR model: Log posterior contour plot with contours at 1 log intervals, for the simplex parameterization with its reference prior.

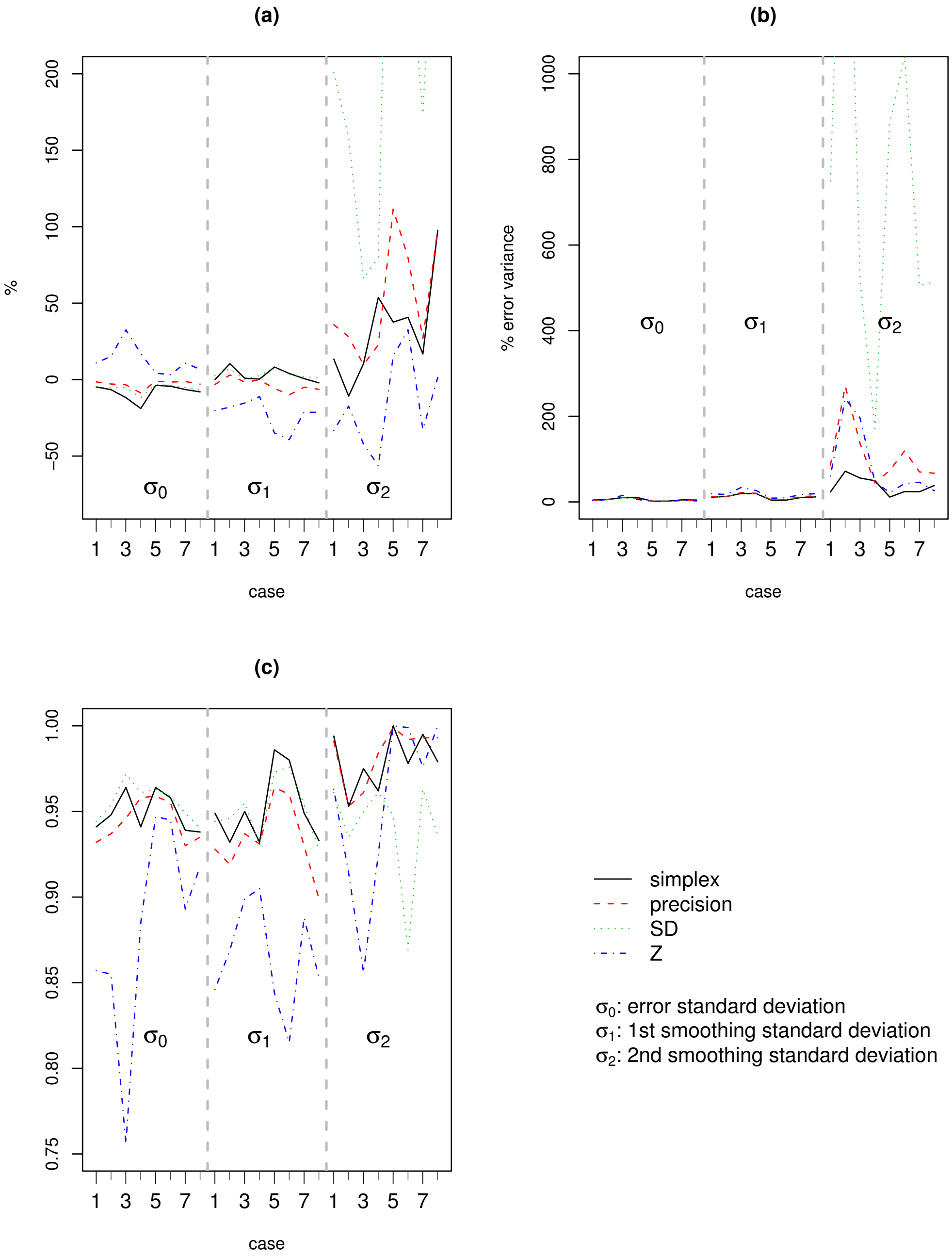


Figure 5: 2NRCAR simulation: Standard deviation bias (as a percent of true standard deviation) and MSE (divided by the true error variance $\frac{1}{\tau_0}$). (a) scaled bias for σ_0, σ_1, and σ_2; (b) scaled MSE for σ_0, σ_1, and σ_2; (c) 95% interval coverage for σ_0, σ_1, and σ_2.

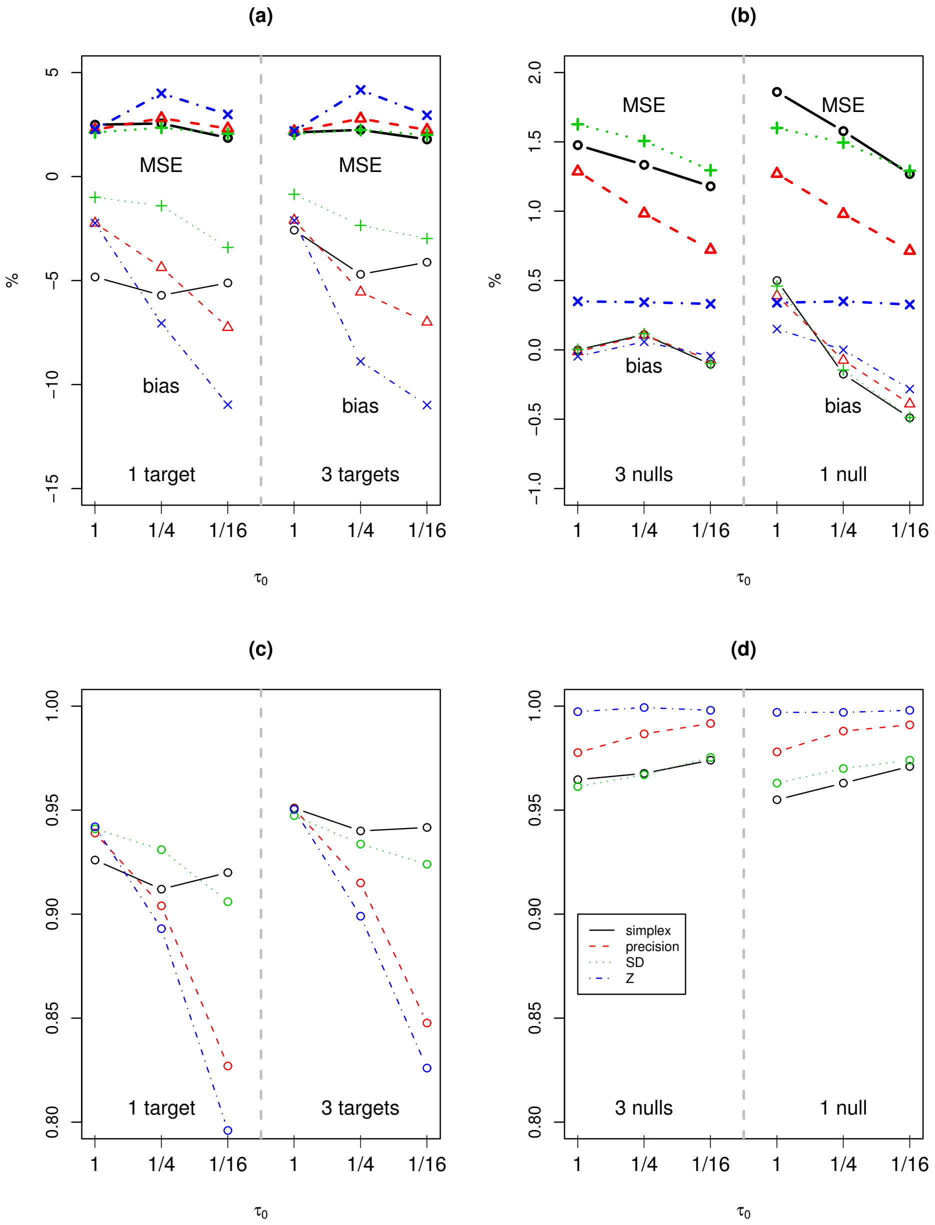


Figure 6: SANOVA simulation: Average bias and MSE as percents of $\frac{1}{\sqrt{\tau_0}}$ and $\frac{1}{\tau_0}$ respectively, for θ_k for truly present interactions (a), and truly absent interactions (b). Within each figure, the upper curves are MSE and the lower curves are biases. 95% interval coverage probability for θ_k for truly present interactions (c), and truly absent interactions (d).

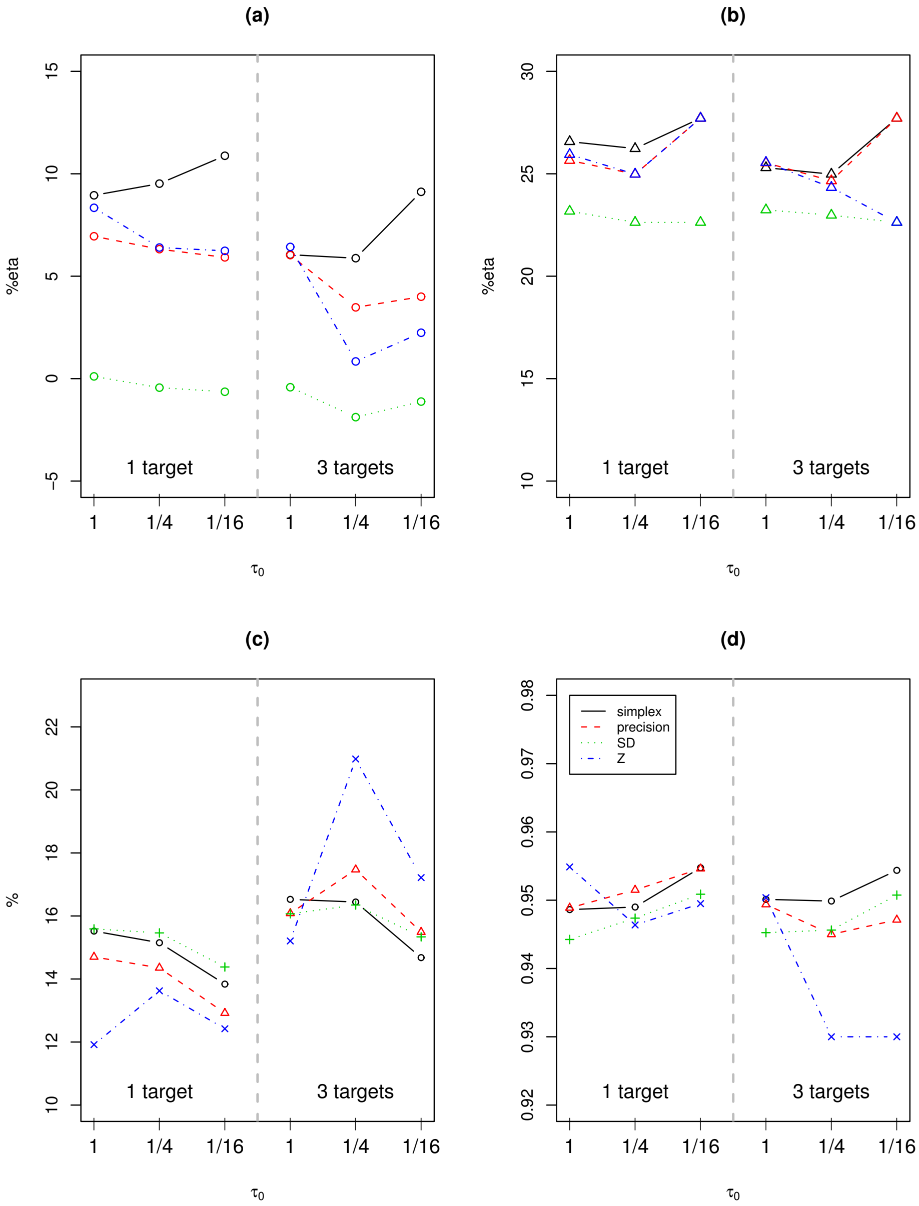


Figure 7: SANOVA simulation: (a) error precision bias as a percent of τ_0; (b) square root of MSE as a percent of τ_0; (c) average cell mean MSE (as a percent of $\frac{1}{\tau_0}$); (d) average cell mean 95% interval coverage probability.

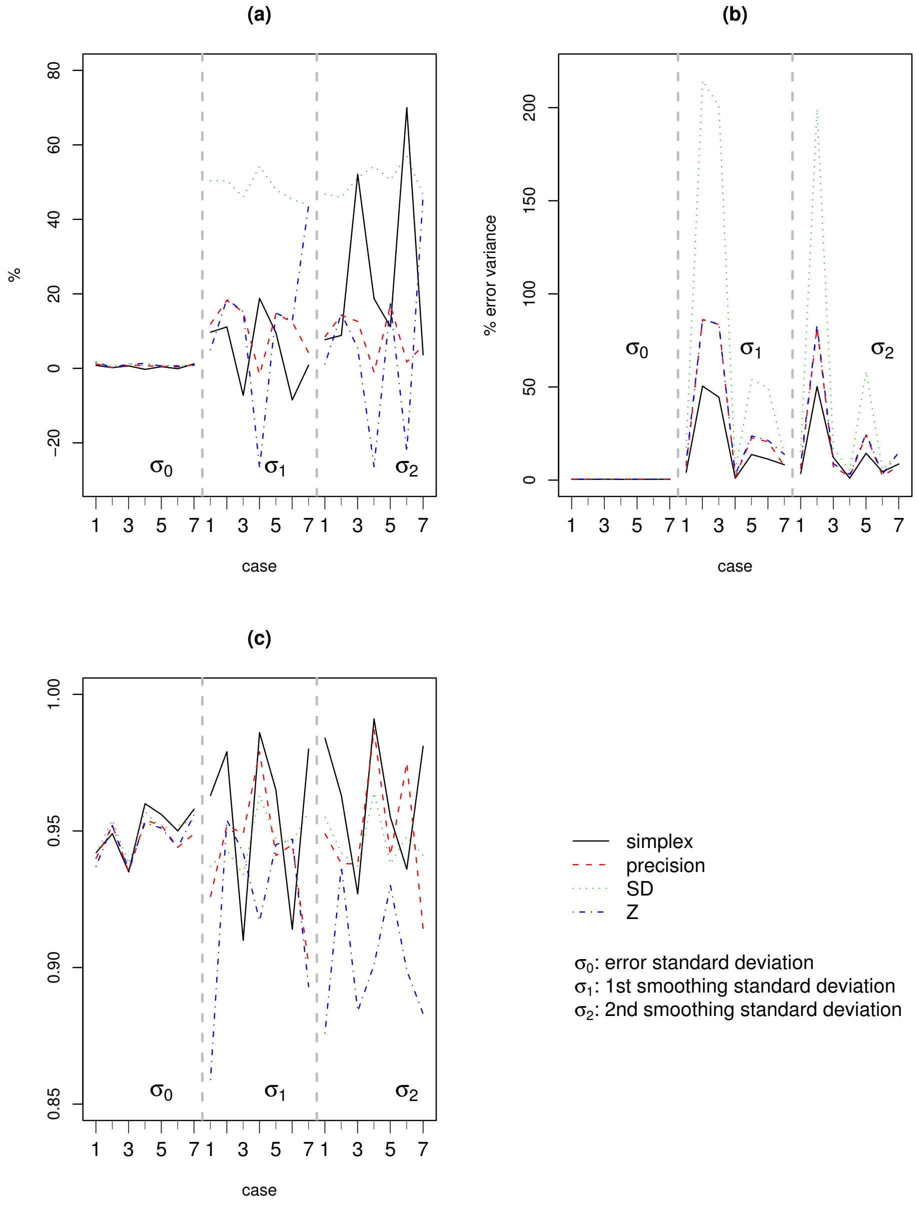


Figure 8: Crossed RE simulation: standard deviation bias (as a percent of true standard deviation) and MSE (divided by the true error variance $\frac{1}{\tau_0}$). (a) scaled bias for σ_0, σ_1, and σ_2; (b) scaled MSE for σ_0, σ_1, and σ_2; (c) 95% interval coverage for standard deviations σ_0, σ_1, and σ_2.

Bayesian Analysis (2007) **2**, Number 3, pp. 557–590

Cluster Allocation Design Networks

Ana Maria Madrigal[*]

Abstract. When planning and designing a policy intervention and evaluation, it is important to differentiate between (future) policy interventions we want to evaluate, F_T, affecting "the world", and experimental allocations, A_T, affecting "our picture of the world". The policy maker usually has to define a strategy that involves policy assignment and recording mechanisms that will affect the (conditional independence) structure of the data available. Causal inference is sensitive to the specification of these mechanisms. Influence diagrams have been used for causal reasoning within a Bayesian decision-theoretic framework that introduces interventions as decision nodes (Dawid 2002). Design Networks expand this framework by including experimental design decision nodes (Madrigal and Smith 2004). They provide semantics to discuss how a design decision strategy (such as a cluster randomised study) might assist the identification of intervention causal effects. The Design Network framework is extended to Cluster Allocation. It is used to assess identifiability when the experimental unit's level is different from the analysis unit's level, and to discuss the evaluation of cluster- and individual-level future policies. Cases of 'pure' cluster (all individuals in a cluster receiving the same intervention) and 'non-pure' cluster (only a subset receiving the policy) are discussed in terms of causal effects. The representation and analysis of a simplified version of a Mexican social policy programme to alleviate poverty (Progresa) is performed as an illustration of the use of Bayesian hierarchical models to make causal inferences relating to household and community level interventions.

Keywords: Cluster allocation, Influence diagrams, Causal inference, Identification of policy effects, DAGs

1 Introduction

Different data sets provide different types of information. Different queries might require different information to obtain answers. When using data for learning, it is important to consider the conditions and circumstances under which the data were collected. The distributions that can be learnt (or not) might vary among apparently similar data sets. This is an important consideration to the analyst before learning model parameters. Consider the case in which we have two data sets that contain records of whether or not children in a population take food nutrition supplements (FS) and whether or not they have gained weight. The first data set comes from a census sample, and the second comes from an experiment where half the children were given food supplements and half of them were not. Suppose we are interested in learning the prevalence of children taking supplements in the population, $p(FS)$. It is clear that learning from the second data set that $p(FS) = 0.5$ only reflects an experimental choice and not a population prevalence, as would be the case if we were to use the first data set (e.g. showing how

[*] University of Warwick, UK mailto:am.madrigal@warwickgrad.net

parents 'naturally' choose to give FS to children). The collecting strategy defines the structure of the data set. A perfect design of experiments will give a more organised layout of the data. An observational study might be more 'disorganised', as data is allowed to arise naturally. In this paper, we focus our attention on a particular type of query related to policy intervention evaluations and discuss which data set structures do or do not let us answer causal queries and extract appropriate causal effects (e.g. the causal effect of FS on children's weight). Discussions about the identifiability of causal effects have been usually phrased as 'Is the causal effect of T on Y identifiable?' (see Pearl 2000; Lauritzen 2001; Dawid 2002). In this paper, the role of data structures is made explicit by phrasing the identifiability question as 'Is the causal effect of T on Y identifiable *from the data available*?'.

Intervention has to do with 'perturbing' the dynamics of a system. If we say that a system consists of components which influence each other and that its dynamics describe the way these components interact with each other in an equilibrium state, some examples of systems might be consumption-expenditure patterns, road traffic in a town or the human body. The system at present has some *pre-intervention* dynamics attached to it. When we intervene a system, by introducing a promotion-advertisement campaign, by adding a red light at a corner, or by giving medicine, we are introducing a new component into a system that will imply new *post-intervention* dynamics. The intervention might have both qualitative effects, modifying the structure of the system (maybe by 'blocking' the interaction between two of its components), and quantitative effects, modifying the value of the components. One of the main interests consists in describing if and how the intervention affects the system. Evaluation of the intervention effects is required and it is usually measured in terms of a response variable, such as sales-awareness, number of accidents, or health condition.

Discussions of causal reasoning have been made usually assuming that the graph representing the system implicitly includes the underlying (experimental) mechanism that is generating the data (see Pearl 2000). Then, in this fixed 'natural' or 'idle' system, whether the future policy intervention F_T effect is identifiable and can be obtained is evaluated. Randomised allocation of treatments to units is a well known practice within medical clinical trials but, because of ethical, social and financial issues, complete randomisation within an experiment designed to evaluate a social policy will usually be unfeasible. Knowing the details of the policy assignment mechanism and a well-planned recording of the data become very relevant issues in order to obtain all the information needed to measure the right 'causal' effects (see Rubin 1978). Influence Diagrams (IDs) are used to represent the system dynamics and interventions graphically; a review of the main features of the framework used is made in Section 2. Our interpretation of causal effects for interventions is Bayesian decision-theoretic, where an intervention on a system is regarded as a decision. Dawid (2002)'s extended influence diagrams are augmented by including 'experimental design' decisions nodes within the set of intervention strategies to create what we call a Design Network (DN), to provide semantics to discuss how a 'design' strategy (such as clustering) might assist the systematic identification of intervention causal effects, to give a taxonomy for design decisions, and to show how these decisions might alter the graphical (conditional independence) structure used to evaluate the causal effect of policy F_T. It is maintained that experimental design decisions are intrinsic to any causal analysis of policy intervention strategies. Design Networks were introduced in Madrigal and Smith (2004) for random allocation, and their main characteristics are presented in Section 3. Design Networks for cluster allocation are discussed in Section 4; the propositions can be derived from the discussion in Appendix A.

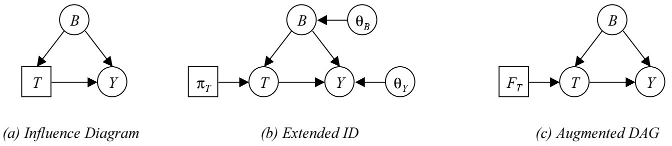
This research was motivated by a Mexican Social Policy Programme (Progresa) whose objective is to alleviate poverty. It consists of a three-stage mechanism to target its eligible population, based on community and household characteristics. The policy involves a collection of interventions at different levels (community, household and individual). All households are recipients of the community-level interventions (e.g. health infrastructure and services). Actions at household and individual level (e.g. extra monetary support and nutritional supplements) affect only 'poor' (eligible) households and vary according to household/individual demographics, so not all units in the community (cluster) are intervened equally. This motivates the discussion about the data structure arising from a cluster allocation, the distinction of 'overall' and 'total' effects, the differences in the inference of cluster- and individual-level interventions, and the nested structures in design and analysis. The design of the study included a *randomised* cluster allocation for treatment and control communities. To illustrate some features of causal analysis in a cluster allocation setting, this paper presents in Section 5, a hierarchical model analysis based on Spiegelhalter (2001). Formulation is performed for the evaluation of cluster- and individual-level interventions based on Progresa data.

2 Intervention Graphical Framework and Causal Inference

Influence diagrams (IDs) have been used for over 20 years to form the framework for both describing (see Howard and Matheson 1981; Oliver and Smith 1990) and also devising efficient algorithms to calculate the effects of decisions (see Jensen 2001) in complex systems which implicitly embody strong conditional independence assertions. However, it is only recently that they have been used to explain causal relationships (Dawid 2000, 2002), and been shown to be much more versatile than Causal Bayesian Networks (Pearl 1993, 1995).

The simplest form of external intervention is when a single variable X is forced to take on some fixed value x'. This is known as an 'atomic intervention' and, following Pearl (2000), it is denoted by $do(X = x')$. The atomic intervention replaces the original mechanism: $p(x \mid pa(x))$ by $p(x \mid pa(x); do(X = x')) = 1$ **if** $X = x'$ where $pa(x)$ denotes the parent nodes of X. This conditioning by intervention formula has appeared in various forms (see Pearl 1993; Spirtes et al. 2000; Robins 1986). It cannot be asserted in general that the effect of *setting* the value of X to x' is the same as the effect of *observing* $X = x'$. Only in limited circumstances (as when the node for X has no parents in the graph) will conditioning by intervention and conditioning by observation coincide. Graphically, interventions are represented by deleting the arrows that enter the intervened node in the original graph, making explicit the fact that when the value is set externally, the parents' values are not relevant post-intervention. Pearl's $do(\cdot)$ corresponds to an external intervention. By recognising interventions as decisions, the Bayesian decision-theoretical framework embeds Pearl's *doing* operation and provides a stronger framework for causal inference. The strong links between decision theory and Pearl's causal model have been discussed by Heckerman and Shachter (2003). Those who are familiar with Bayesian decision theory will find comfort, as I have, in these connections.

Dawid (2002) points out that, traditionally, in IDs conditional distributions are given for random nodes, but no description is supplied of the functions or distributions involved at the decision nodes, which are left arbitrarily at the choice of the decision maker. If we choose to provide some descriptions of the decision rules, then any given specification of the functions or distributions at decision nodes constitutes a decision strategy, π. Decisions determine what we may term the partial distribution, p, of random nodes given decision nodes which is not in general the same as the associated conditional distributions (see Cowell et al. 1999, section 2.3). If E and D denote the set of random events and the set of decisions, respectively, then the full joint specification p_π, consisting of decision strategy π and partial distribution p for all $e \in E$ and $d \in D$ is given by $p_\pi\left(e, d\right)$. The graphical representation of p_π can be made by using *extended* IDs that incorporate non-random parameter nodes $(\theta_e = p(e \mid pa^0\left(e\right))$ and strategy nodes $\left(\pi_d = \pi(d \mid pa^0\left(d\right))\right)$ representing the mechanisms that generate random and decision nodes respectively. Here, $pa^0\left(.\right)$ denotes the set of domain parents of X (i.e. parents in the original non-extended version of the ID). In what he calls *augmented* DAGs, Dawid incorporates intervention nodes F where $F_X = x$ corresponds to 'setting' the value of node X to x (in Pearl's language: $F_X = do(X = x)$), and he introduces a new value $\emptyset$ such that when $F_X = \emptyset$, X is left to have its 'natural' distribution, termed by Pearl the 'idle' system. Figure 1 shows, for a simple case, the usual representation of IDs as well as its extended and augmented versions, for the set (T, B, Y) where $T = (T_1, T_2, .., T_s)$ represents a set of policy variables (treatments), $B = (B_1, B_2, .., B_r)$ is a set of background variables (potential confounders) and Y is a response variable.



(a) Influence Diagram

(b) Extended ID

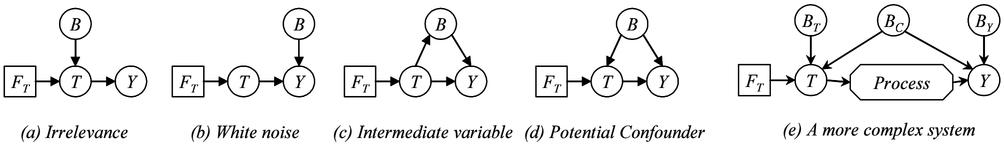
(c) Augmented DAG

Figure 1: Extended influence diagrams and augmented DAGs

Causal reasoning is related to prediction in the face of intervention. It relates to the idea that a variable is a 'cause' if setting this variable to a specific value (by intervention) changes the distribution of the response. Causal enquiries about the 'effect of T on Y' are seen as relating to (comparisons between) the distributions of Y given $F_T = do(T = t')$ for various settings of t'. The intervention node F of the augmented DAG is used as an 'auxiliary' variable to discuss the identifiability of these effects under certain DAG structures. In particular, there is interest in establishing if the causal effect of F_T on Y can be identified and estimated correctly from the available data. The structure of the data available is defined by the set of conditional independencies that are derived from the graph.

