Weighted Chinese restaurant processes and Bayesian mixture models

Albert Y. Lo[1]
Department of Information and Systems Management
The University of Science and Technology
Clear Water Bay, Hong Kong

Lawrence J. Brunner[2]
Department of Statistics
The University of Toronto
Toronto M5S 1A1, Canada

Anthony T. Chan[3]
Department of Information and Systems Management
The University of Science and Technology
Clear Water Bay, Hong Kong

Rev.1.1
July 1998

[1] Supported in part by Hong Kong Research Grants Council Competitive Earmarked Research Grant 674/95.

[2] Supported in part by Natural Sciences and Engineering Research Council of Canada Grant OGPIN 014

[3] Supported in part by (Hong Kong) RGC Direct Allocation Grant 92/93.BM13

Summary This paper discusses numerical methods for Bayesian mixture models. First is a sequential seating algorithm, called the weighted Chinese restaurant process, which uses a Bayesian predictive argument to assign customers (i.e., data) sequentially to tables, resulting in a random partition of the data. The distribution of the weighted Chinese restaurant process concentrates on partitions formed by clusters of data that arise from the same distribution. A Monte Carlo approximation of the posterior distribution of the Bayesian mixture models can be obtained by an iid sample from the weighted Chinese restaurant process. Next, a Gibbs sampler for sampling partitions is discussed. The transition probability function is defined based on the predictive seating probabilities of the weighted Chinese restaurant process. While the sequential weighted Chinese restaurant process seats customers so that the seating tends to result in a clustered partition, the Gibbs sampler weighted Chinese restaurant process moves from one partition to the other and, in the process it shuffles the data into clusters. Ergodic theory for a finite state Markov chain defines a Monte Carlo approximation of the posterior distribution of the Bayesian mixture model. A comparison of experiments is established to show that the Gibbs sampler sampling partitions produce a less variable approximation than a previously studied Gibbs sampler, which sequentially imputes missing values to complete a Gibbs cycle. Approximate seating algorithms, which are easier to implement, are introduced and shown to behave well. Extensions to mixture hazard rate models are discussed. Finally, hierarchical Bayesian mixture models with regression parameters are discussed. A Metropolis–Hastings algorithm nested within a Gibbs sampler weighted Chinese restaurant process defines an appropriate Markov chain Monte Carlo approximation to posterior quantities. Numerical examples are given.

1. Introduction

The Chinese restaurant process (Aldous 1985) takes its name from a seating procedure reportedly witnessed by Jim Pitman in a Chinese restaurant. Before any customers arrive, one table is set up, and other tables are stacked against a wall with their legs folded up. The first customer to arrive is seated at the open table. When the second customer arrives, with some probability he is seated at the table with customer one; otherwise a new table is set up for him. When customer three arrives, he may be seated with customer number one, or with customer number two (or with both, if customer two was seated with customer one), or a third table may be opened for him. The process continues until n customers have been seated. Note that except for the first one to arrive, all customers are seated randomly. Thus different random seating rules define different variants of the Chinese restaurant process and correspond to different ways of choosing a random partition of the integers $\{1, ..., n\}$. The Chinese restaurant process just described is sequential; that is, customers are seated one at a time as they arrive, and once they sit at a table, they remain there. One can also define non–sequential processes, in which customers are moved randomly from table to table. In either case, when the probability distributions used for seating depend on a set of sample data (one observation for each individual in the restaurant), we call it a *weighted* Chinese restaurant process (WCR).

A WCR that generates sequentially a random partition of the data based on prescribed weight functions and a mixing measure is introduced in Section 2. In a (sequential) WCR, the probability that the next customer sits on a table is proportional to the product of the number of accupants in that table and the Bayesian predictive weight" of that table. The WCR distribution assigns high probabilities to partitions formed by clusters of data. The peaking phenomenon of the WCR has a natural interpretation in terms of (Bayesian) prediction.

Section 3 discusses applications of the WCR in approximating the posterior expectations for Bayesian mixture models with Dirichlet processes priors. It also represents the posterior distribution as a weighted average over the WCR. A repeated simulation of the WCR is used as the basis of an iid Monte Carlo method (iidWCR) to approximate posterior quantities. Section 3 also discusses a Gibbs sampler WCR (gWCR) based on the Bayesian predictive argument [See also MacEachern (1994) and the following Remark 6.1.] Numerical examples comparing the iid and gWCRs in Bayesian density estimation are given. The result suggests that the traditional iid Monte Carlo method, which takes into account a full set of sequential predictive distributions, fares almost as well as the vastly popular Gibbs sampler in posterior calculations. A class of approximate WCRs that does not required the computation of predictive weights is proposed. Numerical results show that this class has an edge in approximating the "true" density.

Section 4 discusses the iid sequential imputation algorithm (iidWP) introduced by Liu (1996). This is a weighted sampling of a Blackwell and MacQueen (1973) urn process (or an extended Polya urn process). The corresponding Gibbs sampler imputation algorithm (gWP) discussed by Escobar (1988), Escobar and West (1995), and Brunner (1994, 1995) is discussed. The gWP and gWCR experiemnts are compared theoretically based on Blackwell (1951, 1953)'s comparison of experiments. An interpretation of Lemma 2 in Lo (1984) results in the existence of a transition function from the gWCR to the gWP, implying that a Monte Carlo simulation method based on sampling a gWCR distribution is a Rao−Blackwell improvement of a simulation method based on sampling a gWP distribution. Another consequence of the existence of a transition function is that a posterior distribution of the mixture distribution can be represented as an average over WCR partitions.

Section 5 discusses the application of the WCRs in Bayesian mixture hazard rate models. A representation of the posterior distribution as a weighted average over the WCR is given. Such models are popular in reliability theory and in emission tomography. A convenient reference is Chapter 3 in Snyder and Miller (1991). Approximation methods for the posterior distribution are identical to the Bayesian mixture density discussed in Sections 3 and 4.

The analytic Bayesian solutions for mixture density and hazard rate models (Theorem 5.1 and Theorem 3.1) express posterior quantities as WCR−averages of partitions. In both cases, the WCR distributions peak at clustered partitions and they seem to be a fitting model distribution for statistical cluster analysis. Nevertheless, it is not the objective of this paper to venture into a theory of statistical models for cluster analysis; the latter is explored elsewhere.

Section 6 discusses hierarchical Bayesian mixture models with Dirichlet priors. The posterior distribution of the parameters is represented as a weighted average of a WCR. The analytic solution suggests natural iid and Markov chain Monte Carlo algorithms based on Bayesian predictions for its evaluation. Noted that MacEachern's (1994) Gibbs seating algorithm (see the following Remark 6.1), relies on a "joint", rather than a "predictive" weight in its construction. The Markov chain algorithm is constructed by nesting a Metropolis−Hasting step within a Gibbs sampler cycle for random partitions. Numerical examples are given.

Section 2. Sequential seating: a weighted Chinese restaurant process.

The (unweighted) Chinese restaurant process is a procedure for randomly partitioning the integers {1,...,n} into subgroups that are called tables (or cells). It takes its name from a seating process allegedly witnessed by Jim Pitman in a Bay Area Chinese restaurant he frequented [Aldous (1985); see also Kuo (1986)]. A Chinese restaurant process with parameter $\eta>0$ selects a random partition by sequentially assigning the integers to tables/cells as follows: Customers 1,...,n enter the restaurant in the order written and they are seated one after the other. Initially, all tables in the restaurant are folded up. When customer 1 comes in, a table is set to seat him/her. After customers labelled 1,...,k−1 ($k\geq 2$) are seated, customer k will be seated at an empty table with probability $\eta/[\eta+k-1]$; otherwise, he/she sits at an occupied table with probability proportional to the number of occupants at that table. The seating process will continue until all customers are seated. (In this paper, we only consider restaurants with n or more tables.) The Chinese restaurant process **p** with parameter η has the density

(2.1) $$q(\mathbf{p}|\eta)=\eta^{n(\mathbf{p})}\times\Pi_{1\leq i\leq n(\mathbf{p})}(e_i-1)!/[\eta(\eta+1)\times\ldots\times(\eta+n-1)],$$

where $\mathbf{p}=\{C_1,\ldots,C_{n(\mathbf{p})}\}$ is a partition of {1,...,n} into $n(\mathbf{p})$ tables/cells (i.e., disjoint subsets of {1,...,n}) and $e_{1,\ldots,}e_{n(\mathbf{p})}$ are table sizes.

The seating probabilities are parameters that define a sequential seating process. In the weighted Chinese restaurant process (WCR) case, these seating probabilities are defined in terms of a (prior) mixing measure $\alpha(du)$, the number of customers to be seated n, and a nonnegative and finite (likelihood) weighting function $w_j(u)$ for customer j, j=1,...n. Define the "marginal" weight for a table C by

(2.2) $$\rho(C)\equiv\int\Pi_{j\in C}w_j(u)\alpha(du).$$

If C={k}, $\rho(C)$ is denoted by $\rho(k)$. The marginal weights are assumed to satisfy $\rho(C)<\infty$. Since $\rho(C)=0$ implies that $\rho(r,C)=0$ for $r\notin C$, we define a "predictive" weights (of r with respect to table C) by the ratio

(2.3) $$\rho(r|C)\equiv\rho(r,C)/\rho(C),\ \text{for } r\notin C;\ \equiv 0 \text{ if } \rho(C)=0.$$

If a "posterior distribution" of u given table C is defined by $\pi(du|C)\propto\Pi_{j\in C}w_j(u)\alpha(du)$, $\rho(r|C)$ is the predictive weight of r given C, given by $\int w_r(u)\pi(du|C)$.

By (2.3), for any table C, the marginal weight $\rho(C)$ can be written as a product of predictive weights, by adding customers one at a time, starting from an empty table. The order of seating customers at each table is irrelevant. For example, suppose C has e elements, and $i_1,i_2,\ldots,i_e$ is any ordering of them, the product rule of probability states that

(2.4) $$\rho(C)=\rho(i_1)\times\Pi_{2\leq j\leq e}\rho(i_j|i_1,\ldots,i_{j-1}).$$

The essence of this argument is that the numerator of a term in the product cancels with the denominator of the next term. A similar cancellation reduces the Kaplan–Meier estimator to the empirical distribution function in the absence of incomplete observations.

In WCR, customer r is seated at an empty table with a probability proportional to $\rho(r)$; otherwise, he/she sits at an occupied table with probability proportional to the product $e_i \times \rho(r|C_i)$. More precisely, the WCR algorithm for seating customers 1,...,n is: Set $\lambda(0)=\rho(1)$.

(2.5) Step 1. Assign 1 to the first table with probability $\rho(1)/\lambda(0)=1$.

