Nested Partition Models

Abel Rodriguez and Kaushik Ghosh

Abstract

This paper introduces a flexible class of models for relational data based on a hierarchical extension of the two-parameter Poisson-Dirichlet process. The model is motivated by two different applications: 1) A study of cancer mortality rates in the U.S., where rates for different types of cancer are available for each state, and 2) the analysis of microarray data, where expression levels are available for a large number of genes in a sample of subjects. In both these settings, we are interested in improving estimation by flexibly borrowing information across rows and columns while partitioning the data into homogeneous subpopulations. Our model allows for a novel nested partitioning structure in the data not provided by existing nonparametric methods, in which rows are clustered while simultaneously grouping together columns within each cluster of rows.

1 Introduction

Bayesian nonparametric (BNP) mixture models have become popular in the last decade and have been applied in fields as diverse as finance (Kacperczyk et al., 2003; Rodriguez & Ter Horst, 2009), econometrics (Chib & Hamilton, 2002; Hirano, 2002), epidemiology (Dunson, 2005), genetics (Medvedovic & Sivaganesan, 2002; Dunson et al., 2007a), medicine (Kottas et al., 2002; Bigelow & Dunson, 2007) and auditing (Laws & O'Hagan, 2002). In the simple case where we are interested in estimating a single distribution from an independent and identically distributed sample $y_1, \ldots, y_n$, nonparametric mixtures assume that observations arise from a convolution

$$y_i \sim \int \psi(\cdot|\boldsymbol{\theta}) G(\mathrm{d}\boldsymbol{\theta})$$

where $\psi(\cdot|\boldsymbol{\theta})$ is a parametric kernel indexed by $\boldsymbol{\theta}$, and then place a rich prior on the mixing distribution G, which is often assumed to be almost surely discrete. For example, assuming that G follows a Dirichlet process (DP) prior (Ferguson, 1973; Blackwell & MacQueen, 1973; Ferguson, 1974; Sethuraman, 1994) leads to the well known Dirichlet process mixture (DPM) models (Lo, 1984; Escobar, 1994; Escobar & West, 1995). The discrete nature of G induces a partition of the observations into groups, with observations on each group assumed to be independent and identically distributed samples from the kernel $\psi(\cdot|\boldsymbol{\theta})$ and share a common value for $\boldsymbol{\theta}$. However, unlike traditional clustering models, BNP mixture models automatically choose the number of components in the mixture and provide for a straightforward predictive rule for new observations, simplifying prediction.

In this paper, we consider nested partition models as a mechanism to generate nonparametric and semiparametric priors for *matrix* data. Specifically, let the observations y_{ij} for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, be conveniently arranged in a matrix $\mathbf{Y} = [y_{ij}]$ with I rows and J columns. Typically $\mathbf{Y}$ will be the result from a non-replicated experiment with two crossed factors. For example, y_{ij} might correspond to the mortality count in state i due to cancer type j during a given calendar year, or to the expression level of gene j for patient i in a microarray experiment. More generally, we are interested in relational data, which describes the interactions between members of two or more classes of objects. Focusing for simplicity of exposition on the case of binary relations, if $\mathcal{S}$ and $\mathcal{R}$ are two sets with I and J objects respectively (possibly with $\mathcal{R} = \mathcal{S}$ and $I = J$), we can interpret the value y_{ij} as measuring the strength of the relationship between the i-th object in $\mathcal{S}$ (patient) and the j-th object in $\mathcal{R}$ (gene). In this paper we will focus on hierarchical models of the form

$$y_{ij} \sim \psi(\cdot|\mathbf{x}_{ij}, \boldsymbol{\theta}_{ij}, \boldsymbol{\nu}) \qquad\qquad \boldsymbol{\theta}_{ij} \sim G$$

where G is the unknown distribution of the random effects and $\psi(\cdot|\mathbf{x}_{ij}, \boldsymbol{\theta}_{ij}, \boldsymbol{\nu})$ is the conditional distribution of observation y_{ij} given a vector of predictors $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijq})$, random effects $\boldsymbol{\theta}_{ij}$ and the vector of fixed effects $\boldsymbol{\nu}$. In estimating the random effects, it is desirable to borrow information across observations by exploiting the structure on the data. In particular, for crossed experiments where both rows and columns represent levels of some categorical variable (rather than replicates), we expect that observations that are in either the same row or the same column would have a stronger correlation with each other. For example, in the analysis of microarray experiments it is reasonable to assume *a priori* that expression levels for the same gene across two different subjects have a stronger correlation than two genes from two different subjects. Even more, if a clustering structure is induced in the observations, we would like the same genes to be simultaneously grouped together for all individuals in any given cluster of subjects.

A simple approach to deal with matrix data is to assume that G follows a nonparametric prior, such as a standard Dirichlet process or a Pólya tree. However, such a model implies that observations are conditionally exchangeable, and therefore the correlation structure is independent of the row/column assignment. Specific BNP mixture models for matrix data have previously been considered in the literature. For example, Dunson et al. (2008) introduced the matrix stick-breaking process (MSBP), which uses weights dependent on the location of the observation in the array to construct dependent, row-specific mixing distribution G_i. Therefore, the MSBP allows for clustering within rows but not across columns. The clustering structure generated by other models for nested samples such as the hierarchical Dirichlet process (HDP) (Teh et al., 2006) and the nested Dirichlet process (NDP) (Rodriguez et al., 2008) are similarly difficult to interpret in crossed experiments.

There is a rich literature on BNP models for situations where the mixing distribution G is allowed to depend on covariates. Models based on dependent stick-breaking priors (MacEachern, 1999, 2000) have been developed for spatial and temporal process (Gelfand et al., 2005; Griffin & Steel, 2006b; Duan et al., 2007; Dunson & Park, 2007; Rodriguez & Ter Horst, 2008), as well as for nonparametric analysis of variance (DeIorio et al., 2004) and models for independent and identically distributed distribution (Teh et al., 2006; Rodriguez et al., 2008). Alternative approaches for this problem based on linear combinations of nonparametric processes include Dunson et al. (2007b), Müller et al. (2004) and Griffin & Steel (2006a). Unfortunately, none of these approaches is suitable to model relational data.

Some models for relational data based on Bayesian nonparametric models have been studied in the machine learning community, where the random blocks model (Kemp et al., 2006; Xu et al., 2006) has dominated attention. A recent generalization of the random blocks model is the Mondrian process (Roy & Teh, 2009), which constructs groups of cells by successive random partitions on the previously induced subsets. The nested parition model we discuss in this paper can be conceived as a different generalization of the infinite random blocks model in Kemp et al. (2006) that allows for partitions in set $\mathcal{R}$ to be nested in the partitions of set $\mathcal{S}$.

The rest of the paper is organized as follows: Section 2 briefly reviews product partition models and their connection to Bayesian nonparametric models. Section 3 introduces our nested partition models. Section 4 describes a Pólya urn representation for the nested partition process and describes a computational strategy to fit these models based on Markov chain Monte Carlo algorithms. Section 5 presents two illustrations, one on modeling contingency tables and another one on the analysis of microarray data. Finally, Section 6 discusses additional extensions for the model as well as future directions of work.

2 Partition models and Bayesian nonparametric mixtures

Let $S = \{1, \ldots, n\}$ denote a set of labels identifying observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$. A partition of S, denoted $\eta(S)$, is a collection of subsets $S_1, \ldots, S_L$ (referred to as *clusters* or *groups*) such that $S = \cup_{l=1}^{L} S_l$ and $S_l \cap S_{l'} = \emptyset$. In many cases, it is convenient to represent the partition $\eta(S)$ using a set of indicators $\xi_1, \ldots, \xi_n$ such that $\xi_i = \xi_{i'} = l$ if and only if i and i' belong to the same cluster S_l.

Bayesian partition models are hierarchical models that place a prior on $\mathbf{y}_1, \ldots, \mathbf{y}_n$ by first placing a prior on the set $\eta(S)$ and then, conditional on $\eta(S)$, assume that all observations whose labels belong to cluster S_l are distributed according to some $p_l(\{\mathbf{y}_i : i \in S_l\})$. A well known example is the class of product partition models (Hartigan, 1990; Barry & Hartigan, 1993; Quintana & Iglesias, 2003), which place a prior on $\eta(S)$ such that $\mathsf{Pr}(\eta(S)) \propto \prod_{l=1}^{L} c(S_l)$, where $c(\cdot)$ is a positive cohesion function and

$$p_l(\{\mathbf{y}_i : i \in S_l\}) = \int \left[\prod_{i \in S_l} \psi(\mathbf{y}_i | \boldsymbol{\theta}) \right] G_0(\mathrm{d}\boldsymbol{\theta})$$

where ψ is the conditional distribution of the data given the cluster specific parameters in $\boldsymbol{\theta}$, and G_0 is the prior for $\boldsymbol{\theta}$, which is common to all clusters.

There is a close relationship between Bayesian partition models and nonparametric mixture models (Quintana, 2006). For example, consider a random sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ where $\mathbf{y}_i \sim \psi(\cdot | \boldsymbol{\theta}_i)$ and $\boldsymbol{\theta}_i \sim G$ with $G \sim \mathsf{DP}(\beta G_0)$ following a Dirichlet process. Therefore,

$$G(\cdot) = \sum_{l=1}^{\infty} w_l \delta_{\boldsymbol{\phi}_l}(\cdot) \tag{1}$$

where the atoms $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots$ are independent and identically distributed samples from a baseline measure G_0 defined on $(\mathcal{X}, \mathcal{B})$ and the stick-breaking weights $w_l = u_l \prod_{r<l}(1 - u_r)$, with the stick-breaking ratios $u_1, u_2, \ldots$ being independent and identically distributed from a $\mathsf{Beta}(1, \beta)$. The pattern of ties in $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ induces a distribution on $\eta(S)$ which agrees with that from a product partition model with cohesion function $c(S_l) = \beta(|S_l| - 1)!$.