Definition (Conditional independence) *If X, Y, Z are random variables with a joint distribution $P(\cdot)$, we say that X is conditionally independent of Y given Z under P, if for any possible pair of values (y, z) for (Y, Z) such that $p(x, y) > 0$, $P(x \mid y, z) = P(x \mid z)$. This can be written following Dawid (1979)'s notation as $(X \perp\!\!\!\perp Y \mid Z)_P$.*

The discussion is conducted in terms of the relevance of learning the strategy that gave the value t' to T, namely whether it arose from the original experimental setting $\pi(t \mid b)$ $(F_T = \emptyset)$ or whether it was set externally $(F_T = do(T = t'))$. Conditional independencies of the form $(Y \perp\!\!\!\perp F_T \mid T, \cdot)_{d_E}$ are used for this. Different examples of identifiable and unidentifiable situations are discussed by Pearl (2000), Lauritzen (2001) and Dawid (2002), each with their particular framework and notation. Imagine the set (T, B, Y) is available to us in the data set Δ. Figures 2*(a)* and 2*(b)* show the cases where B is said to be *irrelevant* for Y and where B is said to be *white noise* of Y (with respect to T) respectively. The case where B is an *intermediate* variable between T and Y (i.e. T affects B and B affects Y) is shown in Figure 2*(c)*. In these three structures the definition of absolute non-confounding given by $Y \perp\!\!\!\perp F_T \mid T$ holds (see Dawid 2002, §7). This asserts that the distribution of Y given T will be the same, whether T arose 'naturally' or T is set by intervention. Thus the causal effect can be estimated directly from the data available, Δ, using $p(y \mid t', F_T = do(T = t')) = p(y \mid t', F_T = \emptyset) = p(y \mid t')$. The definition of non-confounding $(Y \perp\!\!\!\perp F_T \mid T)$ does not hold for the structure shown in Figure 2*(d)*. In this latter system, B is said to act as a *confounder*, as it affects both treatment T and response Y. So, in order to obtain the causal effect, we are required to know (or observe) the marginal distribution $p\,(B)$. If this is the case, then the causal effect $p(y \mid F_T = t')$ can be obtained using the 'back-door formula' (Pearl 1993) which 'adjusts' for B such that $p(y \mid F_T = t') = \sum_b p(y \mid t', b)p\,(b)$.



(a) Irrelevance *(b) White noise* *(c) Intermediate variable* *(d) Potential Confounder* *(e) A more complex system*

Figure 2: Possible basic structures

Social policies will usually be more complex systems including all irrelevant (B_T) and white-noise (B_Y) background variables, possible confounders (B_C) and an intermediate process, as shown in Figure 2*(e)*. Most of the examples in social policy interventions F_T involve (a collection of) atomic or contingent interventions. Therefore, the intermediate process might involve both intermediate variables affected by T and possible actions G

that will be triggered when the policy T is done. The 'overall' causal effect will include all direct and indirect effects of $do(T = t')$ on Y.

3 Introducing experimental nodes

3.1 Policy versus Experimental Decisions

When planning and designing a policy intervention and evaluation, the policy maker will have to define a strategy that involves 'policy intervention' actions ($D_T = \{d'\}$) and 'experimental design' actions ($D_E = \{d^*\}$). The former includes decisions related to how the policy is implemented and what (which doses and to whom) will be provided. The latter is related to the evaluation of the policy and includes experimental design decisions that define the (chosen or controlled) conditions under which the study is carried out and the data (Δ) recorded. If $\mathcal{D} = \{d'_1, .., d'_{\mathcal{D}_T}, d^*_1, .., d^*_{\mathcal{D}_\mathcal{E}}\}$ are the components of a particular decision strategy $\pi_\mathcal{D}$, the interest lies in describing $\pi_D\left(\mathcal{D} \mid E\right)$. In this sense, we say that policy intervention actions (D_T) are concerned with intervening 'the world', while experimental design actions (D_E) relate to intervening the statistician's 'view of the world'.

It is important to differentiate between 'choosing a policy' and 'choosing a design', as the goals of these interventions are different. The 'success' of a policy intervention D_T is measured in terms of its efficacy to provoke 'better' values on the response variable Y through its overall effects reflected by $p(Y \mid F_T = do(T = t'); D_E)$. The efficacy of an experimental intervention, D_E, is measured in terms of its ability to isolate the policy effect as much as possible. Making an explicit representation of both types of interventions will assist decisions of the *experimenter* and considerations of the *analyst*, when the aim is to evaluate the causal effect of policy F_T.

When we, as data-collectors, approach the world, the data we collect depend on our way of approaching it. The data we observe in the database (available data, Δ_{d_E}) will reflect the experimental design decisions $D_E = d_E$ made (or deliberately 'not made') at the time of its collection through $p(data \mid D_E)$. Two extreme cases of designed studies might be, on the one hand, the 'perfect' experiment where all factors are controlled, balanced and randomised and, on the other hand, the complete observational study with all the relations that happen in 'natural' conditions (approximated by a census of all population). The available literature discusses broadly the cases for completely experimental data (see, for example, Chaloner and Verdinelli 1995; Wu and Hamada 2000) or completely observational data (e.g. Rosenbaum 2002). Although in the social sciences access to perfect experimental data is usually not feasible, the data is not always completely observational. In some cases, controls are taken at the time of the design/collection of data, which gives rise to partially experimental data. In this work, we consider D_E to include any experimental conditions that might involve a decision by the experimenter (data collector). The choice of 'no control at all' leads to observational data ($\Delta_\emptyset$) which is assumed to be a (degenerate) special type of experimental data.

Experimental design interventions, $D_E = \{M, R(B)\}$, contain the mechanisms $M =$

$\{M_E, M_S, M_T\}$ through which units are selected and assigned to eligible, sample and treatment groups, and the recording mechanism $R(B)$ that determines whether the background variables are observed and available to us in the data Δ_{d_E}. In addition, implementation details which refer to the logistics and how the study will be carried out are important, as they can introduce some biases. A complete description of these mechanisms is presented in Madrigal (2004). In this paper we focus on the treatment assignment mechanisms M_T.

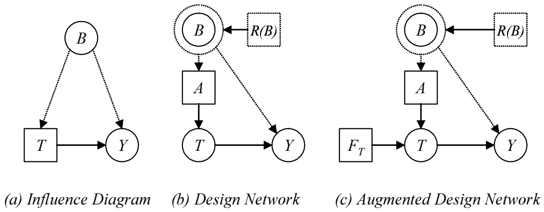
As an example, imagine a policy will be implemented to increase the nutritional state (Y) of 'poor' children in a certain geographical area. Suppose there are two different brands of food supplements (FS) in green or red packages. A decision to sign a contract with the food supplement provider(s) for as long as the policy takes place has to be made. Imagine the policy maker is faced with four possible *policy interventions*: Policy 0 (t'_0): 'Do not give any FS'; Policy 1 (t'_1): 'Give green FS'; Policy 2 (t'_2): 'Give red FS'; and Policy 3 (t'_3):'Give green FS to young babies; and give red FS to older children'. Once a policy is chosen, all children in the target population will be under the same policy. In this case, *policy intervention strategies* (D_T) are defined for the same target population (namely, children in poverty), and the future policy interventions (F_T) are given by t'_0, t'_1, t'_2 and t'_3. When evaluating the policy intervention effect we obtain the 'overall effect' of each of the policies. Although the policy interventions act on children through the actual FS given, it is important to bear in mind that questions 'Is policy t'_i giving better results than policy t'_j?' are different from the question 'Is the green FS working better than the red FS?'. In this case, they will coincide when we are comparing policies t'_1 and t'_2, but to draw conclusions about the effects of green and red FS from a comparison between, say, t'_0 and t'_3 could be dangerous, as in t'_3, the effect of FS is confounded with age. The policy maker, as an experimenter, has to choose the *experimental design strategy* (D_E) used to collect data Δ_{d_E}. This data is used to evaluate policy intervention strategies (D_T) and compare the effects of policies $F_T = do(Policy = t'_s)$. Imagine that policy makers in principle have in mind the implementation of contingent policy t'_3 (against the option of not providing any food supplement at all t'_0). First, the experimental levels $\{t^*\}$ have to be set. These are allocated through action $A_T = do(Policy = t^*)$. Choosing some experimental levels $\{t^*\}$ to be equal to future policy levels $\{t'\}$, such that $t^*_1 = t'_0$ and $t^*_2 = t'_3$, ensures the positivity condition (see Appendix A), and then $\{t^*\} = \{t^*_1, t^*_2\} = \{t'_0, t'_3\}$. Imagine the allocation of policies is done randomly with probability of one half. This random intervention could be expressed as $A_{\theta_T} = do(\theta_T = \theta^*_T)$ such that it fixes $\theta^*_T = p(A_T = do(Policy = t^*_m)) = \frac{1}{2}$ for m=1,2. Policy allocation is randomised and it is defined by the *experimental design strategy*, D_E.

Dawid (2002)'s framework, although open to different strategies for setting the value of a treatment $T = t'$, including randomised or atomic definitions, does not allow us to represent in the same graph and formulae both the atomic (future) policy intervention $F_T \in D_T$ (allocating treatment $T = t'$ with probability one) and the (contingent or randomised) experimental allocation strategy followed when collecting data $A_T \in D_E$ (allocating treatment $T = t^*$ according to θ). Neither does it allow us to represent the impact on the (graphical) data structure of the experimental actions. Therefore, an extension is needed.

3.2 Design Networks: Basics

In its simplified version, let $D_E = \{A, R(B)\}$ where A contains all the policy assignment mechanisms and $R(B)$ contains the recording mechanism, such that $R(B^q) = 1$, for $q = 1, 2...Q$, if variable B^q is recorded and $R(B^q) = 0$ if B^q is either unobservable or not recorded. *Assignment nodes* A and *recording nodes* R can be included in the DAG as decision nodes to create a *design network (DN)*. The design network shows the 'natural' (experimental) mechanisms that generate the data available Δ_{d_E}. In general, no matter whether the data has been collected already or we are planning the design to generate the data, D_E represents decisions to be made at the data collection time.

Consider the set (T, B, Y). For simplicity, suppose that T and Y are univariate, that B does not contain intermediate variables between T and Y (i.e. B consists of pre-intervention variables not affected by T), and that the future policy is an atomic intervention $F_T = do(T = t')$. Figure 3*(a)* shows the usual influence diagram representation of this case, and Figure 3*(b)* gives the corresponding design network. Note that A blocks all the paths going from B to the policy node T. This follows from the assumption that A captures *all* the allocation mechanisms for T that might be influenced by the background variables B, so that A is the only parent of the policy node T in the design network. Recording nodes, $R(B)$, are added for each background variable B^q, introducing the decision to record B^q versus not to record it. A double circle containing a dashed and solid line is given to each background node B^q to show its potential observability. It is assumed that policy T and response variable Y will be recorded. Figure 3*(c)* shows an *augmented design network* in which the future atomic intervention node F_T is added to the design network.



(a) Influence Diagram *(b) Design Network* *(c) Augmented Design Network*

Figure 3: Design Networks

By representing simultaneously the design nodes $D_E = \{A, R(B)\}$ and the future intervention node F_T, the augmented design network is useful to make conclusions about two different tasks involved in policy evaluation and design: (1) the identifiability of the causal effect of T on Y, given a design $d_E \subset D_E$; and (2) the choice of a design strategy d_E to collect data when the interest is to evaluate the effect of intervention F_T. As mentioned before, the identifiability of intervention F_T depends on the data available (determined by mechanisms d_E); and the efficacy of experimental design (d_E) is always to be determined with respect to the effects it tries to isolate (here F_T). Thus, D_E and F_T are and should always be read in the light of each other, and the augmented design network allows us to do that.

For causal reasoning in task (1), to discuss the identifiability of the causal effect of T on Y, we are interested in comparing the relevance of the choice of F_T given particular experimental conditions d_E^*, namely comparing $p(y \mid t', F_T = do(T = t'); D_E = d_E^*)$ and $p(y \mid t', F_T = \emptyset; D_E = d_E^*)$. This provides us with expressions and guidelines for *control via analysis* of possible confounders. On the other hand, in task (2), when planning the data collection by choosing an experimental design, we are interested in the relevance (or irrelevance) of the choice of experimental conditions d_E, (with respect to the identifiability of F_T) for different experimental choices $d_E \subset D_E$. So, we are interested in making comparisons between different settings of d_E^* and then choosing the optimal design from all experimental designs available in D_E. This provides us with guidelines for *control via design*.

Augmented DAGs and the set of conditional independencies derived from them have been used for causal reasoning in task (1) (see Dawid 2002). If two augmented DAGs derived from experimental conditions d_{E1} and d_{E2} share the same conditional independence statements, then they are equivalent for causal reasoning. Assignment actions A might affect the original collection of conditional independence statements. Recording decisions will have an effect on the set of variables that will be available to us through the available (sic) experimental data. Thus $R(B)$ will not introduce any new (in)dependencies in the structure, but will be relevant when discussing the potential identifiability of effects given assignment actions $A = a^*$. Some additional general remarks about the Decision Networks framework can be found in Appendix A.

4 Causal graphical analysis for cluster allocation

Most of the literature in Cluster Randomised Trials (CRTs) has emphasised the fact that Fisher's principle is violated, as the experimental unit does not coincide with the analysis unit, and the difference in levels where the experimental allocation generating data available, Δ, is at cluster level and the analysis is undertaken for a response at individual level. When introducing the need for the evaluation of a future intervention F_T using data generated from a (past) experiment d_E, it is important to acknowledge the fact that the future intervention level might differ from the experimental level. In general, the future intervention ($F_T = do(T = t')$), the experimental allocation ($A_T = do(T = t^*)$) and the response variable (Y) could each be at cluster/individual level and would not necessarily coincide. The experimental level will define the data structure and the conditional independence statements reflected in the 'experimental' causal graph through d_E. The future intervention F_T level will define the 'future' causal graph structure.

The interest could lie in the causal effect at cluster level or at individual level. Responses at cluster level will summarise what is observed at a community level, while responses at individual level are usually more of interest to describe what is the effect in, say, a household within a community. If we are interested in the intervention effect on an individual level response, depending on how the future intervention will be implemented, there are two possible (causal) intervention effects we might be interested to identify: namely, the distribution of the individual outcome given a clustered intervention, $P(Y_{jk} \mid F_{Tj} = do(T_j = t'))$, and the distribution of the individual outcome given an individual intervention, $P(Y_{jk} \mid F_{Tjk} = do(T_{jk} = t'))$. The former would try to estimate the effect of a cluster-level intervention (usually the interest in social policy) and the latter would try to estimate the effect of an individual-level intervention (as could be the goal of many medical trials).

4.1 Cluster design networks

In terms of design decision strategies, a cluster-randomised study implicitly involves two design decisions: (1) the decision of clustering (i.e. to allocate the treatments to clusters of individuals) and (2) the decision of randomising (i.e. to perform the allocation using a random procedure).

When an intervention at cluster level occurs we distinguish between two cases. The first is related to the case in which the intervention affects all individuals in the cluster: for example, when the improvement of health services is undertaken at community level and all families within a community are subject to the same infrastructure. In this case, individuals within the cluster cannot choose not to be affected by the policy intervention. In this paper this case will be referred to as a 'pure-cluster' intervention (and denoted by $d_C = 1$). The second case refers to the situation in which, although an intervention is allocated at cluster level, not all individuals, but only a subset of them within a cluster, will be subject to the intervention. Actions, in this second case, are 'done' at individual-level to individuals within a cluster, and thus individual characteristics might have an influence in the individual's allocation of treatment

An example of the latter is when the intervention affects only eligible individuals. A cluster policy of this type could be seen as: 'all eligible individuals k in cluster j will receive policy t' ' via $A_{Tj} = do(T_j = t')$. So, allocation of policy is done at cluster level and two eligible individuals in the same cluster cannot be allocated different policies (contrary to what would happen if the policy allocation was done at individual level). In the case of Progresa, this will correspond to the case where only poor households are receiving extra money for nutrition and educational grants. These actions are aimed at household-level; however, not all households in a community are poor and therefore not all households within a community receive the same treatment, only the eligible ones. In a more general setting, the individual choice of treatment might depend on some possibly unobserved background variables and not necessarily only on some previously defined (and observed) eligibility criteria. For example, imagine that some health centres are allocated a certain restricted quantity of food supplements to be distributed among families visiting them, but the amount of food supplements is not enough to cover all families. Then, the fact that a family is receiving the food supplement or not could depend on the (unobserved) nurses' choice or on a first-come-first-served basis. In any case, when different units within a cluster do not necessarily receive the same treatment and this is dependent on certain individual-level background variables, the experiment will be referred as a 'non-pure' cluster allocation (and denoted by $d_C = 0$).

4.2 Effects of cluster allocation design decisions

A design network for the general cluster setting for a cluster-level future intervention F_{Tj} is presented in Figure 4. As we are discussing clusters of individuals, naturally variables are not all at the same level and a simple DAG cannot be used without further notation The levels are represented in the graph by squares, following Spiegelhalter's notation (see WinBUGS), meaning that the same graphical structure applies for each of the observations at the same level. Design decisions can then be taken at both individual and cluster level. The structure has been kept similar to that used above, but now we have the situation replicated for the two levels involved. Let cluster j (for $j = 1, 2, ...J$) have K_j units and let T_j and T_{jk} be variables for intervention status (Treatment / Control) at cluster and unit level respectively. Similarly, let B_j and B_{jk} represent the background variables at cluster and unit level and Z_j some recorded cluster-level covariates that might be affected by the policy. Nodes A_j and A_{jk} will correspond to the assignment mechanisms to allocate policy at cluster and individual levels respectively. Action $A_{Tj} = do(T_j = t^*)$ defined in A_j will imply $T_j = 1$ if the value t^* corresponds to the policy taking place in cluster j, and will imply $T_j = 0$ if the value t^* corresponds to 'control'. This will work similarly for the individual-level case. The recording mechanisms could be defined over the set of cluster and individual background variables, $R(B_j)$ and $R(B_{jk})$, respectively. The response Y_{jk} will correspond to that observed for individual k in cluster j.

The decision of running a 'pure' cluster allocation experiment ($d_C = 1$) will imply that the intervention is done equally to all members in the cluster. So, once the treatment for cluster j T_j is fixed by action $A_{Tj} = do(T_j = t^*)$, this fully implies actions $A_{Tjk} = do(T_{jk} = t^*)$, and so the values of the treatments T_{jk} for all K_j individuals in cluster j. Then $t_j = t_{jk} = t_{jk'}$ for all individuals $k, k' = 1, 2, ...K_j$ in cluster j. So, the effect of pure-clustering prohibits individual covariates from influencing the choice of treatment, breaking any links that could be present from B_{jk} (any background individual-level covariates) to T_{jk} in the graph. When 'non-pure' cluster allocation takes place ($d_C = 0$), although the experimental allocation is made at cluster level, individual k within cluster j might be receiving treatment or not depending on some individual-level covariates B_{jk}, and thus t_{jk} might differ from $t_{jk'}$ for $k \neq k'$.

The assignment mechanism nodes A_j and A_{jk} could be expanded. This is not done in Figure 4, to keep the (already complex) graph as simple as possible. The individual assignment mechanism A_{jk} is considered to be dependent on the actual policy that was allocated to cluster j, T_j, and (possibly) on some individual background variables. Thus, the action assigning policy t^* to individual k $A_{Tjk} = do(T_{jk} = t^*)$ is considered to be dependent on T_j and B_{jk} such that $\theta_{Tjk} = p\left(A_{Tjk} = do(T_{jk} = t^*) \mid T_j, B_{jk}\right) = q(T_j, B_{jk})$. The 'pure-cluster' case will imply that the individual assignment mechanism does not depend on individual background variables B_{jk} and
$\theta_{Tjk} = p\left(A_{Tjk} = do(T_{jk} = t^*) \mid A_{Tj} = do(T_j = t^*), d_C = 1\right) = 1$ and the following propo-

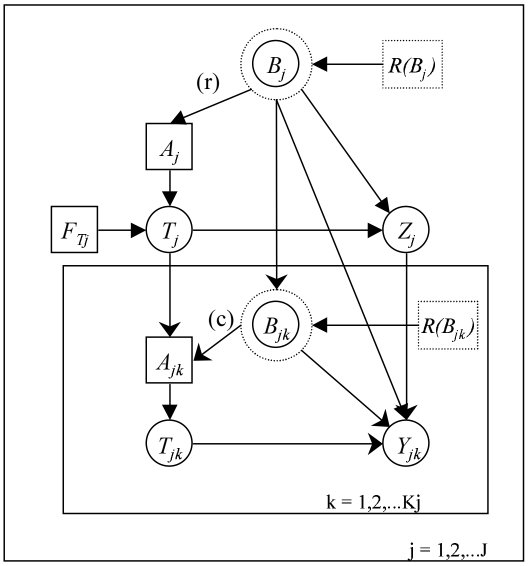


Figure 4: Design Network for cluster allocation and cluster-level future intervention F_{Tj}

sition is thus established.

Proposition (Pure-Clustering) *If the experimental design strategy d_E includes the action of performing a 'pure cluster' experiment such that $\{d_C = 1\} \in d_E$, then the 'structural' effect of $d_C = 1$ on the 'original' set of conditional independencies, is to introduce the set of conditional independencies $(T_{jk} \perp\!\!\!\perp B_{jk})_{d_C=1}$ that will hold on the data Δ_{d_E} generated by d_E.*

Figure 5 includes a close-up of the individual-level plateau in Figure 4, in which the design network has been extended for node A_{jk} and variables Z_{jk} introduced. Now let us refer to the situation when a 'pure' cluster experimental intervention is not feasible, but when we have a non-pure cluster experiment such that, within each cluster j, individual policy allocation follows a deterministic rule based on individual-level observed covariates Z_{jk}. The final policy allocated to an individual through A_{jk} will be a function of Z_{jk}, and any other possible influences on T_{jk} from background variables B_{jk} (other than Z_{jk}) are eliminated. The prevalences of T_{jk} in the experimental data available Δ will depend on the policy allocated to the cluster T_j and Z_{jk} but not on B_{jk}.(e.g. $\theta_{Tjk} = p\left(A_{Tjk} = do(T_{jk} = t^*) \mid T_j, Z_{jk}\right) = q(T_j, Z_{jk}))$ The structure obtained is similar to the stratified allocation presented in Madrigal (2005), and arrow (c) will be deleted when this 'deterministic' allocation takes place.

Proposition *If a 'non-pure cluster' experiment includes a design strategy in which policies at individual level are allocated following a 'deterministic' rule defined by the exper-*

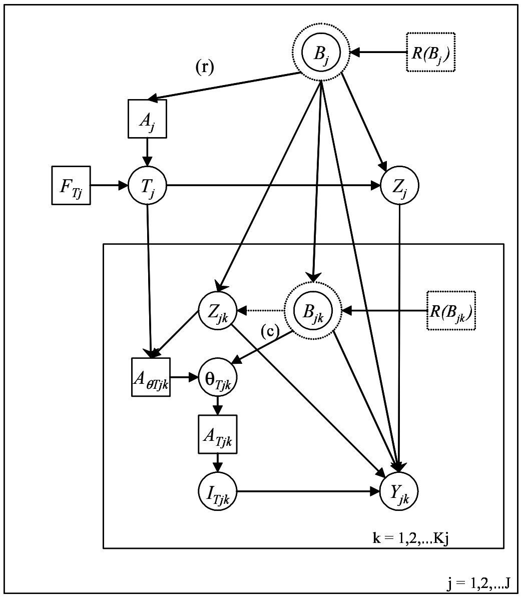


Figure 5: Close-up of individual-level plateau

imenter based on covariates Z_{jk} , then conditional independencies $(T_{jk} \perp\!\!\!\perp B_{jk} \mid Z_{jk})_{d_E}$ are introduced and will hold on the data Δ_{d_E} generated by d_E.

As introduced in Madrigal and Smith (2004), design decision strategy d_E including random allocation of policies (i.e. $A_{\theta_T} = do(\theta_T = \theta_T^*) \in d_E$) might, qualitatively speaking, modify the structure of the data we are to collect (see Appendix A). By allocating the treatments completely at random (i.e. $A_{\theta_{Tj}} = do(\theta_{Tj} = \theta_{Tj}^*)$) we ensure that the treatment received is independent of any background variables B_j that, otherwise, might have an influence on the policy assignment mechanism. Then, when random allocation takes place, arrow (r) from B_j to A_j in Figure 4 disappears and the conditional independence statement $(T_j \perp\!\!\!\perp B_j)_{d_E}$ holds. Again, this probability, θ_{Tj}^*, might depend on possible stratification observed variables. When randomising at cluster level we therefore ensure that the level of treatment that is received by cluster j is independent of the level received by cluster j' (i.e. knowing that cluster j was assigned intervention t^* does not give us any further information about the intervention group at cluster j').

4.3 Identifying cluster-level interventions

The appropriateness and consequences of different design decisions $d_E \subset D_E$ will depend on the goals of the experiment. The case in which the interest is in the effect of a cluster-level future intervention $F_{Tj} = do(T_j = t')$ on a cluster-level response Y_j will degenerate to the one-level case. When the interest lies in an individual-level response Y_{jk}, it can be seen in Figure 4 that, if d_E^* includes a random cluster allocation procedure and arrow (r) is not present, the conditional independencies $(Y_{jk} \perp\!\!\!\perp F_{Tj} \mid T_j)_{d_E^*}$ hold for all j,k. Thus, once the value of the policy assigned to the cluster T_j is known, learning whether the policy status T_j arose from the future policy implemented $F_{Tj} = do(T_j = t')$ or from the 'original' experimental allocation $A_{Tj} = do(T_j = t^*)$ when $F_{Tj} = \emptyset$, is irrelevant. Therefore, direct identifiability of the effect $F_{Tj} = do(T_j = t')$ on Y_{jk} holds and $p(y_{jk} \mid F_{Tj} = do(T_j = t'))$ can be directly obtained from data $\Delta_{d_E^*}$ available as long as $t' \in \{t^*\}$ such that

$$p(y_{jk} \mid F_{Tj} = do(T_j = t'); \Delta_{d_E^*}\ ) = p(y_{jk} \mid T_j = t'; \Delta_{d_E^*}\ )$$

This does not disregard the fact that individuals belonging to the same cluster will have a positive correlation, which must be taken into account in any model used for the analysis and estimation of the effect on an individual-level response.

Again, when non-random cluster allocation is performed as part of the experimental design d_E^*, $(Y_{jk} \perp\!\!\!\perp F_{Tj} \mid T_j)_{d_E^*}$ does not hold anymore, but the conditional independencies $(Y_{jk} \perp\!\!\!\perp F_{Tj} \mid T_j, B_j)_{d_E^*}$ hold for all j,k. Thus, the identifiability of the effect of a future policy $F_{Tj} = do(T_j = t')$ on Y_{jk} will depend on the recordability of cluster-background variables $R(B_j)$ and the causal effect will need to be obtained through an 'adjustment' procedure such that, as before, using the back-door criteria

$$p(y_{jk} \mid F_{Tj} = do(T_j = t')) = \int p(y_{jk} \mid T_j = t', B_j) p(B_j) dB_j.$$

Unless we are ready to assume some prior distribution for $p(B_j)$, the recording of variables B_j as part of the design $(\{R(B_j) = 1\} \in d_E^*)$ are needed to achieve an 'adjusted identification' of the causal effect.

Different recording mechanisms might assist identification. For instance, if we were ready to assume that cluster background variables did not have a direct effect on the individual response, such that arrow from B_j to Y_{jk} was deleted, then all the influence from B_j would be through individual background variables and the observed cluster-level variables Z_j. In this case, conditioning on T_j, B_{jk} and Z_j would be enough and a design able to record these variables will provide identifiability. Thus, if cluster background variables B_j were not accessible to the experiment, this new set of covariates $\{B_{jk}, Z_j\}$ could assist identification.