Step r (r=2,...,n). Given $\mathbf{p}=\{C_1,C_2,...,C_{n(\mathbf{p})}\}$ with table sizes $e_1,...,e_{n(\mathbf{p})}$ from step r−1, calculate $\lambda(r-1)=\rho(r)+\Sigma_{1\le i\le n(\mathbf{p})}e_i\rho(r|C_i)$. Assign r to a new table C_0 with probability $\rho(r)/\lambda(r-1)$; otherwise, r sits at table C_i with probability $e_i\rho(r|C_i)/\lambda(r-1)$, $i=1,...,n(\mathbf{p})$.

The completion of Step n results in a WCR process $\mathbf{p}=\{C_1,C_2,...,C_{n(\mathbf{p})}\}$ with table sizes $e_1,...,e_{n(\mathbf{p})}$, respectively.

The n−step WCR algorithm with the product rule (2.4) operating at each step results in a density of the WCR, $q(\mathbf{p}|\alpha,\mathbf{w})$, given by

Lemma 2.1. $q(\mathbf{p}|\alpha,\mathbf{w})=\phi(\mathbf{p})/(\Lambda_{n-1})$,

where $\phi(\mathbf{p})=\Pi_{1\le i\le n(\mathbf{p})}(e_i-1)!\rho(C_i)$, and $\Lambda_{n-1}=\lambda(0)\times...\times\lambda(n-1)$.

The weighted Chinese restaurant process $q(\mathbf{p}|\alpha,\mathbf{w})$ reduces to the Chinese restaurant process $q(\mathbf{p}|\eta)$ if $w_i(u)\equiv 1$, and the measure $\alpha(.)$ is finite with total mass η. In this case, the predictive weight $\rho(k|C)$ equals the constant one for all nonempty tables C. The Chinese restaurant process and the WCR is connected by

(2.6) $$(n!)\times q(\mathbf{p}|1)\times\Pi_{1\le i\le n(\mathbf{p})}\rho(C_i)=\Lambda_{n-1}\times q(\mathbf{p}|\alpha,\mathbf{w}).$$

The way incoming customers in the WCR are assigned to occupied tables deserves notice as it reveals a Bayesian method of performing (random) cluster analysis for a set of data $\{x_i,\ i=1,...,n\}$ by means of predictions rather than the usual (deterministic) method based on a distance function defined between (groups of) data [Duda and Hart (1973), and Arabie, Hubert, and De Soete (1996)]. Identify the observation x_i with customer i, i=1,...,n, and say that a new customer r is "similar" to the e_i customers at table C_i if $e_i\times\rho(r|C_i)$ is large. The sequential seating WCR builds up a partition by seating customers sequentially, one after the other. It adds customers so that those who are "similar" are likely to sit together. Hence, eventually a partition is likely to be formed by tables occupied by customers who are "similar".

If the mixing measure $\alpha(du)$ has a finite $\alpha(R)$, the WCR takes a neat form: Given $\mathbf{p}=\{C_1,C_2,...,C_{n(\mathbf{p})}\}$ from step r−1 (r=2,...,n), seat r on table C_i with probability $e_i\times\rho(r|C_i)$, $i=0,...,n(\mathbf{p})$, where C_0 denotes an empty table; $\rho(r|C_0)=\int w_r(u)\alpha(du)/\alpha(R)$.

Section 3. A Bayesian mixture model and seating algorithms.

A mixture model is a family of densities defined by

(3.1) $f(x|G) = \int k(x|u)G(du), \ G \in \Theta,$

where the parameter space Θ is the collection of distributions; x and u are points in Euclidean spaces and G is a distribution of u. The kernel $k(.|.)$ is given, and for each u, $k(.|u)$ is a density of x. The kernel densities $\{k(.|u), \text{all } u\}$ are the extreme points of the model (3.1). The model densities have desirable smoothness properties, which sometimes can be characterized via extreme point representations. On the other hand, the mixture model often arises as a result of missing information in the sense that a complete observation (x,u) is not available. Instead, one observes the variable x, which is a randomization of u. Let $x_1,...,x_n|G$ be i.i.d. observations from the mixture density $f(x|G)$. The problem is to estimate G based on a sample $\mathbf{x}=(x_1,...,x_n)$. The classical frequentist approach to this problem has been conveniently summarized in Lindsay (1983, 1995). Here we discuss a Bayesian approach. Assuming a Dirichlet process prior to G [Ferguson (1973)] with shape measure $\alpha(.)$, Lo (1978, 1984) represents explicitly the posterior distribution of G as a mixture of Dirichlet processes [Antoniak (1974)], and the posterior mean of a linear function of G as an average of the partitions of the set $\{1,...,n\}$. The number of partitions of the set $\{1,...,n\}$ is called Bell's number, which increases roughly as the factorial of n. As a result, the exact evaluation of the posterior mean is formidable for sample sizes larger than twelve. This section discusses simulations of random partitions that can be used in Monte Carlo approximations to the stated sum over partitions. To describe it, one needs the notation of a micro–Bayesian system.

The shape measure $\alpha(.)$ of a Dirichlet process is a finite mixing measure with total mass $\alpha(R) \equiv e_0$. u has a (micro–)prior distribution $\pi(du)=\alpha(du)/\alpha(R)$, and $y_1,...,y_n|u$ are iid $k(.|u)$. For a table C, the "marginal function" of $y_j, j \in C$ is defined by $m(y_j, j \in C) = \int \Pi_{j \in C} k(y_j|u)\alpha(du)$. The predictive density of a next observation y given $\{y_j, j \in C\}$ is

$$m(y|C) \equiv m(y, y_j, j \in C)/m(y_j, j \in C); \ m(y|C) \equiv m(y) \text{ if C is empty.}$$

Putting $w_i(u)=k(x_i|u)$, i=1,...,n defines a WCR with density $q(\mathbf{p}|\alpha,\mathbf{k})$. The WCR algorithm also simplifies to: Given $\mathbf{p}=\{C_1,C_2,...,C_{n(\mathbf{p})}\}$ from step r–1 (r=2,...,n), seat r on table C_i with probability $e_i\rho(k|C_i)$, $i=0,...,n(\mathbf{p})$, where C_0 denotes an empty table and $\rho(r|C_0)=\int k(x_r|u)\alpha(du)/e_0$ is the no–sample predictive weight of r. A posterior mean of a linear function of the mixing distribution G(u) is an average of $q(\mathbf{p}|\alpha,\mathbf{k})$n.. In particular, the posterior mean of the mixture density $f(t|G)$ has the representation [Theorem 2 in Lo (1984)]

(3.2) $\hat{f}(t)=E[f(t|G)|\mathbf{x}] \propto \Sigma_{\mathbf{p}}\{m(t)+\Sigma_{1 \le i \le n(\mathbf{p})} e_i m(t|C_i)\} \times \Lambda_{n-1} q(\mathbf{p}|\alpha,\mathbf{k}),$

where Λ_{n-1} is defined in Section 2 (with $w_i(u)=k(x_i|u)$.)

The predictive density $\hat{f}(t)$ is a two–layer mixture of micro–predictive density $m(t|C_i)$s, which are kernel functions with variable bandwidths. This contrasts significantly with the classical kernel estimator [Rosenblatt (1956), Parzen (1962) and Cencov (1962)], which is a one–layer mixture of kernels with a fixed bandwidth.

The case of a statistical deconvolution model deserves a note. This model corresponds to a location or scale mixture model, i.e., $k(x|u)=k(x-u)$ or $k(x|u)=uk(xu)$, and is in general identifiable in the mixing distribution G. As such, it would be of interest to evaluate the posterior mean of G, $\hat{G}(u)=E[G(u)|\mathbf{x}]$. However, $\hat{G}(u)$ has the same expression as (3.2) if one defines $\lambda(n)$ to be $\alpha(u)+\Sigma_{1\le i\le n(\mathbf{p})}e_i\pi(u|C_i)$ at the completing of the WCR algorithm:

$$\hat{G}(u) \propto \Sigma_{\mathbf{p}}\{\alpha(u)+\Sigma_{1\le i\le n(\mathbf{p})}e_i\pi(u|C_i)\}\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k}).$$

Notice that $\pi(.|C_i)$ peaks for a large table C_i. Since $\hat{G}(u)$ is basically an average of $\alpha(u)+\Sigma_{1\le i\le n(\mathbf{p})}e_i\pi(u|C_i)$, it is approximately a mixture of step functions. This contrasts with the maximum likelihood estimator of G(u) which is exactly discrete [Lindsay (1983, 1995)].

3.1 An iidWCR and a gWCR. Kuo (1986) proposed an iid Monte Carlo method to evaluate posterior quantities of the mixture model based on sampling **p** from a (unweighted) Chinese restaurant process $q(\mathbf{p}|\alpha(R))$. One feature of the Chinese restaurant process is that a large table (large e_i) has a higher probability of receiving newcomers and, as a result, it will grow larger still. According to Korwar and Hollander (1973), the number of occupied tables in a Chinese restaurant process is approximately $\alpha(R)\times\log(n)$. The presence of only a very few occupied tables in a random partition results in peaked integrands, the product of which is highly variable. This variability in effect drastically reduces the efficiency of Kuo's method based on sampling a Chinese restaurant process $q(\mathbf{p}|\alpha(R))$ [Ji (1991)]. On the other hand, the weighted Chinese restaurant process accounts for the peaked integrands in the course of the simulation and the problem of highly variable peaked integrands diminishes. An added twist is that the growth of a tabe C_i is controlled by a balance between e_i and $\rho(r|C_i)$: a large table may not continue to grow since $\rho(r|C_i)$ could be small (if r is "far away" from C_i) so that the product $e_i \times \rho(r|C_i)$ is relatively moderate.

At the completion of the nth step of the WCR algorithm, one defines $\lambda(n)=m(t)+\Sigma_{1\le i\le n(\mathbf{p})}e_im(t|C_i)$. In this notation, the mixture density estimate (3.2) becomes

$$\hat{f}(t)=[\alpha(R)+n]^{-1}\times\Sigma_{\mathbf{p}}\Lambda_nq(\mathbf{p}|\alpha,\mathbf{k})/\Sigma_{\mathbf{p}}\Lambda_{n-1}q(\mathbf{p}|\alpha,\mathbf{k}). \quad (3.3)$$

Run the WCR process M times independently to get M iid partitions and compute $\Lambda_{n-1}(m)$, $\Lambda_n(m)$, m=1,...,M (set $x_{n+1}=t$.) The iidWCR approximation to $\hat{f}(t)$ is

$$\hat{f}_M(t)=[\alpha(R)+n]^{-1}\times\Sigma_{1\le m\le M}\Lambda_n(m)/\Sigma_{1\le m\le M}\Lambda_{n-1}(m).$$

An iidWCR approximation to a higher posterior moment of a linear function of G(u) is essentially an extension of the algorithm to more steps. A higher posterior cross moment is a similar sum over partitions of the set $\{1,....,n,n+1,...,n+k-1\}$ where k is the total order of the cross moments. As such, it can be written as an expectation with repect to $q(\mathbf{p'}|\alpha,n+k-1,\mathbf{k'})$ where $\mathbf{p'}$ is a partition of $\{1,...,n+k-1\}$ and $\mathbf{k'}$ has $n+k-1$ components. A WCR, extended to $n+k-1$ steps, provides an appropriate approximation.