The Dirichlet process has limitations as a clustering procedure; for example, the expected number of clusters in the partition tends to grow logarithmically with the number of observations n, and it tends to favor partitions with a small number of large clusters together with a large number of small clusters. More flexible priors on partitions can be obtained from other species sampling models (McCloskey, 1965; Lijoi et al., 2005, 2007; Lee et al., 2009). In the sequel, we focus on the two-parameter Poisson-Dirichlet process (Pitman, 1995; Sudderth & Jordan, 2009).

We say that G follows a Poisson-Dirichlet process with discount α, strength β and baseline G_0, denoted $\mathsf{PY}(\alpha, \beta, G_0)$, if it has the representation in (1) with independent (but not identically distributed) stick-breaking ratios $u_l \sim \mathsf{Beta}(1-\alpha, \beta + l\alpha)$, $0 \leq \alpha < 1$ and $\beta > -\alpha$. Note that by setting $\alpha = 0$, we recover the Dirichlet process as a special case of the Poisson-Dirichlet process. However, for $\alpha > 0$, the growth in the number of clusters follows a power law with exponent α, allowing the model to add new clusters much faster than the Dirichlet process. In our applications, we place priors on α and β, which allow the data to inform us about the appropriate rate of growth in the number of clusters.

There is a straightforward interpretation for the parameters of a Poisson-Dirichlet process. If $G \sim \mathsf{PY}(\alpha, \beta, G_0)$, then for any Borel set $B \in \mathcal{B}$,

$$\mathsf{E}(G(B)) = G_0(B), \qquad \mathsf{Var}(G(B)) = \frac{1-\alpha}{1+\beta} G_0(B)(1 - G_0(B)).$$

Therefore, G_0 and be interpreted as the centering distribution, while α and β contol the amount of variation about this center.

The probability distribution on $\eta(S)$ induced by a Poisson-Dirichlet can be obtained using the generalized Pólya urn representation for the process (Pitman, 1995, 1996). In terms of the indicators $\xi_1, \ldots, \xi_n$, we have that $\xi_1 = 1$ and for $j \geq 2$,

$$\mathsf{Pr}(\xi_j = k | \xi_{j-1}, \ldots, \xi_1) = \sum_{l=1}^{L^{j-1}} \frac{n_l^{j-1} - \alpha}{\beta + j - 1} \delta_l(k) + \frac{\beta + \alpha L^{j-1}}{\beta + j - 1} \delta_{L^{j-1}+1}(k), \tag{2}$$

where the L^{j-1} is the number of clusters observed among the first $j-1$ observations and n_l^{j-1} is the number of observations in cluster l ($l = 1, \ldots, L^{j-1}$) among the first $j-1$ observations. Since the model leading to this Pólya urn assumes exchangeability among the observations, equation (2) also provides the full conditional prior distribution on the cluster assignment, which can be used to develop computational schemes based on Markov chain Monte Carlo methods (Ishwaran & James, 2001). Also, starting with (2), Lijoi et al. (2007) showed that the prior probability of a partition $\eta(\{1,\ldots,I\})$ composed of L clusters with sizes $n_l = |S_l|$ generated by a $\mathsf{PY}(\alpha,\beta,G_0)$ with non-atomic baseline measure G_0 is given by

$$\mathsf{Pr}(L,n_1,\ldots,n_L|\alpha,\beta) = \frac{\prod_{l=1}^{L}(\beta+l\alpha)}{(\beta+1)_{I-1}}\prod_{l=1}^{L}(1-\alpha)_{n_l-1} \tag{3}$$

where $(a)_s = a(a+1)\ldots(a+s-1)$, irrespective of G_0.

3 Nested partition models

We turn our attention now to matrix data. As before, for $i = 1,\ldots,I$ and $j = 1,\ldots,J$, let $y_{ij} \sim \psi(\cdot|\boldsymbol{\theta}_{ij},\boldsymbol{\nu})$ where ψ is parametric kernel indexed by $\boldsymbol{\theta}_{ij}$. When modeling gene expression data, y_{ij} might denote the expression level for gene j in subject i, and we might take $\boldsymbol{\theta}_{ij} = (\mu_{ij},\sigma_{ij})$ and $y_{ij} \sim \mathsf{N}(\mu_{ij},\sigma_{ij}^2)$. Our goal is to generate a nonparametric mixture model for this data. However, for interpretation and efficiency purposes, we want to generate a very specific clustering structure. First, we want to identify groups of subjects with similar overall expression patterns. In addition, within each group of subjects, we want to identify co-regulated genes with similar expression levels.

In the sequel, let $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1},\ldots,\boldsymbol{\theta}_{iJ})'$ be the vector of parameters corresponding to row (e.g., subject) i. In order to cluster subjects, it is natural to assume $\boldsymbol{\theta}_i \sim F$, where $F \sim \mathsf{PY}(\alpha,\beta,\mathbf{H})$, i.e.,

$$F = \sum_{k=1}^{\infty}\pi_k\delta_{\boldsymbol{\phi}_k}, \tag{4}$$

is a (random) discrete distribution such that $\pi_k = v_k\prod_{s<k}(1-v_s)$ with $v_k \sim \mathsf{Beta}(1-\alpha,\beta+k\alpha)$, and $\boldsymbol{\phi}_k$ are independent and identically distributed from some baseline measure $\mathbf{H}$.

To obtain cluster-specific partitions for the columns, we need to carefully choose the baseline measure $\mathbf{H}$ generating the unique, cluster-specific atoms $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots$. Writing $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kJ})'$, one option is to assume that, for $j = 1, \ldots, J$, $\phi_{kj} \sim G_k$ where $G_1, G_2, \ldots$ is a sequence of independent and identically distributed random distributions such that $G_k \sim \mathsf{PY}(\gamma, \epsilon, G_0)$. Using (2), this leads to a baseline measure

$$\mathbf{H}(\boldsymbol{\phi}) = \prod_{j=1}^{J} H_j(\phi_j | \phi_{j-1}, \ldots, \phi_1) \tag{5}$$

where

$$H_j(\phi_j | \phi_{j-1}, \ldots, \phi_1) = \sum_{l=1}^{L^{j-1}} \frac{n_l^{j-1} - \gamma}{\epsilon + j - 1} \delta_{\phi_l^*} + \frac{\epsilon + \gamma L^{j-1}}{\epsilon + j - 1} G_0, \tag{6}$$

$\phi_1^*, \ldots, \phi_{L^{j-1}}^*$ are the L^{j-1} distinct values among $\phi_1, \ldots, \phi_{j-1}$ and n_l^{j-1} is the number of components among $\phi_1, \ldots, \phi_{j-1}$ to be assigned to the l-th cluster. We call the process generating the matrix of random effects $\boldsymbol{\Theta} = [\boldsymbol{\theta}_{ij}]$ a nested partition model (NPM), and write

$$\boldsymbol{\Theta} \sim \mathsf{NPM}(\alpha, \beta, \gamma, \epsilon, G_0).$$

Notationwise, the NPM is somewhat reminiscent of the hierarchical Dirichlet process (HDP) (Teh et al., 2006) and the nested Dirichlet process (NDP) (Rodriguez et al., 2008). However, the nested partition model is quite different. In particular, note that both the HDP and the NDP assume exchangeability of *cells* within each row, rather than exchangeability of columns within clusters of rows, as in the NPM (see Figure 1).

The NPM is closely related to two nonparametric models for relational data recently introduced in the machine learning literature — the infinite random blocks model (Kemp et al., 2006) and the Mondrian process (Roy & Teh, 2009). However, there are differences in the partitions generated by each of the models (see Figure 2). In particular, it is clear from our stylized graphs that the random blocks models is but a special case of the NPM where

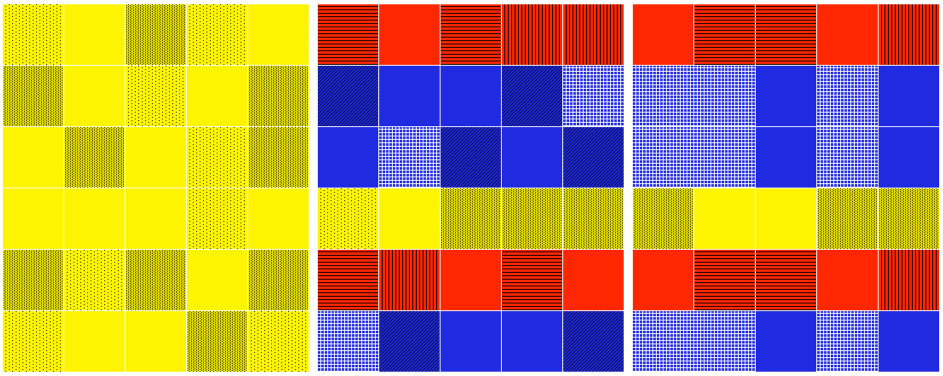


Figure 1: Sample clustering structure induced in a 6×5 matrix by three different nonparametric priors: (a) the HDP, (b) the NDP and (c) the NPM respectively. Colors are used to represent clusters of rows, while patterns are used to represent clusters of cells. The HDP (left panel) only clusters cells across rows; the NDP (center panel) clusters rows and, within each cluster of rows, clusters cells without regard to which column they belong to; the NPM (right panel) clusters rows, and within clusters of rows, clusters columns of cells together.

partitions within each group of rows have been constrained to be equal. In addition, there are important differences in interpretation between the NPM and the Mondrian process: by successively parititioning the matrix along both dimensions, the Mondrian process can generate very flexible partitions of the matrix cells; however, the resulting blocks cannot be interpreted in terms of nested row/column clusters.