4.3.1 Bayesian hierarchical models

In a hierarchical setting, data within each cluster j is assumed to depend on parameters θ_j, which in turn are assumed to be drawn from some population distribution with parameters ψ. In an initial model, the response y_{jk} for individual k in cluster j is assumed to have a Normal distribution, such that

$$\begin{aligned} y_{jk} &\sim N(\mu_{jk}, \sigma^2) \\ \mu_{jk} &= u_j \end{aligned} \tag{1}$$

and cluster-specific random effects (u_j) are assumed to have a Normal distribution with mean ϕ_j and variance σ_u^2, such that

$$\begin{aligned} u_j &\sim N(\phi_j, \sigma_u^2) \\ \phi_j &= \alpha + \beta T_j \end{aligned} \tag{2}$$

where T_j represents the treatment given to the jth cluster. There are many potential elaborations to this basic model (see Spiegelhalter 2001; Turner et al. 2001). The priors that need to be specified for this model are $p(\alpha)$, $p(\beta)$, $p(\sigma^2)$, $p(\sigma_u^2)$. Making causal assumptions and a graphical representation of all influences present in the particular system analysed could assist recognition of possible confounders and thus assist both the experimenter's decisions for control via design and the analyst's decisions for control via analysis.

When cluster allocation is done randomly, if we are ready to assume linear relations, the two-level Bayesian hierarchical model as specified in equations (1) and (2) could be used to estimate the effect of F_{Tj} on Y_{jk}, and coefficient beta can 'safely' be given a causal interpretation as an 'overall' effect. For the non-random case, the analysis will need the conditioning on the 'relevant' background variables. The inclusion of individual-level and cluster-level covariates in the analysis could be done directly by including them in equations (1) and (2) respectively. The conclusions just derived hold for both 'pure' cluster and 'non-pure' cluster allocations. The 'overall' causal effect will correspond to a 'total effect' when a 'pure' cluster allocation ($d_C = 1$) is done. However, this will not be the case for $d_C = 0$, where the 'total effect' cannot be obtained. To make the difference between 'overall' and 'total' effects clearer, the interactions of individuals in a cluster have to be considered, and this is discussed below in Section 4.5.

4.4 Identifying individual interventions from clustered data

Consider the case where the main interest is in obtaining the individual-level causal effect, namely $P(Y_{jk} \mid F_{Tjk} = do(T_{jk} = t'))$ from data that is clustered. If randomised allocation could take place at individual level, then it could be directly identified from the experimental data Δ, as individual random allocation will break the possible influence of cluster background variables on the policy allocated to the individual. Suppose that it is not feasible to randomise at individual level, but to intervene clusters is possible. The design network for this case will basically coincide with that shown in Figure 4, but in this case we assume that the future policy will consist of an individual intervention F_{Tjk}, and that we are interested in identifying effects at the individual level.

From the design network in Figure 6 it can be seen that $(Y_{jk} \perp\!\!\!\perp F_{Tjk} \mid T_{jk})$ does not hold even if arrows (r) and (c) are deleted from the graph. So, the effect of policy intervention $F_{Tjk} = do(T_{jk} = t')$ on Y_{jk} cannot be identified directly from the data and some adjustment will be needed.

When a 'non-pure' cluster allocation takes place in the experiment, individual policy assignment will depend on both the policy allocated T_j and individual background

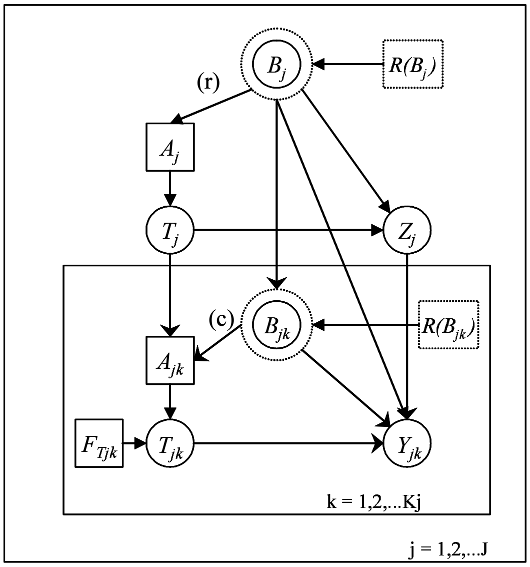


Figure 6: Design Network for cluster allocation and individual-level future intervention F_{Tjk}

variables B_{jk}. Conditioning on this set of variables, the irrelevance of F_{Tjk} is gained such that $(Y_{jk} \perp\!\!\!\perp F_{Tjk} \mid T_{jk}, T_j, B_{jk})$ Thus the recording of B_{jk} is needed in order to obtain adjusted identifiability through

$$p(y_{jk} \mid F_{Tjk} = do(T_{jk} = t'); d_C = 0) = \int p(y_{jk} \mid T_{jk} = t', T_j = t^*, B_{jk}) p(T_j = t^*, B_{jk}) dB_{jk} dT_j \ ,$$

where, if randomisation did not take place at cluster level, T_j and B_{jk} are not independent (both having B_j as an ancestor) and their joint distribution is needed. If recording at individual level for B_{jk} is not undertaken, the causal effect will be unidentifiable.

When 'pure' cluster allocation is done, and as a result arrow (c) is deleted, then there are no individual level confounders and all possible confounders will be at cluster-level. So, 'pure cluster' assignment might improve identification of individual intervention effects, in particular, when randomisation at cluster level is feasible or in the case when cluster-level confounders (B_j) are easier to observe and/or control than individual-level confounders (B_{jk}).

As shown above, the 'overall' effect of a future cluster-level intervention at cluster level F_{Tj} can be identified from the experimental data when policies in the experiment are allocated randomly at cluster level. Something similar happens when the assignment is not carried out randomly, but cluster background variables are recordable and an 'adjustment' measure is needed. Moreover, for $d_C = 1$, the overall effect will coincide with the participants' total effect. Thus, if the indirect effect due to interaction among neighbours is negligible, as could be the case when vitamin supplements are administered to children in Progresa, the total effect measured will be equal to the direct (personal) effect. Although this might not be always true for social policies, this might be the case for some treatments in a medical setting in which, for example, a drug is supposed to act in an individual regardless of his interaction with other people in the cluster. In this case, a 'pure cluster' design could assist identification of individual interventions. Therefore, a 'good' (randomised or controlled) cluster design might be able to provide more information than a 'bad' (with unobserved confounders) individual design. The distinction of overall, total, direct and indirect effects will be discussed next.

4.5 Overall versus total effects

As Koepsell (1998) states, 'just as infectious agents can be spread from person to person, transmission of attitudes, norms and behaviours among people who are in regular contact can result in similar responses'. So, when people interact or communicate, their response to an intervention can be explained (and partitioned) in terms of direct ('personal') effect and indirect ('neighbours') effect. So, interventions may affect the whole population, not just those who participate (or were subject to interventions).

The fact that all individuals in a group follow the same policy, or are encouraged to take a particular action, has thus an additional 'interaction effect'. This is so as individuals interact with each other, creating a domino effect. In the case of Progresa we have, for example, the fact that mothers talk! Thus, if a mother is encouraged to take children to the health centre for food supplements, besides her possible individual motivation, the fact that other mothers in the village are encouraged as well, creates an additional effect on her (i.e. if everybody is doing it, there is an extra motivation to do it, and being the only one not doing it will be rare and possibly socially penalised).

If in a cluster, not all individuals are allocated the same intervention, then the effects of interventions can be classified, following Hayes et al. (2000)'s definition, according to the 'intervention status of the individual' as participants (treated) or nonparticipants (controls). Those who participate receive both a direct $\left(DE_{(P)}\right)$ and an indirect effect $\left(IE_{(P)}\right)$, which combine to form the total effect $\left(\lambda_{(P)} = DE_{(P)} + IE_{(P)}\right)$. The non-participants receive only an indirect effect, $IE_{(NP)}$, so their total effect contains only those indirect effects $(\lambda_{(NP)} = IE_{(NP)})$. The indirect effects received by participants and non-participants may differ in magnitude, so an index is used to distinguish them:

	Participants (P)	Non-participants (NP)
Total effects	$\lambda_{(P)} = DE_{(P)} + IE_{(P)}$	$\lambda_{(NP)} = IE_{(NP)}$

If we are ready to assume that these effects are equal for all individuals in a cluster, then the overall effect observed in a cluster will correspond to the weighted average of the effects on participants and non-participants such that

$$\text{Overall effect} = w_{(P)}\lambda_{(P)} + w_{(NP)}\lambda_{(NP)} \tag{3}$$

where $w_{(P)}$ and $w_{(NP)}$ are just weights that will be functions of the number of participants and non-participants in the cluster (or in terms of the 'coverage' -% of participants- of the experiment). So, the overall effect will include a combination of direct and indirect effects. In particular, expression (3) could be extended to be re-written as

$$\begin{aligned}\text{Overall effect} &= w_{(P)}\left(DE_{(P)} + IE_{(P)}\right) + w_{(NP)}IE_{(NP)} \\ &= w_{(P)}DE_{(P)} + \left(w_{(P)}IE_{(P)} + w_{(NP)}IE_{(NP)}\right)\end{aligned}$$

In the case where indirect effects, for both participants and non-participants, are assumed to be negligible such that $IE_{(P)} \approx 0$ and $IE_{(NP)} \approx 0$, then the overall effect will be approximately proportional to the direct effect such that

$$\text{Overall effect} \approx w_{(P)}DE_{(P)}.$$

In the case of a 'pure cluster' experiment, either all individuals are participants (treated cluster) or all are non-participants (control cluster). In this situation, contamination within clusters is completely avoided, and in control clusters no intervention indirect effects are observed (i.e. $IE_{(NP)} = 0$). For the treated clusters, all individuals are participants, and overall intervention effect of the cluster will coincide with the total participant effect, denoted by $\tau_{(P)}$: namely,

$$\begin{aligned}\text{Overall effect (control cluster)} &= 0 \cdot \lambda_{(P)} + 1 \cdot \lambda_{(NP)} = \textit{total effect}_{(NP)} = 0 \\ \text{Overall effect (treated cluster)} &= 1 \cdot \lambda_{(P)} + 0 \cdot \lambda_{(NP)} = \textit{total effect}_{(P)} \quad = \tau_{(P)}\end{aligned}$$

and therefore

$$\text{Overall effect}_{(d_C=1)} = \text{total effect }_{(P)} = \tau_{(P)} = DE_{(P)} + IE_{(P)}$$

Individually randomised trials typically aim to measure the direct effect, $DE_{(P)}$. By contrast, CRTs measure the total effect $\tau_{(P)}$ if all individuals participate, but otherwise they measure the overall effect, which will vary according to intervention coverage and the characteristics of the population.

If individuals are naturally clustered, the magnitude of the indirect effect of an intervention is likely to be important in deciding whether a trial should be individually - or cluster- randomised. Indirect effects, due to interaction, will be included in the outcome measure. As a consequence, if the main interest is in measuring only the direct effect that a possible drug/treatment, say, has on an individual and it is known that indirect effects could be relevant, then CRTs might not be the best option as they will measure the overall effect instead of the direct effect.

In assessing the value of intervention it is important to take into account their indirect as well as direct effects. In some cases it may be better to avoid intervention if the coverage needed to make it beneficial is too high to be realistically achievable. In addition, it may be desirable to separate the overall effect into its direct and indirect components. Methods for measuring direct and indirect effects separately have mostly been developed in the context of vaccination (see Hayes et al. 2000; Longini et al. 1998). Standard CRT designs measure the overall effect of intervention, and this is often the most useful measure for policy makers because it includes all the components, both direct and indirect, which a population would experience if a cluster policy were to be implemented.

It should be clear that the stable-unit-treatment-value-assumption (SUTVA, as labeled in Rubin 1980), which implies that the response of the unit does not depend on which treatment was applied to other units, does not hold when units interact and the indirect neighbours effects are not negligible. However the Bayesian predictive decision-theoretic approach that is followed in this paper does not require this assumption, as would be the case in Rubin's counterfactual approach. The counterfactual model for causal inference could lead to ambiguities and pitfalls, as discussed by Dawid (2000).

5 Progresa effect example using hierarchical models

In this section a hierarchical model analysis based on Spiegelhalter (2001) is performed for the evaluation of cluster- and individual-level interventions based on Progresa data. In the programme, communities were randomly allocated either to a treatment or a control group. The community level interventions G_1 (such as the improvement of health services and educational talks) are received by all households in a 'treatment' community. In addition, all eligible (poor) households that belong to a 'treatment' community receive household interventions, such as financial support, G_2. The data recorded includes a census of eligible and non-eligible households for (treated and control) communities selected for the study.

Let T_j be the cluster treatment indicator, such that $A_{T_j} = do(T_j = 1)$ if community j was allocated to Progresa programme and $A_{Tj} = do(T_j = 0)$ if it was allocated to control, so we have

$$T_j = \begin{cases} 1 & \text{if community Treatment} \\ 0 & \text{if community Control} \end{cases}$$

Let E be an indicator variable denoting eligibility status. Then $E_{jk} = 1$ if household k in community j is eligible and $E_{jk} = 0$ if non-eligible. In Progresa, $E_{jk} = 1$ corresponds to a poor household. Thus,

$$E_{jk} = \begin{cases} 1 & \text{if `poor' household} \\ 0 & \text{if `non-poor' household} \end{cases}$$

So, household k in community j will be allocated household-level Progresa interventions T_{jk} (e.g. financial support) through an allocation A_{jk}, in which a household is given extra money if, in addition to belonging to a treatment cluster, the household is 'eligible'. It will not be given extra money if either it is not eligible or if it belongs to a control community. If we denote by P_{jk}, the indicator variable for 'Progresa participant', such that $P_{jk} = 1$ if household k in cluster j receives economical support and $P_{jk} = 0$ if not, then, P_{jk} is defined as

$$P_{jk} = \begin{cases} 1 & \text{if } T_j = 1 \text{ and } E_{jk} = 1 \\ 0 & \text{otherwise} \end{cases}$$

From the general formulation of the Design Network for cluster allocation presented above, a simplified version of Progresa's experimental design containing the main features is shown in Figure 7, where Y_{jk} represents the response of household k in community j for $k = 1, 2, ...K_j$; B_j represents the background variables that are shared by all individuals in community j. Background variables at individual level were not added to keep the graph simple, but could be easily incorporated. As allocation at cluster level was done randomly in Progresa, no arrow is drawn from B_j to T_j. If G_1 denotes Progresa's cluster-level intervention (action) corresponding to health services and talks (i.e. the 'encouragement' that communities receive to improve their nutrition) and G_2 denotes Progresa's household-level action of giving financial support to poor households, then note that the action $do(T_j = 1)$ will trigger both atomic cluster-level intervention G_1 and contingent (on T_j and E_{jk}) individual-level intervention G_2.

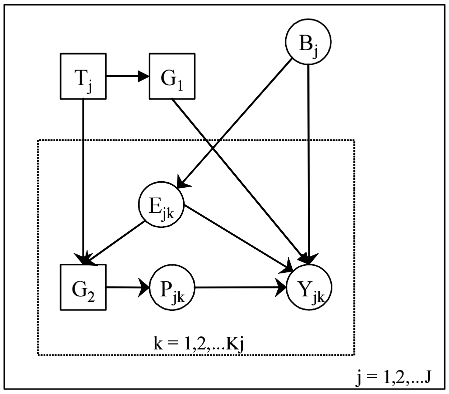


Figure 7: Progresa experimental Design Network for basic nodes

5.1 Cluster-level intervention effect

Imagine that we are interested in the effect of Progresa on the total food consumption y_{jk}, measured in terms of the amount of money spent on food in a household. The complete data set, including poor and non-poor households, consists of 20,589 households in 500 clusters. To begin with, assume we are interested in measuring the overall effect of Progresa intervention $F_{Tj} = do(T_j = t')$. A hierarchical model following the setting presented in Section 4.3 (equations (1) and (2)) is used to estimate this effect. This model was run in BUGS using vague priors following Spiegelhalter (2001) with

$$\begin{aligned} p(\alpha) &\sim Uniform(-10000, 10000) \\ p(\beta^*) &\sim Uniform(-10000, 10000) \\ p(\sigma^{-2}) &\sim Gamma(0.01, 0.01) \\ p(\sigma_u^{-2}) &\sim Gamma(0.01, 0.01) \end{aligned}$$

We encountered no difficulties in convergence of this model. The analysis is based on a sample of 10,000 iterations following a burn-in of 5,000.

The posterior inference (means and 95% intervals) for the parameters involved is presented in Table 1. As can be seen from the table, the posterior mean for beta

Parameter	Mean	95% interval
β^*	39.48	$(20.36, 58.33)$
α	444.9	$(435.2, 454.6)$
σ^{-2}	$2.102E-5$	$(2.061E-5, 2.143E-5)$
σ_u^{-2}	$1.054E-4$	$(1.052E-41, .216E-4)$

Table 1: Posterior distributions of parameters for cluster-level intervention

is $E\left[\beta^* \mid \Delta\right] = 39.48$ with a 95% interval of (20.36,58.33). In this case β^* gives the cluster-level total overall effect of Progresa intervention on the food expenditure in a household. So, β^* contains a summary of the effects of the programme (through G_1 and G_2) on Y for all the population in a cluster, by averaging participants (receiving G_1 and G_2) and non-participants (only receiving G_1). Depending on the aims of the study this total overall effect might be the relevant causal effect of interest. In such a case, it could be said that the causal effect of Progresa is to increase, on average, the food expenditure of a household by 39.48 Mexican Pesos. This in relation to the average food expenditure for a household in a control community that will be of $E\left[\alpha \mid \Delta\right] = 444.9$ Mexican Pesos.

5.2 Individual-level effect

Now imagine that we are interested in obtaining an estimate of the 'causal' effect of the individual-level intervention $F_{G_2} = do(G_2 = g_2' = q(poor))$ of giving financial support to poor households. The allocation of G_2 depends on the cluster-allocated policy T_j (Treatment/Control) and on the eligibility condition E_{jk} of a household defined as 'poor': both are assumed to have an effect on the household expenditure level and thus act as confounders in this case. So, to identify the individual-level effect of F_{G_2}, it is needed to control by including these two confounders in the analysis.

The hierarchal model used for the cluster-level effect above can be extended to include covariates T_j and E_{jk} at cluster and individual level respectively. The 'Progresa participants' status of a household P_{jk} acts as an indicator variable of the presence of economic support provided by G_2. We include here the household size Z_{jk} as an individual covariate to illustrate the possible inclusion of other covariates in the model. Household size will have an influence on the total expenses of the household Y, and it is neither affected by the policy nor affecting (at least directly) policy allocation. So now this is considered part of the white noise (with respect to T and P) at the recorded individual level. Equations (1) and (2) can be substituted by

$$\begin{aligned} y_{jk} &\sim N(\mu_{jk}, \sigma^2) \\ \mu_{jk} &= u_j + \beta_2 P_{jk} + \delta E_{jk} + \gamma Z_{jk} \end{aligned} \tag{4}$$

$$\begin{aligned} u_j &\sim N(\phi_j, \sigma_u^2) \\ \phi_j &= \alpha + \beta_1 T_j \end{aligned} \tag{5}$$

We chose to use the same priors to estimate this model, as before. Thus all the coefficients (namely $\alpha, \beta_1, \beta_2, \delta$ and γ) were given vague uniform distributions a priori. Again, we encountered no difficulties in obtaining convergence and the sample simulated is the same size as before.

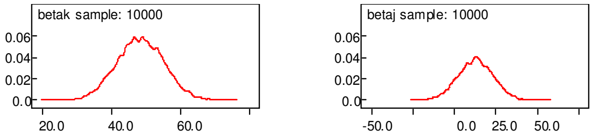
The posterior means and 95% intervals for all the parameters are given in Table 2. The individual-level effect here will be measured by the coefficient of P_{jk}, namely β_2, whose posterior mean is given by $E[\beta_2 \mid \Delta] = 45.71$ with a 95% interval of (33.95, 57.49). In general β_2 will isolate the effect of G_2 (from the effect of G_1) and we could say that the effect of a policy F_{G2} that provides economic support according to the poverty level of a household will increase, on average, the food expenditure of a participant household by 45 Mexican Pesos (regardless of the presence or not of a secondary action G_1). However, β_2 implicitly includes possible indirect effects resulting from the interaction of participant households with non-participant households in a community.

Parameter	Mean	95% interval
α	473.5	(461.7 , 485.5)
β_1	16.7	(-3.848 , 36.63)
β_2	45.71	(33.95 , 57.49)
δ	-34.14	(-43.54 , -24.69)
γ	30.56	(29.54 , 31.6)

Table 2: Posterior distributions of parameters for individual-level intervention

A second reading of this analysis could consider the case in which the total effect of Progresa (defined by G_1 and G_2) is split in its effect on food expenditure, due to the community-level action of educational talks (G_1) and the economic support provided at household level to poor people (G_2). Then, β_1 becomes a parameter of interest containing the direct effect of G_1 (i.e. the effect of Progresa on food expenditure that is not due to economic support) and P_{jk} is regarded as an intermediate variable. In this case, and following the reasoning of path analysis (Bollen 1989; Pearl 2000), we can see that the total overall effect of Progresa β^* could be written as $\beta^* = \beta_1 + \lambda\beta_2^*$ where λ will contain information about the prevalence of participants within a treated community. The total overall effect at household level is here denoted by β_2^* $(= \beta_2 + \delta b_{(E)P})$ where, as before, β_2 represents the direct individual-level effect and $\delta b_{(E)P}$ the confounding effect, which in this case has been controlled via analysis.. In this case it can be seen that, although the posterior mean for β_1 has a value of 16.7, the 95% posterior interval includes the value zero. So we cannot assert that the direct effect of the cluster-level intervention G_1 was different from zero. A more careful analysis, possibly including more 'white noise' covariates at cluster level, might provide narrower intervals for the coefficients. However, given that this response variable is measured in money terms, the main effect of the programme could be expected to be due to the increase of income of the participant households derived from G_2. This might not be true for other response variables.

We can notice that one could be tempted to offer a causal interpretation to the



(a) *(b)*

Figure 8: Posterior distribution for *(a)* β_2 and *(b)* β_1

coefficient of eligibility (δ) that distinguishes between poor and non-poor households. However, the data available was not properly selected in order to isolate the relationship between poorness and food expenditure. The level of poorness in Progresa was obtained through a discriminant function that depends on many household level covariates that could also affect the response, becoming confounders for this coefficient. Therefore, if we are interested in interpreting δ, a data set including these covariates will be needed. In this analysis, household size is known to be part of the variables used to define the poorness of a household. Its inclusion or not in the model (analysis not shown), although 'transparent' for β_1 and β_2 (given E_{jk}), will have an important effect on the posterior mean $E\left[\delta \mid \Delta\right]$. Although, in this case the posterior mean seems to be significantly different from zero and has the 'correct' sign (i.e. it should be expected that poor people spend less money than non-poor people), this superficial conclusion might still be subject to unrecorded confounding.

6 Conclusion

The primary contribution of this paper is to expand on Dawid (2002)'s model for causality reasoning within the Bayesian decision-theoretic framework: to 'adapt' it to policy analysis, to include experimental nodes, to allow intervention nodes 'do' parameters nodes, to discuss the relevance (or irrelevance) of experimental design and to include interventions at different levels (clusters) of units. Observational data is considered a degenerate type of experimental data. In addition, there was a need to create some notation to describe the mechanisms derived from choices, such choices as the experimenter might make when choosing the 'lens of the camera' to picture the world. These choices affect the characteristics (units, variables and distributions) of the database. The inspection of influence diagrams and, in particular, the augmented DAGs derived from them, has been shown to be useful to decide if the data available is sufficient for obtaining consistent estimates of the target causal effect of policy intervention $F_T = do(T = t')$. If so, we can derive a closed-form expression for the target quantity in terms of distributions of available quantities. If it is not sufficient, this framework can help suggest a set of observations and experiments that, if performed, would render a consistent estimate feasible. Design Networks expand the IDs framework to address explicitly experimental design and provide the semantics to discuss how design can assist identification, and when and how one can identify causal effects. Incorporating nodes for experimental design decisions is useful in demonstrating their impact on the graphical structure and on the 'data structure' derived from it. Certain policy assignment mechanisms, such as randomised cluster allocation, will add 'extra' independencies to the ID, defining a new collection of conditional independencies. To make a causal inference of F_T, it is important that we consider the mechanisms producing the data (d_E). Furthermore, we need to differentiate between policy interventions we want to evaluate, $F_T \in D_T$, and experimental allocations, $A_T \in D_E$. The relevance of D_E in assisting the identification and comparison of different mechanisms d_E^* in terms of identifiability can be addressed using DN for diverse types of assignment. Design networks were introduced for cluster allocation, and Spiegelhalter (2001)'s Bayesian hierarchical model was extended to include causal interpretation and used to illustrate a causal analysis of a simplified version of the Progresa programme.

When cluster allocation is done randomly, two-level Bayesian hierarchical models could be used to obtain the effect of F_{Tj} on Y_{jk} and the relevant coefficient can 'safely' be given a causal interpretation as an 'overall' effect. For the non-random case the analysis will need the conditioning on the 'relevant' background variables. Cluster allocation might help identifying individual-policy effects in certain cases.

Design of experiments within the Bayesian decision theoretic approach has been studied broadly in the literature; however, not in terms of causal reasoning and identifiability. In most of the literature on Bayesian experimental design the discussions have been limited to a) a set of options which include the choice of the levels of treatment and the number of repetitions within each level, and b) to a utility function usually defined in terms of minimising the posterior variance of estimates (or maximising entropy), which is an important issue. However, an experiment that overlooks identification could lead to the wrong conclusions if causal analysis is of interest. Causal inference imposes an extra criterion for the evaluation of the designs. This work extends the on-going discussion to a more general setting where the set of options is extended to include decisions about policy allocation and recording mechanisms and where the utility function is allowed to include a measurement of identifiability.

Appendix A. Design Networks: General remarks

By allowing the 'idle' system in Dawid (2002) to refer to any experimental system, the list of propositions in this section could be derived directly or are analogous to the results presented in Dawid (2002).

In general, we will say that the 'causal' effect of T on Y is identifiable *directly* from the available (experimental) data collected under $D_E = d_E$, if learning the value of F_T (i.e. learning if the future policy was set to a value or left to vary 'naturally') does not provide any 'extra' information about the response variable Y given the value of T and experimental conditions $D_E = d_E$ (*i.e.* if $(Y \perp\!\!\!\perp F_T \mid T)_{D_E}$) then $p(y \mid t', F_T = do(T = t'); D_E = d_E) = p(y \mid t', F_T = \emptyset; D_E = d_E)$. Note that this will hold (or not) regardless of $R(B)$.

Definition (Direct identifiability) *The 'causal' effect of* T *on* Y *is identifiable* directly *from available (experimental) data collected under $D_E = d_E$, if $(Y \perp\!\!\!\perp F_T \mid T)_{d_E}$. Then*

$$p(y \mid F_T = do(T = t'); D_E = d_E) = p(y \mid t'; D_E = d_E).$$

This will imply that the conditional distribution $p(y \mid t')$ that is extracted from data generated according to d_E can be used directly to estimate the target causal effect $p(y \mid F_T = do(T = t'))$ regardless of the recordability or the actual values of B.

Let $d_{E1} = \{A = a_1; R(B) = r_1\}$ and $d_{E2} = \{A = a_2; R(B) = r_2\}$ be two experimental design interventions.

Proposition *If two experimental DNs under allocations defined by $A = a_1$ and $A = a_2$ share the same conditional independencies $S_{a_1} = S_{a_2}$, and $(Y \perp\!\!\!\perp F_T \mid T)_a$ holds for $a = a_1, a_2$ then experiments d_{E1} and d_{E2} share the same 'direct identifiability' status for the causal effect of T on Y defined by intervention F_T for any recording mechanisms r_1 and r_2 Thus, the choice of assignment mechanism (between a_1 and a_2) is said to be irrelevant to obtaining direct identification.*

For instance, if a_1= pure random allocation with probability θ_1^* and a_2= pure random allocation with probability θ_2^*, such that $0 < \theta_T^* < 1$ for all t^*, both assignments lead to direct identifiability. Then, regardless of the background variables recorded, the choice between a_1 and a_2 is irrelevant for identification purposes. Both allocations might be different in terms of a balanced sample and the variance and efficacy of the estimates, but this is regarded as a secondary goal of the choice of experiment.