A fine point of the WCR is that it depends on the sequential order in which customers enter the restaurant. Numerical studies based on the location mixture model and location–scale mixture of normals model show that sorting the data in descending and ascending order and Siegel–Tukey ranking the data produce almost identical $\hat{f}_M(t)$ for nested sample sizes n=10, 50, 150, and 300. These numerical results suggest that the dependence of WCR on the ordering of the data is minor.

The Markov chain Monte Carlo method can also be used to approximate posterior quantities. A posterior mean of a linear function of G has an representation of the form

(3.4) $\Sigma_{\mathbf{p}}h(\mathbf{p})w(\mathbf{p})$,

where $w(\mathbf{p}) \propto \phi(\mathbf{p})=\Pi_{1\le i\le n(\mathbf{p})}[e_i-1)!\rho(C_i)]=\Lambda_{n-1}q(\mathbf{p}|\alpha,n,\mathbf{k})$

is a probability distribution of $\mathbf{p}$. For example, in the case of $\hat{f}(t)$ in (3.3), $h(\mathbf{p})$ is $\{m(t)+\Sigma_{1\le i\le n(\mathbf{p})}e_i m(t|C_i)\}/[\alpha(R)+n]$; likewise for $\hat{G}(u)$. The aim is to simulate a Markov chain sequence of partitions $\mathbf{p}_0,\mathbf{p}_1,...\mathbf{p}_M,...$, which has a stationary distribution $w(\mathbf{p})$. According to the law of large numbers of an ergodic Markov chain [see for example Chung (1967)], the average of $h(\mathbf{p}_1),...,h(\mathbf{p}_M)$ approximates $\Sigma_{\mathbf{p}}h(\mathbf{p})w(\mathbf{p})$ with an error of $O(M^{-1/2})$. The Gibbs sampler method [Geman and Geman (1984); see also Tanner and Wong (1987) and Gelfand and Smith (1990)] translated into the present situation dictates that to move from one state (a partition) to the next (partition), one reseats the n customers one by one using prediction. Suppose a present $\mathbf{p}_0$ is given. One performs a cycle step to move from $\mathbf{p}_0$ to the next state $\mathbf{p}_1$ as follows: Remove j from $\mathbf{p}_0$ to obtain a "skip–j" partition denoted by $\tilde{\mathbf{p}}$. Given $\tilde{\mathbf{p}}$, which is a partition of $\{1,...,n\}-\{j\}$, we can reseat j to get a refreshed $\mathbf{p}_0$ according to a predictive seating probability distribution [defined in (4.2) below]. Use this method to reseat j=1,...,n successively, and in the order written, completing a cycle. The random partition obtained at the completion of one cycle is the new state $\mathbf{p}_1$. The cycling process repeats (using $\mathbf{p}_1$ as $\mathbf{p}_0$) to generate $\mathbf{p}_2$, and so on, to get the sequence of random partitions $\{\mathbf{p}_k\}$.

MacEachern (1994) proposed such a skip–j algorithm for a hierarchical location mixture of normals model. His method, discussed in the following Remark 6.1, depends on a "joint weight" that blurs the predictive nature of the underlying cluster process. Here we bring in the predictive weight of the WCR to reseat j to the skip–j partition. Suppose the present partition is $\mathbf{p}_0$. Suppose the last integer to be reseated is n and the delete–n partition is $\tilde{\mathbf{p}}=\{\tilde{C}_i, i=1,...,n(\tilde{\mathbf{p}})\}$ with table sizes $\tilde{e}_i$, $i=1,...,n(\tilde{\mathbf{p}})$. From the WCR algorithm, the seating (conditional) probability for the event that n seats in $\tilde{C}_i|\tilde{\mathbf{p}}$ is proportional to $e_i\times\rho(n|\tilde{C}_i)$, $i=1,...,n(\tilde{\mathbf{p}})$. (The proportionality constant depends on $\tilde{\mathbf{p}}$ and n.) The next result states that a similar seating probability prevails if j, rather than n, is the last customer to be seated.

Lemma 3.1. Given $\tilde{\mathbf{p}}=\{\tilde{C}_i, i=1,...,n(\tilde{\mathbf{p}})]$, which partitions $\{1,...,n\}-\{j\}$ with table sizes $\{\tilde{e}_i, i=1,...,n(\tilde{\mathbf{p}})\}$, the conditional probability that j sits at table $\tilde{C}_i|\tilde{\mathbf{p}}$ is proportional to $\tilde{e}_i\times\rho(j|\tilde{C}_i)$.

Proof. Note that $\tilde{\mathbf{p}}$ is a function of $\mathbf{p}$, and $\tilde{\mathbf{p}}$ differs from $\mathbf{p}$ only at the table containing j. The (joint) density of $\mathbf{p}$ (and $\tilde{\mathbf{p}}$) is proportional to $\phi(\mathbf{p})=\Pi_{1\le r\le n(\mathbf{p})}(e_r-1)!\rho(C_r)$ [see Lemma 2.1.] Suppose j sits at table $\tilde{C}_i$, where $i\in\{0,1,...,n(\tilde{\mathbf{p}})\}$. The product rule $\rho(C_i)=\rho(j|\tilde{C}_i)\times\rho(\tilde{C}_i)$ yields

$$\phi(\mathbf{p})=\tilde{e}_i\times\rho(j|\tilde{C}_i)\times\phi(\tilde{\mathbf{p}}),$$

where $\phi(\tilde{\mathbf{p}})=\Pi_{1\le i\le n(\tilde{\mathbf{p}})}(\tilde{e}_i-1)!\rho(\tilde{C}_i)$ is a function of $\tilde{\mathbf{p}}$ only and is a proportionality constant. ‖

The WCRs cluster the data by adding customers one after the other so that those who are "similar" are likely to sit together. Hence, a partition is likely to be formed by tables occupied by customers that are "similar". The gWCR accomplishes clusters differently. An initial partition is chosen quite arbitrarily. Customers will be removed and then reseated at tables one after the other. Repeatedly reseating the customers results in tables occupied by customers who are "similar". In short, the gWCR shuffles the customers one by one so that eventually "similar" customers are likely to sit together.

Customer r is "similar" to those at table C_i if the seating probability $e_i\times\rho(r|C_i)$ is large, and "similar" means closeness somewhat in the usual distance sense [Duda and Hart (1973), Arabie, Hubert, and De Soete (1996)]. For a location mixture of normals with variance 1, $\rho(r|C_i)\propto$ a $N(\text{ave}\{x_j,j\in C_i\},1)$ density evaluated at x_r, and a location and scale mixture of normals model yields $\rho(r|C_i)\propto$ a t–density evaluated at x_r where the df$\approx e_i$, location$\approx$ $\text{ave}\{x_j,j\in C_i\}$, and precision$\approx 1/\text{variance}\{x_j,j\in C_i\}$. In both cases, "similar" means essentially that $|x_r-\text{ave}\{x_j,j\in C_i\}|$ is small. A contrast is provided by a scale mixture of normals, which has $\rho(r|C_i)\propto$ a t–density evaluated at x_r where df$\approx e_i$, location 0, and precision $1/\text{ave}\{x_j^2,j\in C_i\}$. Here "similar" means that the ratio $x_r^2/\text{ave}\{x_j^2,j\in C_i\}$ is small.

The WCRs can be specialized to apply for mixture models that generate unimodal mixture densities. A list includes the scale–mixture of exponentials [Jewell (1982)], of uniforms [Brunner and Lo (1989, 1997)], and of normals. Finite mixture models [Everitt and Hand

(1981); Titterington, Smith, and Markov (1985); Diebolt and Robert (1994)] form a finite dimensional subset of the mixture model considered in this paper. The posterior quantities that result from assuming a finite mixture model are also sums over partitions and corresponding WCR algorithms can be constructed by a properly chosen discrete $\alpha(.)$. A more interesting example is provided by mixture models with multimodal mixture densities. Conjugate priors [see for example De Groot (1970)] exist for mixtures of text–book parametric models, and we shall use them to accomplish explicit expressions for seating probabilities. The (micro–)posterior distribution with respect to a non–conjugate prior can be obtained by reweighting the posterior density with respect to a conjugate prior with a weight being the ratio of the conjugate and the non–conjugate prior densities.

In this section we illustrate the WCRs in Bayesian density estimation using the location–scale mixture of normals model: $k(x|u)$ is $\tau k(\tau(x-s)$ where $u=(\tau,s)$ is two dimensional and $k(.)$ is a standard normal density. The idea is that by allowing G to carry mass at τ close to zero, the consistent behavior of a shrinking kernel in the classical kernel density estimator can be captured. Suppose $\alpha(du)/\alpha(R)$ is gamma–normal $(a,1/b; m,1/t)$: τ is gamma $(a,1/b)$ and $\mu|\tau$ is normal $(m,1/(\tau t))$. [The mean of gamma $(a,1/b)$ is a/b.]. The micro–posteriors are gamma–normal $(a_i,1/b_i; m_i,1/t_i)$ where $t_i=t+e_i$, $m_i=(mt+\bar{x}_i e_i)/t_i$, $a_i=a+e_i/2$, and $b_i=b+2^{-1}[\Sigma_{j\in C(i)}(x_j-\bar{x}_i)^2+(m-\bar{x}_i)^2/(t^{-1}+e_i^{-1})]$. The micro–sample predictive density $m(x|C_i)$ is a t–density with degrees of freedom $2a_i$, location m_i, and precision $(a_i/b_i)t_i/(t_i+1)$. See for example, DeGroot (1970). The interplay between the prior parameters a, b, m and t is subtle. Inspection of the micro–predictive t–density suggests that if t is very large, $m_i\approx m$ and $\hat{f}(t)$ will be approximately unimodal with a mode close to m. On the other hand, a small t results in a predictive t–density centered at $\bar{x}_i$ and reveals the data structure better. Such a choice of t results in a $\hat{f}$ closer to the "true" density. To keep matters simple for discussion, we assume a and $|m|$ are moderate, say, bounded by 2. If b is large, the precision $(a_i/b_i)t_i/(t_i+1)$ for the predictive t–densities will be small, resulting in a flat $\hat{f}(t)$; we shall avoid a large b.

Two "true" densities are chosen to be sampled. A U(0,1) density is selected to test the resolution of the WCR approximations as it has sharp discontinuities at 0 and 1. Another parent density is a three–peak mixture of normals,

$$0.6N(-5,0.25)+0.25N(0,1)+0.15N(5,2.25), \tag{3.5}$$

which serves to test the peak–detecting ability of the WCR approximations. In the numerical results that follow, unless otherwise stated, $\alpha(R)=2.5$ and the Monte Carlo sample size is M=1000. Figure 3.1 displays the Bayesian density estimate based on iidWCR (dotted) and gWCR (solid) for sampling from a U(0,1) density based on n=400 and 2000 observations. The micro−prior parameters are set to be a=1.5, m=0, and t=0.0005. The three rows display results based on b=1.5 (first row), 0.005, and 0.0005 (last row), respectively. Computer printouts indicate that with the micro−prior parameters defining the first row routinely yield partiions of very few tables (less than 5) with one dominatingly large table and a few very small tables. This results in the unimodal iidWCR approximations appearing in the first row. Figure 3.2 displays iidWCR and gWCR approximations for the three−peak density (3.5) based on n=300 observations. The figures show that on the average, the gWCRs stay closer to the "true" density.