We proceed now to discuss some of the properties of the NPM. The correlation between entries in the matrix can be easily be obtained (see Appendix

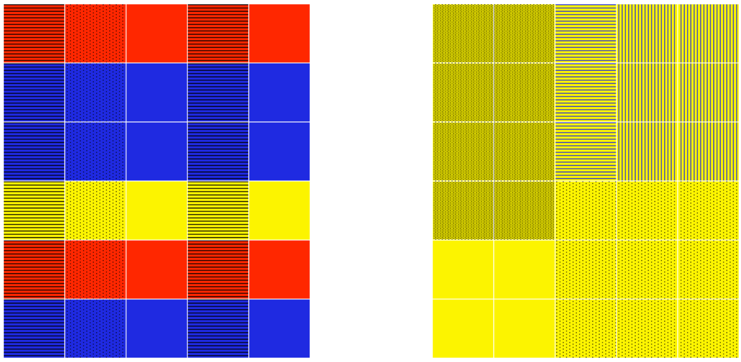


Figure 2: Sample clustering structure induced in a 6×5 matrix by the random blocks model and the Mondrian process. Again the differences between the models are clear: The vertical cuts induced by the random blocks models (left panel) are shared by all clusters of rows (making it a special case of the NPM) while the Mondrian process (right panel) producess patches of clustered cells but does not provide by itself a clustering structure across rows and columns as the NPM does.

A), yielding

$$\mathsf{Cor}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) = \begin{cases} \frac{1-\alpha}{1+\beta} & i \neq i', j = j' \\ \frac{1-\gamma}{1+\epsilon} & i = i', j \neq j' \\ \frac{1-\alpha}{1+\beta}\frac{1-\gamma}{1+\epsilon} & i \neq i', j \neq j' \end{cases}$$

As desired, $\mathsf{Cor}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'})$ is smaller if both $i \neq i'$ and $j \neq j'$, in accordance to what we expect the structure in the data to be. Some interesting models arise as limiting cases of the NPM. As both $\gamma, \epsilon \rightarrow 0$ we have $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}_{ij'}$ for all $j = j'$. Therefore, the model clusters rows assuming that the random effect is the same for all columns on each cluster of rows. On the other hand, if $\epsilon \rightarrow \infty$ then $\boldsymbol{\theta}_{ij} \neq \boldsymbol{\theta}_{ij'}$ almost surely, and the model does not borrow information across columns. Similarly letting $\alpha, \beta \rightarrow 0$ induces a model where all rows are assumed to be in a single cluster, and columns are clustered together, while $\beta \to \infty$ leads to a model where no information is borrowed across rows.

Another interesting propertiy of the NPM is that it defines an exchangeable prior on arrays, in the sense of Hoover (1979), Aldous (1981) and Diaconis & Janson (2007). A distribution on the entries of an array is said to be *separately exchangeable* if the distribution is invariant to *separate permutations* on each of its dimensions. Similarly, a distribution on arrays that is invariant when *the same permutation* is applied to all dimensions is said to be *jointly exchangeable*. Note that a squared array that is separately exchangeable is always jointly exchangeable, but not the other way around.

Theorem 1 *Let* $\boldsymbol{\Theta} \sim \mathsf{NPM}(\alpha, \beta, \gamma, \epsilon, G_0)$ *be a random matrix with* I *rows and* J *columns. Then the joint distribution on* $\boldsymbol{\Theta}$ *induced by the nested partition model,* $p(\boldsymbol{\Theta})$*, is separately exchangeable. That is, for any two pairs of permutations* $\pi_1(1 : I)$*,* $\pi_2(1 : J)$ *and* $\nu_1(1 : I)$*,* $\nu_2(1 : J)$ *we have* $p(\boldsymbol{\Theta}_{\nu_1(1:I),\nu_2(1:J)}) = p(\boldsymbol{\Theta}_{\pi_1(1:I),\pi_2(1:J)})$*, where* $\boldsymbol{\Theta}_{\pi_1(1:I),\pi_2(1:J)}$ *is the matrix obtained by permuting the rows of* $\boldsymbol{\Theta}$ *using* π_1 *and its columns using* π_2*.*

The proof is straightforward, and relies on the exchangeability of the joint distribution of draws from a two-parameter Poisson-Dirichlet process (Pitman, 1995). Also, the result holds more generally if the Poisson-Dirichlet process is replaced by any other species sampling model. However, note that the model is not invariant to transpositions of the array, which implies that choosing the ordering of the variables (if not of the observations) is important. In the examples we discuss in Section 5 such ordering arises naturally from the problem. More broadly, this means that the NPM is going to be most useful for modeling directed relations.

4 Pólya urn representation and computational methods

We can derive a Pólya urn representation for the NPM, which in turn leads to a fairly straightforward computational algorithm for inference. Since $\boldsymbol{\theta}_i \sim$

F and $F \sim \mathsf{PY}(\alpha, \beta, \mathbf{H})$, exchangeability ensures that the full conditional distribution for $\boldsymbol{\theta}_i$ is given by

$$\boldsymbol{\theta}_i | \{\boldsymbol{\theta}_1,\ \ldots,\ \boldsymbol{\theta}_n\} \backslash \boldsymbol{\theta}_i \sim \sum_{k=1}^{K^{-i}} \frac{n_k^{-i} - \alpha}{\beta + I - 1} \delta_{\boldsymbol{\phi}_k} + \frac{\beta + \alpha K^{-i}}{\beta + I - 1} \mathbf{H}, \tag{7}$$

where $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{K^{-i}}\}$ are the K^{-i} unique values among $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\} \backslash \boldsymbol{\theta}_i$, n_k^{-i} is the number of values in the set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\} \backslash \boldsymbol{\theta}_i$ that are equal to $\boldsymbol{\phi}_k$, and $\mathbf{H}$ is given in equations (5) and (6). Similarly, conditional exchangeability within each cluster of rows implies that the full conditional distribution for the entries of $\boldsymbol{\phi}_k$ is given by

$$\boldsymbol{\phi}_{kj} | \{\boldsymbol{\phi}_{k1}, \ldots, \boldsymbol{\phi}_{kJ}\} \backslash \boldsymbol{\phi}_{kj} \sim \sum_{l=1}^{L_k^{-j}} \frac{m_{kl}^{-j} - \gamma}{\epsilon + J - 1} \delta_{\boldsymbol{\phi}^*_{kl}} + \frac{\epsilon + \gamma L_k^{-j}}{\epsilon + J - 1} G_0 \tag{8}$$

where $\{\boldsymbol{\phi}^*_{k1}, \ldots, \boldsymbol{\phi}^*_{kL_k^{-j}}\}$ are the L_k^{-j} unique values among $\{\boldsymbol{\phi}_{k1}, \ldots, \boldsymbol{\phi}_{kJ}\} \backslash \boldsymbol{\phi}_{kj}$ and m_{ku}^{-j} is the number of values in the set $\{\boldsymbol{\phi}_{k1}, \ldots, \boldsymbol{\phi}_{kJ}\} \backslash \boldsymbol{\phi}_{kj}$ that are equal to $\boldsymbol{\phi}^*_{ku}$.

Equations (7) and (8) can be used to develop a collapsed (marginal) sampler (MacEachern, 1994; Escobar, 1994; Escobar & West, 1995; Neal, 2000). First, we consider the case where the baseline measure G_0 is conjugate to the kernel $\psi(\cdot|\boldsymbol{\theta})$. To develop the sampler, we introduce a collection of indicators $\{\zeta_i\}$ such that $\zeta_i = k$ if and only if $\boldsymbol{\theta}_i = \boldsymbol{\phi}_k$, as well as indicators $\{\xi_{kj}\}$ such that $\xi_{kj} = l$ if and only if $\boldsymbol{\phi}_{kj} = \boldsymbol{\phi}^*_{kl}$. This collapsed sampler sequentially samples the parameters in the model from its full conditional distributions to generate (after a suitable burn-in period) a dependent sample from the posterior distribution of interest.