Proposition *If direct identifiability holds for a_1, i.e. $(Y \perp\!\!\!\perp F_T \mid T)_{a1}$, but not for a_2, then the choice between $d_{E1} = \{a_1, r\}$ and $d_{E2} = \{a_2, r\}$ is not irrelevant for direct identifiability.*

An example of this is when a_1 = pure random allocation and $a_2 = \emptyset$. Although naturally it could be observed that $(T \perp\!\!\!\perp B)_\emptyset$ holds, direct identifiability will usually not hold for $a_2 = \emptyset$. So, the choice between performing a randomised experiment and observing the original mechanism is not irrelevant for the isolation of effects and their direct identification.

Direct identifiability of the causal effect implies assuming $(Y \perp\!\!\!\perp F_T \mid T)_{d_E}$, which is a very strong assumption that usually will not hold when observational studies or imperfect experiments take place. However, we might be ready to assume that for a set $B^* \subseteq B$ where $B^* \perp\!\!\!\perp F_T$, conditional on B^* the learning of F_T is irrelevant for the response, such that $(Y \perp\!\!\!\perp F_T \mid T, B^*)_{d_E}$ and then

$$p(y \mid t', B^*, F_T = do(T = t'); D_E = d_E) = p(y \mid t', B^*, F_T = \emptyset; D_E = d_E)$$

so we could 'substitute' the future intervened probability with the 'natural experimental' distribution available from the data.

Definition (Conditional identifiability) *The 'causal' effect of T on Y* conditional *on B^* is identifiable directly from the available (experimental) data collected under $D_E = d_E$, if $(Y \perp\!\!\!\perp F_T \mid (T, B^*))_{d_E}$. Then*

$$p(y \mid F_T = do(T = t'), B^*; D_E = d_E) = p(y \mid t', B^*; D_E = d_E).$$

Notice that conditional identifiability alone does not imply that procedures like the back-door formula can be used to calculate the overall effect of T on Y, which needs condition $(B^* \perp\!\!\!\perp F_T)_{d_E}$ to hold as well.

Proposition *If direct identifiability does not hold for a_1, $d_{E1} = \{a_1, r_1\}$, then the choice of the recording mechanism $R(B) = r_1$ in the experimental design defined by $d_{E1} = \{a_1, r_1\}$, is relevant for obtaining 'adjusted' identifiability.*

When identifiability cannot be obtained directly from the data defined by D_E, identifiability can still hold for a particular configuration of $R(B)$. Then, we say that the causal effect is identifiable through an '*adjustment*' procedure, and this leads to another definition.

Definition (Adjusted identifiability) *The 'causal' effect of T on Y is identifiable through an* 'adjustment' *procedure if*

$$p(y \mid t', F_T = do(T = t'); D_E = d_E) = h(y, t', B^* \mid D_E = d_E)$$

such that $R(B^*_q) = 1$ for all $B^*_q \in B^* \subseteq B$ and h is a function of known probabilistic distributions of recorded variables under d_E.

If we had a complete picture of the systems, then we could observe all background variables B and their influences and no unobserved or latent variables would exist. Then, $R(B^q) = 1$ would be plausible for all q and we would always be able to find a combination $R(B^*)$ such that $p(y \mid F_T = do(T = t'); R(B^*))$ would be identifiable through an adjustment procedure. However, our vision as experimenters willing to collect data is much narrower and is restricted to a partial view in which not all background variables are accessible and not all settings r are accessible. Nevertheless, we can still choose among different settings of $R(B)$. The design network representation permits us to evaluate identifiability for different choices of the recording mechanism $R(B)$. In consequence, it could assist the experimenter to choose among a possible set of recording settings, r, in order to assist identification of the effect of interest. In a first raw classification, recording mechanisms could be classified into those for which adjusted identifiability holds (h exists) and those for which the effect remains unidentifiable. Different recordings might have further consequences in the inference of causal effect; however, in terms of identifiability, the choice between two recordings that ensure adjusted identifiability is irrelevant. Thus we have,

Proposition *Let $d_{E1} = \{a, r_1\}$ and $d_{E2} = \{a, r_2\}$ be two experimental conditions such that direct identifiability does not hold for the policy assignment mechanism defined by $A = a$, and where r_1 and r_2 represent recording mechanisms in which collections B_1^* and B_2^* are recorded respectively. If functions h_1 and h_2 of known probabilistic distributions can be found for both recordings r_1 and r_2, then d_{E1} and d_{E2} are said to be equivalent for adjusted identifiability and the choice between recording mechanisms r_1 and r_2 is irrelevant for identifiability.*

In the case where neither h_1 nor h_2 can be found, the choice of r_1 and r_2 is also irrelevant, but in this case both recordings produce non-identifiability. However, when h_1 exists, but h_2 does not, then d_{E1} and d_{E2} do not share the same identifiability status, as the target causal effect of future intervention F_T can be obtained through an adjustment procedure for d_{E1} but it is not identifiable under d_{E2}.

When 'adjustment' is needed, some closed-forms for the function h have been given. The 'back-door' criterion (Pearl 1993), the 'front-door' formula (Pearl 1995) and the 'G-computation' formula (Robins 1986) are examples of criteria and formulae that imply the use of background variables to obtain 'adjusted' estimates and are all particular cases of functions h. A broader discussion of these criteria under different approaches can be found in Pearl (2000), Dawid (2002) and Lauritzen (2001).

If we can assume we are in a situation represented by a system in which potential confounders exist, pure random allocation will provide a data-generating mechanism that ensures direct identifiability of the effect of interest. In this case, we are performing *control via design* of the potential confounders that might be affecting the choice of policy, like politicians' preferences to benefit some particular communities. A generating mechanism that can only provide identifiability through an 'adjustment' formulation will correspond to a situation in which potential confounders have to be *controlled via analysis*.

Even if functions h_1 and h_2 can be found for d_{E1} and d_{E2} and adjusted identifiability can be obtained, further considerations are necessary when choosing an experiment. If r_1 and r_2 are such that $B_1 \subseteq B_2$ then d_{E1} will be generally preferred to d_{E2}, as recording a larger data set implies a more costly implementation and storage. In this sense, we would like the set of recorded variables to be minimal, but sufficient for identifiability. Definitions of sufficient sets have been made (see Lauritzen 2001; Dawid 2002; Pearl 2000). Functions h_1 and h_2 might be found for sets $B_1 \neq B_2$ where neither of them is a subset of the other. In any case, functions h_1 and h_2 might not have the same form and particular estimates might not be equally efficient when derived from h_1 than when derived from h_2, reflecting the loss of information associated with our restricted partial views determined by r_1 and r_2. An example of this, for the front-door formulation, can be found in Lauritzen (2001).

When direct or adjusted identifiability holds, the design d_E is ignorable. However, as Rubin (1978) notes, not all ignorable mechanisms can yield data from which inferences for causal effects are insensitive to prior specifications. Direct identifiability gives a situation where effects are insensitive to the specification of prior distributions of the data. However, this will not hold for adjusted identifiability where the causal effect is dependent on the prior distribution of background variables, $P(B)$.

A.1 The positivity condition

In general, the set $\{t^*\}$ of intervention-values assigned through A_T is not necessarily the same as the set of future-policy-values $\{t'\}$ defined by F_T. In order to be able to evaluate the causal effects of intervention $F_T = do(T = t')$, we need treatment t' to be observed under experimental conditions d_E. So, we need $p(t' \mid B^*, F_T = \emptyset; D_E = d_E) > 0$. This requires that treatment assignment mechanism A_T includes t' as one of its allocated values. In other words, this requires that $t' \in \{t^*\}$. In a prospective study, this condition will usually hold. However, when data has been already collected, we might face the case where $t' \notin \{t^*\}$. In this case, we would only be able to use the data available if we could make some parametric assumptions for $p(Y \mid T, \cdot)$ before the policy effects can be identified. In general, if all the relevant information needed to evaluate the causal effect is encoded in a function $\tau(\eta)$ of η, and the experimental data provides us with information about $\lambda(\eta)$, it will suffice if $\tau(\eta) \subseteq \lambda(\eta)$. In this case, predictively,

$$p(y \mid F_T = do(T = t'); d_E) = \int_\eta p(y \mid F_T = do(T = t'), \tau(\eta))p(\lambda(\eta) \mid \Delta_{d_E}; d_E)d\eta.$$

If different policies represent categorical variables, this could be difficult. In the FS example, imagine the two supplements provided in the experiment through assignment $A_T = do(T = t^*)$ are from different brands, say brand A (t^*_A) and brand B (t^*_B), and the future policy consists of providing a food supplement from brand C (t'_C). The data available, no matter how the actual assignment was made (random or not), will hardly be useful to conclude anything about the effect of food supplement C. However, imagine, that we have a measure in terms of the calorie intake that each supplement provides and that $t^*_A = 100$ kcal and $t^*_B = 300$ kcal, and we know that supplement C has 200 kcal, then if we are ready to assume that η contains a summary of the effect on weight per each increase of one kcal, then we would be able to estimate its effect.

A.2 Choice of experimental design

The problem of choosing an experiment has been set in Bayesian decision theory using decision trees (see Lindley 1971; Bernardo and Smith 1994). A DN could be viewed as its corresponding ID, allowing us to represent influences between decisions and random nodes. Optimality can be defined in various ways, and qualities for the distributions of estimators, such as minimum variance, are desirable (see Chaloner and Verdinelli 1995). Here we focus on the isolation of the target causal effect and thus on its identification. The efficacy of experimental design interventions D_E could then be measured in terms of making the (causal) effects of $F_T = do(T = t')$ identifiable and then two (or more) experiments can be compared in these terms, and among the experimental decisions D_E we choose the one with highest utility. ‘Pure’ (i.e. non-stratified) individual random allocation contrasted with the ‘no experiment’ choice (i.e. observational data) is used to introduce this procedure. When the policy assignment is done through random allocation, two control actions are performed: randomisation and intervention. So treatment t^* is done, $A_T = do(T = t^*)$, according to a probability distribution θ_T^* totally fixed and controlled by the experimenter through $A_{\theta_T} = do(\theta_T = \theta_T^*)$. Node A might be expanded to show explicitly the mechanisms underlying the assignment and the new independencies that might be introduced. This expansion involves parameter and intervention nodes that are included in an *augmented-extended design network*.

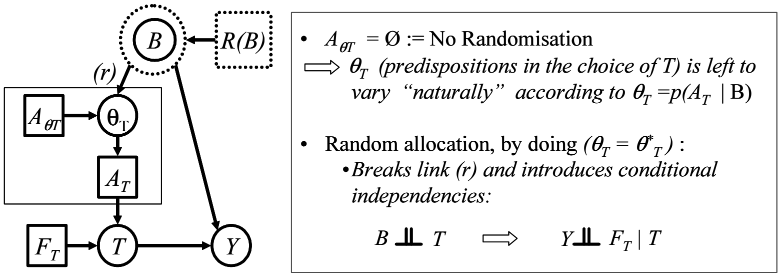
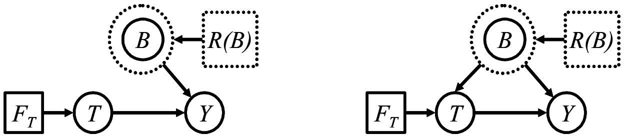

B
R(B)
(r)
$A_{\theta T}$
θ_T
A_T
F_T
T
Y
• $A_{\theta T}$ = Ø := No Randomisation
⟹ θ_T (predispositions in the choice of T) is left to vary "naturally" according to $\theta_T = p(A_T \mid B)$
• Random allocation, by doing $(\theta_T = \theta^*_T)$:
•Breaks link (r) and introduces conditional independencies:
$B \perp\!\!\!\perp T \Longrightarrow Y \perp\!\!\!\perp F_T \mid T$

Figure 9: Augmented - Extended DN for random allocation

Experimental actions 'do' parameter nodes. Random allocation breaks the link (r) and therefore two experimental structures arise from this choice. For each design strategy $d_E^* \subset D_E$ taken we can obtain an *experimental DN* from which independencies can be easily read. These experimental DNs define the data structure or data pattern.



Experimental DN for Random Allocation

Experimental DN for No Random Allocation

Figure 10: Experimental DN

Imagine we establish that the utilities associated with obtaining direct identifiability, adjusted identifiability and unidentifiability are given by U_D, U_A and U_U respectively. Then for the pure random allocation vs observational case, the four possible combinations of $(A, R(B))$ are shown in Table 3. Both experimental decisions that include random allocation, $A_{\theta_T} = do(\theta_T = \theta_T^*)$, have the same utility associated in terms of identifiability and are equivalent in these terms. However, performing an experiment (randomising and/or recording) will typically involve an associated cost that is not included here. The fact that $U_D \neq U_A$ (and actually we consider $U_D > U_A$) is due to the fact that the recording of B will increase the cost of the experiment and that $p(y \mid F_T = t', d_E = 3)$ is sensitive to the specification of prior distributions of the data. The choice of θ_T^* could have an effect on the efficacy of the estimators as it could affect the balance of the experiment, but the actual value θ_T^* does not affect the graph independence structure and the identifiability status derived from it.

	Experimental Decisions		Design Consequence	Utility(d_E)
D_E	A_{θ_T}	$R(B)$	$p(y \mid F_T = t'; d_E)$	U
1	random	1	direct identifiability	U_D
2	random	0	direct identifiability	U_D
3	$\emptyset$	1	adjusted identifiability	U_A
4	$\emptyset$	0	No identifiable	U_U

Table 3: Choice of experimental design for pure random example

A.3 An influence diagram for policy analysis

Figure 11 shows an influence diagram of the (simplified version of the) complete system for policy analysis. As before, the policy variable is denoted by a decision node T that has been augmented to make explicit policy intervention F_T. When the policy is defined through policy intervention decisions D_T, it can contain a collection of actions G that are triggered when intervention $F_T = do(T = t')$ takes place. Actions G can be contingent on a set of observed variables Z and are children nodes of T. The definition of possible structures and correspondent formulae for the calculation of the overall effect of intervention F_T in Y through actions G have been discussed in Madrigal (2004). The policy assignment mechanisms are contained in decision node A, which could be influenced by some background variables B. Both, the policy assignment mechanisms, A, and the recording mechanisms of B, $R(B)$, are defined as part of the experimental decisions D_E.

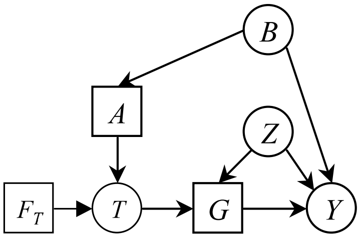


Figure 11: Complete ID for policy analysis

This simple structure shows how the two sets of decisions (namely, policy intervention decisions D_T and experimental decisions D_E) could be represented in the same graph. A more realistic graph should include some possible links between the background variables B and the variables Z in which actions G are contingent on, and possibly some type of influence of Z in the policy assignment mechanism. It is important to use the policy makers' expertise and knowledge to be able to represent in the influence diagram a most accurate version of the 'real' system with all possible influences. This will assist our causal inferences conclusions and help the choice of actions.

References

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons. 584

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons. 578

Chaloner, K. and Verdinelli, I. (1995). "Bayesian Experimental Design: A Review." *Statistical Science*, 10: 273–304. 562, 584

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. New York, NY: Springer-Verlag. 560

Dawid, A. P. (1979). "Conditional Independence in Statistical Theory." *Journal os the Royal Statistical Society (series B)*, 41: 1–31. 561

— (2000). "Causal Inference Without Counterfactuals." *Journal of the American Statistical Association*, 95(450): 407–424 (C/R: p424–448). 559, 575

— (2002). "Influence diagrams for causal modelling and inference." *International Statistical Review*, 70(2): 161–189. 557, 558, 559, 560, 561, 563, 565, 579, 580, 583

Hayes, R. J., Alexander, N. D. E., Bennett, S., and Cousens, S. N. (2000). "Design and Analysis Issues in Cluster-randomized Trials of Interventions against Infectious Diseases." *Statistical Methods in Medical Research*, 9(2): 95–116. 573, 574

Heckerman, D. and Shachter, R. (2003). "Discussion in Pearl, J. Statistics and Causal Inference: A Review." *Test.*, 12(2): 327–331. 560

Howard, R. A. and Matheson, J. E. (1981). "Influence Diagrams." *Principles and Applications of Decision Analysis. Menlo Park, CA.* 559

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs.*. New York: Springer. 559

Koepsell, T. P. (1998). "Epidemiologic Issues in the Design of Community Intervention Trials." In Brownson, R. C. and Petitti, D. B. (eds.), *Applied Epidemiology, Theory and Practice*. New York: Oxford University Press. 573

Lauritzen, S. L. (2001). "Causal Inference from Graphical Models." In Barndorff-Nielsen, O. E., Cox, D. R., and Klüppelberg, C. (eds.), *Complex stochastic systems*, 63–107. Chapman & Hall Ltd. 558, 561, 583

Lindley, D. V. (1971). *Making Decisions*. Chichester: Wiley-Interscience. 584

Longini, J., Ira M., Sagatelian, K., Rida, W. N., and Halloran, M. E. (1998). "Optimal Vaccine Trial Design When Estimating Vaccine Efficacy for Susceptibility and Infectiousness from Multiple Populations." *Statistics in Medicine*, 17: 1121–1136 (Corr: 1999 V18 p890). 574

Madrigal, A. M. (2004). "Evaluation of Policy Interventions under Experimental Conditions Using Bayesian Influence Diagrams." Ph.D. thesis, University of Warwick,UK. 563, 586

— (2005). "Design Networks, Policy Interventions and Causal Inference." Technical Report 448, University of Warwick,UK. 568

Madrigal, A. M. and Smith, J. Q. (2004). "Causal Identification in Design Networks." In R. Monroy, E. E. (ed.), *MICAI 2004: Advances in Artificial Intelligence, (Conference Proceedings)*, 517–526. LNAI 2972, Springer-Verlag. 557, 559, 569

Oliver, R. M. and Smith, J. Q. (1990). *Influence Diagrams, Belief Nets and Decision Analysis*. John Wiley & Sons. 559

Pearl, J. (1993). "Comment on "Graphical Models"." *Statistical Science*, 8: 266–269. 559, 561, 583

— (1995). "Causal Diagrams for Empirical Research." *Biometrika*, 82: 669–688 (Disc: p688–710). 559, 583

— (2000). *Casuality: Models, Reasoning, and Inference*. Cambridge University Press. 558, 559, 561, 578, 583

Robins, J. (1986). "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period - Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modeling*, 7: 1393–1512. 559, 583

Rosenbaum, P. (2002). *Observational Studies.*. New York: Springer Verlag. 562

Rubin, D. B. (1978). "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics*, 6: 34–58. 558, 583

— (1980). "Comments on "Randomization Analysis of Experimental Data: The Fisher Randomization Test"." *Journal of the American Statistical Association*, 75: 591–593. 575

Spiegelhalter, D. J. (2001). "Bayesian Methods for Cluster Randomised Trials with Continuous Responses." *Statistics in Medicine*, 20: 435–452. 559, 571, 575, 580

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. Cambridge, MA: MIT Press. 559

Turner, R. M., Omar, R. Z., and Thompson, S. G. (2001). "Bayesian Methods of Analysis for Cluster Randomised Trials with Binary Outcome Data." *Statistics in Medicine*, 20: 453–472. 571

Wu, C.-F. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization.* John Wiley & Sons. 562

Acknowledgments

This work was partially funded by CONACYT. The author thanks PROGRESA for giving access to databases for this research.

Bayesian Analysis (2007) **2**, Number 3, pp. 591–610

Bayesian Hierarchical Multiresolution Hazard Model for the Study of Time-Dependent Failure Patterns in Early Stage Breast Cancer

Vanja Dukić[*] and James Dignam[†]

Abstract. The multiresolution estimator, developed originally in engineering applications as a wavelet-based method for density estimation, has been recently extended and adapted for estimation of hazard functions (Bouman et al. 2005, 2007). Using the multiresolution hazard (MRH) estimator in the Bayesian framework, we are able to incorporate any *a priori* desired shape and amount of smoothness in the hazard function. The MRH method's main appeal is in its relatively simple estimation and inference procedures, making it possible to obtain simultaneous confidence bands on the hazard function over the entire time span of interest. Moreover, these confidence bands properly reflect the multiple sources of uncertainty, such as multiple centers or heterogeneity in the patient population. Also, rather than the commonly employed approach of estimating covariate effects and the hazard function separately, the Bayesian MRH method estimates all of these parameters jointly, thus resulting in properly adjusted inference about any of the quantities.

In this paper, we extend the previously proposed MRH methods (Bouman et al. 2005, 2007) into the hierarchical multiresolution hazard setting (HMRH), to accommodate the case of separate hazard rate functions within each of several strata as well as some common covariate effects across all strata while accounting for within-stratum correlation. We apply this method to examine patterns of tumor recurrence after treatment for early stage breast cancer, using data from two large-scale randomized clinical trials that have substantially influenced breast cancer treatment standards. We implement the proposed model to estimate the recurrence hazard and explore how the shape differs between patients grouped by a key tumor characteristic (estrogen receptor status) and treatment types, after adjusting for other important patient characteristics such as age, tumor size and progesterone level. We also comment on whether the hazards exhibit nonmonotonic patterns consistent with recent hypotheses suggesting multiple hazard change-points at specific time landmarks.

Keywords: Multiresolution models, Bayesian survival analysis, hazard estimation

[*] Department of Health Studies, University of Chicago, Chicago, IL, mailto:vdukic@health.bsd.uchicago.edu

[†] Department of Health Studies, University of Chicago, Chicago, IL, mailto:jdigman@health.bsd.uchicago.edu

1 Introduction

In survival analysis, because the hazard function $h(t)$ often exhibits unstable behavior making it difficult to reliably discern patterns of change or make comparisons between groups, aggregates of the hazard over time are more frequently used. The cumulative hazard $H(t)$, or more commonly, functions of $H(t)$ such as survival or cumulative incidence functions, are used for summary and inference on failure risk. While useful and easy to interpret for most purposes, these summaries can partially obscure important patterns in the hazard of failure over time. Alternatively, examination of the hazard function itself in detail can reveal important properties of the failure process (Aalen and Gjessing 2001). Generally, some type of smoothed estimate of the hazard function is used to characterize its shape, which is indicative of how failure risk (in the population) changes with respect to some time origin. While a variety of approaches toward hazard estimation have been proposed, methodological challenges remain for both estimation and associated statistical inferential procedures. In particular, flexible estimation and modeling approaches are needed, because in contrast to the hazard functional form in most parametric survival models, the hazard may exhibit complex non-unimodal shape with ‘change-points’ that may reveal important information about the process under study.

In this article, we investigate the hazard of disease recurrence among women treated for breast cancer and followed over several years. The data originate from large multicenter randomized clinical trials evaluating the effect of hormonal or cytotoxic chemotherapy treatment agents administered after surgery (referred to generally as adjuvant therapy) in women with early stage breast cancer. As we describe in the next section, there is considerable interest in the patterns of recurrence hazards after breast cancer diagnosis and initial treatment, both to gain biological insights and to better manage the disease in a clinical setting. We accommodate the biologically plausible situation whereby different subgroups of women with specific disease features may have distinct functional forms of failure hazard that are non-proportional to each other, by constructing a joint model for all separate subgroup hazards, while keeping the effects of some factors (such as age or tumor size) common across all strata and proportional within each stratum.

To model the recurrence hazards, we apply the semiparametric multi-resolution hazard (MRH) estimator recently presented in work by Bouman and colleagues (Bouman et al. 2005, 2007). Employing a piece-wise constant prior for the hazard rate which is constructed in a tree-based and self-consistent manner, the MRH approach permits flexible modeling with the ability to incorporate a variety of *a priori* assumptions about the shape and smoothness of hazard functions in each of several defined strata. Furthermore, Bayesian modeling allows us to easily address specific hypotheses concerning the timing of peaks in the risk of failure and how these may differ with respect to key biologic and clinical parameters in breast cancer.

In the next section, we describe some of the key questions of interest in modeling of the breast cancer recurrence hazard, and the data source for this study. In Section 3 we review the basics of the multiresolution hazard model, and present the extension to the hierarchical multiresolution (HMRH) setting. In Section 4 we provide technical details of the Markov chain Monte Carlo (MCMC) model implementation. Section 5 presents the analysis. We conclude with a discussion of the findings in relation to the broader literature on breast cancer hazards as well as plans for future work on this problem.

2 Recurrence Risk after Surgery for Early Stage Breast Cancer

Over the past few decades, there has been appreciable progress in therapeutic strategies for early stage (i.e., localized and operable, as opposed to metastatic) breast cancer, with a well-developed array of treatment options. Due to increased screening vigilance and disease awareness, currently over 75% of women diagnosed have early stage tumors. Despite this progress, the clinical course of breast cancer after diagnosis remains heterogeneous from patient to patient and thus highly unpredictable for individuals. Thus, a significant clinical challenge is deciding which and how much adjuvant therapy is needed, and determining the magnitude of recurrence risk over extended post-treatment follow-up. Characteristics that prospectively identify which women are at greater or lesser risk of treatment failure are needed to aid in individually tailoring therapy for optimal disease management. Answers may also lie partly in gaining a better understanding of the intermediate and long-term clinical course of the disease, identifying patterns that can portend time periods of increased recurrence risk.

Apparent patterns in breast cancer recurrence hazard are readily observable from large cohorts of patients systematically followed over time. It is well known that risk of recurrence remains elevated for a long period of time after initial diagnosis and tumor removal, and there is longstanding interest in the the prospect of "cure" after sufficient time tumor-free has been achieved (Berg and Robbins 1966). Some long-term follow-up studies have suggested that a finite but lengthy"dormancy period" exists (possibly over 20 years), whereby tumor recurrences may still appear (Gordon 1990; Demicheli et al. 1996; Karrison et al. 1999). Studies examining the shape of the recurrence hazard consistently show a sharp peak 12-24 months after initial diagnosis and treatment, followed by a decline over time, although the hazard remains persistently elevated relative to individuals in the population never having had breast cancer (Saphner et al. 1996; Hess et al. 2003). This pattern stands in sharp contrast to the recurrence hazard for several other major cancers (e.g., colon, lung), where most recurrences appear within the first five years after discovery and removal of the primary tumor, followed by a period in which the hazard of recurrence and death from the disease resemble that of the population at large.

2.1 Data: Randomized Clinical Trials for Early Stage Breast Cancer

The National Surgical Adjuvant Breast and Bowel Project (NSABP) is a U.S. National Cancer Institute sponsored and funded multi-center cancer clinical trials group that has investigated a spectrum of treatments for breast and colorectal cancers since the late 1950s. The group consists of more than 5,000 participating physicians, nurses, and other research specialists located at over 1,000 medical centers, university and community hospitals, oncology practice groups, and health maintenance organizations in North America. The group has enrolled more than 60,000 women and men in clinical trials for breast and colorectal cancer.

While use of systemic adjuvant therapy began in the mid 1970s for women with tumors that had spread to the axillary lymph nodes, those with so-called lymph node negative breast cancer continued to be treated by surgery only, as they were considered at sufficiently favorable prognosis so as not to require further interventions. Because a significant fraction of such women *do* suffer recurrence and eventually die from the disease, beginning in the 1980's the NSABP began a series of clinical trials among patients with axillary lymph node negative breast cancer. These trials were serially designed, beginning with comparisons surgery alone to post-surgical hormonal or cytotoxic adjuvant therapy regimens, and following benefits seen for the latter, subsequently comparing different adjuvant therapy regimens. These studies provide a unique view to the long-term prognosis of women with early stage breast cancer, and are an ideal data source for examining factors influencing the hazard of disease recurrence.