3.2. An approximate WCR. The WCR is defined through a predictive weight $\rho(r|C)$ which is an averaging of the kernel by a micro−posterior distribution at a table. In the case that such an averaging operation is hard to evaluate explicitly, one could replace the micro−posterior distribution $\pi(du|C)$ by its first−order approximation: a point mass probability at the posterior mean (or its relative, the mle) in the averaging operation. That is, define a predictive weight by

(3.6) $\rho(r|C)=k(x_r|\hat{u}_C)$

where $\hat{u}_C$ is the posterior mean, mle or a umvu (whenever definable) estimator for the micro−model of table C. The seating algorithms are defined with no changes. [The only difference is that $\rho(C)$ [i.e., (2.4)] depends on the order the customers arrived at table C as the product rule may not apply. This dependence will not matter much if the sample size n is not too small.] We illustrate the case that $\hat{u}_C$ is the maximum likelihood estimator with respect to table C. Here, a slight inconvenience is that $\hat{u}_C$ may not be definable for tables of small sizes. (By comparison, a posterior mean is usually well−defined for all table sizes.) In any event, if $\hat{u}_C$ exists for the micro−model on table C, the mle predictive weight is defined to be $\rho(r|C)=k(x_r|\hat{u}_C)$; otherwise it is $\rho(r|C)=k(x_r|u_0)$ where u_0 is a pre−determined initial missing value. We call the resulting WCRs the WCR.mle. We examine the performance of the WCR.mle for the density estimation using the location and scale mixture of normals model. An initial kernel $k(.|m_0,\sigma_0)$ is chosen; the m_0, σ_0 are matched with the chosen micro−prior parameters a,b,m,t so that m_0=m and σ_0^2=1/precision=(b/a)×(t+1)/t. Set a=1.5, b=0.005, m=0.0 and t =0.0005. Figure 3.3 displays gWCR approximations using the U(0,1) data with b=1.5 (first row), 0.005, and 0.0005, so that σ_0=44.7, 2.58 and 0.81, respectively. Compared with Figure 3.1, the gWCRs reveal the "true" density better than the iidWCRs. Figure 3.4 displays the gWCR.mle, the iidWCR.mle approximations for the density function where data are n=300 observations from the three peak density (3.5). Compared with the WCRs, the WCR.mles perform surprisingly well. In fact, a case can be made that the WCR.mles outperform the original WCRs in approximating the "true" parameter. The WCR.mles are more robust against changes in prior parameter values. If the mle parameter $N(m_0, \sigma_0^2)$ gives rise to a density that is "orthogonal" to the data likelihood, the seating probability for the second customer is very small, and this creates a numerical problem that results in an incomplete execution of the computer codes. The missing of the iidWCR.mle graph corresponding to σ_0^2 =0.07 illustrates a case in point.

Numerical examples indicate that the WCR.mles run more quickly than the WCRs. The latter require managing and sampling from t−densities which is more time consuming than sampling from normal densities for the WCR.mles. At M=1000, the CPU times in seconds on a Sun SPARC−20 running compiled C code are approximately 360 secs (gWCR; b=0.0005), 405 secs (iidWCR; b=0.0005), 49 secs (gWCR.mle; σ_0=0.81), 50 secs (iidWCR.mle; σ_0=0.81). At M=1, both iidWCR (b=0.0005) and iidWCR.mle (σ_0=0.81) run for less than one second. We do not list the time for gWCRs for M=1 as they run initially very slow.

The approximate predictive weight (3.6) could be used when non−conjugate priors $\alpha(du)$ are used for the model $\{k(.|u), \text{all } u\}$. In the case that conjugate priors do not exist, say $k(.|u)$ is Cauchy with location u, we recommend the use of $\rho(r|C)=k(x_r|\hat{u}_C)$ where $\hat{u}_C$ is the sample medium at table C.

Section 4. Posterior distributions and imputation algorithms.

Liu (1996), studying binomial mixture of normals, proposed a sequential sampling scheme that samples a Blackwell–MacQueen (1973) urn sequence weighted by the binomial kernel. The extension to a general kernel k(.|.), called a weighted Polya process (WP), goes as follows. Assume the mixture model (3.1) and G has a Dirichlet process prior $\mathcal{D}(dG|\alpha)$ [Ferguson (1973)] with shape measure $\alpha(.)$, the posterior distribution of G is an average with respect to the distribution of an extended Polya sequence as follows [see the derivation in Lo (1984)]: $\mathbf{u}=(u_1,...,u_n)$:

(4.1) $$\pi(dG,d\mathbf{u}|\mathbf{x}) \propto \mathcal{D}(dG|\alpha+\Sigma_i\delta_{u_i})\Pi_i k(x_i|u_i)\mu(d\mathbf{u}|\alpha),$$

where $$\mu(d\mathbf{u}|\alpha)=\Pi_{1\le i\le n}(\alpha+\Sigma_{1\le j\le i-1}\delta_{u_j})(du_i).$$

The normalized $\mu(d\mathbf{u}|\alpha)$, $\mu(d\mathbf{u}|\alpha)\times\Gamma(\alpha(R))/\Gamma(\alpha(R)+n))$, is the distribution of the extended Polya sequence [Blackwell and MacQueen (1973)]. The sequential WP algorithm works as follows. Set $\kappa(0)=m(x_1)$ and let u_1 has distribution $\pi(du|x_1)$. Given $u_1,u_2,...,u_{k-1}$, calculate $\kappa(k-1)=m(x_k)+k(x_k|u_1)+...+ k(x_k|u_{k-1})$. Let u_k equal u_j with probability $k(x_k|u_j)/\kappa(k-1)$, $j=1,...k-1$; otherwise, u_k has distribution $\pi(du|x_k)$. The WP distribution is

$$p(d\mathbf{u}|\alpha,\mathbf{k}) =[\Pi_i k(x_i|u_i)\mu(d\mathbf{u}|\alpha)]\times[K_{n-1}]^{-1},$$

$K_{n-1}=\kappa(0)\times...\times\kappa(n-1)$. The posterior mean of $f(t|G)$ is

(4.2) $$\hat{f}(t) \propto \int\{m(t)+\Sigma_{1\le i\le n} k(t|u_i)\}\times K_{n-1}\times p(d\mathbf{u}|\alpha,\mathbf{k}).$$

Similar to the iidWCR, an iidWP samples from $p(d\mathbf{u}|\alpha,\mathbf{k})$ independently. The WP algorithm is a straightforward simulation of the "missing value" u_i based on k(.|.) and $\alpha(.)$. The computation of predictive weights is limited to the marginal weight m(t); computation of predictive weights such as m(t|C) is replaced by sampling from $\pi(du|x_i)$. There is a trade off: the WP does not account for the data reduction part, i.e., ties, of the missing value u_is. Another issue is a prior–sensitivity problem. At the kth step of the WP algorithm, u_k would be an observation from $\pi(du|x_k)$, or else it is one of the previous $u_1,...,u_{i-1}$. Since $\pi(du|x_k)$ differs from $\pi(du)$ only by one observation, the u_is are close to a sample from the micro–prior $\pi(du)$. This implies that if the data likelihood and the micro–prior $\pi(du)$ are approximately "orthogonal" (for a lack of a better word), the iidWP would be more sensitive to the prior. The coefficient of variations (CV) of the WCR and WP iid weights for the three–peaked data in Figure 3.2 are listed below.

Table 4.1. The coefficient of variations for the iid weights in Figure 3.2 wcr (wp)

13.14 (30.01)	17.91 (24.4)
24.92 (31.62)	2.13 (20.75)
31.63 (31.12)	22.48 (30.13)

The smaller CVs for the iidWCR weights seems to signal that it provides a more credible approximation to the posterior mean than the iidWP [Kong, Liu, and Wong (1997)]. Notice that the approximation with smallest CV provides a best fit to the "true" density.

Similar to the gWCR, a Gibbs sampler WP (gWP) can be defined to approximate posterior expectation with respect to the missing value representation (4.1) [see Escobar (1988)]. This approach was discussed by Escobar and West (1995) for a location mixture of normals model and by Brunner (1994, 1995) in general mixture models. Specifically, the stated Markov chain moves from a present state $\mathbf{u}_0$ to the next state $\mathbf{u}_1$ by going through a skip–j prediction cycle. The skip–j predictive probability distribution is defined by: Given $\{u_i, i\neq j\}$, u_j equal u_i with probability $k(x_j|u_i)/m(x_j)+\Sigma_{i\neq j}k(x_j|u_i)$, $i\neq j$; otherwise, u_j has distribution $\pi(du|x_j)$. The stationary distribution of the gWP is

(4.3) $$v(d\mathbf{u}) \propto K_{n-1}\times p(d\mathbf{u}|\alpha,\mathbf{k}) \text{ where } \int v(d\mathbf{u})=1.$$

Similar to the iidWP, the gWP enjoys the flexibility of not having to compute the explicit predictive weights $\rho(j|C)$ [except the marginal weight $\rho(j)=m(x_j)$; note that sampling from $\pi(du|x_j)$s is also required.] However, the lack of data reduction remains a theoretical defect which, according to Blackwell's notion of comparison of experiments [Blackwell (1951, 1953); see also Strassen (1965) and De Groot (1970)], can not be eliminated. To see this, one requires to establish a conditional distribution of $\mathbf{u}|\mathbf{p}$ where the marginal distribution of $\mathbf{u}$ is $v(d\mathbf{u})$ and that of $\mathbf{p}=C_1,\ldots,C_{n(\mathbf{p})}$ is

$$w(\mathbf{p}) \propto \Pi_{1\leq i\leq n(\mathbf{p})}[e_i-1)!\rho(C_i)],$$

i.e., the stationary distribution of the gWCR chain.

Lemma 4.1. Suppose the (marginal) distributions are $\mathbf{u}\sim v(d\mathbf{u})$ and $\mathbf{p}\sim w(\mathbf{p})$. There exists a conditional distribution of $\mathbf{u}|\mathbf{p}$ given by

(i) $t_i|\mathbf{p}$ has distribution $\pi(dt|C_i)$; $i=1,\ldots,n(\mathbf{p})$,

(ii) $t_1,\ldots,t_{n(\mathbf{p})}|\mathbf{p}$ are independent,

(iii) for $i=1,\ldots,n(\mathbf{p})$, duplicate t_i a total of e_i times and, with an abuse of notation, denote them by u_j, $j\in C_i$; $i=1,\ldots,n(\mathbf{p})$.