First, note that the unique $\boldsymbol{\phi}^*_{kl}$s are conditionally independent given the indicators. Therefore, updating them is straightforward as the full conditional distributions are given by

$$\boldsymbol{\phi}^*_{kl} | \cdots \sim \left[\prod_{\{(i,j):\zeta_i = k, \xi_{\zeta_i j} = l\}} \psi(\mathbf{y}_{ij} | \boldsymbol{\phi}^*_{kl}, \boldsymbol{\nu}) \right] G_0(\boldsymbol{\phi}^*_{kl}).$$

Next, we show how to update the partition of columns within each cluster of rows. Since G_0 is conjugate to ψ, we can explicitly integrate out the $\boldsymbol{\phi}_{kl}^{*}$s, and the full conditional posterior distribution for ξ_{kj} is given by

$$\mathsf{Pr}(\xi_{kj} = l | \cdots) = \sum_{u=1}^{L_k^{-j}} q_{ku}\delta_u(l) + q_{k,L_k^{-j}+1}\delta_{L_k^{-j}+1}(l)$$

where

$$q_{ku} \propto (m_{ku}^{-j} - \gamma)\frac{\int \left[\prod_{(r,s)\in\Omega_k^j \cup \Omega_{ku}^{-j}} \psi(\mathbf{y}_{rs}|\boldsymbol{\phi})\right] G_0(\mathrm{d}\boldsymbol{\phi})}{\int \left[\prod_{(r,s)\in\Omega_{ku}^{-j}} \psi(\mathbf{y}_{rs}|\boldsymbol{\phi})\right] G_0(\mathrm{d}\boldsymbol{\phi})}$$

for $u \leq L_k^{-j}$, and

$$q_{k,L_k^{-j}+1} \propto (\epsilon + \gamma L_k^{-j}) \int \left[\prod_{(r,s)\in\Omega_k^j} \psi(\mathbf{y}_{rs}|\boldsymbol{\phi})\right] G_0(\mathrm{d}\boldsymbol{\phi}).$$

In the previous expression, $\Omega_k^j = \{(r,s) : \zeta_r = k, s = j\}$ is the set of observations in the k-th cluster of rows that fall in column j, and $\Omega_{ku}^{-j} = \{(r,s) : \zeta_r = k, \xi_{ks} = u, s \neq j\}$ is the set of observations in the k-th cluster of rows that fall in the u-th cluster of columns, possibly excluding those in the j-th column.

Now, we discuss the updates to the partition of rows. In principle, a similar Pólya urn sampler to the one discussed above could be used to sample ζ_i given $\zeta_1, \ldots, \zeta_I \backslash \zeta_i$. However, even if G_0 is conjugate to ψ, computing $\int \prod_{i\in\Lambda} \psi(\mathbf{y}_i|\boldsymbol{\phi})\mathbf{H}(\mathrm{d}\boldsymbol{\phi})$ involves a sum with a number of terms that grows very fast with I, making the operation unwieldy for moderately large I. To avoid this problem, we explicitly condition on the indicators $\{\xi_{kj}\}$, but integrate out the $\boldsymbol{\phi}_{kl}^{*}$. More concretely, we update ζ_i by sampling

$$\mathsf{Pr}(\zeta_i = k | \cdots) = \sum_{u=1}^{K^{-i}} w_u^{-i}\delta_u(k) + w_{K^{-i}+1}\delta_{K^{-i}+1}(k),$$

where

$$w_u^{-i} \propto (n_u^{-i} - \alpha) \prod_{l=1}^{L_u} \frac{\int \left[\prod_{(r,s) \in \Lambda_{ul}^{i} \cup \Lambda_{ul}^{-i}} \psi(\mathbf{y}_{rs} | \boldsymbol{\phi}) \right] G_0(\mathrm{d}\boldsymbol{\phi})}{\int \left[\prod_{(r,s) \in \Lambda_{ul}^{-i}} \psi(\mathbf{y}_{rs} | \boldsymbol{\phi}) \right] G_0(\mathrm{d}\boldsymbol{\phi})}$$

for $u \leq K^{-i}$, and

$$w_{K^{-i}+1} \propto (\beta + \alpha K^{-i}) \prod_{l=1}^{L_{K^{-i}+1}} \int \left[\prod_{(r,s) \in \Lambda_{K^{-i}+1,l}^{i}} \psi(\mathbf{y}_{rs} | \boldsymbol{\phi}) \right] G_0(\mathrm{d}\boldsymbol{\phi}).$$

$\Lambda_{ul}^{i} = \{(r,s) : \xi_{us} = l, r = i\}$, $\Lambda_{ul}^{-i} = \{(r,s) : \zeta_r = u, \xi_{us} = l, r \neq i\}$ have interpretations analogous to Ω_k^j and Ω_{kl}^{-j}, and $\xi_{K^{-i}+1,1}, \ldots, \xi_{K^{-i}+1,J}$ is randomly sampled according to (2), i.e., $\xi_{K^{-i}+1,1} = 1$ and

$$\xi_{K^{-i}+1,j} | \xi_{K^{-i}+1,j-1}, \ldots, \xi_{K^{-i}+1,1} \sim \sum_{u=1}^{L_{K^{-i}+1}^{j-1}} \frac{m_{K^{-i}+1,u}^{j-1} - \gamma}{\epsilon + j - 1} \delta_u(l) + \frac{\epsilon + \gamma L_{K^{-i}+1}^{j-1}}{\epsilon + j - 1} \delta_{L_{K^{-i}+1}^{j-1}+1}(l).$$

If $\zeta_i = K^{-i} + 1$ (which implies that it has been assigned to a new component of its own) then we retain the $(\xi_{K^{-i}+1,1}, \ldots, \xi_{K^{-i}+1,J})$ used to compute $w_{K^{-i}+1}$ as the column indicators corresponding to this new component.

Since the quadruplet $(\alpha, \beta, \gamma, \epsilon)$ controls the prior distribution on the partitions, it is important to learn about these parameters from the data. For (α, β), this can be done by treating (3) as the full conditional likelihood, which can be combined with a prior distribution to generate the full conditional posterior distribution. In particular, in our applications we consider a prior $p(\alpha, \beta) = p(\alpha)p(\beta|\alpha)$ such that $\alpha \sim \mathsf{Beta}(a_\alpha, b_\alpha)$ and $\log(\beta + \alpha) | \alpha \sim \mathsf{N}(a_\beta, b_\beta^2)$. Since the resulting posterior does not follow any standard distribution, we need a Metropolis-Hastings step to update these parameters. In the applications discussed in Section 5, we employ independent Gaussian random walks for α

and β. A similar approach can be used to sample the pair (γ, ϵ), with the difference that the conditional likelihood is now formed as the product of distributions of the form (3) over K clusters.

Finally, the fixed effects $\boldsymbol{\nu}$ can updated by sampling from its full conditional distribution given by

$$\boldsymbol{\nu} | \cdots \sim \left[\prod_{i=1}^{I} \prod_{j=1}^{J} \psi(\mathbf{y}_{ij} | \boldsymbol{\phi}^*_{\zeta_i \xi_{\zeta_i j}}, \boldsymbol{\nu}) \right] p(\boldsymbol{\nu}).$$

This full posterior is model-specific, but for appropriate choices of $p(\boldsymbol{\nu})$ it will typically have a standard form.

If G_0 is not a conjugate prior to the kernel ψ, we can still use a very similar algorithm that does not integrate out the component specific parameters, such as the no-gaps algorithms (MacEachern & Müller, 1998; Neal, 2000).

4.1 Missing data and predictions

It is common in many applications for entries in the data matrix to be missing. For example, in gene expression experiments, local defects in the array might lead to missing spots or in the case of cancer mortality data, not all states might report statistics for the same set of cancer types. Additionally, very low mortality counts that are below a certain threshold are often truncated, since such numbers are inherently unreliable. In such situations, we will usually be interested in analyzing the full data set (including the available information contained in rows and columns with some missing or truncated entries), and in providing predictions for the missing values in the sample by borrowing strength from other members of the cluster.

Since we employ Markov chain Monte Carlo algorithms to fit the nested partition model, handling missing data and providing predictions for the missing entries is straightforward using data-augmentation approaches that are already standard in the Bayesian literature. Indeed, by construction, the observations are conditionally independent from each other given the row and column indicators, the cluster-specific random effects $\{\boldsymbol{\phi}^*_{kl}\}$, and the fixed effects $\boldsymbol{\nu}$. Therefore, for the pairs (i, j) where y_{ij} is missing in our original dataset, we can augment our sampler by generating $y_{ij} \sim \psi(\cdot|\mathbf{x}_{ij}, \boldsymbol{\phi}^*_{\zeta_i,\xi_{\zeta_i,j}}, \boldsymbol{\nu})$ conditionally on the current value of these parameters. For the case of truncated observations, we can sample from the corresponding truncated distribution. In turn, sampling for the parameters given the observations and imputed values can proceed along the lines discussed earlier without the need for any modification. The resulting sampler converges to the posterior distribution of interest, and point and interval predictions for the missing/truncated values can be obtained through summaries of the samples of their posterior distributions.

4.2 Summarizing the posterior distributions

Once a sample from the posterior distribution has been obtained by means of the MCMC sampler described above, computing posterior summaries for the random effects $\{\boldsymbol{\theta}_{ij}\}$, the fixed effects $\boldsymbol{\nu}$, the shape parameters α, β, γ and ϵ, and any missing/truncated value is straightforward. In particular, point estimators that are optimal with respect to squared error loss functions can be obtained by computing a simple average of the sampled values, while uncertainty in the point estimator can be evaluated by computing posterior credible intervals from the quantiles of the sample.

The situation is not as simple when summarizing the clustering structure generated by the NPM, which is described by the indicators $\{\zeta_i\}$ and $\{\xi_{kj}\}$. In general, we will be interested in providing point estimators for the partition of both rows and columns, and in providing some mechanism to asses the uncertainty associated with these partitions. One option is to report the partition with the largest posterior probability as the point estimator. However, since the size of the space of partitions grows exponentially with the number of observations being clustered, MAP estimators are notably difficult to obtain in this setting. As an alternative, we focus on extending the method described Lau & Green (2006) to deal with the nested clustering structure generated by our model.