A key determinant of both expected prognosis and potential choice of adjuvant therapy type is the presence and quantity of estrogen receptors (ER) on the tumor. From 1982-1988, patients were accrued according to ER status into one of two trials conducted in parallel: Protocol B-13 randomized 760 patients with ER-negative (ER-) tumors to no further treatment after surgery (384) or to 12 cycles of cytotoxic chemotherapy treatment with methotrexate and 5-fluorouracil (376). Protocol B-14 randomized 2,892 patients with ER-positive (ER+) tumors to placebo (1,453) or the estrogen antagonist drug tamoxifen (1,439) after surgery. Primary findings were first obtained in 1989, showing a significant reduction in breast cancer recurrence risk for patients receiving the adjuvant therapy regimens (Fisher et al. 1989b,a). Longer follow-up eventually revealed a survival advantage for those who received adjuvant therapy (Fisher et al. 1996b,a). Further details of the trial designs and findings can be found in the published primary reports. Follow-up continues to date, with mean follow-up of over 15 years and over 900 recurrence events observed.

Primary endpoints for the trial were overall survival, defined as time from surgery to death from any cause, and disease-free survival, defined as time to first breast cancer recurrence at any local, regional, or distant anatomic site, occurrence of a tumor in the opposite breast, occurrence of other second primary cancers, or death prior to these events (that is, time to first event of any kind). In this study, we model the cause-specific hazard for breast recurrence, defined as time to breast cancer recurrence, treating the other event types as censored observations. We do this because modeling the cause-specific hazard for breast cancer events only may have more clinical relevance, and furthermore, with the exception of endometrial cancer, which occurs in less than 1.5% of patients but is more frequent among women taking tamoxifen, hazard rates for non-breast cancer events are essentially equal between the two treatment groups. As in previous studies modeling these data (Fisher et al. 1989c; Bryant et al. 1997), we examine tumor size, tumor progesterone receptor level, menopausal status, and age as covariates potentially associated with the recurrence hazard.

3 The Multiresolution Hazard Model

In this section, we review the multiresolution approach to modeling hazard rate in a semiparametric Bayesian context developed and described in more detail in Bouman et al. (Bouman et al. 2005, 2007). As will be explained below, the method relies on the clever tree-based construction of the prior for the hazard rate, that ultimately yields a resolution-invariant and self-consistent prior for an arbitrarily fine piece-wise constant approximation to the hazard rate. The parameterization of the prior tree uniquely defines not only the prior expectations of the hazard rate in each of the intervals, but it also determines the amount of correlation and smoothness in hazard between the intervals.

3.1 Approaches to Hazard Modeling

One of the most common approaches to assessing the impact of factors on the hazard of failure is the Cox proportional hazard model. In this model individual covariates X affect baseline hazard $h_{\text{base}}(t)$ via $\exp(\mathbf{X}'\boldsymbol{\beta})$ (Cox 1972). This approach is readily adapted to modeling the cause-specific hazard (Prentice et al. 1978). In the typical application of the Cox model, covariate effects for the relative hazard are the primary focus, with the baseline hazard function treated as a nuisance parameter. As we have indicated, however, interest in our particular study here lies precisely in the estimation of the hazard functions. Specifically, we are interested in the estimation of a separate hazard rate function within certain strata based on treatment type and tumor characteristics, so that we might compare the shape of the hazards, while simultaneously estimating and performing inference about other covariate effects that are reasonably assumed to be common over strata.

Non-parametric methods for extracting hazard function estimates have been proposed for the Cox model (Gray 1990, 1992), primarily for the purpose of performing model diagnostics (e.g., changes in the effects of covariates over time) and correct functional forms for covariates in relation to failure hazard. A more general hazard regression approach that involves partitioning the time axis more specifically focused on hazard estimation within covariate strata (Gray 1996). There are many other approaches and variations (see Andersen et al. (1993) for detailed review), many of which provide estimates of functionals ($S(t)$, etc.) after the model is fit, but most still focus on covariates and model checking, rather than on efficient hazard function estimation *per se*. We discuss the properties and justify the MRH approach to hazard modeling more in the next section.

3.2 Multiresolution Model for the Baseline Hazard

Multiscale models for estimation of a discretized intensity function were developed for astrophysics applications by Kolaczyk (Kolaczyk 1999; Nowak and Kolaczyk 2000). Details of how this methodology has been adapted to hazard estimation are summarized in the following, while a more extensive discussion of its theoretical properties can be found in Bouman et al. (2005, 2007).

In summary, the multiresolution hazard (MRH) approach yields an estimate of the baseline (i.e., estimate from which covariate-specific curves can be generated) survival function based on the multiresolution baseline hazard estimate. It consists of first choosing the "time resolution" – a set of time points $\{t_0, t_1, ..., t_J\}$ – and then estimating the underlying baseline survival at those points, $S_{\text{base}}(t_j)$, based on the the cumulative hazard $H_{\text{base}}(t)$ and its discrete increments $d_j \equiv H_{\text{base}}(t_j) - H_{\text{base}}(t_{j-1}) = \int_{t_{j-1}}^{t_j} h_{\text{base}}(s)ds$, where $h_{\text{base}}(t)$ represents the baseline hazard rate at time t. For those times t such that $t_{j-1} < t < t_j, j = 1, \ldots, J$, a piecewise-constant hazard rate is assumed.

The MRH model is thus a semiparametric model which is able to estimate the baseline hazard rate $h_{\text{base}}(t)$ at the resolution times t_j, along with covariate effects $\boldsymbol{\beta}$. For convenience, we set the number of time intervals J equal to 2^M, where $M > 0$. In general, these intervals need not be of equal length – one can choose any resolution, though in practice we would recommend one such that there are multiple failure times observed in almost all intervals. Furthermore, the resolution should be chosen so that the average (or total) hazard rates within its intervals are clinically meaningful. However, in cases when prior information and clinical input about the resolution are vague, the optimal number of intervals could be chosen via model selection criteria, such as the DIC (Spiegelhalter et al. 2002). Note that the failure times after t_J become right-censored at t_J; hence, J should also be chosen so that a relatively small fraction of failure times is censored as a result.

After fixing the resolution, the discretized hazard is modeled in a way that allows us to incorporate the prior belief about the shape and smoothness of the true underlying hazard function. Following the notation in Bouman et al. (Bouman et al. 2005, 2007), we denote the total cumulative baseline hazard $H(t_J)$ as $H_{0,0}$, and the hazard increments d_1 as $H_{M,0}$, d_2 as $H_{M,1}, \ldots,$ and d_J as $H_{M,2^M-1}$. We then build the multiresolution hazard tree by recursively defining $H_{m-1,p} = H_{m,2p} + H_{m,2p+1}$, for $m = 1, \ldots, M$, and $p = 0, \ldots, 2^{m-1} - 1$. We refer to m as the level of resolution and p as the position within that level. Thus, at the top of this hazard tree we have the total cumulative hazard $H(t_J)$, which we split into finer components with each additional level of the tree, until we finally end at the the bottom of the tree (the highest level of resolution) with the hazard increments $d_1, \ldots, d_J$. If we further define $R_{m,p} \equiv H_{m,2p}/H_{m-1,p}$, we can parametrize the hazard increments by $H_{0,0}$ and the "splits" $R_{1,0}, \ldots, R_{M,2^{M-1}-1}$ (denoted $\mathbf{R}_{m,p}$). For example, when $M = 3$ (implying $J = 8$) we have: $d_1 = H_{0,0}R_{1,0}R_{2,0}R_{3,0}$, $d_2 = H_{0,0}R_{1,0}R_{2,0}(1 - R_{3,0}), \ldots, d_8 = H_{0,0}(1 - R_{1,0})(1 - R_{2,1})(1 - R_{3,3})$.

It is important to note that the piecewise-constant hazard assumption has been employed multiple times in the literature (for example, Walker and Mallick (1997). However, the uniqueness of the MRH model lies in its clever construction of the tree-based prior for a piece-wise constant function, so that the prior essentially does not depend on the final resolution level M (i.e., it is invariant to the height of the tree). More precisely, integrating out higher-level parameters, one would obtain the exact same prior as if that level and its parameters had simply not been considered in the first place. To aid with the understanding of the MRH model, a simple diagram of the two-level multiresolution prior is given in Figure 1.

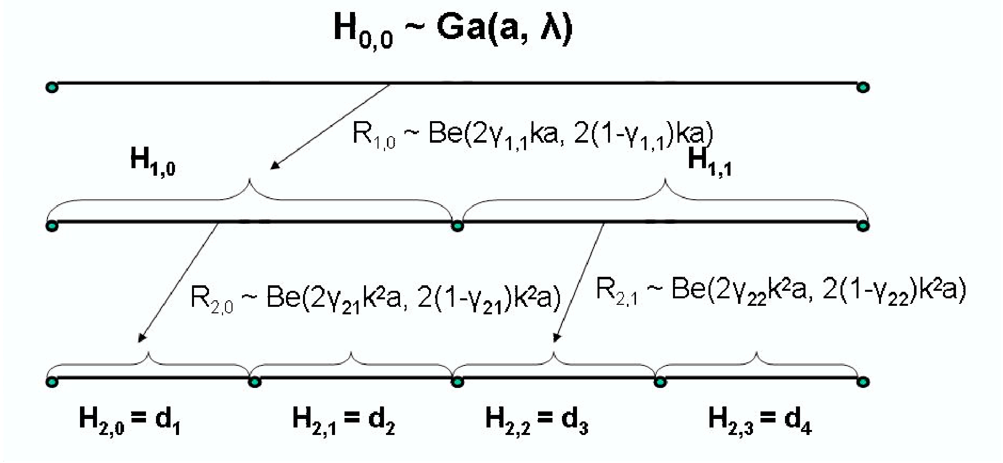

$H_{0,0} \sim Ga(a, \lambda)$
$R_{1,0} \sim Be(2\gamma_{1,1}ka, 2(1-\gamma_{1,1})ka)$
$H_{1,0}$
$H_{1,1}$
$R_{2,0} \sim Be(2\gamma_{21}k^2a, 2(1-\gamma_{21})k^2a)$
$R_{2,1} \sim Be(2\gamma_{22}k^2a, 2(1-\gamma_{22})k^2a)$
$H_{2,0} = d_1$
$H_{2,1} = d_2$
$H_{2,2} = d_3$
$H_{2,3} = d_4$

Figure 1: Diagram illustrating the multiresolution prior for the hazard rate function, with 2 levels (i.e., with resolution 2).

As in Nowak and Kolaczyk (Nowak and Kolaczyk 2000), we place a beta prior on each $R_{m,p}$ and a gamma prior on H. The shape parameters of each of these beta priors and the hyperparameters for H determine the prior expectations of the hazard increments, $d_j^*, j = 1, \ldots, J$. Furthermore, to allow for extra smoothness in the multiresolution prior, Bouman et al. (2007) introduce a multiplier for the shape parameters of the beta priors at each additional level of the hierarchy, denoted by k. Proceeding in this fashion, the priors for H and each $R_{m,p}$ when $M = 3$ and $J = 8$ are:

$$\begin{array}{rc}
H \sim & \mathcal{G}a(a, \lambda), \\
R_{1,0} \sim & \mathcal{B}e(2\gamma_{1,0}ka, 2(1-\gamma_{1,0})ka), \\
R_{2,p} \sim & \mathcal{B}e(2\gamma_{2,p}k^2a, 2(1-\gamma_{2,p})k^2a), \quad p = 0, 1 \\
R_{3,p} \sim & \mathcal{B}e(2\gamma_{3,p}k^3a, 2(1-\gamma_{3,p})k^3a), \quad p = 0, 1, 2, 3.
\end{array}$$

Note that under this prior structure, $E(R_{m,p}) = \gamma_{m,p}$, which, because H and $\mathbf{R}_{m,p}$ are independent *a priori*, easily allows one to choose $\gamma_{m,p}$, λ and a so that the prior expectation of the baseline hazard in each time interval j is any value d_j^* desired. This particular formulation of the gamma-beta tree also determines the prior correlation of the d_j as a function (among other things) of k and a. Specifically, when $k = 0.5$, the baseline hazard increments d_j are *a priori* independent gamma random variables. Choosing k less or greater than 0.5 yields, respectively, negative (rougher hazard) or positive (smoother hazard) prior correlation among the d_j's (Bouman et al. 2005, 2007). Smoother hazard may in particular be employed in problems with much censoring. It is also possible to treat k as a hyperparameter and estimate it jointly with other parameters.

It is well known that MRH models, because they are based on a tree-like structure, may have a blocky correlation pattern; for example, it is possible that two neighboring hazard increments are less correlated than those further apart which happen to share more ancestral split parameters. Bouman et al. (Bouman et al. 2005) propose placing a hyperprior on the hazard H shape parameter a to even out the prior correlations among hazard increments and bypass this rather counterintuitive property.

3.3 Hierarchical Multiresolution Hazard Model

It may sometimes be of interest or necessity to relax the proportional hazards assumption, permitting different baseline hazards in particular groups. These strata-specific hazards may be treated as fixed, or as random (infinite dimensional) strata-specific parameters. This could be desirable in particular when we have data from multiple centers or multiple studies, when one needs to allow for differences in baseline hazards due to unobserved covariate processes, or to account for correlation within subjects from the same strata.

In breast cancer, it is well-known that women with ER- and ER+ tumors have different expected prognosis due to association of ER with both tumor pathology and clinical characteristics such as patient age (Hess et al. 2003). Because we are primarily interested in estimating the hazard shapes, we wish to avoid imposing any proportionality constraint on ER in the model. In any case, Figure 2, showing recurrence-free survival curves by ER and treatment, clearly illustrates deviation from proportional hazards between ER- and ER+ patients. While proportionality appears to hold better between treatment groups within ER categories (Fig. 2), we also wish to permit a different hazard shape according to treatment type, as biologic hypotheses concerning the action of adjuvant therapy would suggest the possibility of different shapes (Skipper 1971). Thus, we define strata defined by the ER by treatment group combinations (i.e., ER- surgery only, ER- chemotherapy, ER+ surgery only (placebo), ER+ tamoxifen).

More specifically, in this model we allow the hazard for each of the strata to be *a priori* an independent and identically distributed random MRH variable. For each stratum $s, s \in \{1, 2, 3, 4\}$, we draw the stratum cumulative hazard H_s from a $\mathcal{G}a(a_s, \lambda_s)$, and then draw the stratum set of splits $\mathbf{R}_{m,p,s}$. For this reason we will call this model the hierarchical MRH or the HMRH for short.

To complete the prior for the hazard rate, we place a zero-truncated Poisson (ZTP) hyperprior with mean μ_a on each a_s: $e^{-\mu_a}{\mu_a}^a/[a!(1 - e^{-\mu_a})]$, and we allow each scale

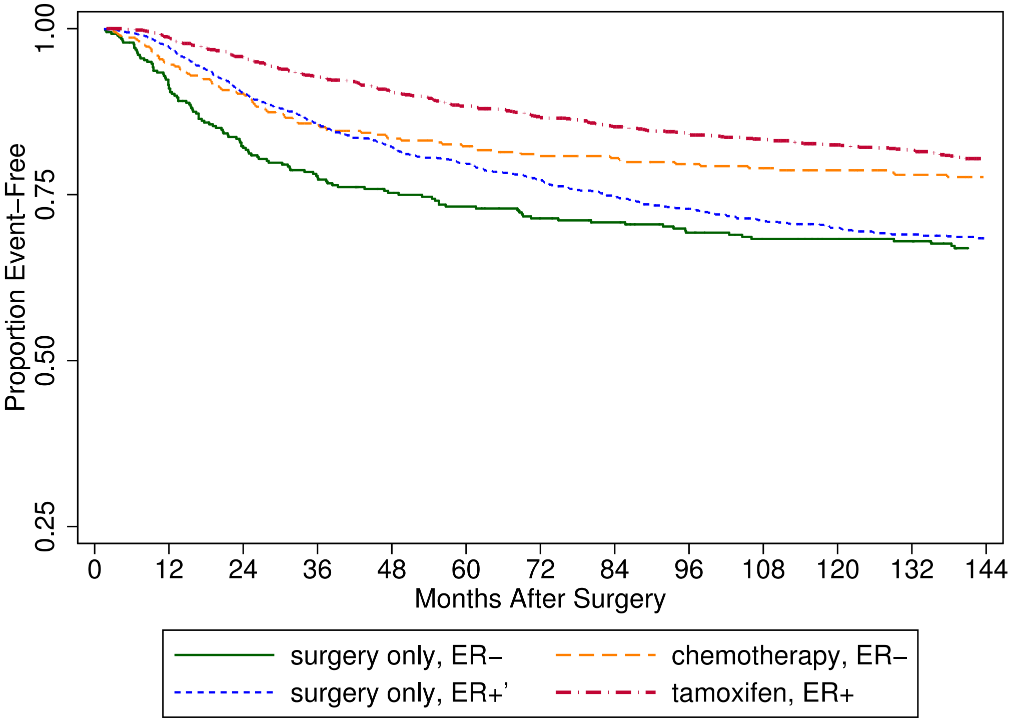


Figure 2: Kaplan-Meier estimates of recurrence-free survival among early stage breast cancer patients from NSABP clinical trials. Failures continue to occur in all groups many years after initial diagnosis and surgery.

parameter λ_s, the parameter of the total cumulative hazards $H_s(t_J)$, to follow an exponential distribution with mean μ_λ. The parameters k_s can be also given exponential priors with mean μ_k, but they could also be fixed if specific smoothness (positive or negative correlation) is desired *a priori*.

The continuous-time complete-data likelihood is based on the proportional-hazards model: $h_s(t|\mathbf{X}, \boldsymbol{\beta}) = \exp(\mathbf{X}'\boldsymbol{\beta})h_{\text{base},s}(t)$, where s denotes the stratum group. For a patient i in stratum s, if the failure time $T_{i,s} \in [0, t_J]$ is observed without censoring, the likelihood function is:

$$L_{i,s}(\boldsymbol{\beta} \mid T_{i,s}, \mathbf{X}_{i,s}) = \exp(\mathbf{X}'_{i,s}\boldsymbol{\beta})h_{\text{base},s}(T_{i,s})S_{\text{base},s}(T_{i,s})^{\exp(X'_{i,s}\boldsymbol{\beta})}. \tag{1}$$

When an observation is right-censored, i.e., $T_{i,s} > t_{cens}$ for some $t_{cens} \leq t_J$, we have:

$$L_{i,s}(\boldsymbol{\beta} \mid T_{i,s}, \mathbf{X}_{i,s}) = S_s(t_{cens}|\mathbf{X}_{i,s}, \boldsymbol{\beta}) = S_{\text{base},s}(t_{cens})^{\exp(X'_{i,s}\boldsymbol{\beta})}. \tag{2}$$

Here, each h_s is a priori distributed as an MRH variable, with independent total hazard and split parameters. The vector of covariates $\mathbf{X}_{i,s}$ contains patient and disease characteristics (age, tumor size, tumor progesterone receptor level, etc.) for patient i in stratum s.

4 Markov chain Monte Carlo

The Gibbs sampling algorithm used to simulate the parameter posterior and its sequence of full conditional posterior distributions for parameters from all strata is outlined below. The details of the simpler model are provided in Bouman et al. (2005, 2007).

The likelihood for patient i in stratum s, whose failure (or censoring) time is $T_{i,s}$, the censoring indicator $\delta_{i,s}$, and the covariates $\mathbf{X}_{i,s}$, is

$$\begin{aligned} L(T_{i,s} \quad | \quad & \delta_{i,s}, \boldsymbol{\beta}, H_s, \mathbf{R}_{m,p,s}, \mathbf{X}_{i,s}) = \\ & \left[\exp(\mathbf{X}'_{i,s}\boldsymbol{\beta}) h_{\text{base},s}(T_{i,s})\right]^{\delta_{i,s}} \exp(-\exp(\mathbf{X}'_{i,s}\boldsymbol{\beta}) H_{\text{base},s}(\min(T_{i,s}, t_J))). \quad (3) \end{aligned}$$

The log-likelihood for all $N = \sum_{s=1}^{4} N_s$ patients is thus:

$$\delta' \left[\mathbf{X}\boldsymbol{\beta} + \mathbf{F_s}\boldsymbol{\Pi}\tilde{\mathbf{R}}\right] - \sum_{s=1}^{4}\sum_{i=1}^{N_s} \exp(\mathbf{X}'_{i,s}\boldsymbol{\beta}) H_{\text{base},s}(\min(T_{i,s}, t_J)) \qquad (4)$$

where δ is the vector of censoring indicators for all patients, $\mathbf{X}$ is the $N \times L$ matrix of covariates and $\boldsymbol{\Pi}$ is the $2^M \times (2^{M+1} - 1)$ multiresolution tree matrix. In that matrix, the (i, j) element is 1 when $j = 1$ or $i \in [1 + (j \bmod 2^{\mathrm{m}}), \ldots, 2^{\mathrm{M-m}} + (\mathrm{j} \bmod 2^{\mathrm{m}})]$, $\mathrm{m} = \lfloor \log_2(\mathrm{j}) \rfloor$, and 0 otherwise. $\tilde{\mathbf{R}}$ is the multiresolution log-parameter vector $(\log(H), \log(R_{1,0}), \log(1 - R_{1,0}), \ldots, \log(R_{M,2^M-1}), \log(1 - R_{M,2^M-1}))$. $\mathbf{F}$ is an $N \times 2^M$ matrix for which the (i, j) element is 1 if the ith patient (among all patients in all strata put together) has $T_i \in (t_{j-1}, t_j]$, and 0 otherwise; patients with $T_i > t_J$ have $F_{i,j} = 0, j = 1, \ldots, J$ (see Bouman et al. (2005, 2007) for details).

The Gibbs sampler steps (Geman and Geman 1984) for the parameters H_s, $R_{m,p,s}$, λ_s, a_s, and k_s, for all strata s are the same:

1. Sample H_s from its full conditional density:
   $\pi(H_s | \lambda_s, a_s, \mathbf{R}_{m,p,s}) = \mathcal{G}a\left((a_s + \sum_{i=1}^{N_s} \delta_{i,s}), 1/\left[(1/\lambda_s) + \sum_{i=1}^{N_s} \mathcal{F}(T_{i,s})\right]\right)$,
   with mean $\mu = (a_s + \sum_{i=1}^{N_s} \delta_{i,s}) / \left[(1/\lambda_s) + \sum_{i=1}^{N_s} \mathcal{F}(T_{i,s})\right]$,
   where $\mathcal{F}(T_{i,s}) = H_s(\min(T_{i,s}, t_J))/H_s(t_J)$ is a function of the $T_{i,s}$ and $\mathbf{R}_{m,p,s}$.

2. Sample each $R_{m,p,s}$ from the full conditional $\pi(R_{m,p,s} | k_s, a_s, H_s)$

3. Sample k from $\pi(k_s | a_s, \mathbf{R}_{m,p,s}, \boldsymbol{\beta})$, λ_s from $\pi(\lambda_s | H_s, a_s, \boldsymbol{\beta})$, and
   a_s from $\pi(a_s | H_s, \mathbf{R}_{m,p,s}, \lambda_s, \boldsymbol{\beta})$.

4. Sample $\boldsymbol{\beta}$ from $\pi(\boldsymbol{\beta} | H_s, \mathbf{R}_{m,p,s})$.

Similarly as in Bouman et al. (2005, 2007), the full conditional posterior distributions for H_s are gamma, while the full conditional distributions for each $R_{m,p,s}$ and $\boldsymbol{\beta}$ are log-concave and therefore easy to sample from (using Gilks and Wild (1992) algorithm, for example). On the other hand, the full conditional distributions for hyperparameters λ_s, a_s, and k_s are in general more difficult to sample from as they are not log-concave. We recommend following Bouman et al. (2005) who use the rejection Metropolis sampling (see Gilks et al. (1995)).

5 Analysis of the Recurrence Hazard after Breast Cancer

5.1 Covariate Effect Estimation

The 16- and 32-bin multiresolution model with the "flat" prior hazard rate for each stratum (with all $\gamma_{m,p,s}$ and all k_s set to 0.5), was fit using output from Gibbs sampler chains with 12,000 iterations each, with the first 2,000 iterations of each chain discarded as burn-in. Every 5^{th} iteration was retained to reduce correlation between adjacent draws. The Gelman-Rubin diagnostics, performed separately for each parameter, were used to establish convergence.

Table 1: Posterior Credible Intervals for Predictor Effects, 16-bin model

	Tumor Size	Standardized PGR	Standardized Age
2.5%	0.0148	-0.151	-0.281
50%	0.0199	-0.066	-0.217
97.5%	0.0251	0.007	-0.149

Table 1 gives marginal 95% posterior credible intervals for covariates considered. Tumor size was measured in millimeters (mm), ranging from 0 to 60mm. Progesterone receptor concentration (PGR) was standardized using the sample mean of 139.17 and standard deviation of 294.59. Age was standardized using the sample mean of 53.27 and standard deviation of 10.42 years. Larger tumor size and younger age at diagnosis were found associated with increased recurrence hazard: within each stratum, an increase of 10.42 years (1 standard deviation) in age resulted approximately in 20% reduction, while an increase of 1mm in tumor size resulted in approximately 2% increase in recurrence hazard. A higher concentration of progesterone receptors on the tumor is weakly indicative of lesser failure risk. While menopausal status is generally an important factor in breast cancer prognosis, here it was only marginally associated with recurrence hazard after stratification by estrogen receptor status and inclusion of age at diagnosis in the model, and so was omitted from further consideration. The direction and magnitude of these covariate effects are essentially consistent with other prognostic factor studies of these patients.

5.2 Hazard Function Estimation

Figure 3 displays by treatment/ER strata the median posterior estimates for the 16-bin baseline hazard increments (corresponding to constant 11.3-month hazard function values). This 16-dimensional vector is a discrete approximation to the baseline hazard rate $h(t)$, estimated via the hazard increments $d_j = \int_{t_{j-1}}^{t_j} h(s)ds$. In Figure 4, we show a smoothed version of the same hazard estimates. Smoothing was performed via the median-spline method, with a hazard value of zero included at time zero for each stratum, to reflect the fact that patients are considered cancer-free immediately after surgery and essentially do not fail until some time has elapsed.

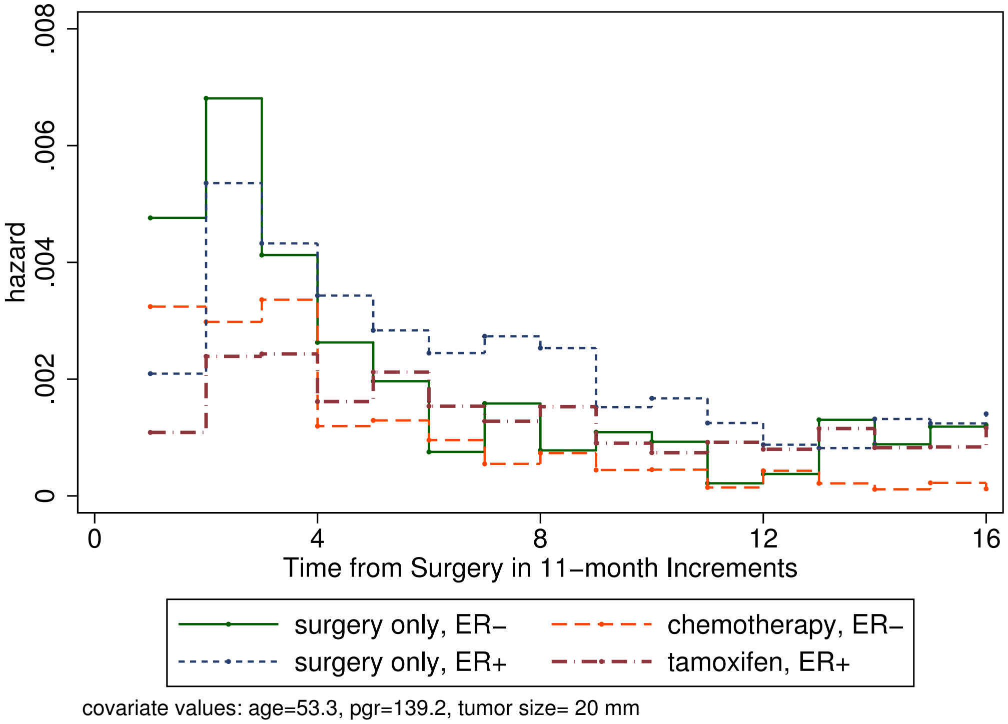


Figure 3: Discrete recurrence hazard increments for the 16-bin HMRH model. Horizontal sections represent the hazard value within the approximate 11-month increment in time from surgery.