Proof. It suffices to show that

(4.4) $$\int\Pi_{1\leq i\leq n}h_i(u_i)v(d\mathbf{u})=\Sigma_{\mathbf{p}}\{\Pi_{1\leq i\leq n(\mathbf{p})}\int[\Pi_{j\in C_i}h_j(u)]\pi(du|C_i)\}w(\mathbf{p})$$

for all nonnegative functions $h_i(u_i)$s. Putting $g_i(u_i)=h_i(u_i)k(x_i|u_i)$, $i=1,..,n$ in Lemma 2 in Lo (1984) results in

(4.5) $$\int\Pi_{1\leq i\leq n}h_i(u_i)K_{n-1}\times p(d\mathbf{u}|\alpha,\mathbf{k})$$
$$=\Sigma_{\mathbf{p}}\Pi_{1\leq i\leq n(\mathbf{p})}[e_i-1)!\int[\Pi_{j\in C}h_j(u)k(x_j|u)]\alpha(du)]$$

$$=\Sigma_{\mathbf{p}}\Pi_{1\le i\le n(\mathbf{p})}\{\int[\Pi_{j\in C_i}h_j(u)]\pi(du|C_i)\times(e_i-1)!\times\rho(C_i)\}.$$

Putting $h_i\equiv 1$ results in an equality for the two normalization constants

$$\int K_{n-1}\times p(d\mathbf{u}|\alpha,\mathbf{k})=\Sigma_{\mathbf{p}}\Pi_{1\le i\le n(\mathbf{p})}\{(e_i-1)!\times\rho(C_i)\}.$$

Devide both sides of (4.4) by the normalization constants, respectively, to get (4.4). ‖

Lemma 4.1 and expansion (4.1) combine to yield

(4.6) $G|(\mathbf{p},\mathbf{u})$ has distribution $\mathsf{D}\ (dG|\alpha+\Sigma_{1\le i\le n}\delta_{u_i})$.

Lemma 4.1 and (4.6) complete the description of the posterior distribution of G (and **u**), and the n–folded integral in (4.1) can be strengthened.

Theorem 4.1. The joint posterior distribution of G and the missing value u_is have the following representation

$$\pi(dG,d\mathbf{u}|\mathbf{x}) = \Sigma_{\mathbf{p}}\mathsf{D}\ (dG|\alpha+\Sigma_{1\le i\le n(\mathbf{p})}e_i\delta_{u_i})[\Pi_{1\le i\le n(\mathbf{p})}\pi(du_i|C_i)]w(\mathbf{p})$$

where $w(\mathbf{p}) \propto \phi(\mathbf{p})=\Pi_{1\le i\le n(\mathbf{p})}[e_i-1)!\rho(C_i)]$.

Let us examine the implication of Lemma 4.1 and Theorem 4.1 in terms of comparing Monte Carlo Gibbs sampler experiments. Suppose one wishes to evaluate a posterior integral which is expressed by the missing value representation (4.1), i.e., $\xi=\int T(\mathbf{u})v(d\mathbf{u})$. Using Theorem 4.1 this integral can be expressed alternativelyby $\xi=\Sigma_{\mathbf{p}}S(\mathbf{p})w(\mathbf{p})$, and necessarily $S(\mathbf{p})=E[T(\mathbf{u})|\mathbf{p}]$. Since $S(\mathbf{p})$ is a Rao–Blackwell improvement of $T(\mathbf{u})$, it is less variable in the following sense.

Corollary 4.1. For any convex function c(.), $\int c(T(\mathbf{u}))v(d\mathbf{u})\ge\Sigma_{\mathbf{p}}c(S(\mathbf{p}))w(\mathbf{p})$.

The objective of gWP is to simulate $\mathbf{u}\sim v(d\mathbf{u})$ and evaluate $T(\mathbf{u})$, and the objective of gWCR is to simulate $\mathbf{p}\sim w(\mathbf{p})$ and evaluate $S(\mathbf{p})$. These objectives are achieved upon the Markov chains converging to stationarity. The foregoing argument dictates that $T(\mathbf{u})$ is more variable than $S(\mathbf{p})$, and hence **u** beats **p**, at least eventually.

The application of this result requires the explicit computation of $S(\mathbf{p})=E[T(\mathbf{u})|\mathbf{p}]$. This does not present difficulty in evaluating posterior cross–moments of G (or of u_is) discussed in the paragraph after expression (3.3). An example is given by the Gibbs sampler approximation of $\xi=\hat{f}(t)=E[f(t|G)|\mathbf{x}]$. From (3.2) and (4.2),

$$T(\mathbf{u})=m(t)+\Sigma_{1\le i\le n}k(t|u_i) \text{ and } S(\mathbf{p})=m(t)+\Sigma_{1\le i\le n(\mathbf{p})}e_i m(t|C_i).$$

Another example is given by the approximation of missing values **u**. The Rao–Blackwell improved $S(\mathbf{p})=E[\mathbf{u}|\mathbf{p}]=(\int u\pi(du|C_1),...,\int u\pi(du|C_{n(\mathbf{p})})$ beats $T(\mathbf{u})=\mathbf{u}$. The integral $\int u\pi(du|C)$ is a micro–posterior mean which is easily evaluated if one uses conjugate micro–priors.

Remark 4.1. In the unweighted case, **p** has a marginal density $q(\mathbf{p}|\alpha(R))$ and **u** has a marginal distribution $\mu(d\mathbf{u}|\alpha)\times\Gamma(\alpha(R))/\Gamma(\alpha(R)+n))$. Korwar and Hollander (1973) constructed a conditional distribution of $\mathbf{u}$ given $\mathbf{p}$ where (i) is replaced by (i') $u_i|\mathbf{p}$ has distribution $\alpha(du)/\alpha(R)$, for $i=1,...,n(\mathbf{p})$. The existence of the conditional distribution $\mathbf{u}|\mathbf{p}$ in (i') supports Kuo's (1986) initial proposal to use the (unweighted) Chinese restaurant process rather than the Blackwell and MacQueen (1973) urn to perform an iid Monte Carlo approximation to posterior quantities. Unfortunately, the contribution from the peaked integrands dominates and blurs the otherwise observable improvements.

Remark 4.2. A referee points out that counting the ties of missing values in a gWP chain automatically produces the required gWCR partitions; see West, Muller and Escobar (1994). While the comparison of experiments result favors the directly simulated gWCR partition, the method based on counting ties may be useful if the seating probabilities are difficult to calculate and sampling from the micro–posteriors $\pi(du|x_i)$s does not cause problems.

Section 5. The mixture hazard rate model.

The likelihood function of a hazard rate point process model is proportional to

(5.1) $[\Pi_{1\leq i\leq n}\, r(x_i)]\, \exp\{-\int_I Y(s)r(s)ds\}$

where r(s) is a hazard rate, I is the interval in which the point process is being observed, x_i, i=1,...,n are the uncensored failure times, and Y(s) is a left continuous integer valued function of the data. The likelihood function of the multiplicative point process models [Aalen (1981)] involves a product of these likelihood factors, each of which can be treated independently (by a Bayesian), and creates no additional complexities [Lo and Weng (1989)]. The hazard rate model (4.1) and its multiplicative extension include point process models such as life testing models with censored data, Poisson models and competing risk models, among other point process models. In a mixture hazard rate model, the hazard rate r(.) depends on a kernel k(.|.) and a mixing measure $\mu(du)$ on the "missing" variable u such that

(5.2) $r(s|\nu)=\int k(s|u)\nu(du);$

$k(.|.)\geq 0$ and satisfies integrability conditions in both variables. Often, it is convenient to assume that for each u, k(.|u) is a density. In some cases, the kernel k(.|.) generates a hazard rate that has desirable smoothness properties. For example, the scale mixture of uniforms generates monotone hazard rates, and the scale mixture of exponentials generates "completely monotone" hazard rates [see page 20 in Feller (1971)]. On the other hand, the mixture hazard rate model often arises as a result of missing information. The most renowned example is perhaps the emission tomography model where the data are in fact from a Poisson point process with mixture hazard rates [Chapter 3 in Snyder and Miller (1991)].

In their discussion of the Bayesian mixture hazard rate model, Lo and Weng (1989) argue that the likelihood function (5.1) [and (5.2)] looks like a gamma density in $\nu(.)$, and suggest a weighted gamma process prior for $\nu(.)$. A random measure $\gamma(.)$ is a gamma process with a (σ–finite) shape measure $\alpha(.)$ if (i) $\gamma(.)$ is an "independent increment" process, and (ii) for each A, $\gamma(A)$ is a gamma $(\alpha(A),1)$ random variable. The theory of Dirichlet process can be understood via a gamma process in the sense that $\gamma(.)/\gamma(R)$ is a Dirichlet process with a finite shape measure $\alpha(.)$. The random process $\nu(.)$ defined by $\nu(A)=\int_A\beta(u)\gamma(du)$ is called a weighted gamma process with shape $\alpha(.)$ and multiplier $\beta(.)\geq 0$; its distribution is denoted by $\mathcal{G}\ (d\nu|\alpha,\beta)$. See Lo (1982) for the calculus of weighted gamma processes. In this gamma process prior setting, hazard rates, which are scale mixtures of uniform kernels, were considered by Dykstra and Laud (1981) in life testing models. Theorem 4.1 in Lo and Weng (1989) represents the posterior distribution of ν for model (5.1) and (5.2) as a WP mixture of gamma processes, given by

(5.3) $$\pi(d\nu,d\mathbf{u}|\mathbf{x}) \propto \mathsf{G}\ (d\nu|\alpha+\Sigma_i\delta_{u_i},\beta^*)\Pi_i k^*(x_i|u_i)\mu(d\mathbf{u}|\alpha),$$

where $\beta^*(u)=\beta(u)/[1+\beta(u)\int_I Y(t)\times k(t|u)dt]$,

$k^*(t|u)=\beta^*(u)k(t|u)$,

and $\mu(d\mathbf{u}|\alpha)$ is defined in (4.1); the index i runs through the indice of uncensored observations x_is. The mixture of gamma processes (5.3) resembles (4.1), where $k^*(x_i|u)$ plays the role of the likelihood weight $k(x_i|u)$. The WCRs and WPs for a hazard rate model can be defined by a change to the *–notation: $w_i(u)=k^*(x_i|u)$ in the definition of the WCR; the definitions of Λ^*_{n-1}, Λ^*_n, $m^*(t|C_i)$ and $q(\mathbf{p}|\alpha,\mathbf{k^*})$ follow. The posterior mean of $r(t|\nu)=E[r(t|\nu)|\mathbf{x}]$ is [Theorem 4.2 in Lo and Weng (1989)], $\lambda_n{}^*(t)=m^*(t)+\Sigma_{1\le i\le n(\mathbf{p})}e_i m^*(t|C_i)$,

(5.4) $$\hat{r}(t)=\Sigma_{\mathbf{p}}\Lambda^*_n q(\mathbf{p}|\alpha,\mathbf{k^*})/\Sigma_{\mathbf{p}}\Lambda^*_{n-1}q(\mathbf{p}|\alpha,\mathbf{k^*}).$$

This expression is identical to the expression of $\hat{f}(t)$ in (3.3), except for the factor $[\alpha(R)+n]^{-1}$, which has been incorporated in k* through β*. Sampling the WCR from $q(\mathbf{p}|\alpha,\mathbf{k^*})$ results in an iid Monte Carlo approximation to $\hat{r}(t)$ and to posterior moments of ν. The Gibbs chains for WP and WCR are defined similarly. One could also incorporate a regression model here by letting $w_j(u)\equiv k_j(x_j|u)$, which depends on a regression variable z_j.