If we were interested only in clustering rows, we could follow Lau &

Green (2006) directly and introduce a loss function of the form

$$L(\{\hat{\zeta}_i\}, \{\zeta_i\}) = \sum_{i=2}^{I} \sum_{i'=1}^{i-1} \left[a_1 \mathbf{1}_{(\zeta_i \neq \zeta_{i'})} \mathbf{1}_{(\hat{\zeta}_i = \hat{\zeta}_{i'})} + a_2 \mathbf{1}_{(\zeta_i = \zeta_{i'})} \mathbf{1}_{(\hat{\zeta}_i \neq \hat{\zeta}_{i'})} \right] \tag{9}$$

where $\hat{\zeta}_i$ is the point estimator for ζ_i and $\mathbf{1}_A$ denotes the indicator function of the set A. Note that this utility function, which was originally proposed in Binder (1978), is invariant under label switching on both the true indicators and their point estimators. The constants a_1 and a_2 control the loss caused by the two types of errors considered by this loss function: a_1 is the loss incurred by clustering together two observations that belong to separate clusters, while a_2 is the loss produced by putting in separate groups two observations that in reality are clustered together. Therefore, when a_1 is much larger than a_2, our point estimators tend to consist of a relatively large number of groups, while if a_1 is much smaller than a_2 we tend to favor partitions with a very small number of clusters. Since the point estimate might depend on the choice of these constants, a sensitivity analysis can provide insights into the uncertainty associated with the point estimator provided.

Lau & Green (2006) show that minimizing the expected loss under (9) is equivalent to maximizing the expected utility function

$$U(\{\hat{\zeta}\}) = \sum_{i=2}^{I} \sum_{i'=1}^{i-1} \mathbf{1}_{(\hat{\zeta}_i = \hat{\zeta}_{i'})} \left(\rho_{ii'} - C\right) \tag{10}$$

where $\rho_{ii'} = \mathsf{Pr}(\zeta_i = \zeta_j | \mathbf{Y})$, the posterior probability that rows i and i' are clustered together, can be easily obtained from the Monte Carlo samples and $C = \frac{a_2}{a_1 + a_2}$ is the relative cost of misclassification of two items as not being in the same cluster. Maximization of this expected utility function can be accomplished through an iterative algorithm were labels are updated one at a time in order to conditionally optimize the value of the utility function.

In the sequel we employ this estimation procedure in two steps. In the first step, we apply the original Lau & Green (2006) approach as described at the beginning of this Section to obtain a point estimator $\hat{\zeta}$ for the partition of the rows. Then, we rerun our MCMC algorithm keeping the row clusters fixed at $\zeta_j = \hat{\zeta}_j$, and apply again the Lau & Green (2006) algorithm within each cluster of rows but using the point estimates for the pairwise clustering probabilities obtained from this new MCMC run.

5 Illustrations

5.1 Contingency tables

As a first illustration of the NPM we analyzed data on cancer mortality in the United States that occurred during the year 2000. Mortality data in the US are based on death certificates that are filed by certifying physicians and is collected and maintained by the National Center for Health Statistics (`http://www.cdc.gov/nchs`) as part of the National Vital Statistics System. Accurate and timely counts of cancer mortality are very useful in the cancer surveillance community for puposes of efficient resource allocation and planning. Estimation of current and future cancer mortality broken down by geograhic area (state) and tumor have been discussed in recent articles (Tiwari et al., 2004; Ghosh & Tiwari, 2007; Ghosh et al., 2007, 2008).

We used the SEER*Stat software (`http://seer.cancer.gov/seerstat`) provided by The Surveillance, Epidemiology and End Results (SEER) program (`http://seer.cancer.gov`) of the National Cancer Institute to access the data on cancer mortality. The particular SEER*Stat database used in this paper is titled "Mortality-Cancer, Total U.S. (1950-2000), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2003." The data was obtained in the form of a 2-way table classified according to the following criteria: (a) the state where the death occurred (this has 51 possible values, including District of Columbia listed alphabetically in the form of a 2-letter abbreviation) indicating the row and (b) the type of tumor (there were 25 different categories) indicating the column. Along with the mortality counts in each of the 51×25 cells, we also obtained the population counts for 2000 in each of the states using the SEER*Stat software. The corresponding SEER*Stat database used is titled

"Populations -Total US (1990-2002), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released November 2004." Our primary goal is to cluster the states into groups and, within each group of states, cluster the tumors into groups. In addition, we would like to be able to impute any missing/truncated values and provide error bands on the imputed values.

The mortality data were modeled as follows. Let y_{ij} be the number of deaths due to cancer j in state i and let n_i be the population of state i ($1 \leq i \leq 51$, $1 \leq j \leq 25$.) We assume that n_i's are fixed and completely known. Our assumed model is given by $y_{ij}|\lambda_{ij} \sim \mathsf{Poi}(n_i \times 10^{-5} \times \lambda_{ij})$ and $\lambda_{ij} \sim \mathsf{NPM}(\alpha, \beta, \gamma, \epsilon, G_0)$, where $G_0 \equiv \mathsf{G}(a, b)$. We choose $a = 0.54$, $b = 0.016$, $\alpha \sim \mathsf{Beta}(1, 1)$, $\log(\beta + \alpha)|\alpha \sim \mathsf{N}(0, 1)$, $\gamma \sim \mathsf{Beta}(1, 1)$ and $\log(\epsilon + \gamma)|\gamma \sim \mathsf{N}(0, 1)$. The choice of a and b were made based on empirical Bayes estimates. Results are based on 30,000 iterations after a burn-in of 30,000. Figure 3 gives heatmaps for the resulting clustering of rows (states) for three different relative cost structures. In particular, $C = .5$ gave rise to 10 clusters of states, which is presented in Table 1.

The results provide interesting insights into cancer mortality patterns in the US, as these seem to cluster roughly along geographical lines. Cluster 2 is the biggest and consists of 11 states mostly from the rural south and the midwest. Similarly, cluster 8 consists entirely of states from the midwest and northern plains. Cluster 10 consist entirely of members from the northeast as does cluster 4. Two of the most populous states are in cluster 9 while cluster 6 consists of states with some of the tallest mountains in the US.

In order to better understand the way states are clustered by the model, we show in Figure 4 the clustering of cancers for $C = .5$ for state clusters 2, 5 and 8. We can see some common patterns for all three clusters of states, for example, the mortality rate of oral cancer (Ora) and melanoma (Mel) always cluster together, as do the rates of thyroid cancer (Thy) and Hodgkin lymphoma (Hod). However, there are also important differences, for example, note that in cluster 8, the mortality rate for corpus and uterus cancer (Cor) is similar enough to the mortality rates for oral cancer and melanoma for all three to be clustered in a single group, whereas corpus and uterus cancer does

Table 1: Clustering of the states for the cancer mortality data for $C = .5$.

Cluster	Members
1	AZ, GA, WA, ID, WY
2	AL, AR, IN, IA, KY, LA, MS, MO, OH, OK, TN
3	FL, RI, WV
4	DC, ME, MA, PA
5	DE, KS, MD, MI, NV, NC, OR, SC, VA
6	AK, CO, UT
7	CT, IL, NJ, NY
8	MN, MT, NE, ND, SD, WI
9	CA, HI, NM, TX
10	NH, VT

not seem to cluster with any other type of cancer in the other two groups of states. Also, for cluster 5 we see that esophagus (Eso) and brain cancer (Bra) rates are clustered together, while for the other two clusters of states esophagus clusters with bladder cancer instead. In general, we can see that most of the common cancers such as lung, breast and prostate cluster individually, but the rarer cancers tend to cluster together with a pattern that is specific to each cluster of states. Therefore, the model allows us to improve estimation of mortality rates for rare cancers by selectively borrowing information across cancer types and states.

Now we turn our attention to prediction and missing value imputation. Cell entries with values below 25 in the 2-way table of cancer mortality counts are often deemed unreliable in the cancer surveillance community. Our analysis treated these values as censored observations and imputed them as part of the MCMC algorithm. The results presented in Table 2. In addition to providing a way of validating the sparse counts, our method also provides a measure of their reliability. Note that in all the cases that we have presented, the reported value falls inside the 95% prediction interval.

Finally, to ascertain the robustness of our model we also re-ran our analysis asssuming some of the cell entries were actually missing. The missing

Table 2: Imputed mortality counts < 25 in the 2000 cancer mortality figures.

State and tumor	Reported value	Imputed value	
		Median	95% interval
NM, Testis	2	2	(0, 6)
MN, Thyroid	22	21	(14, 24)
WY, Larynx	7	7	(2, 13)
NV, Hodgkin Lymphoma	6	8	(3, 15)

Table 3: Imputed missing values in the 2000 cancer mortality figures.

State and tumor	Reported value	Imputed value	
		Median	95% interval
AK, Breast	61	65	(50, 82)
TX, Brain	844	903	(836, 970)
NY, Melanoma	450	485	(440, 532)

entries were chosen at random and they were imputed using the methods described earlier. The results of their imputation is given in Table 3. As before, the 95% interval contains the reported figure.

5.2 Gene expression data

In our second illustration we analyze gene expression data associated with lymph node positivity status (LNPos) in human breast cancer; this dataset has been previously analyzed in Pittman et al. (2004) and Hans et al. (2007). The data comprises of information on 4512 genes on 148 subjects, of which 100 are low-risk (node-negative) and 48 are high-risk (node-positive). Due to computational constrains, our analysis focuses on a subset of 500 genes.