Several notable features are seen. First, all four groups have the distinctive hazard peak around 12-24 months after surgery, with the ER- groups experiencing the peak a bit earlier. This pattern is similar to that noted by Hess et al. (2003) in their study of recurrence hazards by ER status. Second, the peak is greatest for the ER- patients

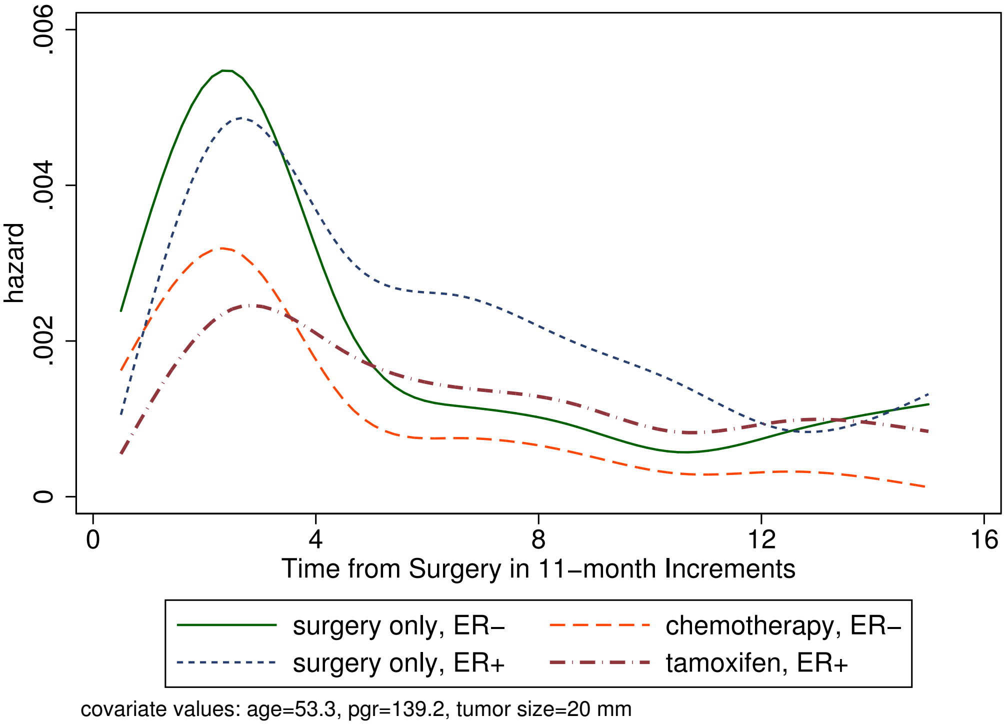


Figure 4: Smoothed recurrence hazards for the 16-bin model.

receiving surgery only, and is substantially reduced by chemotherapy, being lower than untreated patients from the more favorable ER+ group. The lowest peak is among ER+ patients randomized to tamoxifen. Interestingly however, at longer follow-up times the hazard in this group is no lower than that of chemotherapy treated and even untreated ER- patient groups, both of which have smaller hazard than untreated ER+ patients. Ultimately, the ER- chemotherapy treated group has the lowest recurrence hazard.

Figure 5 shows pointwise posterior credible intervals, based on 2.5% and 97.5% estimated posterior percentiles, for the four strata. With the exception of time points around the hazard peak, credible intervals for the four hazard estimates tend to overlap, particularly at longer follow-up times. Thus, we currently cannot reliably conclude whether there is a crossover of failure hazard among the groups at later time points. We should note that in other analyses collapsing across one or the other stratification factors (ER or treatment) and treating ER or treatment as covariates in modeling, large treatment effects within ER groups were apparent, while differences between ER groups within treatment modalities (surgery, adjuvant) were large initially but attenuated over time, substantively violating the proportional hazards assumption, as is evident in Figure 2.

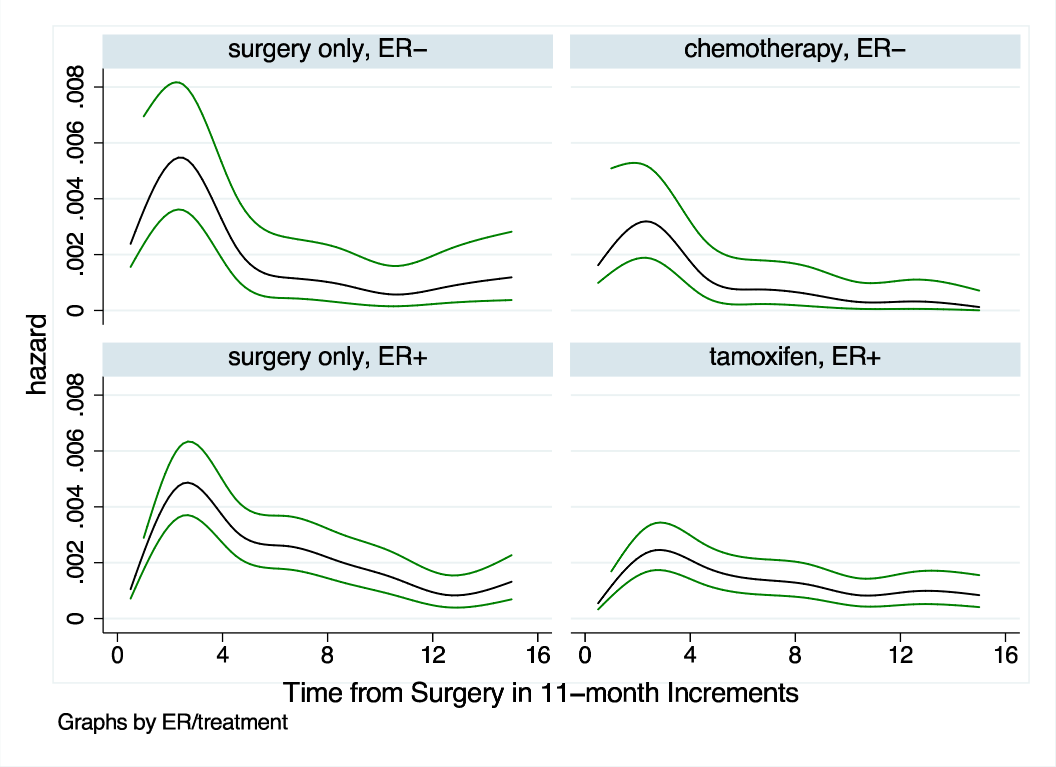


Figure 5: Credible intervals for the recurrence hazards from Figure 4 by ER and treatment strata.

5.3 Sensitivity Analysis

We now turn to the question of robustness of our model results. An alternate model was based on 32 time intervals of length just under 6 months each, instead of the 16 intervals of approximate length of 1 year. The purpose of this sensitivity check was twofold: first, to compare the estimates of the hazard rates for each of the four groups under these alternative resolutions; and second, to assess the impact of additional resolution level onto the estimates of the covariates (age, tumor progesterone receptor level, and tumor size). Note that in this data set, we could not reasonably use a finer resolution than the 32-bin choice for two reasons: 1) although this is a large cohort of patients, prognosis is relatively favorable and so failures are few if the intervals (bins) are very small, and 2) follow-up visits do not happen more frequently than once every 3 to 6 months, and while recurrence events can occur in continuous time, recurrence events that are not clinically apparent will be detected at these visits, causing clustering of events (and consequently artificial periodicity in the hazard function estimate) if very small time bins are used.

Table 2: Posterior Credible Intervals for Predictor Effects, 32-bin model

	Tumor Size	Standardized PGR	Standardized Age
2.5%	0.0150	−0.147	−0.281
50%	0.0200	−0.066	−0.215
97.5%	0.0250	0.009	−0.147

Based on the invariance properties of the MRH prior, one can expect that some of that invariance is preserved in the posterior as well; and we observe this to a large degree. For example, a 16-bin hazard estimate obtained by fitting the 32-bin model and then aggregating the neighboring hazard increments is almost indistinguishable from the original 16-bin model hazard estimate. In addition, hazard rate plots using the 32-bin model and the 16-bin model are very similar as well; as expected, the 32-bin estimates are slightly more variable, but all hazard shapes are very similar (not shown).

With respect to covariate effect estimates, the effects of increasing the resolution is minimal. Compared to Table 1, the effects shown in Table 2 differ by a negligible amount.

6 Discussion

We have illustrated the application of a flexible extension of the familiar Cox proportional hazards model to jointly estimate covariate effects and separate hazard rate functions for several patient strata. This approach allows us to incorporate covariate effects and perform inference related to shapes and change-points in the hazard over time, our primary interest in the problem of recurrence after early stage breast cancer. The estimation and examination of the hazard functions directly reveals important patterns not readily apparent from quantities such as the survival functions. However, the hazard function remains a difficult quantity to draw robust inference from, as even in this large dataset, estimates suggest potentially important differences in shape, but variability estimates preclude any definitive conclusions pending additional analyses, as discussed below.

The observation that among those patients with initially higher risk disease (i.e., those with ER- tumors), the fraction escaping the early failure risk go on to have substantially lower long-term failure risk than those with initially more favorable prognosis (ER+ tumors), has significant implications in both clinical management and considerations regarding further developments of adjuvant therapies. In fact, there has been much recent interest in the development of ‘switching’ strategies whereby women with ER+ tumors discontinue tamoxifen and begin use of other hormonal treatments, in order to extend and improve on the benefit of this treatment modality. Currently, little is known about what factors might be key to optimizing the switching strategy. For ER- patients, newer chemotherapy and molecularly targeted agents that act on specific tumor vulnerabilities may offer the best opportunity for a *bona fide* cure once early failure is avoided.

In addition to the well-known initial wave of failures following surgery, in recent years investigators have suggested that additional reproducible patterns are manifest in the recurrence hazard in the intermediate to long-term follow-up period. The notion of a bimodal or "double-peaked" recurrence hazard has been proposed, where after the first period of increased failure risk at 1-5.3.0 years post-surgery, the decline in failure hazard is followed by a second peak centered roughly around 8 years (Demicheli et al. 1996, 2001; Baum et al. 2005). A number of cancer biologic hypotheses have been put forth regarding the meaning and cause of a possible double-peaked failure pattern. For example, it has been conjectured that growth kinetics perturbed by surgery may contribute to the first wave in failure hazard, while heterogeneous disseminated tumor cells that require more time to become established may account for the latter peak in failure (Demicheli et al. 2001; Baum et al. 2005). This idea may seem to harken back to the naive concept that cancer surgery "spreads" cancer, but the influence of surgery on growth kinetics does have foundation in substantive biologic theory (Fisher et al. 1983). However, before any such interpretations of the hazard shape can be made or gain further credibility, more rigorous analytic methods such as those proposed here must be applied. Furthermore, it may be difficult to uniquely ascribe such a pattern to specific biologic phenomena, because other circumstances, such as the existence of patient mixtures due to unrecognized factors present at diagnosis or apparent hazard spikes caused by clustering of failures in time due to discovery of subclinical disease around certain time landmarks (e.g., mandatory 5-year post-diagnosis screen) would be expected to produce similar patterns. Nonetheless, this intriguing concept merits further investigation.

Our future work on this problem will involve the inclusion of data from trials conducted subsequent to those included here. As the trials are designed in a hierarchical fashion, these studies share some treatment arms in common with the current data, but also include newer treatment regimens. Extension of the model to more data sources will involve incorporation of 'trial' effects to allow for heterogeneity among common treatment arms. The inclusion of additional data will permit a more thorough exploration of changes in the hazard over time and more robust inference (including "collapsibility" of neighboring intervals in some regions), due to the considerably larger sample size that will result in narrower bounds on estimated hazards. This analysis will also have greater biologic and clinical relevance as we explore more recent drug regimens designed to reduce recurrence risk in women with breast cancer.

References

Aalen, O. and Gjessing, H. (2001). "Understanding the shape of the hazard rate: A process point of view." *Statistical Science*, 16: 1–22. 592

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statistical Methods Based on Counting Processes*. Berlin: Springer–Verlag. 595

Baum, M., Demicheli, R., Hrushesky, W., and Retsky, M. (2005). "Does surgery unfavourably perturb the "natural history" of early breast cancer by accelerating the appearance of distant metastases?" *European Journal of Cancer*, 41: 508–515. 606

Berg, J. and Robbins, G. (1966). "Factors influencing short and long term survival of breast cancer patients." *Surgical and Gynecologic Obstetrics*, 122: 1311–1316. 593

Bouman, P., Dignam, J., Dukic, V., and Meng, X. (2007). "A multiresolution hazard model for multi-center survival studies: Application to tamoxifen treatment in early stage breast cancer." *Journal of the American Statistical Association*, in press. 591, 592, 595, 596, 597, 598, 600, 601

Bouman, P., Dukic, V., and Meng, X. (2005). "Bayesian multiresolution hazard model with application to an AIDS reporting delay study." *Statistica Sinica*, 15: 325–357. 591, 592, 595, 596, 598, 600, 601

Bryant, J., Fisher, B., Gunduz, N., Costantino, J., and Emir, B. (1997). "S-phase fraction combined with other patient and tumor characteristics for the prognosis of node-negative, estrogen-receptor-positive breast cancer." *Breast Cancer Research and Treatment*, 51: 239–253. 594

Cox, D. (1972). "Regression models and life tables." *Journal of the Royal Statistical Society - Series B*, 34: 187–220. 595

Demicheli, R., , Valagussa, P., and Bonadonna, G. (2001). "Does surgery modify growth kinetics of breast cancer micrometastases?" *British Journal of Cancer*, 85: 490–492. 606

Demicheli, R., Abbattista, A., Miceli, R., Valagussa, P., and Bonadonna, G. (1996). "Time distribution of the recurrence risk for breast cancer patients undergoing mastectomy: further support about the concept of tumor dormancy." *Breast Cancer Research and Treatment*, 41: 177–185. 593, 606

Fisher, B., Constantino, J., Redmond, C., Poisson, R., Bowman, D., Couture, J., Dimitrov, N., Wolmark, N., Wickerham, D., and Fisher, E. (1989a). "A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors." *The New England Journal of Medicine*, 320: 479–484. 594

Fisher, B., Dignam, J., Bryant, J., DeCillis, A., Wickerham, D., Wolmark, N., J, J. C., Redmond, C., Fisher, E., Bowman, D., Deschenes, D., Dimitrov, N., Margolese, R., Robidoux, A., Shibata, H., Terz, J., Paterson, A., Feldman, M., Farrar, W., Evans, J., and Lickley, H. (1996a). "Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor positive tumors." *Journal of the National Cancer Institute*, 88: 1529–1542. 594

Fisher, B., Dignam, J., Mamounas, E., Costantino, J., Wickerham, D., Redmond, C., Wolmark, N., Dimitrov, N., Bowman, D., Glass, A., Atkins, J., Abramson, N., Sutherland, C., Aron, B., and Margolese, R. (1996b). "Sequential methotrexate and fluorouracil for the treatment of node-negative breast cancer patients with estrogen receptor-negative tumors: eight-year results from National Surgical Adjuvant Breast and Bowel Project (NSABP) B-13 and first report of findings from NSABP B-19 comparing methotrexate and fluorouracil with conventional cyclophosphamide, methotrexate, and fluorouracil." *Journal of Clinical Oncology*, 14: 1982–1992. 594

Fisher, B., Gunduz, N., and Saffer, E. (1983). "Influence of the interval between primary tumor removal and chemotherapy on kinetics and growth of metastases." *Cancer Research*, 43: 1488–1492. 606

Fisher, B., Redmond, C., Dimitrov, N., Bowman, D., Legault-Poisson, S., Wickerham, D., Wolmark, N., Fisher, E., Margolese, R., and Sutherland, C. (1989b). "A randomized clinical trial evaluating sequential methotrexate and fluorouracil in the treatment of patients with node-negative breast cancer who have estrogen-receptor-negative tumors." *The New England Journal of Medicine*, 320: 473–478. 594

Fisher, B., Redmond, C., Wickerham, D., Wolmark, N., Bowman, D., Couture, J., Dimitrov, N., Margolese, R., Legault-Poisson, S., and Robidoux, A. (1989c). "Systemic therapy in patients with node-negative breast cancer. A commentary based on two National Surgical Adjuvant Breast and Bowel Project (NSABP) clinical trials." *Annals of Internal Medicine*, 111: 703–712. 594

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6: 721–741. 600

Gilks, W., Best, N., and Tan, K. (1995). "Adaptive rejection Metropolis sampling." *Applied Statistics*, 44: 455–472. 601

Gilks, W. and Wild, P. (1992). "Adaptive rejection sampling for Gibbs sampling." *Applied Statistics*, 41: 337–348. 601

Gordon, N. (1990). "Application of the theory of finite mixtures for the estimation of 'cure'." *Statistics in Medicine*, 9: 397–407. 593

Gray, R. (1990). "Some diagnostic methods for Cox regression models through hazard smoothing." *Biometrics*, 46: 93–102. 595

— (1992). "Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis." *Journal of the American Statistical Association*, 87: 942–951. 595

— (1996). "Hazard rate regression using ordinary nonparametric regression smoothers." *Journal of Computational and Graphical Statistics*, 5: 190–207. 595

Hess, K., Pusztai, L., Buzdar, A., and Hortobagyi, G. (2003). “Estrogen receptors and distinct patterns of breast cancer relapse.” *Breast Cancer Research and Treatment*, 78: 105–118. 593, 598, 602

Karrison, T., Ferguson, D., and Meier, P. (1999). “Dormancy of mammary carcinoma after mastectomy.” *Journal of the National Cancer Institute*, 91: 80–85. 593

Kolaczyk, E. (1999). “Bayesian multiscale models for Poisson processes.” *Journal of the American Statistical Association*, 94: 920–933. 596

Nowak, R. and Kolaczyk, E. (2000). “A statistical multiscale framework for Poisson inverse problems.” *IEEE Transactions on Information Theory*, 46: 1811–1825. 596, 597

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). “The Analysis of Failure Times in the Presence of Competing Risks.” *Biometrics*, 34: 541–554. 595

Saphner, T., Tormey, D., and Gray, R. (1996). “Annual hazard rates of recurrence for breast cancer after primary therapy.” *Journal of Clinical Oncology*, 14: 2738–2746. 593

Skipper, H. (1971). “Kinetics of mammary tumor cell growth and implications for therapy.” *Cancer*, 28: 1479–1499. 598

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). “Bayesian measures of model complexity and fit (with discussion).” *Journal of the Royal Statistical Society - Series B*, 64: 583–616. 596

Walker, S. and Mallick, B. (1997). “Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing.” *Journal of the Royal Statistical Society - Series B*, 59: 845–860. 597

Acknowledgments

Support for this research was partially provided by the Susan G. Komen Breast Cancer Foundation and Public Health Service grant NCI-U10-CA-69651 from Institutes of Health, Department of Health and Human Services.

Bayesian Analysis (2007) **2**, Number 3, pp. 611–634

A Spatially-adjusted Bayesian Additive Regression Tree Model to Merge Two Datasets

Song Zhang,[*] Ya-Chen Tina Shih[†] and Peter Müller[‡]

Abstract. Scientific hypotheses of interest often involve variables that are not available in a single survey. This is a common problem for researchers working with survey data. We propose a model-based approach to provide information about the missing variable. We use a spatial extension of the BART (Bayesian additive regression tree) model. The imputation of the missing variables and inference about the relationship between two variables are obtained simultaneously as posterior inference under the proposed model. The uncertainty due to imputation is automatically accounted for. A simulation analysis and an application to data on self-perceived health status and income are presented.

Keywords: BART, CART, Missing variables, Spatial model, Survey

1 Introduction

We consider the problem of inference about the relationship of two variables reported in two different datasets. This is a common problem for researchers working with survey data. Scientific hypotheses of interest often involve variables that are not available in a single survey. Specifically, we are interested in inference on how a variable z is affected by another variable y, when there is no such dataset that collects z and y simultaneously. Instead, z is only reported in dataset D_1 and y is only collected in dataset D_2.

Many model-based methods have been developed to deal with missing data problems, including maximum likelihood (ML) methods, multiple imputation (MI) methods, weighted estimating equations (WEE), and fully Bayesian (FB) methods. See Little (1992), Horton and Laird (1999), Schafer and Graham (2002), Ibrahim et al. (2005) and the references therein for detailed discussions. There are some assumptions associated with each of these methods. Many ML methods assume a large sample size so that the ML estimates are approximately unbiased and normally distributed. The likelihood function is assumed to arise from a parametric model of the complete data. Finally, ML methods usually require the missing at random (MAR) assumption (Rubin 1976). MI methods also rely on large-sample approximation and assume a parametric form for the joint model of the observed and missing data. They require some assumption about the distribution of missingness, although not limited to MAR. WEE methods are extensions of generalized estimating equations (GEE). Two models need to be specifed:

[*] Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, mailto:songzhang@mdanderson.org

[†] Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, http://gsbs.uth.tmc.edu/tutorial/shih.html

[‡] Department of Biostatistics, University of Texas MD Anderson Cancer Centeri, Houston, TX, http://odin.mdacc.tmc.edu/~pm/

One regression model for the data, and the other describing the missingness mechanism. WEE methods are considered to be doubly robust because the estimates of the regression parameters remain consistent as long as one of the two models is correctly specified. The MAR assumption and a large sample size are required. FB methods do not require a large sample size. Specifying a joint probability model, however, they require assumptions about the sampling model for the data and about the missingness mechanism. In summary, all the above methods regard missingness as a probabilistic phenomenon. In contrast, in the following discussion, missingness is not random. The variable y is missing for all records in D_1.

The most commonly applied method to borrow information from one dataset (i.e., y in D_2) to provide information not collected in another dataset (i.e., D_1) is the use of census-based socioeconomic status (SES) characteristics to supplement individual-level data, such as medical records, claims or registries (Gornick et al. 1996; Geronimus and Bound 1998; Devesa and Diamond 1983). The census-based approach obtains aggregate statistics of SES variables at certain geographic levels (e.g., census track, county, or zip code) and uses these aggregate numbers as proxy measures of SES in individual-level data. It has been used extensively in studies of health disparities. For more examples, see Mandelblatt et al. (1991), Kraus et al. (1986) and Byrne et al. (1994).

Geronimus and Bound (1998) cautioned that although the census-based approach is easy to execute, these aggregate measures should not be interpreted as if they were micro-level variables. The approach has several limitations. It requires detailed residential information to be collected in D_1. If due to privacy concerns this information is not collected or is not detailed enough (for example, only state code is available), then the method breaks down. The method only makes use of geographic information. Other individual-level covariates are ignored. For example, if we are interested in imputing missing income in D_1, then information such as age, gender, education, occupation could be very informative. Finally, the true value of the missing variable in D_1 may not match the neighborhood profile. This uncertainty is usually ignored.

In this paper we propose to approach the problem in the framework of Bayesian hierarchical modeling. A spatially adjusted Bayesian additive regression tree (SBART) is defined to impute the missing variable in D_1 based on individual-level covariates as well as geographic information. SBART is an extension of the BART model. The idea of BART is to model an unknown function as a mixture of tree models. Each tree is a priori constrained to have a simple structure. It only contributes a small portion to the overall model. Chipman, George and McCulloch (2006a) demonstrated that the sum over all trees provides a sufficiently rich model to incorporate both direct effects and interaction effects of different orders. SBART extends BART by incorporating spatial random effects. Correlation among neighboring areas is utilized to improve inference. Our method implements a full probability model with likelihood and priors. The imputation of the missing variable and the inference about the relationship between the two variables are obtained simultaneously as posterior inference under the model, and the uncertainty due to imputation is accounted for automatically. Unrelated to the problem of merging datasets that we consider here, a similar spatial extension of the

BART model has been developed independently in current work by Chipman, George, McCulloch and Musio (2006b).

The outline of the paper is as follows. Section 2 introduces notation and presents the Bayesian hierarchical model. A simulation study is conducted in Section 3. We illustrate our method with a data analysis example in Section 4. Finally, Section 5 discusses some limitations of our method as well as some possible extension.

2 A Spatial BART Model

Let I be the number of spatial units at the finest level of detail recorded in both datasets. This could be, for example, census tract, zip code area or county.

In dataset D_1, let m_i denote the number of subjects from area i $(i = 1, \cdots, I)$. The sample size of D_1 is $m = \sum_{i=1}^{I} m_i$. For the jth subject from area i, we are interested in the relationship between variables z_{ij} and y_{ij}, where z_{ij} but not y_{ij} is recorded in dataset D_1. We use v_{ij} to denote a vector of other individual-level covariates reported in D_1.

The variable y_{ij} that is missing in D_1 is recorded on a different set of individuals in dataset D_2. For notational ease, we use the variable name x_{ij} rather than y_{ij}, to distinguish the fact that these variable values are recorded in D_2 rather than D_1. Similarly, for the vector of variables v_{ij}, we use w_{ij} rather than v_{ij} for those variables recorded in D_2. We assume w_{ij} and v_{ij} to be consistent, i.e., they record the same variables and use the same coding for the values. Because it would be unusual for all covariates recorded in D_1 and D_2 to be consistent, we only assume that after suitable pre-processing a subset of the covariates can be considered consistent across the two datasets. Let n_i be the number of subjects from area i, so $j = 1, \ldots, n_i$ in D_2. Then $n = \sum_{i=1}^{I} n_i$ is the sample size of D_2.

We define $\boldsymbol{Z} = \{z_{ij}, i = 1, \cdots, I, j = 1, \cdots, m_i\}$. Similarly we use $\boldsymbol{Y}$, $\boldsymbol{V}$, $\boldsymbol{X}$ and $\boldsymbol{W}$ to denote the vector of all y_{ij}, v_{ij}, x_{ij} and w_{ij}, respectively.

We describe in words how the proposed approach facilitates learning about the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$ with $\boldsymbol{Y}$ missing. We assume that $(\boldsymbol{Y}, \boldsymbol{V})$ (in D_1) and $(\boldsymbol{X}, \boldsymbol{W})$ (in D_2) arise from the same model M. We use the posterior for the parameters in M, obtained conditional on $(\boldsymbol{X}, \boldsymbol{W})$ to impute the missing $\boldsymbol{Y}$ conditional on $\boldsymbol{V}$. Finally, the regression of $\boldsymbol{Z}$ on the imputed $\boldsymbol{Y}$ approximates the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$. By integrating with respect to $\boldsymbol{Y}$, the marginal posterior distribution of the regression parameter $\boldsymbol{\beta}$ accounts for the variability induced by the imputation. The described learning process is complicated by the need to specify a joint probability model for $(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{X} \mid \boldsymbol{V}, \boldsymbol{W})$. Details are described later.

For the learning process to work we make the key assumption that $(\boldsymbol{X}, \boldsymbol{W})$ and $(\boldsymbol{Y}, \boldsymbol{V})$ are independent samples from the same model. This assumption ensures that we can apply what we have learned from $(\boldsymbol{X}, \boldsymbol{W})$ to $(\boldsymbol{Y}, \boldsymbol{V})$. For example, this assumption is satisfied if both D_1 and D_2 are representative samples from the U.S. population.