The following Lemma 5.1 specifies a conditional distribution of $\mathbf{u}|\mathbf{p}$ where the shape measure α is σ–finite. It suffices to assume the setting in the general case (Section 2): α is σ–finite, $w_i(.)\ge 0$, and for each table C, ρ(C) is finite. The marginal distribution of **u** is $v(d\mathbf{u})$ and that of **p** is $w(\mathbf{p})$ where

$$v(d\mathbf{u}) \propto [\Pi_i w_i(u_i)]\Pi_{1\le i\le n}(\alpha+\Sigma_{1\le j\le i-1}\delta_{u_j})(du_i)$$

$$w(\mathbf{p}) \propto \Pi_{1\le i\le n(\mathbf{p})}(e_i-1)!\rho(C_i).$$

Lemma 5.1. For all non–negative functions g_i,

$$\int\{\Pi_{1\le i\le n}g_i(u_i)\}\times v(d\mathbf{u})=\Sigma_{\mathbf{p}}\{\Pi_{1\le i\le n(\mathbf{p})}\int\Pi_{j\in C_i}g_j(u)\pi(du|C_i)\}\times w(\mathbf{p}).$$

Proof. Since α is σ–finite, there exists a sequence of sets A_k such that $\alpha(A_k)$ is finite and $\alpha(A_k)$ increases to α(R). Recall $\phi(\mathbf{p})=\Pi_{1\le i\le n(\mathbf{p})}(e_i-1)!\rho(C_i)$. The equality

(5.5) $$\int\Pi_i[w_i(u)(\alpha+\Sigma_{1\le j\le i-1}\delta_{u_j})(du_i)]=\Sigma_{\mathbf{p}}\phi(\mathbf{p})$$

is true if α(.) is restricted to A_k [Lemma 2 in Lo (1984)]. As k increases to infinity, both sides increase to an identical finite limit since ρ(C) is finite. Next,

$$\int\{\Pi_{1\le i\le n}[g_i(u_i)\ w_i(u_i)]\}\times\Pi_i[(\alpha+\Sigma_{1\le j\le i-1}\delta_{u_j})(du_i)]$$

$$=\Sigma_{\mathbf{p}}[\Pi_{1\le i\le n(\mathbf{p})}(e_i-1)!]\times\Pi_{1\le i\le n(\mathbf{p})}\int\Pi_{j\in C_i}[g_j(u)w_j(u)]\alpha(du)$$

$$=\Sigma_{\mathbf{p}}\{\Pi_{1\le i\le n(\mathbf{p})}\int\Pi_{j\in C_i}g_j(u)\pi(du|C_i)\}\times\phi(\mathbf{p});$$

the last equality follows from the product rule of probability. Put $g_j(u)=1$ in the last equality to conclude that (5.5) is true for a σ–finite α. ‖

The equality in Lemma 5.1 defines a conditional distribution exactly as Lemma 4.1, and the missing value representation for a posterior distribution (5.3) can then be strengthened to

Theorem 5.1. For the Bayesian mixture hazard rate model with posterior distribution (5.3),

$$\pi(d\mu, d\mathbf{u}|\mathbf{x}) = \Sigma_{\mathbf{p}} G \quad (d\mu|\alpha+\Sigma_{1\le i\le n(\mathbf{p})} e_i \delta_{u_i}, \beta^*)[\Pi_{1\le i\le n(\mathbf{p})} \pi(du_i|C_i)] w^*(\mathbf{p}),$$

where $w^*(\mathbf{p}) \propto \Lambda_{n-1} q(\mathbf{p}|\alpha, \mathbf{k^*})$.

It follows that Corollary 4.1 is also valid for a mixture hazard rate model, and the simulation of $\mathbf{p}$~$w^*(\mathbf{p})$ beats the simulation of $\mathbf{u}$~$v^*(d\mathbf{u})$ in terms of average convex loss functions.

6. Mixture models in the presence of additional parameters.

We shall discuss the case of mixture density as the mixture hazard rate case amounts to a change of notation. The data $\mathbf{x}=(x_1,...,x_n)$ are assumed to have a joint density

(6.1) $$f(\mathbf{x}|\theta,G)=\Pi_{1\le i\le n}\int k_i(x_i|u,\theta)G(du)$$

where, for each i and u, $k_i(.|u)$ is a density. Allowing θ to be a (possibly multi–dimensional) regression parameter [Bunke (1985), Brunner (1995)] allows for regression analysis based on different kernel k_is. The model seems flexible enough to include Poisson, binomial, normal, nonlinear and multivariate regression models as special cases. The prior on the pair (θ,G) is specified by the following:

(6.2) (θ,λ) has a distribution $\pi(d\theta,d\lambda)$ and $G|\theta,\lambda$ is a Dirichlet process with shape measure $\alpha_{\theta,\lambda}(.)$ denoted by $\mathcal{D}\ (dG|\alpha_{\theta,\lambda})$.

The parameter λ is a mixing parameter and it generates the so–called hierarchical Bayesian mixture models [Antoniak (1974)]. Antoniak (1974, Sections 4 and 5) derived explicit expressions for posterior expectations for a sample size of two. The parameter θ plays the role of a regression parameter. The following representation describes the joint distribution of $(\theta,,\lambda,G)$ given $\mathbf{x}$. Fix (θ,λ), use $k_i(x_i|u,\theta)$ and $\alpha_{\theta,\lambda}$ in the definition of the WP and WCR in Section 2 to define $\pi_{\theta,\lambda}(du_i|C_i)$, $\rho_{\theta,\lambda}(C_i)$, and $\phi_{\theta,\lambda}(\mathbf{p})$.

Theorem 6.1. The joint distribution of $(\theta,\lambda,G,\mathbf{u})$ given the data $\mathbf{x}$, is given by

$$\pi(d\theta,d\lambda,dG,d\mathbf{u}|\mathbf{x})$$
$$\propto \Sigma_{\mathbf{p}}\mathcal{D}\ (dG|\alpha_{\theta,\lambda}+\Sigma_{1\le i\le n(\mathbf{p})}e_i\delta_{u_i})[\Pi_{1\le i\le n(\mathbf{p})}\pi_{\theta,\lambda}(du_i|C_i)]\}\phi_{\theta,\lambda}(\mathbf{p})\pi(d\theta,d\lambda).$$

Proof. The joint distribution of $(\theta,\lambda,G,\mathbf{u})$ given the data $\mathbf{x}$ is determined by,

(6.3) $$\pi(d\theta,d\lambda,dG,d\mathbf{u}|\mathbf{x})$$
$$\propto f(\mathbf{x}|\theta,G)\mathcal{D}\ (dG|\alpha_{\theta,\lambda})\pi(d\theta,d\lambda)$$
$$\propto \mathcal{D}\ (dG|\alpha_{\theta,\lambda}+\Sigma_{1\le i\le n)}\delta_{u_i})[\Pi_i k(x_i|\theta,u_i)][\Pi_{1\le i\le n}(\alpha_{\theta,\lambda}+\Sigma_{1\le j\le i-1}\delta_{u_j})(du_i)]\pi(d\theta,d\lambda).$$

Fix (λ,θ) and apply Theorem 4.1 to conclude that the last expression is proportional to

$$\mathcal{D}\ (dG|\alpha_{\theta,\lambda}+\Sigma_{1\le i\le n(\mathbf{p})}e_i\delta_{u_i})[\Pi_{1\le i\le n(\mathbf{p})}\pi_{\theta,\lambda}(du_i|C_i)]\phi_{\theta,\lambda}(\mathbf{p})\pi(d\theta,d\lambda). \quad \|$$

In hierarchical Bayesian mixture models, it is often assumed that $\mathcal{D}\ (dG|\alpha_{\theta,\lambda})=\mathcal{D}\ (dG|\alpha_\lambda)$ which is independent of θ. However, this does not simplify posterior calculations as $\Pi_{1\le i\le n(\mathbf{p})}\pi_{\theta,\lambda}(du_i|C_i)$ in Theorem 6.1 still depends on θ through the kernel $k_i(x_i|u,\theta)$s.

Theorem 6.1 and (6.3) state that given $\mathbf{x}$, the posterior distribution of $(\theta,\lambda,\mathbf{p},\mathbf{u},G)$ can be specified as follows:

(6.4) (i) (θ,λ) has distribution $\pi(d\theta,d\lambda)$ and $\mathbf{p}|(\theta,\lambda)$ has distribution $w_{\theta,\lambda}(\mathbf{p}) \propto \phi_{\theta,\lambda}(\mathbf{p})$,

(ii) $u_i|(\theta,\lambda,\mathbf{p})$ has distribution $\pi_{\theta,\lambda}(du|C_i)$; $i=1,...,n(\mathbf{p})$,

(iii) $u_1,...,u_{n(\mathbf{p})}|(\theta,\lambda,\mathbf{p})$ are independent,

(iv) for $i=1,...,n(\mathbf{p})$, duplicate u_i a total of e_i times and, with an abuse of notation, denote them by u_j, $j \in C_i$; $i=1,...,n(\mathbf{p})$,

and (v) $G|(\theta,\lambda,\mathbf{p},\mathbf{u})$ has distribution $\mathcal{D}$ $(dG|\alpha+\Sigma_{1\le i\le n}\delta_{u_i})$.

It follows than a posterior expectation of a function of (θ,λ,G) (and $\mathbf{u}$) can be represented as $\xi=E[T(\theta,\lambda,\mathbf{u})]$ where the expectation E is with respect to the coherent system specified by (i) to (v). The Rao–Blackwell improvement of the integrand (estimator) $T(\theta,\lambda,\mathbf{u})$ is

(6.5) $$S(\mathbf{p})=E[T(\theta,\lambda,\mathbf{u})|\mathbf{p}].$$

Jensen's inequality for conditional distribution states that

Corollary 6.1. If c(.) is a convex function, $E[c(T(\theta,\lambda,\mathbf{u}))]\ge E[c(S(\mathbf{p}))]$ where $w(d\theta,d\lambda,\mathbf{p})$ is defined by (i) above.