Our goal is to identify groups of subjects with similar overall expression profiles while simultaneously identifying co-regulated genes within each group of subjects. We analyzed the data using a nested partition mixture model with Gaussian kernels. More concretely, we let y_{ij} be the expression level of gene j for subject i, and set

$$y_{ij} \sim \mathsf{N}(\mu_{ij}, \sigma_{ij}^2), \qquad \mathbf{\Theta} = [\mu_{ij}, \sigma_{ij}^2] \sim \mathsf{NPM}(\alpha, \beta, \gamma, \epsilon, G_0),$$

where $G_0 = \mathsf{NIG}(\mu_0, \kappa_0, \varphi_0, \varsigma_0)$ is a normal inverse gamma distribution. We let $\mu_0 = 0$, $\varphi_0 = 1$, and assigned priors $\kappa_0 \sim \mathsf{G}(1,1)$, $\varsigma_0 \sim \mathsf{G}(1,1)$, $\alpha \sim \mathsf{Beta}(1,1)$, $\log(\beta + \alpha)|\alpha \sim \mathsf{N}(0,1)$, $\gamma \sim \mathsf{Beta}(1,1)$ and $\log(\epsilon + \gamma)|\gamma \sim \mathsf{N}(0,1)$. The algorithm described in Section 4 was used to fit the model. Inference is based on 10,000 iterations obtained after a burn-in period of 5,000. Visual inspection of the trace plots did not reveal any obvious mixing or convergence problems.

The resulting clustering for the subjects is shown in Figure 5 for three cost structures. The clustering of rows is quite robust, with $C = 0.2$ and $C = 0.5$ producing the same 17 groups, and $C = 0.8$ differing only on the assignment of one single subject. Therefore, all further analysis was performed conditional on the clustering induced by $C = 0.5$. Cluster sizes vary dramatically, with the largest clusters having 20, 19 and 16 subjects respectively, and the three smallest clusters being singletons. The largest cluster is particularly interesting as 13 out of its 20 subjects are lymph positive. This cluster is characterized by two small groups of corregulated genes (see table 4), with the rest of the genes not showing any discernible pattern. The first cluster comprises 12 overexpressed genes, including STARD8 (D80011) which is a tumor growth inhibitor, and XRCC6 (AF052148), which is involved in the repair of double-strand break and transposition damage. Therefore, it is reasonable to assume that all elements in this cluster are involved in a common pathway that seems to be protective against the development of tumors. The second group comprises 19 underexpressed genes including EAN57 (Z82180), which has been shown to be associated with breast cancer recurrence (Huang et al., 2003), ZMIZ2, which interacts with androgen receptor (AR) and enhances AR-mediated transcription, and H2A Histonine (Z80776), which is over expressed in prostate cancer (Ernst et al., 2002). Hence, this cluster seems to correspond to genes that play a role in the occurrence of cancer. Overall, the expression pattern for this cluster of subjects seems to suggest patients who, in spite of their lymph-node status, are less predisposed to suffer from cancer.

6 Discusssion

A straightforward extension of the NPM involves allowing a different γ and ϵ parameter for each cluster of rows. This provides additional flexibility without a significant increase in computational expense. Another interesting extension of the NPM involves the use of other species sampling models, such as those described in Lijoi et al. (2005) and Lee et al. (2009), instead of the Poisson Dirichlet process we employed in this paper.

As a simple alternative to the NPM model, nested clustering problem can in principle be solved by sequentially applying traditional clustering procedures, such as hierarchical or K-means clustering. However, our approach has a number of theoretical and practical advantages. First of all, our approach automatically accounts for uncertainty in the number of clusters and allows us to estimate these parameters without any additional computational cost. Second, sequential clustering approaches ignore the uncertainty in the clustering at the lowest levels of the sequence when clustering at the higher levels. Finally, such approaches cannot deal with missing data or generate predictions for new observations, and cannot be easily incorporated into hierarchical models.

7 Acknowledgements

We would like to thank Adrian Dobra for providing access to LNPos dataset.

A Correlation structure

Note that

$$\mathsf{E}(\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{i'j'}) = \mathsf{E}\left\{\sum_{k=1}^{\infty}\sum_{k'=1}^{\infty}\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty}\mathsf{Pr}(\zeta_i = k, \zeta_{i'} = k', \xi_{kj} = l, \xi_{k'j'} = l')\boldsymbol{\phi}^*_{kl}\boldsymbol{\phi}^*_{k'l'}\right\}$$

Table 4: List of clustered genes and their descriptions.

GeneBank Accession Number	Description
D80011	Human mRNA for KIAA0189 gene
AL080191	Homo sapiens mRNA
AF052177	Homo sapiens clone 24510 mRNA sequence
AF070536	Homo sapiens clone 24566 mRNA sequence
S83374	Glutamate transporter II variant B/HBGT IIB
M90357	Human basic transcription factor 3a (BTF3a) gene
W27762	Homo sapiens cDNA 37c6
AF052148	Homo sapiens clone 24507 mRNA sequence
AJ001481	Homo sapiens mRNA for DUX1 protein
AA492299	Homo sapiens cDNA ng80e03.s1
U57843	Human phosphatidylinositol 3-kinase delta catalytic subunit mRNA
M16653	Human pancreatic elastase IIB mRNA, complete cds
AC005329	Homo sapiens chromosome 19, cosmid R34382
Z80776	Homo sapiens H2A/g gene
Z82180	Human DNA sequence from clone E81G9 on chromosome 22
AL022165	dJ71L16.4 (putative Chondroitin 6-Sulfotransferase like protein)
AF054910	Homo sapiens testicular tektin B1-like protein mRNA
AB015330	Homo sapiens HRIHFB2007 mRNA, partial cds
AJ011654	Homo sapiens mRNA for triple LIM domain protein
AA524802	Homo sapiens cDNA nh33h11.s1
AC004794	Homo sapiens chromosome 19, cosmid F20569
AF060865	Homo sapiens chromosome 16 zinc finger protein ZNF210
AC004410	Homo sapiens chromosome 19, fosmid 39554
AC004523	Homo sapiens chromosome 19, cosmid F22329
Z30643	Homo sapiens mRNA for chloride channel (putative)
D63789	Homo sapiens DNA for SCM-1beta precursor, complete cds
AI200373	Homo sapiens cDNA, 3'end, qf98c03.x1
X95289	Homo sapiens mRNA for HCGIX protein
X73079	Homo sapiens encoding Polymeric immunoglobulin receptor
D89094	Homo sapiens mRNA for 3,5 -cyclic GMP phosphodiesterase

When $i \neq i'$ and $j \neq j'$,

$$\mathsf{Pr}(\zeta_i = k, \zeta_{i'} = k', \xi_{ij} = l, \xi_{i'j'} = l') = \mathsf{Pr}(\zeta_i = k)\mathsf{Pr}(\zeta_{i'} = k')\mathsf{Pr}(\xi_{kj} = l)\mathsf{Pr}(\xi_{k'j'} = l')$$

and therefore

$$\begin{aligned}
\mathsf{E}(\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{i'j'}) &= \sum_{k=1}^{\infty}\sum_{k'=1}^{\infty}\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty} \mathsf{E}(w_k w_{k'} \pi_{kl}\pi_{k'l'})\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{kl}\boldsymbol{\phi}^*_{k'l'}) \\
&= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \mathsf{E}(w_k^2)\mathsf{E}(\pi_{kl}^2)\mathsf{E}_{G_0}(\boldsymbol{\phi}^{*2}_{kl}) - \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \mathsf{E}(w_k^2)\mathsf{E}(\pi_{kl}^2)\left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{kl})\right\}^2 + \\
&\qquad \sum_{k=1}^{\infty}\sum_{k'=1}^{\infty}\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty} \mathsf{E}(w_k w_{k'}) E(\pi_{kl}\pi_{k'l'})\left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{kl})\right\}^2 \\
&= \mathsf{Var}_{G_0}(\boldsymbol{\phi}^*_{11}) \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \mathsf{E}(w_k^2)\mathsf{E}(\pi_{kl}^2) + \left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{11})\right\}^2 \\
&= \frac{(1-\alpha)}{(1+\beta)}\frac{(1-\gamma)}{(1+\epsilon)}\mathsf{Var}_{G_0}(\boldsymbol{\phi}^*_{11}) + \left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{11})\right\}^2
\end{aligned}$$

because

$$\sum_{k=1}^{\infty} \mathsf{E}(w_k^2) = \mathsf{Pr}(\zeta_1 = \zeta_2) = \frac{1-\alpha}{1+\beta} \text{ and } \sum_{k=1}^{\infty} \mathsf{E}(\pi_{kl}^2) = \mathsf{Pr}(\xi_{k1} = \xi_{k2}) = \frac{1-\gamma}{1+\epsilon}.$$

Similarly, when $i = i'$ and $j \neq j'$,

$$\mathsf{Pr}(\zeta_i = k, \zeta_{i'} = k', \xi_{kj} = l, \xi_{k'j'} = l') = \begin{cases} \mathsf{Pr}(\zeta_i = k)\mathsf{Pr}(\xi_{kj} = l)\mathsf{Pr}(\xi_{kj'} = l') & k = k' \\ 0 & k \neq k' \end{cases}$$

and therefore

$$\begin{aligned}\mathsf{E}(\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{i'j'}) &= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty}\mathsf{E}(w_k\pi_{kl}\pi_{k'l'})\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{kl}\boldsymbol{\phi}^*_{kl'})\\ &= \sum_{l=1}^{\infty}\sum_{l'=1}^{\infty}\mathsf{E}(\pi_{1l}\pi_{1l'})\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{1l}\boldsymbol{\phi}^*_{1l'})\\ &= \mathsf{Var}_{G_0}(\boldsymbol{\phi}^*_{11})\sum_{l=1}^{\infty}\mathsf{E}(\pi_{1l}\pi_{1l'}) + \left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{11})\right\}^2 = \frac{1-\gamma}{1+\epsilon}\mathsf{Var}_{G_0}(\boldsymbol{\phi}^*_{11}) + \left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{11})\right\}^2 .\end{aligned}$$

Finally, when $i \neq i'$ and $j = j'$,

$$\begin{aligned}\mathsf{E}(\boldsymbol{\theta}_{ij}\boldsymbol{\theta}_{i'j'}) &= \sum_{k=1}^{\infty}\mathsf{E}(w_k^2)\mathsf{E}_{G_0}(\boldsymbol{\phi}^{*2}_{k1}) - \sum_{k=1}^{\infty}\mathsf{E}(w_k^2)\left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{kl})\right\}^2 +\\ &\qquad \sum_{k=1}^{\infty}\sum_{k'=1}^{\infty}\sum_{l=1}^{\infty}\sum_{l'=1}^{\infty}\mathsf{E}(w_k w_{k'})E(\pi_{kl}\pi_{k'l'})\left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{kl})\right\}^2\\ &= \mathsf{Var}_{G_0}(\boldsymbol{\phi}^*_{11})\sum_{k=1}^{\infty}\mathsf{E}(w_k^2) + \left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{11})\right\}^2 = \frac{1-\alpha}{1+\beta}\mathsf{Var}_{G_0}(\boldsymbol{\phi}^*_{11}) + \left\{\mathsf{E}_{G_0}(\boldsymbol{\phi}^*_{11})\right\}^2 .\end{aligned}$$

References

ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis* **11**, 581–598.