2.1 The Sampling Model

The proposed approach is model-based. We start the model construction with assumed sampling models for $\boldsymbol{Z}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$. In the following description, we use $N(m, s^2)$ to denote a normal distribution with moments m and s^2. We assume that a sampling model $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ is available for z_{ij}, conditional on v_{ij} and assumed values for y_{ij}, and indexed by a set of parameters Φ. For example, if z_{ij} is continuous, we can assume a linear regression model with z_{ij} being the dependent variable, y_{ij} and v_{ij} defining the design matrix, and Φ including the regression coefficients and variance parameter. If z_{ij} is ordinal, an ordinal probit model may be used. Specific examples of $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ are used in the simulation study and the case study.

The model $p(x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2)$ describes the relationship between x_{ij} and w_{ij}. Specifically, we assume

$$x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2 \sim N\left(f(w_{ij}) + \theta_i, \sigma^2\right), \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_I)'$ is a vector of random spatial effects, $f(w_{ij})$ is an unknown function associating x_{ij} with w_{ij}, and σ^2 is the residual variance. We represent the mean function $f(w_{ij})$ as a BART model. Since the additional random effects θ_i introduce the desired spatial correlation among neighboring areas, we refer to model (1) as the spatially-adjusted Bayesian additive regression tree (SBART) model.

For reference, and to introduce notation for later use, we give a brief review of the BART model. See Chipman et al. (2006a) for details. We begin with the notation for a single tree model. Let T denote a tree. Its nodes can be divided into two categories, interior nodes and terminal nodes. A splitting rule is defined at each interior node. We limit splitting rules to binary splits. Each rule consists of a splitting variable and a splitting value. The splitting value is a threshold on the splitting variable that defines the splitting rule. Starting from the root, an individual with covariates w_{ij} selects branches in the tree according to the splitting rules until it is assigned to a terminal node. Suppose that there are K terminal nodes. We define $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)'$, with μ_k being assigned to the kth terminal node. The tree maps each covariate vector w_{ij} into one element of $\boldsymbol{\mu}$. A single tree model is denoted by the pair $(T, \boldsymbol{\mu})$, and the association between μ_k and w_{ij} through a tree T is written as $\mu_k = g(w_{ij}, T, \boldsymbol{\mu})$.

The BART model defines a summation of such tree models, as

$$f(w_{ij}) = g(w_{ij}, T_1, \boldsymbol{\mu}_1) + g(w_{ij}, T_2, \boldsymbol{\mu}_2) + \cdots + g(w_{ij}, T_L, \boldsymbol{\mu}_L),$$

where L is the total number of trees that form the BART. We usually assign a large value for L (e.g., $L = 200$) to encourage flexibility. On the other hand, to avoid overfitting, the BART model includes a strong prior on each tree to keep its effect small, effectively making each tree into a "weak learner". But overall, the sum of trees provides a sufficiently rich model to fit a variety of functions. For example, μ_k represents an interaction effect if its assignment involves more than one component of w_{ij} (i.e., more than one splitting variable). Furthermore, because $f(w_{ij})$ can be based on trees of different sizes, the BART model can incorporate both direct effects and interaction effects of different orders. SBART extends BART by incorporating an additional spatial effect into the conditional mean of x_{ij} given w_{ij}.

BART is closely related to ensemble methods that combine a set of tree models. Examples of ensemble methods include boosting, bagging and random forests. Boosting (Freund and Schapire 1997; Friedman 2001) fits a sequence of trees. Each tree is fit conditional on data variation that is not explained by the other trees. Bagging (Breiman 1996; Clyde and Lee 2001) and random forests (Breiman 2001) construct a large number of independent trees through data randomization and stochastic search. The methods then use an average of the trees to improve prediction. Ensemble methods are not derived as coherent inference under a probability model. In contrast, BART is a model-based approach that reports inference as the summary of a full probabilistic description of all relevant uncertainties. Bayesian single tree models have been developed by Chipman et al. (1998) and Denison et al. (1998). Compared with single tree models, the sum-of-trees models provide vastly more flexibility by easily incorporating additive effects. Chipman et al. (2006a) provided a posterior Markov chain Monte Carlo (MCMC) simulation scheme for the BART model. They demonstrated that the proposed MCMC simulation has good mixing properties.

The third part of the top-level sampling model is an assumed model for y_{ij} conditional on the observed covariate vector v_{ij}. We assume the same model as for the regression of x_{ij} on w_{ij}:

$$y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2 \sim N\left(f(v_{ij}) + \theta_i, \sigma^2\right),$$

with $f(\cdot)$ defined by the SBART model as before.

2.2 The Prior Model

We complete the Bayesian hierarchical model with priors $p(\Phi)$, $p(f)$, $p(\boldsymbol{\theta})$ and $p(\sigma^2)$, for Φ, f, $\boldsymbol{\theta}$ and σ^2, respectively. We assume a priori independence.

The choice of $p(\Phi)$ depends on the particular form of $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$. For example, in a linear regression model, conjugate priors are technically convenient choices. That is, normal priors for the regression coefficients and an inverse Gamma prior for the residual variance.

The BART model in (1) is indexed by $\{(T_l, \boldsymbol{\mu}_l), l = 1, \cdots, L\}$. We use

$$p(f) = \prod_{l=1}^{L} p(T_l, \boldsymbol{\mu}_l) = \prod_{l=1}^{L} \Big\{ p(T_l) \cdot p(\boldsymbol{\mu}_l \mid T_l) \Big\}.$$

Following Chipman et al. (2006a), we define $p(T_l)$ by three factors, corresponding to a node being non-terminal, the selection of the splitting variable for a non-terminal node, and the choice of the splitting value conditional on a chosen splitting variable. The probability that a node at depth d is nonterminal, is assumed to be

$$\alpha(1+d)^{-\gamma},$$

where $\alpha \in (0,1)$ and $\gamma \in [0,\infty)$ are two hyper-parameters reflecting our prior belief about the tree. For example, if we believe that the depth of the tree should be small, we can assign a big value for γ, so that the probability decays fast with d. Chipman et al. (2006a) proposed $\alpha = 0.95$ and $\gamma = 2$ as default values, which implies that with prior probability 0.05, 0.55, 0.28, 0.09 and 0.03, the tree has 1, 2, 3, 4, and ≥ 5 terminal nodes, respectively. A natural choice for the selection of the splitting variable, conditional on a node being non-terminal, is a uniform prior over all available variables. A default choice for the distribution of the splitting value is a uniform distribution over the set of available splitting values. Finally, we define a prior for $\boldsymbol{\mu}_l$. Let μ_{lk} be the kth element of $\boldsymbol{\mu}_l$. Conditional on T_l, we assume i.i.d. normal priors for μ_{lk}. The mean and variance of the normal prior are specified in such a way that each tree is constrained to be a weak learner, and it plays a small role in the overall fit. More details can be found in Chipman et al. (2006a), Section 3.2.

For the spatial random effects $\boldsymbol{\theta}$ we use a conditionally autoregressive (CAR) prior. The key idea of the CAR model is simple. It formalizes the notion that each area is similar to its neighbors. Specifically, we define $p(\boldsymbol{\theta})$ by the set of conditional distributions

$$p(\theta_i \mid \theta_{(-i)}, \rho, \delta^2) = N\Big(\frac{\rho}{h_i}\sum_{j\neq i} c_{ij}\theta_j,\ \frac{1}{h_i}\delta^2\Big),\ i = 1, \cdots, I, \tag{2}$$

where $\theta_{(-i)}$ denotes all the elements of $\boldsymbol{\theta}$ except θ_i; ρ is a parameter with range $(-1,1)$; δ^2 is the variance component; $c_{ij} = 1$ $(i \neq j)$ if area i and area j are neighbors, and $c_{ij} = 0$ otherwise, including $c_{ii} = 0$; and $h_i = \sum_{j=1}^{I} c_{ij}$ is the total number of neighbors for area i. The joint distribution $p(\boldsymbol{\theta})$ implied by (2) is

$$p(\boldsymbol{\theta} \mid \rho, \delta^2) = N\Big(0,\ \delta^2(\boldsymbol{H} - \rho\boldsymbol{C})^{-1}\Big), \tag{3}$$

where $\boldsymbol{C} = (c_{ij})$ is an $I \times I$ adjacency matrix, and $\boldsymbol{H}$ is an $I \times I$ diagonal matrix with h_i being the diagonal elements. Model (2) specifies that given random effects from all the other areas, the distribution of θ_i only depends on its neighbors. When $\rho = 0$, the variance matrix in (3) is diagonal, implying that θ_i are independent. When $\rho = 1$, the conditional mean of θ_i in (2) equals the average of its neighbors. However, $\rho = 1$ implies that $\boldsymbol{H} - \rho\boldsymbol{C}$ is singular. That is, the covariance matrix of $\boldsymbol{\theta}$ does not exist. Sun et al. (1999) specified $-1 < \rho < 1$ as a smoothing or spatial correlation parameter. It can be thought of as a measure of spatial association. For more discussion of CAR models, see Cressie (1993) page 407, Besag et al. (1991), Clayton and Kaldor (1987) and Whittle (1954).

We complete the prior model with probability models for the hyper-parameters σ^2, ρ and δ^2. Chipman et al. (2006a) assumed $p(\sigma^2)$ to be an inverse chi-square distribution $\sigma^2 \sim \nu\lambda/\chi^2_\nu$, where ν is the degree of freedom. This is a special case of the inverse Gamma distribution. The key idea to specify the hyper-parameters ν and λ is to first obtain a preliminary estimate $\hat{\sigma}^2$ by exploratory data analysis (for example, through linear regression of x_{ij} and w_{ij}), and then specify ν and λ such that $\hat{\sigma}^2$ matches the qth quantile of $p(\sigma^2)$. The default setting recommended by Chipman et al. (2006a) is $(\nu, q) = (3, 0.90)$. Finally, we define prior distributions for the parameters ρ and δ^2 in the CAR model. It is natural to assume that the spatial effects are positively correlated. We therefore assume ρ to be uniform between 0 and 1, i.e., $U(0,1)$. We assume $p(\delta^2)$ to be an inverse Gamma distribution, denoted by $IG(a_\delta, b_\delta)$, with density function

$$p(\delta^2) \propto \frac{1}{(\delta^2)^{a_\delta+1}} \exp(-\frac{b_\delta}{\delta^2}).$$

Here a_δ and b_δ are fixed hyperparameters.

For reference, we state the joint probability model on the data $\boldsymbol{Z}$, $\boldsymbol{Y}$, $\boldsymbol{X}$ and the parameters:

$$p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) \cdot p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) \cdot p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) \\ \cdot p(\Phi) \cdot p(f) \cdot p(\boldsymbol{\theta} \mid \rho, \delta^2) \cdot p(\sigma^2) \cdot p(\rho) \cdot p(\delta^2), \quad (4)$$

where

$$\begin{aligned} p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) &= \prod_{i=1}^{I} \prod_{j=1}^{m_i} p(z_{ij} \mid y_{ij}, v_{ij}, \Phi), \\ p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) &= \prod_{i=1}^{I} \prod_{j=1}^{n_i} p(x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2), \\ p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) &= \prod_{i=1}^{I} \prod_{j=1}^{m_i} p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2). \end{aligned}$$

We are interested in the inference on Φ given all observations, namely $p(\Phi \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$. Carrying out the desired inference requires integration with respect to $\boldsymbol{Y}$ and the other parameters. This integration does not have a closed form solution. We set up MCMC simulation and obtain inference based on random samples from the posterior distribution of Φ. Details of the sampling scheme can be found in the Appendix. By integrating out $\boldsymbol{Y}$, $p(\Phi \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ automatically accounts for the variability induced by the imputation. A byproduct of this process is the imputation of the missing variable $\boldsymbol{Y}$, which can be obtained as random samples from $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$.

3 A Simulation Study

We conduct a simulation study to examine the performance of the proposed approach. We define $I = 99$ spatial areas, with an assumed spatial structure (adjacency matrix $\boldsymbol{C}$) equal to that of the 99 counties in the state of Iowa. We also assume $n_i = 4$ and $m_i = 2$ for $i = 1, \cdots, I$. Thus we have sample size $n = 396$ and $m = 198$.

The simulated data are generated as follows. We assume covariate vectors w_{ij} and v_{ij} to be of dimension 10. Each of the 10 elements is generated from independent $U(0,1)$ distribution. We generate the simulation truth for the spatial random effects $\boldsymbol{\theta}$ from a

$N(\mathbf{0}, \delta^2(\boldsymbol{H} - \rho \boldsymbol{C})^{-1})$ distribution, using $\rho = 0.3$ and $\delta = 1$. The mean function $f(u)$ is evaluated as

$$f(u) = 10\sin(\pi u_1 u_2) + 20(u_3 - 0.5)^2 + 10u_4 + 5u_5, \tag{5}$$

where u_i is the ith element of $u = (u_1, \cdots, u_{10})'$. The same function was used in simulation in Friedman (1991) and Chipman et al. (2006a). The added variables together with the interactions and nonlinearities make it difficult to fit the model by standard parametric methods. Conditional on the covariates w_{ij}, we generate x_{ij} by

$$x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2 \quad \sim \quad N(f(w_{ij}) + \theta_i, \sigma^2),$$

using $\sigma = 0.2$. Similarly, we generate y_{ij} conditional on v_{ij},

$$y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2 \quad \sim \quad N(f(v_{ij}) + \theta_i, \sigma^2).$$

Thus x_{ij} and y_{ij} only depend on the first 5 elements of w_{ij} and v_{ij}, respectively.

Finally, z_{ij} is generated by

$$z_{ij} \mid y_{ij}, v_{ij}, \boldsymbol{\beta}, \tau^2 \sim N(h(v_{ij}, y_{ij}, \boldsymbol{\beta}), \tau^2), \tag{6}$$

where we assume $\tau = 0.2$, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_6)' = (3, -3, -2.5, -1, 1.5, 2, 1)'$, and

$$h(v_{ij}, y_{ij}, \boldsymbol{\beta}) = \beta_0 + v_{ij4}\beta_1 + v_{ij5}\beta_2 + v_{ij6}\beta_3 + v_{ij7}\beta_4 + v_{ij8}\beta_5 + y_{ij}\beta_6.$$

Here v_{ijk} denotes the kth element of v_{ij}. The simulation model for z_{ij} is a linear regression model. We assume that part of the covariates (v_{ij4}, v_{ij5}) are involved in the generation of y_{ij} and others $(v_{ij6}, v_{ij7}, v_{ij8})$ are not. Matching the earlier notation $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$, we have $\Phi = (\boldsymbol{\beta}, \tau^2)$, where $\boldsymbol{\beta}$ is the vector of regression coefficients and τ^2 is the variance parameter.

Conditional on the simulated data $(\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V})$, but pretending that $\boldsymbol{Y}$ is missing, we generate a Monte Carlo sample from the posterior distribution $p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ under model (4). See the Appendix for details of the posterior simulation.

We repeat the described simulation $K = 100$ times. For the kth simulation, we save the simulation truth $\boldsymbol{Y}^{(k)}$ and $\boldsymbol{\beta}$, the imputed values $\hat{\boldsymbol{Y}}^{(k)}$, and the estimated effects $\hat{\boldsymbol{\beta}}^{(k)}$. We obtain $\hat{\boldsymbol{Y}}^{(k)}$ and $\hat{\boldsymbol{\beta}}^{(k)}$ as marginal posterior expectations under $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ and $p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$, respectively. The mean squared error (MSE) for $\boldsymbol{Y}$ is defined as

$$MSE_Y = \frac{1}{Km} \sum_{k=1}^{K} \left\{ \sum_{i,j} (\hat{y}_{ij}^{(k)} - y_{ij}^{(k)})^2 \right\}.$$

Similarly, for $\boldsymbol{\beta}$ we define

$$MSE_{\beta_p} = \frac{1}{K} \sum_{k=1}^{K} \left\{ (\hat{\beta}_p^{(k)} - \beta_p)^2 \right\}, \; p = 0, 1, \cdots, 6.$$

For comparison we record results under two different models.

Table 1: MSE from Simulation to Compare SBART and BART

	(a)	(b)
β_0	0.0055	0.0062
β_1	0.0078	0.0154
β_2	0.0017	0.0045
β_3	0.0029	0.0030
β_4	0.0048	0.0048
β_5	0.0079	0.0090
β_6	0.0136	0.0321
$\boldsymbol{Y}$	0.596	3.864

Column (a) under SBART; Column (b) under BART.

M1: Model (4) with a $U(0,1)$ prior for ρ, an $IG(0.001, 0.001)$ prior for δ^2, and a CAR prior for $\boldsymbol{\theta}$. This is the proposed SBART model.

M0: Model (4) with $\boldsymbol{\theta} = 0$. This is a BART model without spatial adjustment. Under the BART model, the priors $p(\boldsymbol{\theta} \mid \rho, \delta^2)$, $p(\rho)$ and $p(\delta^2)$ are not needed.

The remaining prior choices include a normal prior for $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}) = N(\boldsymbol{0}, 100\boldsymbol{I}_6)$, and an inverse Gamma prior for τ^2, $p(\tau^2) = IG(0.001, 0.001)$. Here $\boldsymbol{0}$ is a vector of $0's$ and $\boldsymbol{I}_6$ is an identity matrix of dimension 6. For the hyper-parameters in $p(f)$ and $p(\sigma^2)$, we use the default setting recommended by Chipman et al. (2006a).

Table 1 compares the MSE from models **M1** and **M0**. The results suggest that when spatial correlation is present, incorporating spatial effects improves the estimation of regression coefficients. This is particularly true for β_6, the coefficient of the missing variable, which is of primary interest. In the simulation, the MSE of β_6 is reduced from 0.0321 to 0.0136. A byproduct of the proposed approach is the inference about the missing variable, which might be of interest to researchers by itself. Monte Carlo sample averages evaluate posterior means and provide point estimates of the missing variables. Other summaries characterize the uncertainty of the imputation. Table 1 shows that incorporating spatial effects greatly improves the imputation of the missing variable. The MSE for $\boldsymbol{Y}$ is reduced from 3.864 to 0.596. This improvement can also be seen in Figure 1, where we plot $\boldsymbol{Y}^{(k)}$ versus $\hat{\boldsymbol{Y}}^{(k)}$ from one simulation.

The estimated spatial correlation parameter $\hat{\rho}^{(k)}$ has a mean 0.414 and a standard deviation 0.091, suggesting a slight overestimation of ρ. The histogram of $\hat{\rho}^{(k)}$ is plotted in Figure 2. We also plot $\boldsymbol{\theta}^{(k)}$ against $\hat{\boldsymbol{\theta}}^{(k)}$, the true and estimated values of $\boldsymbol{\theta}$, respectively, from one simulation in Figure 3. The fact that the points fall around the 45 degree line suggests that the method successfully recovers the spatial pattern.

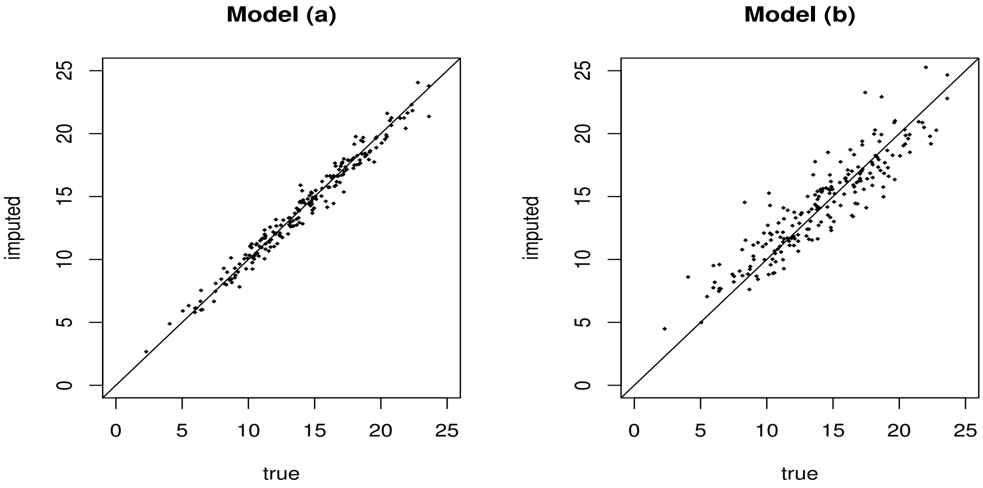


Figure 1: Simulation example. The imputation of $\boldsymbol{Y}$ under **M1** and **M0** (under one simulation). **M1** uses the SBART model. **M0** uses the BART model without spatial random-effects.

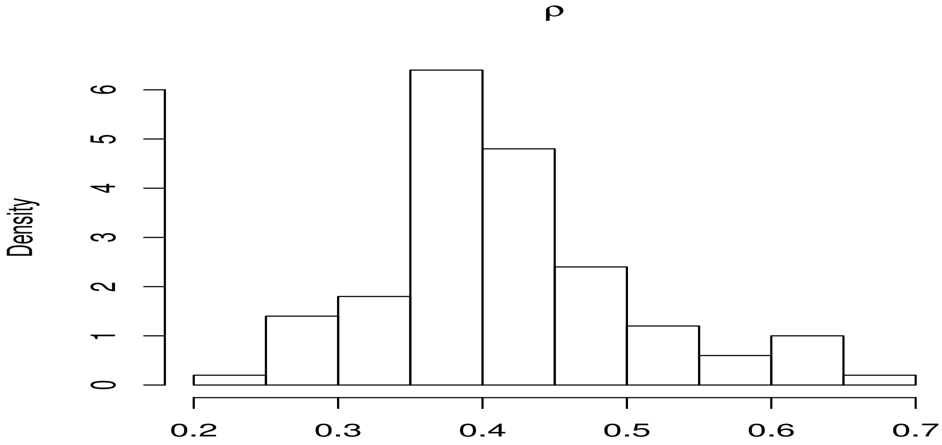


Figure 2: Simulation example. Histogram of $p(\hat{\rho}^{(k)} \mid data)$.

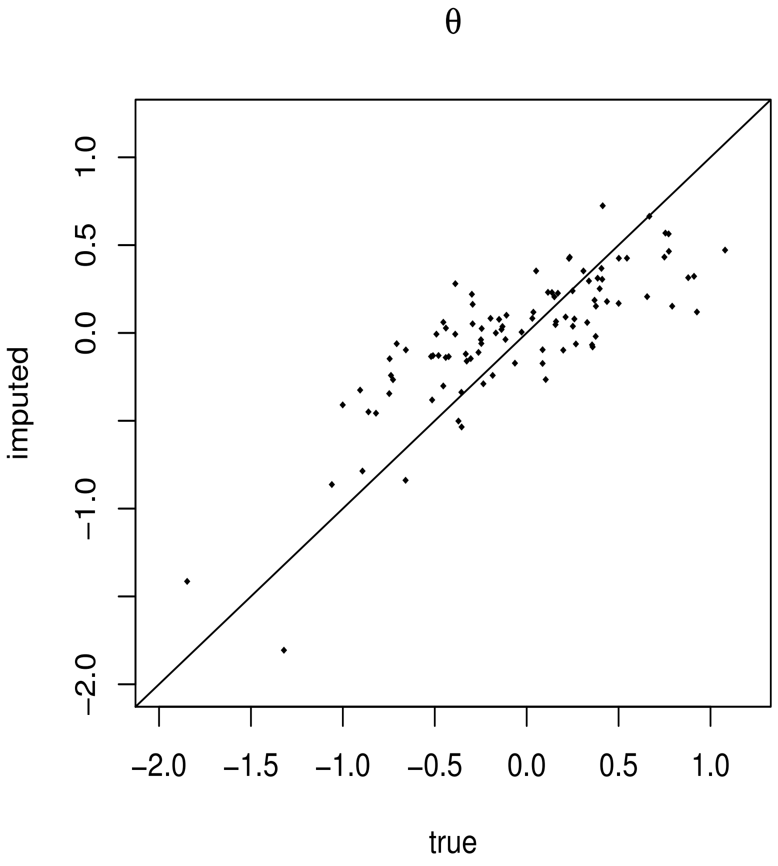


Figure 3: Simulation example. Simulation truth and imputed values of $\boldsymbol{\theta}$.

4 Joint Inference with the CPS and SIPP Surveys

We evaluate the proposed approach with real survey data. In this evaluation, we apply our method to explore the relationship between self-perceived health status and income using two different surveys. One survey includes data on health status, income, and other variables $(\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{V})$. The second survey reports income and other variables $(\boldsymbol{X}, \boldsymbol{W})$.

We implement inference through the proposed approach *without* using the observed values of income $\boldsymbol{Y}$ in the first survey. That is, we carry out the analysis pretending that we did not have income $(\boldsymbol{Y})$ information in the first survey.

For comparison, we also implement inference *with* the observed $\boldsymbol{Y}$ values. Using data from the first survey only, we implement posterior simulation in the model

$$p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) \cdot p(\Phi), \tag{7}$$

and summmarize $p(\Phi \mid \boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{V})$. By comparing the inference with $\boldsymbol{Y}$ missing versus inference conditional on $\boldsymbol{Y}$, we will validate the proposed model.

4.1 The Datasets

We let D_1 be a dataset extracted from the 2001 Current Population Survey (CPS), March Supplement. The variable $\boldsymbol{Z}$ is self-perceived health status with values 1 to 5, where 1 denotes the best health status and 5 denotes the poorest health status. The variable $\boldsymbol{Y}$ is defined to be total personal income. We are interested in the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$. The set of individual-level covariates are denoted by $\boldsymbol{V}$, which include age, race, gender, education, health insurance coverage, marital status, employment, industry and occupation. The dataset D_2 comes from the 2001 Survey of Income and Program Participation(SIPP), where total personal income $\boldsymbol{X}$ and the other covariates $\boldsymbol{W}$ are collected. Both CPS and SIPP report income, denoted as $\boldsymbol{Y}$ in CPS and $\boldsymbol{X}$ in SIPP. We pretend, however that $\boldsymbol{Y}$ is missing in D_1 to illustrate and validate the proposed method. CPS and SIPP are two independent surveys that each collects information from a representative sample of the U.S. civilian noninstitutional population. It is therefore reasonable to assume that $(\boldsymbol{Y}, \boldsymbol{V})$ and $(\boldsymbol{X}, \boldsymbol{W})$ arise from the same model.

CPS reports annual income while SIPP collects the information of monthly income. To make the income variables consistent between two datasets, we scale them to a common range of 0 to 1. Furthermore, personal income is known to be heavily skewed to the right, which makes the normal assumption in (1) inappropriate. We carry out a square root transformation to mitigate the problem. Thus eventually $\boldsymbol{Y}$ and $\boldsymbol{X}$ denote the square root of the scaled personal income.

The finest available spatial area in both datasets is metropolitan statistical area (MSA), which is defined as a core area that contains a substantial population nucleus, together with adjacent communities having a high degree of social and economic integration with that core. MSAs comprise one or more entire counties. In D_1 and D_2 there are altogether $I = 239$ MSAs. The original datasets from CPS and SIPP have more than 90,000 and 260,000 records, respectively. For this illustrative analysis, we obtain D_1 and D_2 by randomly sampling 10,000 observations from each of the two original datasets.