For a hierarchical location mixture of normals model (θ is a constant), Escobar and West (1995) proposes a Gibbs sampler for a mixture of WPs to simulate the missing value u_i. MacEachern (1994) proposes a Gibbs sampler based on sampling a partition **p** which has a mixture of WCR distribution (the mixing distribution being the posterior distribution of $\lambda|\mathbf{x}$) and he also presents some numerical studies of his method. This corresponds to the case that $T(\lambda,\mathbf{u})=u_i$ and according to Corollary 6.1, the Rao–Blackwell improved $S(\mathbf{p})=E[u_i|\mathbf{p}]$ beats u_i. [Note however that MacEachern (1994) essentially uses skip–j partitions rather than the full-information partition in his numerical examples.] The conditional distribution of $\mathbf{u}|\mathbf{p}$ can be read off from (i) to (v).

Except in isolated cases, sampling a partition **p** which has a mixture of WCR distribution has difficulties; Example 6.2 is a case in point. Theorem 6.1 suggests a more natural Monte Carlo method for evaluating posterior expectations. According to Theorem 6.1, the posterior expectation can be written concisely as a mixture of three variables $(\theta,\lambda,\mathbf{p})$ with a mixing distribution $w(d\theta,d\lambda,\mathbf{p})$ specified by (i) above. The issue is to sample the triple $(\theta,\lambda,\mathbf{p})$ from $w(d\theta,d\lambda,\mathbf{p})$ sequentially.

The iidWCR is easy to implement. Select (θ,λ) from $\pi(d\theta,d\lambda)$. Next, given (θ,λ), simulate a WCR partition **p** with $\alpha_{\theta,\lambda}(.)$ playing the role of $\alpha(.)$. The usual conjugate prior and posterior analysis operates on each table [for a given (θ,λ)], and simulation can be done efficiently if $\alpha_{\theta,\lambda}(du)$ is a conjugate prior for the model $k_i(.|u,\theta)$. Calculate $\Lambda(n-1)$ [which depends on (θ,λ)]. Repeat to get $(\theta_1,\lambda_1,\mathbf{p}_1),...,(\theta_M,\lambda_M,\mathbf{p}_M)$, and $\Lambda_1(n-1),...,\Lambda_M(n-1)$. An iidWCR approximation to $E[h(\theta,\lambda,G)|\mathbf{x}]$ is

(6.6) $$\Sigma_{1\le k\le M}L(\theta_k,\lambda_k,\mathbf{p}_k)\Lambda_k(n-1)/\Sigma_{1\le k\le M}\Lambda_k(n-1),$$

where $L(\theta,\lambda,\mathbf{p})=\iint h(\theta,\lambda,G)\mathcal{D}\ \ (dG|\alpha_{\theta,\lambda}+\Sigma_{1\le i\le n(\mathbf{p})}e_i\delta_{u_i})[\Pi_{1\le i\le n(\mathbf{p})}\pi_{\theta,\lambda}(du_i|C_i)]$.

The gWCR requires the construction of a Markov chain $(\theta_1,\lambda_1,\mathbf{p}_1),\ldots,(\theta_M,\lambda_M,\mathbf{p}_M),\ldots$ with a stationary distribution $w(d\theta,d\lambda,\mathbf{p})$ and uses $\Sigma_{1\le k\le M}L(\theta_k,\lambda_k,\mathbf{p}_k)/M$ rather than the weighted average (6.6) as an estimator. The Markov chain moves from a present state $(\theta_0,\lambda_0,\mathbf{p}_0)$ to the next state $(\theta_1,\lambda_1,\mathbf{p}_1)$ as follows:

(6.7) Step 1. Given a present state $(\theta_0,\lambda_0,\mathbf{p}_0)$, use a skip–j cycle to move $\mathbf{p}_0$ to $\mathbf{p}_1$; the seating probability is defined with $\alpha_{\theta_0,\lambda_0}(.)$ and $k_i(x_i|u,\theta_0)$ playing the role of $\alpha(.)$ and $k(x_i|u)$.

Step 2. Given $(\theta_0,\lambda_0,\mathbf{p}_1)$, sample (θ,λ) from $(\theta,\lambda)|\mathbf{p}_1$ to get (θ_1,λ_1).

The sampling of $(\theta,\lambda)|\mathbf{p}$ may not be easy. If (θ,λ) has a density $\pi'(\theta,\lambda)$. The conditional density of (θ,λ) given $\mathbf{p}$ is a weighted $\pi'(\theta,\lambda)$,

(6.8) $\phi_{\theta,\lambda}(\mathbf{p})\pi'(\theta,\lambda)/\iint\phi_{\theta,\lambda}(\mathbf{p})\pi'(\theta,\lambda)d\theta d\lambda$.

First is a case where a direct simulation from $(\theta,\lambda)|\ \mathbf{p}$ is possible.

Example 6.1. A hierarchical location mixture of normals. This is the hierarchical Bayesian mixture model considered by MacEachern (1994); see also Escobar and West (1995). Here $k_i(.|u)$ is $N(u,1/\tau_i)$ where τ_is are known. The prior can be specified by $\pi(d\theta,d\lambda)$ which degenerates at a constant for the θ factor and is $N(m_0,1/t_0)$ for the λ factor. $G|\lambda$ is a Dirichlet process with an expectation a $N(\lambda,1/\tau_0)$ distribution. Given λ, the micro–predictive density of x_r given table C, $m_\lambda(x_r|C)$, $r\notin C$, is $N(\mu_c,1/\tau_c)$ where

(6.9) $\mu_c(\lambda)=(\tau_0\lambda+\Sigma_{j\in C}\tau_j x_j)/(\tau_0+\Sigma_{j\in C}\tau_j)$, $\tau_c{}^{-1}=\tau_0{}^{-1}+(\tau_0+\Sigma_{j\in C}\tau_j)^{-1}$.

The conditional density of $\lambda|\mathbf{p}$ is proportional to $\phi_\lambda(\mathbf{p}_1)\exp\{-(t_0/2)(\lambda-m_0)^2\}$, which simplifies to a $N(\mu(\mathbf{p}),1/\tau(\mathbf{p}))$ density where

$\tau(\mathbf{p})=\Sigma_{0\le i\le n(\mathbf{p})}t_i$, and $\mu(\mathbf{p})=\Sigma_{0\le i\le n(\mathbf{p})}t_i m_i/\tau(\mathbf{p})$,

$m_i=\Sigma_{j\in C_i}x_j\tau_j/\Sigma_{j\in C_i}\tau_j$, and $t_i{}^{-1}=\tau_0{}^{-1}+(\Sigma_{j\in C_i}\tau_j)^{-1}$, $i=1,\ldots,n(\mathbf{p})$.

Since $\lambda|\mathbf{p}$ is an explicit $N(\mu(\mathbf{p}),1/\tau(\mathbf{p}))$, use the average of $\{E[h(\lambda_m)|\mathbf{p}_m]$, $m=k+1,\ldots,k+M\}$ to approximate the posterior mean $E[h(\lambda)|\mathbf{x}]$.

Remark 6.1. MacEachern (1994) proposed a seating probability for j seats at table $\tilde{C}_i$ ($\tilde{C}_0$ is the empty table) proportional to (essentially) a $[1+n(\mathbf{p})]$–folded multiple integral

$\tilde{e}_i\times\iint\Pi_{0\le q\le n(\mathbf{p})}\Pi_{j\in\tilde{C}_q}k_r(x_r|u_q)\alpha_\lambda(du_q)\pi(d\lambda)$, $i=0,\ldots,n(\tilde{\mathbf{p}})$.

This quantity can be called a "joint" weight. Given λ, the independence of u_qs (see Lemma 4.1) reduces the inner $n(\mathbf{p})$–folded integral to an $n(\mathbf{p})$–folded product of single integrals

(6.10) $\tilde{e}_i\times\int\Pi_{0\le q\le n(\mathbf{p})}\rho_\lambda(\tilde{C}_q)\pi(d\lambda)$, $i=0,\ldots,n(\tilde{\mathbf{p}})$, where $\rho_\lambda(\tilde{C}_q)=\int\Pi_{r\in\tilde{C}_q}k_r(x_r|u)\alpha_\lambda(du)\}$.

These expressions do not involve the predictive weight $\rho_\lambda(j|C_i)$, and in view of the averaging by $\pi(d\lambda)$ in (6.10), sampling a partition based on this seating probability is equivalent to sampling a partition that has a mixture of WCR distribution. However, a mixture of WCR distribution usually has complicated seating probabilities [see MacEachern (1994)]. Furthermore, averaging out λ implies that the evaluation of a conditional expectation given $\mathbf{p}$ is rarely easy (see the discussions following Example 6.2).

The problem of sampling from $(\theta,\lambda)|\mathbf{p}$ is often non−trivial even for natural mixture models equipped with micro−conjugate priors. Note that $\rho_{\theta,\lambda}(C_i)$ in (6.7) is the mixture density of the observations x_j, $j\in C_i$, with mixing distribution $\alpha_{\theta,\lambda}(du)$, and $\rho_{\theta,\lambda}(C_i)$ is often summarized by a micro−sufficient statistic of a fixed dimension (the Cauchy kernel is notably excluded), say s_i for the ith table. The model is the following: $s_i|\theta,\lambda$ are independent and nonidentically distributed with densities that can be read from the expression for $\rho_{\theta,\lambda}(C_i)$; the prior density for (θ,λ) is $\pi'(\theta,\lambda)$. Sampling the posterior density of (θ,λ) given the data s_i, $i=1,...,n(\mathbf{p})$ can be done via the rejection method [see for example Chibb and Greenberg (1995)]. However, it may take several rejections to produce an acceptable (θ,λ) since the expected number of rejection is a prior number that is difficult to choose.

Another approach is to nest a Metropolis−Hastings rejection step within a Gibbs cycle [Hastings (1970), Chib and Greenberg (1995); see also Brunner (1995)]. This procedure replaces Step 2 in (6.7) by a randomization step:

(6.11) Step 2a. Sample (θ^*,λ^*) from $\pi'(\theta,\lambda)$. Let $(\theta_1,\lambda_1)=(\theta^*,\lambda^*)$ with probability $\min\{\phi_{\theta^*,\lambda^*}(\mathbf{p}_1)/\phi_{\theta_0,\lambda_0}(\mathbf{p}_1),1\}$; otherwise $(\theta_1,\lambda_1)=(\theta_0,\lambda_0)$.

This algorithm accepts the new value (θ^*,λ^*) as (θ_1,λ_1) if $\phi_{\theta^*,\lambda^*}(\mathbf{p}_1)$ dominates $\phi_{\theta_0,\lambda_0}(\mathbf{p}_1)$, and it retains a positive probability of accepting the new value (θ^*,λ^*) even if $\phi_{\theta^*,\lambda^*}(\mathbf{p}_1)$ is dominated by $\phi_{\theta_0,\lambda_0}(\mathbf{p}_1)$. The last step allows the algorithm to avoid being caught at a local maximum $\phi_{\theta_0,\lambda_0}(\mathbf{p}_1)$.