BARRY, D. & HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association* **88**, 309–319.

BIGELOW, J. L. & DUNSON, D. B. (2007). Posterior simulation across nonparametric models for functional clustering. *Journal of the Royal Statistical Society, Series B.* .

BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65**, 31–38.

BLACKWELL, D. & MACQUEEN, J. B. (1973). Ferguson distribution via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355.

CHIB, S. & HAMILTON, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67–89.

DEIORIO, M., MÜLLER, P., ROSNER, G. L. & MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

DIACONIS, P. & JANSON, S. (2007). Graph limits and exchangeable random graphs. ArXiv:0712.2749v1.

DUAN, J. A., GUINDANI, M. & GELFAND, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **,** *in press*.

DUNSON, D. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association* **100**, 618–627.

DUNSON, D. B., HERRING, A. H. & MULHERI-ENGEL, S. A. (2007a). Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the Royal Statistical Society , In press*.

DUNSON, D. B. & PARK, J.-H. (2007). Kernel stick-breaking processes. Technical report, Institute of Statistics and Decision Sciences - Duke University.

DUNSON, D. B., PILLAI, N. & PARK, J.-H. (2007b). Bayesian density regression. *Journal of the Royal Statistical Society, Series B.* **69**, 163–183.

DUNSON, D. B., XUE, Y. & CARIN, L. (2008). The matrix stick-breaking process: Flexible Bayes meta analysis. *Journal of American Statistical Association* **113**.

ERNST, T., HERGENHAHN, M., KENZELMANN, M., COHEN, C. D., BONROUHI, M., WENINGER, A., KLÄREN, R., GRÖNE, E. F., WIESEL, M., GÜDEMANN, C., KÜSTER, J., SCHOTT, W., STAEHLER, G., KRETZLER, M., HOLLSTEIN, M. & GRÖNE, H.-J. (2002). Decrease and gain of gene expression are equally discriminatory markers for prostate carcinoma. A gene expression analysis on total and microdissected prostate tissue. *American Journal of Pathology* **160**, 2169–2180.

ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

ESCOBAR, M. D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association* **90**, 577–588.

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

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2** , 615–629.

GELFAND, A. E., KOTTAS, A. & MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

GHOSH, K., GHOSH, P. & TIWARI, R. C. (2008). Comment on “The nested Dirichlet process” by Rodriguez, Dunson and Gelfand. *Journal of the American Statistical Association* **103**, 1147–1149.

GHOSH, K. & TIWARI, R. C. (2007). Prediction of U.S. cancer mortality counts using semiparametric Bayesian techniques. *Journal of the American Statistical Association* **102**, 7–15.

GHOSH, K., TIWARI, R. C., FEUER, E. J., CRONIN, K. & JEMAL, A. (2007). Predicting U.S. cancer mortality counts using state space models. In *Computational Methods in Biomedical Research*, Eds. R. Khattree & D. N. Naik, pp. 131–151. Boca Raton, FL: Chapman & Hall / CRC.

GRIFFIN, J. E. & STEEL, M. F. J. (2006a). Nonparametric inference in time series problems. In *Valencia Statistics 8*.

GRIFFIN, J. E. & STEEL, M. F. J. (2006b). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* **101**, 179–194.

HANS, C., DOBRA, A. & WEST, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* **102**, 507–516.

HARTIGAN, J. A. (1990). Partition models. *Communications in Statistics - Theory and Methods* **19**, 2745 – 2756.

HIRANO, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* **70**, 781–799.

HOOVER, D. (1979). Relations on probability spaces and arrays of random variables. Technical report, Institute for Advanced Study, Princeton, NJ.

HUANG, E., CHENG, S. H., DRESSMAN, H., PITTMAN, J., TSOU, M. H., HORNG, C. F., BILD, A., IVERSEN, E. S., LIAO, M., CHEN, C. M., WEST, M., NEVINS, J. R. & HUANG, A. T. (2003). Gene expression predictors of breast cancer outcomes. *The Lancet* **361**, 1590–1596.

ISHWARAN, H. & JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

KACPERCZYK, M., DAMIEN, P. & WALKER, S. G. (2003). A new class of Bayesian semiparametric models with applications to option pricing. Technical report, University of Michigan Business School.

KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. & UEDA, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 22nd Annual Conference on Artificial Intelligence*.

KOTTAS, A., BRANCO, M. D. & GELFAND, A. E. (2002). A nonparametric Bayesian modeling approach for cytogenetic dosimetry. *Biometrics* **58**, 593–600.

LAU, J. W. & GREEN, P. (2006). Bayesian model based clustering procedures. Technical report, Department of Mathematics, University of Bristol.

LAWS, D. J. & O'HAGAN, A. (2002). A hierarchical Bayes model for multilocation auditing. *Journal of the Royal Statistical Society, Series D* **51**, 431–450.

LEE, J., MÜLLER, P., TRIPPA, L. & QUINTANA, F. A. (2009). Defining predictive probability functions for species sampling models. Technical report, Pontificia Universidad Católica de Chile.

LIJOI, A., MENA, R. H. & PRÜNSTER, I. (2005). Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Statistical Inference for Stochastic Processes* **8**, 283–309.

LIJOI, A., MENA, R. H. & PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics* **12**, 351–357.

MACEACHERN, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commnunications in Statistics, Part B - Simulation and Computation* **23**, 727–741.

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.

MACEACHERN, S. N. (2000). Dependent Dirichlet processes. Technical report, Ohio State University, Department of Statistics.

MACEACHERN, S. N. & MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

MCCLOSKEY, J. W. (1965). *A Model for the Distribution of Individuals by Species in an Environment*. Ph.D. thesis, Michigan State University.

MEDVEDOVIC, M. & SIVAGANESAN, S. (2002). Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

MÜLLER, P., QUINTANA, F. & ROSNER, G. (2004). Hierarchical meta-analysis over related non-parametric Bayesian models. *Journal of Royal Statistical Society, Series B* **66**, 735–749.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

PITMAN, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, Eds. T. S. Ferguson, L. S. Shapeley & J. B. MacQueen, pp. 245–268. Hayward, CA:IMS.

PITTMAN, J., HUANG, E., DRESSMAN, H., HORNG, C. F., CHENG, S. H., TSOU, M. H., CHEN, C. M., BILD, A., IVERSEN, E. S., HUANG, A. T., NEVINS, J. R. & WEST, M. (2004). Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences* **101**, 8431–8436.

QUINTANA, F. & IGLESIAS, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society, Series B.* **65**, 557–574.

QUINTANA, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference* pp. 2407–29.

RODRIGUEZ, A., DUNSON, D. B. & GELFAND, A. E. (2008). The nested Dirichlet process, with discussion. *Journal of American Statistical Association* **103**, 1131–1144.

RODRIGUEZ, A. & TER HORST, E. (2008). Bayesian dynamic density estimation. *Bayesian Analysis* **3**, 339 – 366.

RODRIGUEZ, A. & TER HORST, E. (2009). Measuring expectations in options markets: An application to the S&P500 index. *Quantitative Finance* .

ROY, D. M. & TEH, Y. W. (2009). The Mondrian process. In *Proceedings of the Conference on Neural Information Processing Systems*.