4.2 Model Specification

Health status $\boldsymbol{Z}$ is an ordinal categorical variable. We construct an ordinal probit model $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi)$. We define the probit model by introducing a latent normal random variable

$$\eta_{ij} \mid \boldsymbol{\beta}, \tau^2 \sim N(\beta_0 + v_{ij1}\beta_1 + v_{ij2}\beta_2 + v_{ij3}\beta_3 + y_{ij}\beta_4, \tau^2).$$

For given values of η_{ij} and a set of cut points $c_1, \cdots, c_4$, we set

$$z_{ij} \mid \eta_{ij} \quad = \quad \begin{cases} 1, & \text{if } \eta_{ij} \le c_1, \\ r, & \text{if } c_{r-1} < \eta_{ij} \le c_r \text{ for } r = 2, 3, 4, \\ 5, & \text{if } \eta_{ij} > c_4, \end{cases} \tag{8}$$

where $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_4)'$. See, for example, Johnson and Albert (1999) for a discussion of Bayesian inference in ordinal regression models, including the latent variable construction used here. The latent variable η_{ij} is assumed to arise from a linear regression model with covariates being personal income y_{ij}, health insurance coverage v_{ij1}, gender v_{ij2}, and age v_{ij3}, and $\boldsymbol{\beta}$ is the corresponding coefficient vector. Income and age are continuous; age ranges from 18 to 84; gender is binary with 0 indicating male and 1 indicating female; health insurance coverage is binary with 0 indicating covered and 1 indicating not covered. We define $\boldsymbol{\eta}$ to be the collection of η_{ij}, and $\Phi = (\boldsymbol{\eta}, \boldsymbol{\beta}, \tau^2)$. The cutpoints $(c_1, \cdots, c_4)$ are specified as fixed. Random cutpoints would provide more flexibility. For example, Johnson and Albert (1999) jointly update the cutpoints and the latent probit variable. However, the choice of the sampling model for $\boldsymbol{Z} \mid \boldsymbol{Y}$ is not directly related to the missing data problem. We assume fixed cutpoints to keep the model simple and keep the discussion focused.

The models $p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2)$ and $p(x_{ij} \mid w_{ij}, f, \boldsymbol{\theta}, \sigma^2)$ are defined in (1). We complete the model with priors for $(\rho, \delta^2, \boldsymbol{\beta}, \tau^2)$. We assume diffuse priors, a uniform prior for ρ, $p(\rho) = U(0, 1)$, an inverse Gamma prior for δ^2, $p(\delta^2) = IG(0.001, 0.001)$, independent normal priors for β_p, $p(\beta_p) = N(0, 100)$, $p = 0, \cdots, 4$, and an inverse Gamma prior for τ^2, $p(\tau^2) = IG(0.001, 0.001)$. We use default values recommended in Chipman et al. (2006a) for the hyper-parameters of $p(f)$ and $p(\sigma^2)$.

4.3 Implementation Details

Some practical issues arise in the application to real data. First, in fitting the model $p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2)$, we can use the entire vector of v_{ij}. There is no need for formal variable selection. As pointed out by Chipman et al. (2006a), the BART model is a nonparametric Bayesian regression approach which uses dynamic random basis elements that are dimensionally adaptive. Variable selection is already part of the model. In contrast, $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ is a generalized linear model and inference can be sensitive to correlation among the covariates (y_{ij}, v_{ij}). Like any other regression analysis, the specification of $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ requires a good understanding of the research questions to identify the relevant covariates. Importantly, high linear correlation among (y_{ij}, v_{ij}) complicates interpretation and should be avoided. With y_{ij} missing, we use (x_{ij}, w_{ij}) instead to check for linear correlation among the covariates.

Another issue concerns a bias in the inference on Φ induced by the imputation of y_{ij}. Figure 4 clearly shows a shrinkage effect. An ideal imputation would have a scatter plot falling around the 45 degree line. In Figure 4 the range of the imputed values is much narrower compared with that of the true values. Chipman et al. (2006a) observed similar shrinkage in a simulation study, which they attributed to extreme extrapolation. That is, when we make prediction outside the observed data, because of lack of information, the prior takes over and the imputed values are shrunk towards the center. We believe, however, that the cause of shrinkage in Figure 4 is more than extreme extrapolation. If the shrinkage arises from extrapolation alone, then it should have equal effect on both extremes. In Figure 4, we see more shrinkage on the higher incomes than on the lower incomes. From this observation, we hypothesize that the shrinkage is caused by a violation of the normality assumption in model (1). If personal income is heavily skewed to the right, then the square root transformation does not suffice to achieve normality, and extremely high incomes are not correctly imputed.

We propose to address the issue of shrinkage through the following two steps. First we carry out a preliminary analysis using model (4). We compare the distribution of imputed income $\hat{\boldsymbol{Y}}$ based on $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ with the observed income distribution from D_2. We use a deterministic adjustment to match some features of these two distributions. For example, in this study we construct a linear transformation of the imputed values, $t(\hat{y}_{ij}) = a\hat{y}_{ij} + b$, such that some selected quantiles (for example, the 10th and 90th quantiles) of $t(\hat{y}_{ij})$ match those of $\boldsymbol{X}$, the incomes observed in D_2. In the second step, we replace $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ in model (4) by

$$p^*(z_{ij} \mid y_{ij}, v_{ij}, \Phi) \equiv p(z_{ij} \mid t(y_{ij}), v_{ij}, \Phi), \tag{9}$$

and proceed with the final analysis. Because $t(y_{ij})$ is a one-to-one transformation of y_{ij}, $p(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ and $p^*(z_{ij} \mid y_{ij}, v_{ij}, \Phi)$ define the same conditional distribution. But the latter provides a better calibrated estimation of Φ by adjusting for the effect of shrinkage. See Foster and Stine (2004) for more discussion about calibration.

Effectively, the proposed two steps use the SBART model to impute the rank of the missing income variable, and use an observed distribution to set specific values. This approach is valid because both CPS and SIPP are conducted by the US Census Bureau to collect information from representative samples of the US population.

This adjustment can be automated in each MCMC iteration, where we readjust the values of a and b such that the selected quantiles of $t(y_{ij}^{(k)})$ match the corresponding quantiles in the empirical distribution of $\boldsymbol{X}$. We conducted a simple simulation study to assess the performance of the automated adjustment. Because the shrinkage effect

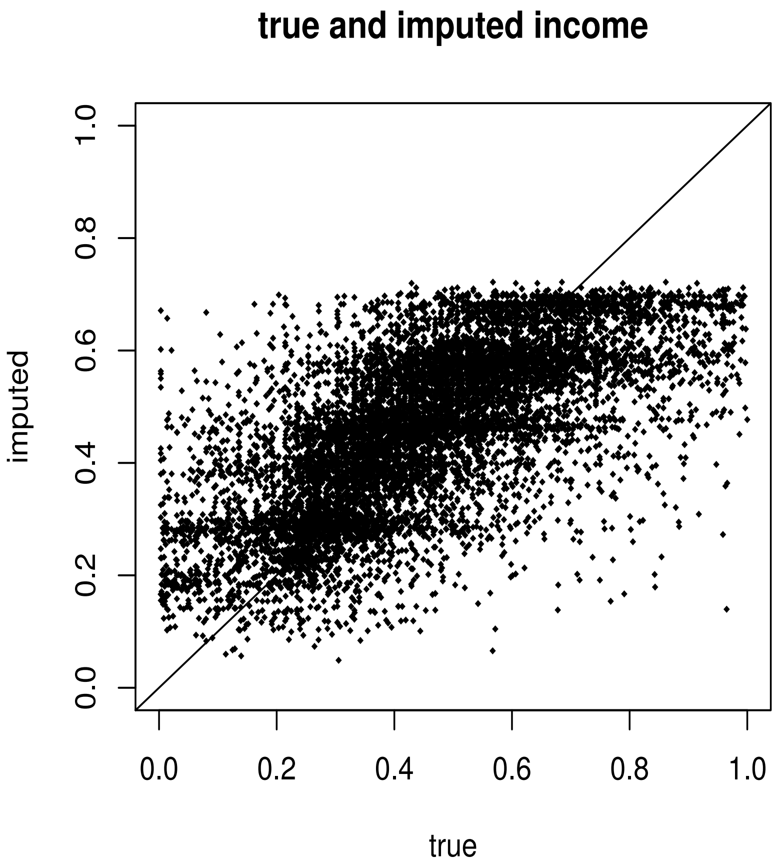


Figure 4: CPS survey: True and imputed income. Income is scaled between 0 and 1. Note the severe shrinkage in the imputed income.

Table 2: MSE from Simulation to Check Adjustment

	(a)	(b)
β_6	0.0078	0.0106
$\boldsymbol{Y}$	0.0238	0.0240

Column (a) with automated adjustment; Column (b) without adjustment.

is more obvious when the sample size is large, we set $n = m = 2000$. The simulation truth is similar to the model assumed in Section 3, except that we drop the spatial component $\boldsymbol{\theta}$ to facilitate computation, and the residual effects in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed to have Student t distribution with 3 degree of freedom. The MSE of the estimated regression coefficients and imputed $\boldsymbol{Y}$ are presented in Table 2. Because Table 2 is based on simulations with a larger sample size and a simpler model, the MSE are much smaller than those in Table 1. Our primary interest is in β_6, the regression coefficient of $\boldsymbol{Y}$. Without adjustment, the shrinkage effect leads to overestimation of β_6. With the automated adjustment, the MSE of β_6 is reduced from 0.0106 to 0.0078.

The simulation indicates that the adjustment can provide better calibrated estimates when there is some shrinkage effect induced by imputation of the missing variable. However, we caution that such an adjustment for shrinkage is ad hoc, and it relies heavily on the assumption that D_1 and D_2 are representative samples of the same population. Researchers should carefully check this assumptions before implementing the approach.

4.4 Results

Table 3 lists the posterior means and standard deviations of the regression coefficients $\boldsymbol{\beta}$ under three inference approaches, which are implemented by MCMC simulation. One set of inference summaries is based on true income and model (7). This serves as the gold standard. The second set of inferences is based on missing income and model (4). The third set is based on missing income and model (9). Both model (4) and model (9) are SBART models, the difference being that model (9) adjusts for the shrinkage effect while model (4) does not. Table 3 shows that if we ignore the shrinkage effect, model (4) will lead to a conclusion that overstates the effect of income. The posterior means based on model (7) and (9) are similar, suggesting that our method successfully merges information from two datasets and provides a good estimate of the relationship between self-perceived health status and income. Due to the uncertainty induced by imputing the missing income, the standard deviations under model (9) are slightly larger. The estimated regression coefficients suggest that subjects with higher income tend to have a better self-perceived health status. Women generally report better self-perceived health. Additionally, younger age and health insurance coverage are associated with better self-perceived health status. We plot the imputed income based on samples from $p(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{V}, \boldsymbol{W})$ versus the true income in Figure 4. The spatial correlation parameter ρ has a posterior mean 0.362 and standard deviation 0.242, indicating a moderate spatial correlation. A histogram of the samples from its posterior distribution is plotted in Figure 5.

Table 3: Real Data, Posterior Mean (Standard deviation) of $\boldsymbol{\beta}$

	model (7)	model (4)	model (9)
Intercept	-4.157(0.073)	-3.982(0.075)	-4.019(0.075)
Health insurance	0.865(0.059)	0.624(0.069)	0.619(0.069)
Sex	-0.194(0.047)	-0.316(0.054)	-0.320(0.055)
Age	0.057(0.001)	0.052(0.001)	0.052(0.001)
Income	-2.513(0.126)	-3.392(0.216)	-2.677(0.167)

Model (7) uses true income; Model (4) uses SBART to impute "missing" income without adjusting for shrinkage; Model (10) uses SBART to impute "missing" income and adjusts for shrinkage.

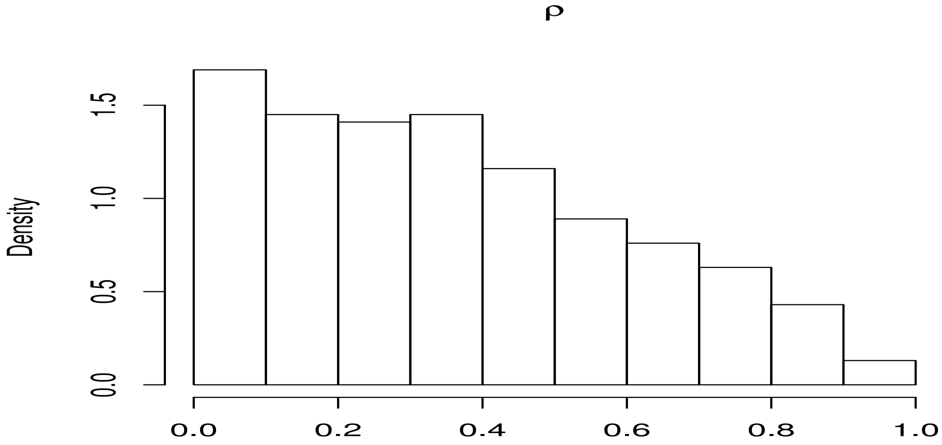


Figure 5: CPS and SIPP surveys. Histogram of $p(\rho \mid data)$.

Besides comparing our results with those based on the complete data, we also compare with results from a census-based approach, which supplements missing individual-level variables with aggregate information based on the neighborhood socioeconomic profile. With MSA being the finest available spatial area, we could supplement missing y_{ij} with average personal income from MSAi. However, compared with the average by census block or census track, the average by MSA is much coarser and would result in a large imputation error. To achieve a fairer comparison with the proposed method we instead proceed as follows. In the CPS dataset, about 41.5% of the records contain county codes. To investigate the performance of census-based methods with finer area units, we create D_1^* by randomly sampling 10,000 observations from those that have county code in the original CPS dataset. We then replace the missing income with county median income (denoted by $\widetilde{\boldsymbol{Y}}$) from the US census. Conditioning on $(\boldsymbol{Z}, \widetilde{\boldsymbol{Y}}, \boldsymbol{V})$ we report inference on Φ under model (7). This is the result from the census-based method. Table 4 lists the posterior means (standard deviations) of $\boldsymbol{\beta}$ from three procedures: (a) based on model (7) and true income; (b) the proposed method, based on model (9) with missing income; (c) census-based method, based on model (7) and median income at county level. Because Table 3 is based on D_1 while Table 4 is based on D_1^*, the estimates in the two tables do not match exactly. The estimates from the proposed method are close to those based on true incomes. This is not the case for the estimates based on imputation by county median income. The estimated coefficients of health insurance coverage and income are quite different from those based on true incomes. Most strikingly, the estimated coefficient of sex switches the sign. This could lead to very misleading conclusions. In summary, Table 4 shows that our model provides an improvement of the census-based method. This is true even though we have improved the latter by using county median income while keeping our proposed method at the MSA-level, a coarser spatial area.

Table 4: Comparing with Census-Based Method

	(a)	(b)	(c)
Intercept	-4.220(0.073)	-4.129(0.075)	-4.380(0.076)
Health insurance	0.841(0.059)	0.734(0.066)	1.116(0.057)
Sex	-0.088(0.047)	-0.165(0.056)	0.148(0.046)
Age	0.054(0.001)	0.052(0.002)	0.052(0.001)
Income	-2.230(0.127)	-2.335(0.164)	-1.673(0.248)

Column (a) uses true income; column (b) uses SBART to impute missing income and adjusts for shrinkage; column (c) uses county median income as imputation.

5 Discussion

We have developed an approach that allows researchers to borrow information across surveys and investigate hypotheses that cannot be considered using only one dataset alone. The proposed method is flexible and fully model-based. The key assumption is that $(\boldsymbol{Y}, \boldsymbol{V})$ and $(\boldsymbol{X}, \boldsymbol{W})$ are independent samples from the same model. This assumption allows researchers to apply the knowledge learned from $(\boldsymbol{X}, \boldsymbol{W})$ to $(\boldsymbol{Y}, \boldsymbol{V})$. This facilitates imputation of the missing $\boldsymbol{Y}$. By specifying a flexible SBART model, the proposed method does not make restrictive assumptions about the specific model for $(\boldsymbol{X}, \boldsymbol{W})$.

In the simulation study and the data analysis example we have assumed parametric models for the regression of $\boldsymbol{Z}$ and $\boldsymbol{Y}$. This parametric form, however, is not a requirement for the proposed approach. It is unrelated to the missingness of $\boldsymbol{Y}$. Alternatively, a non-parametric regression model could be used. The only caveat is that the increased uncertainty induced by the imputation of $\boldsymbol{Y}$ might make meaningful data analysis with a non-parametric model difficult.

The proposed imputation of the missing variable is a data-driven procedure. That is, in each MCMC iteration, we have a large number of trees such that each contributes a small portion of the conditional mean. Therefore it is difficult to evaluate the relationship between the missing variable and individual covariates. It is not a critical issue if the primary interest is to explore the relationship between $\boldsymbol{Z}$ and $\boldsymbol{Y}$, instead of $\boldsymbol{Y}$ and $\boldsymbol{V}$. If the researchers are interested in the the marginal effect of a single predictor, partial dependence plots might be a useful tool. See Friedman (2001) and Chipman et al. (2006a) for details.

Appendix: MCMC Sampling Schemes

We use MCMC posterior simulation to implement inference in model (4). See, for example, Gamerman (1997) for a review of MCMC methods. In the following discussion we use $[U \mid \cdots]$ to indicate that the random variable U is updated conditional on the currently imputed values of all other parameters. The transition probability for the implemented MCMC is defined by the following steps.

Step 1. Updating Φ.

$$[\Phi \mid \cdots] \propto p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi) \cdot p(\Phi).$$

The updating of Φ depends on the specific form of $p(\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{V}, \Phi)$, which in our example is either a linear regression model or an ordinal probit model. There are well established methods to update parameters in such models. For example, see Gelman et al. (2003) and Albert and Chib (1993).

Step 2. Updating f and σ^2.

$$[f, \sigma^2 \mid \cdots] \propto p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) p(f) p(\sigma^2). \tag{10}$$

If we define $x^*_{ij} = x_{ij} - \theta_i$ and $y^*_{ij} = y_{ij} - \theta_i$, then (10) is equivalent to

$$[f, \sigma^2 \mid \cdots] \propto \prod \left\{ p(x^*_{ij} \mid w_{ij}, f, \sigma^2) \right\} \prod \left\{ p(y^*_{ij} \mid v_{ij}, f, \sigma^2) \right\} p(f) p(\sigma^2), \tag{11}$$

with

$$\begin{aligned} p(x^*_{ij} \mid w_{ij}, f, \sigma^2) &= N(f(w_{ij}), \sigma^2), \\ p(y^*_{ij} \mid v_{ij}, f, \sigma^2) &= N(f(v_{ij}), \sigma^2). \end{aligned}$$

Note that (11) is exactly a BART model with x^*_{ij} and y^*_{ij} being the dependent variable, and the updating algorithm can be found in Chipman et al. (2006a) Section 4.

Step 3. Updating $\boldsymbol{\theta}$.

$$[\boldsymbol{\theta} \mid \cdots] \propto p(\boldsymbol{X} \mid \boldsymbol{W}, f, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{Y} \mid \boldsymbol{V}, f, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta} \mid \rho, \delta^2).$$

Define $e_{ij} = x_{ij} - f(w_{ij})$ and $s_{ij} = y_{ij} - f(v_{ij})$, and use $\boldsymbol{e}$ and $\boldsymbol{s}$ to denote the collection of e_{ij} and s_{ij}, respectively. We find

$$\begin{aligned}[\boldsymbol{\theta} \mid \cdots] \quad \propto \quad & \exp\Big\{-\frac{(\boldsymbol{e}-\boldsymbol{U}_x\boldsymbol{\theta})'(\boldsymbol{e}-\boldsymbol{U}_x\boldsymbol{\theta}) + (\boldsymbol{s}-\boldsymbol{U}_y\boldsymbol{\theta})'(\boldsymbol{s}-\boldsymbol{U}_y\boldsymbol{\theta})}{2\sigma^2}\Big\} \\ \cdot \quad & \exp\Big\{-\frac{1}{2\delta^2}\boldsymbol{\theta}'(\boldsymbol{H}-\rho\boldsymbol{C})\boldsymbol{\theta}\Big\},\end{aligned}$$

where $\boldsymbol{U}_x$ and $\boldsymbol{U}_y$ are the design matrix of $\boldsymbol{\theta}$ corresponding to $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. We can show that $[\boldsymbol{\theta} \mid \cdots]$ is a normal distribution with variance $[(\boldsymbol{U}_x'\boldsymbol{U}_x + \boldsymbol{U}_y'\boldsymbol{U}_y)/\sigma^2 + (\boldsymbol{H}-\rho\boldsymbol{C})/\delta^2]^{-1}$ and mean $[(\boldsymbol{U}_x'\boldsymbol{U}_x + \boldsymbol{U}_y'\boldsymbol{U}_y)/\sigma^2 + (\boldsymbol{H}-\rho\boldsymbol{C})/\delta^2]^{-1}(\boldsymbol{U}_x'\boldsymbol{e} + \boldsymbol{U}_y'\boldsymbol{s})/\sigma^2$.

Step 4. Updating ρ and δ^2.

$$[\rho, \delta^2 \mid \cdots] \propto p(\boldsymbol{\theta} \mid \rho, \delta^2) p(\rho) p(\delta^2).$$

CAR is a widely used spatial model and the posterior sampling of ρ and δ^2 has been discussed extensively in literature. For example, see He and Sun (2000).

Step 5. Updating $\boldsymbol{Y}$. We update $\boldsymbol{Y}$ one element at a time, i.e.,

$$[y_{ij} \mid \cdots] \propto p(z_{ij} \mid y_{ij}, v_{ij}, \Phi) p(y_{ij} \mid v_{ij}, f, \boldsymbol{\theta}, \sigma^2).$$

Under model (6), a linear regression model, we have

$$[y_{ij} \mid \cdots] \propto \exp\Big\{-\frac{1}{2\tau^2}(z_{ij} - h_{ij}^* - y_{ij}\beta_6)^2\Big\} \exp\Big\{-\frac{1}{2\sigma^2}(y_{ij} - f(v_{ij}) - \theta_i)^2\Big\},$$

where $h_{ij}^* = \beta_0 + v_{ij4}\beta_1 + v_{ij5}\beta_2 + v_{ij6}\beta_3 + v_{ij7}\beta_4 + v_{ij8}\beta_5$. We can show that $[y_{ij} \mid \cdots]$ is normal with variance $(\beta_6^2/\tau^2 + 1/\sigma^2)^{-1}$ and mean

$$\Big(\frac{\beta_6^2}{\tau^2} + \frac{1}{\sigma^2}\Big)^{-1}\Big(\frac{1}{\tau^2}\beta_6(z_{ij} - h_{ij}^*) + \frac{1}{\sigma^2}(f(v_{ij}) + \theta_i)\Big).$$

Under model (8), an ordinal probit model, we have

$$[y_{ij} \mid \cdots] \propto \exp\Big\{-\frac{1}{2\tau^2}(\eta_{ij} - h_{ij}^{\triangle} - y_{ij}\beta_4)^2\Big\} \exp\Big\{-\frac{1}{2\sigma^2}(y_{ij} - f(v_{ij}) - \theta_i)^2\Big\},$$

where $h_{ij}^{\triangle} = \beta_0 + v_{ij1}\beta_1 + v_{ij2}\beta_2 + v_{ij3}\beta_3$. Thus $[y_{ij} \mid \cdots]$ is normal with variance $(\beta_4^2/\tau^2 + 1/\sigma^2)^{-1}$ and mean

$$\Big(\frac{\beta_4^2}{\tau^2} + \frac{1}{\sigma^2}\Big)^{-1}\Big(\frac{1}{\tau^2}\beta_6(z_{ij} - h_{ij}^{\triangle}) + \frac{1}{\sigma^2}(f(v_{ij}) + \theta_i)\Big).$$

References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88: 669–679. 629

Besag, J., York, J., and Molli, A. (1991). "Bayesian image restoration, with two applications in spatial statistics." *Annals of the Institute of Statistical Mathematics*, 43: 1–20, (Disc: pp21–59). 616

Breiman, L. (1996). "Bagging predictors." *Machine Learning*, 24: 123–140. 615

— (2001). "Random Forests." *Machine Learning*, 45: 5–32. 615

Byrne, C., Nedelman, J., and Luke, R. (1994). "Race, socioeconomic status, and the development of end-stage renal disease." *American Journal of Kidney Diseases*, 23(1): 16–22. 612

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). "Bayesian CART Model Search." *Journal of the American Statistical Association*, 93: 935–948, (C/R: P948–960). 615

— (2006a). "BART: Bayesian Additive Regression Trees." Technical report, Department of Mathematics and Statistics, Acadia University, Canada. 612, 614, 615, 616, 618, 619, 623, 624, 629

Chipman, H. A., George, E. I., McCulloch, R. E., and Musio, M. (2006b). "Spatial BART." *Abstract at the Valencia/ISBA Eighth World Meeting on Bayesian Statistics, Valencia*. 613

Clayton, D. and Kaldor, J. (1987). "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping." *Biometrics*, 43: 671–681. 616

Clyde, M. and Lee, H. (2001). "Bagging and Bayesian Bootstrap." In Richardson, T. and Jaakkola, T. (eds.), *Artificial Intelligence and Statistics 2001*, 169–174. 615

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley and Sons. 616

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). "A Bayesian CART Algorithm." *Biometrika*, 85: 363–377. 615

Devesa, S. and Diamond, E. (1983). "Socioeconomic and racial differences in lung cancer incidence." *American Journal of Epidemiology*, 118(6): 818–831. 612

Foster, D. P. and Stine, R. A. (2004). "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy." *Journal of the American Statistical Association*, 99(466): 303–313. 624

Freund, Y. and Schapire, R. (1997). "A decision-theoretic generalization of online learning and an application to boosting." *Journal of Computer and System Sciences*, 55: 119–139. 615

Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines." *The Annals of Statistics*, 19: 1–67, (Disc: P67–141). 618

— (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, 29(5): 1189–1232. 615, 629

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall Ltd. 629

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman & Hall. 629

Geronimus, A. and Bound, J. (1998). "Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples." *American Journal of Epidemiology*, 148(5): 475–486. 612

Gornick, M., Eggers, P., Reilly, T., Mentnech, R., Fitterman, L., Kucken, L., and Vladeck, B. (1996). "Effects of race and income on mortality and use of services among Medicare beneficiaries." *New England Journal of Medicine*, 335(11): 791–799. 612

He, Z. and Sun, D. (2000). "Hierarchical Bayes estimation of Hunting success rates with spatial correlations." *Biometrics*, 56(2): 360–367. 630

Horton, N. J. and Laird, N. M. (1999). "Maximum Likelihood Analysis of Generalized Linear Models with Missing Covariates." *Statistical Methods in Medical Research*, 8: 37–50. 611

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005). "Missing-data Methods for Generalized Linear Models: A Comparative Review." *Journal of the American Statistical Association*, 100(469): 332–346. 611

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. Springer-Verlag Inc. 623

Kraus, J., Fife, D., Cox, P., Ramstein, K., and Conroy, C. (1986). "Incidence, severity, and external causes of pediatric brain injury." *American Journal of Diseases of Children*, 140(7): 687–693. 612

Little, R. J. A. (1992). "Regression with Missing X's: A Review." *Journal of the American Statistical Association*, 87: 1227–1237. 611

Mandelblatt, J., Andrews, H., Kerner, J., Zauber, A., and Burnett, W. (1991). "Determinants of late stage diagnosis of breast and cervical cancer: the impact of age, race, social class, and hospital type." *American Journal of Public Health*, 81(5): 646–649. 612

Rubin, D. B. (1976). "Inference and Missing Data." *Biometrika*, 63: 581–590. 611

Schafer, J. L. and Graham, J. W. (2002). "Missing Data: Our View of the State of the Art." *Psychological Methods*, 7(2): 147–177. 611

Sun, D., Tsutakawa, R. K., and Speckman, P. L. (1999). "Posterior distribution of hierarchical models using CAR(1) distributions." *Biometrika*, 86: 341–350. 616

Whittle, P. (1954). "On stationary processes in the plane." *Biometrika*, 41: 434–449. 616