Example 6.2. A location model with a scale mixture of normals error. A scale mixture of normals is heavy−tailed and could be an appropriate model for data with outliers. This model is also a stamping ground for regression problems. Here $k_i(.|\theta,u)$ is $N(\theta z_i,\Sigma_i/u)$ where θ is the "regression" parameter. To simplify the discussion, we assume that all variables are univariate and the predictor z_i and covariance matrix Σ_i are a constant 1. The prior specification is: θ and G are independent, θ is $N(\mu_0,1/t_0)$ and $G|\theta$ is Dirichlet with shape probability a gamma (a,1/b) distribution; $\alpha(R)>0$. Given θ, the micro−posteriors are also gamma: $\pi_\theta(du|C_i)$ is gamma $(a_i,1/b_i(\theta))$ where

(6.12) $a_i=a+e_i/2$ and $b_i(\theta)=b+2^{-1}[\Sigma_{j\in C_i}(x_j-\bar{x}_i)^2]+e_i(\bar{x}_i-\theta)^2]$ for all θ.

The micro–predictive density $m_\theta(x_r|C_i)$ is a t–density with degrees of freedom $2a_i$, location θ, and precision $2a_i/b_i(\theta)$, evaluated at x_r. The implementation of the iidWCR is rather straight forward and was discussed. A gWCR requires an additional simulation from the conditional density of $\theta|\mathbf{p}$ proportional to $\phi_\theta(\mathbf{p})\exp\{-t_0(\theta-\mu_0)^2\}$. This conditional distribution is the posterior distribution of θ under the following sampling plan: θ has density $\pi'(\theta)$, $\bar{x}_i|\theta$ is a t–density with df$=2a_i-1$, location=θ, and precision $e_i(2a_i-1)/b_i(\bar{x}_i)$, $i=1,...,n(\mathbf{p})$. From another viewpoint, the posterior density of $\theta|\mathbf{p}$, say $\pi'(\theta|\mathbf{p})$, is (proportional to) a product of independent yet non–identically distributed t–densities and $\pi'(\theta)$. Given $\theta=\theta_0$, the Gibbs cycle gives $\mathbf{p}=\mathbf{p}_1$. To get the next value of $\theta=\theta_1$, sample $\theta=\theta^*$ from $N(\mu_0,1/t_0)$. Accept θ^* as θ_1 with probability $\min\{\phi_{\theta^*}(\mathbf{p}_1)/\phi_{\theta_0}(\mathbf{p}_1),1\}$; otherwise keep θ_0 as θ_1. In the present setting,

(6.13) $\phi_{\theta^*}(\mathbf{p}_1)/\phi_{\theta_0}(\mathbf{p}_1)=\Pi_{1\le i\le n(\mathbf{p}_1)}[b_i(\theta_0)/b_i(\theta^*)]^{a_i}$.

Proceed with the next gWCR cycle to simulate $\mathbf{p}_2|\theta_1$. Repeat to get a sequence of states $(\theta_1,\mathbf{p}_1)$, $(\theta_2,\mathbf{p}_2)$,..., which will be the basis of the Markov chain Monte Carlo approximations.

Suppose one wishes to approximate the posterior distribution of the location parameter, say $\xi=\int h(\theta)\pi(d\theta|\mathbf{x})$. Gelfand and Smith (1990) pointed out that the average of conditional expectations $E[h(\theta_m)|\mathbf{p}_m]$, $m=k+1,...,k+M$ (k is the warm–up time) is a Rao–Blackwell improvement to the average of $h(\theta_m)$, $m=k+1,...,k+M$. This is good as long as the conditional expectation $E[h(\theta_m)|\mathbf{p}_m]$ can be evaluated explicitly (see Example 6.1). However, for hierarchical Bayesian mixture models, this is the exception rather than the norm (see Example 6.2). Often, the computation of $E[h(\theta_m)|\mathbf{p}_m]$ requires a numerical routine such as Newton's method, which drastically reduces the benefit of conditioning. For this reason, we shall use the "marginal values" $h(\theta_{k+1})$, $h(\theta_{k+2})$,..., rather than the conditional expectations, as the marginal values are already available in a realization of the Markov chain.

Figure 6.1 plots the cumulative distribution function of $(\theta_{k+1},\theta_{k+2},...)$ which approximates the posterior distribution of θ; data are n=300 observations from standard normal (column 1) and standard Cauchy (column 2) densities. Prior parameters are set at a= 1.5, $\mu_0=0.5$, $t_0=0.05$, and $\theta_0=4.22810$ which is an observation from $N(\mu_0,1/t_0)$.

The warm–up time has little effect in this nested algorithm. When the parameter b decreases to 0.005, a value which was found to having good resolution in approximating a U(0,1) density, the cumulative distribution of θ (dashed) becomes unacceptable. Table 6.1 summarizes the numerical result.

Table 6.1		normal		Cauchy	
warm−up	b	mean	SD	mean	SD
k=0	1.5	−0.0876	0.10691	0.07403	0.13984
	0.005	−0.0084	0.10976	0.17813	0.19549
k=2,000	1.5	−0.0390	0.04690	0.10091	0.08859
	0.005	0.00472	0.01877	0.07193	0.01050
k=5,000	1.5	−0.0428	0.06649	0.08993	0.08338
	0.005	0.03163	0.02847	0.04476	0.00671
k=50,000	1.5	−0.03396	0.05079	0.12245	0.05832
	0.005	0.02497	0.03916	0.09724	0.04156

To estimate the mixture density, we note that given θ and $\mathbf{p}$, an estimate of the density is given by [see (3.2)]

(6.14) $$\hat{f}_\theta(t|\mathbf{p})=\Sigma_{0\le i\le n(\mathbf{p})}[e_i/(e_0+n)]m_\theta(t|C_i).$$

Figure 6.2 plots the average of $\hat{f}_\theta(t|\mathbf{p})$s evaluating at $(\theta,\mathbf{p})=(\theta_{k+1},p_{k+1}),\ldots,(\theta_{k+M},\mathbf{p}_{k+M})$, which is a Markov chain approximation of $E[f(t|\theta,G)|\mathbf{x}]$. The convergence is fast and the warm−up time has no effect.

Acknowledgements. The authors are indebted to Martin Law, Man−Wai Ho, Mike K. P. So, Ms. Chi−Man Tsui and Amy C. Yip for their inputs to this project. The referees and associate editor contributed useful comments and criticisms.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics,* **6**, 701–726.

Aldous, D.J. (1985). Exchangeability and Related Topics. Lecture Notes in Mathematics. 1117 Springer–Verlag.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics,* **2**, 1152–1174.

Arabie, P., Hubert, L. J. and De Soete, G. (1996). Clustering and Classification. *World Scientific Publications,* River Edge, N.J., USA

Blackwell, D. (1951). Comparison of experiments. *Proc. 2nd Berkeley Symp. Math. Statist. Prob.* Univer. Calif. Press, Berkeley, Calif., 93–102.

Blackwell, D. (1953). Equivalent comparison of experiments. *Ann. Math. Statist.,* **24**, 265–272.

Blackwell, D. and MacQueen, J. (1973). Fergusion distribution via Polya urn schemes. *The Annals of Statistics,* **1**, 353–355.

Brunner, L. J. (1994). Using the Gibbs sampler to simulate from the Bayes estimate of a decreasing density. *Communications in Statistics,* **24**, 215–226.

Brunner, L.J. (1995). Bayesian linear regression with error terms that have symmetric unimodal densities. *Journal of Nonparametric Statistics,* **4**, 335–348.

Brunner, L.J. and Lo, A.Y. (1989). Bayes method for a symmetric and unimodal density and its mode. *The Annals of Statistics,* **17**, 1550–1566.

Brunner, L.J. and Lo, A.Y. (1994). A Bayes approach to directional data. *Canadian Journal of Statistics,* **22**, 401-412.

Bunke, O. (1985). Bayesian estimators in semiparametric problems. Preprint Nr. 102, Sektion Mathematik, Humboldt–Universitat Zu Berlin.

Censov, N.N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.,* **3**, 1559–1562.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis–Hastings Algorithm. *American Statistician,* **49**, 327–335.

Chung, K. L. (1967). Markov chains with stationary transition probabilities. Springer–Verlag.

Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian Sampling. *JRSS* B, **56**, 363–375.

De Groot, M. (1970). *Optimal Statistical Decisions* McGraw–Hill, Inc.

Duda, R.O and Hart, P.E. (1973). *Pattern Classification and Scene Analysis,* John Wiley and

Sons.

Dykstra, R.L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability. *The Annals of Statistics,* **9**, 356–367.

Escobar, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Ph. D. Dissertation, Yale University, Department of Statistics.

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

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions.*Chapman and Hall, London

Feller, W. (1971). *An Introduction to Probability and its Applications Vol. II.* John Wiley and Sons.

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

Ferguson, T.S. (1983). Bayesian density estimation by mixture of Normal distributions. *Recent Advances in Statistics* Academic Press, 278–302.

Geman, A., 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.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling–based approaches to calculating marginal densities. *Journal of the American Statistical Association,* **85**, 398–409.

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

Jewell, N.P. (1982). Mixtures of exponential distributions. *The Annals of Statistics,* **10**, 479–484.

Ji, Wenyun (1991). The evaluation of Bayes estimates for mixture models. Thesis, Master of Arts degree in Statistics. SUNY at Buffalo.

Korwar, R.M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Statistics,* **1**, 705–711.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM J. Sci. Statist. Comput.,* **7**, 60–71.

Lindsay, B. (1983). The geometry of mixture likelihoods, I and II. *The Annals of Statistics,* **11**, 86–94 and 783–792.

Lindsay, B. (1995). Mixture Models: Theory, Geometry and Applications. *CBMS–NSF regional Conference series in Probability and Statistics,* Vol. **5**.

Liu, J. (1996). Nonparametric Hierarchical Bayes via Sequential Imputations. *The Annals of Statistics*, **24**, 910–930.

Lo, A.Y. (1978). Bayesian nonparametric density methods. Technical Report, Department of Statistics, University of California at Berkeley.

Lo, A.Y. (1982). Bayesian nonparametric statistical inference for Poisson point process. *Z. Wahr. verw. Gebiete.*, **59**, 55–66.

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

Lo, A.Y. and Weng, C.S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Instit. Statist. Math.*, **41**, 227–245.

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

Parzen, E. (1962). On the estimation of a probability density and mode. *Ann. Math. Statist.*, **33**, 1065–1076.

Robert, C.P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science*, **10**, 231–253.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–835.

Snyder, D.L. and Miller, M.I. (1991). *Random Point Processes in Time and Space.* Springer–Verlag New York, Inc.

Strassen, V. (1965). The existence of probability measures with given marginals. *Ann. Math. Statist.*, **36**, 423–439.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions* Wiley, New York.

Tanner, M. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **52**, 528–550.

West, M., Muller, P., and Escobar, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation". *Aspect of Uncertainty: A tribute to D.V. Lindley*, Edited by P.R. Freeman and A. F. M. Smith. John Wiley.