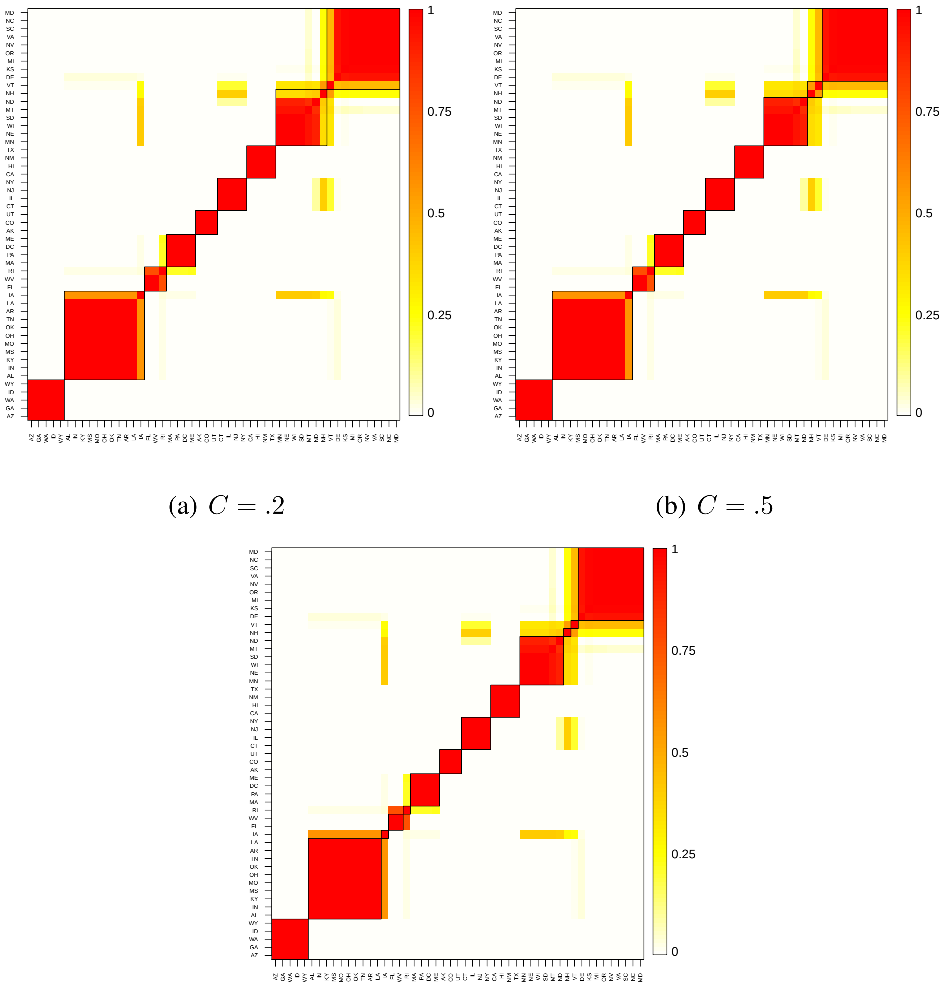
SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

SUDDERTH, E. B. & JORDAN, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems 21*, Eds. D. Koller, D. Schuurmans, Y. Bengio & L. Bottou.

TEH, Y. W., JORDAN, M. I., BEAL, M. J. & BLEI, D. M. (2006). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

TIWARI, R. C., GHOSH, K., JEMAL, A., HACHEY, M., WARD, E., THUN, M. J. & FEUER, E. J. (2004). A new method for predicting US and state-level cancer mortality counts for the current calendar year. *CA: A Cancer Journal for Clinicians* **54**, 30–40.

XU, Z., TRESP, V., YU, K. & KRIEGEL, H.-P. (2006). Infinite hidden relational models. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*.

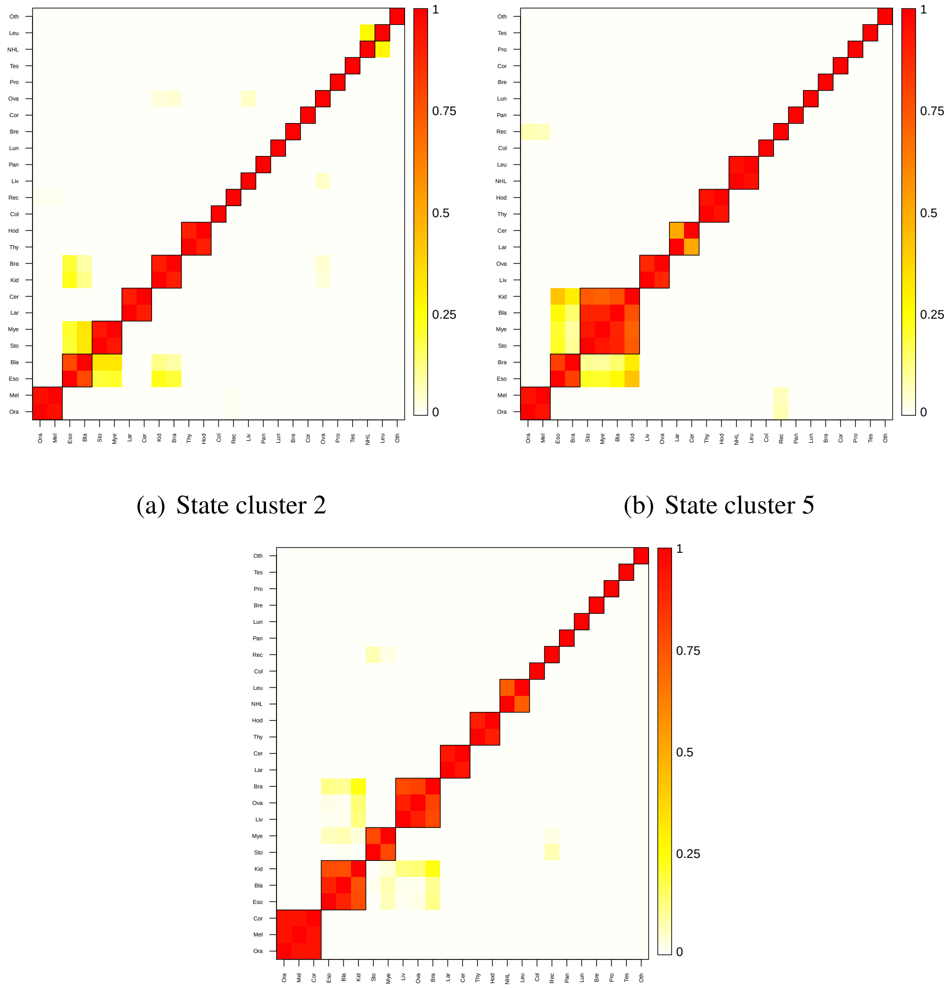


(a) $C = .2$

(b) $C = .5$

(c) $C = .8$

Figure 3: Heatmap showing the clustering of the 50 states and the District of Columbia in the cancer mortality data resulting from three different cost structures. C is the proportion of cost in misclassifying two units into the same cluster. $C = 0.2$, $C = 0.5$ and $C = 0.8$ result in 9, 10 and 13 clusters of states respectively.

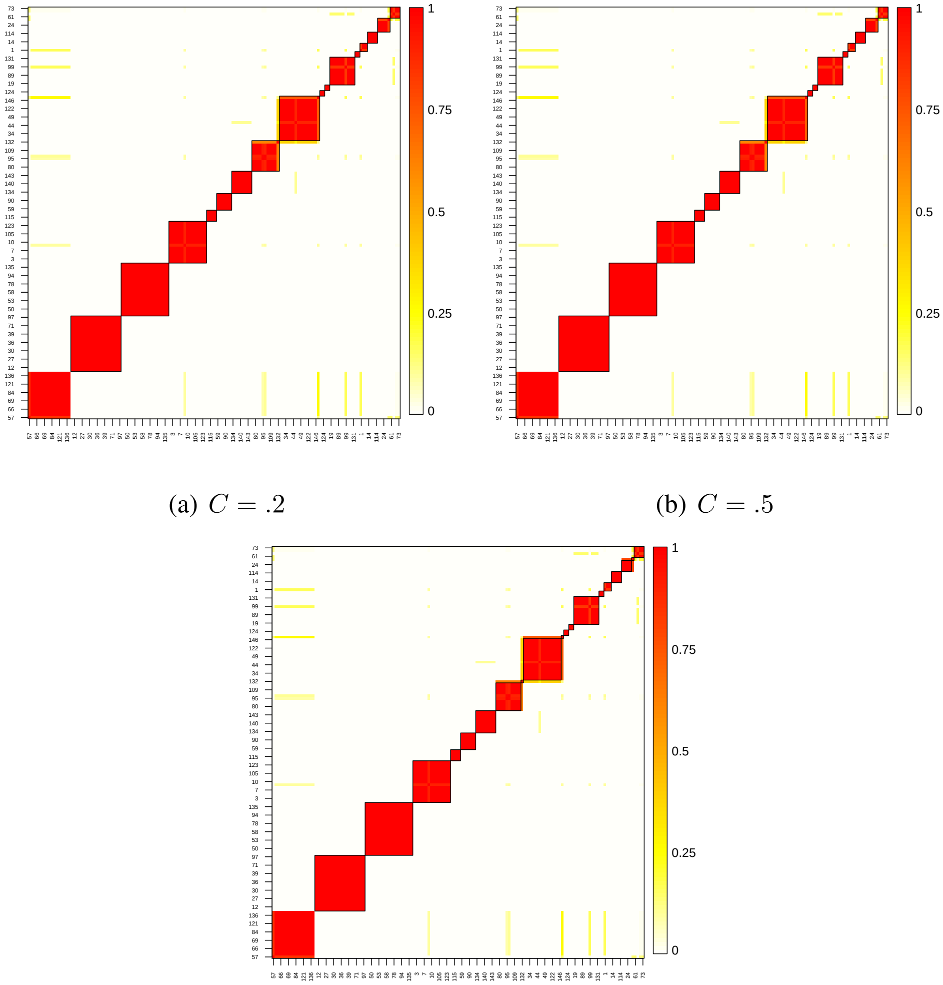


(a) State cluster 2

(b) State cluster 5

(c) State cluster 8

Figure 4: Heatmap showing the clustering of the 25 different types of cancer for 3 selected clusters of states, using $C = 0.5$.



(a) $C = .2$

(b) $C = .5$

(c) $C = .8$

Figure 5: Heatmap showing the clustering of the 148 subjects in the microarray data resulting from three different cost structures. C is the relative cost of misclassifying two units into the same cluster. $C = 0.2$ and $C = 0.5$ give rise to 17 clusters, while $C = 0.8$ give rise to 20 clusters.