(*) Handbook of Mixture Analysis

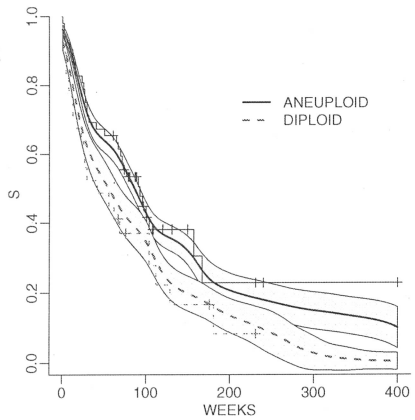


**FIGURE 6.4**
Estimated survival function $E(F_x \mid \boldsymbol{y})$ by tumor type (solid black and dashed red curves). The grey shaded bands around the estimated survival functions show pointwise $\pm 1.0$ posterior standard deviation bounds. The piecewise constant lines plot the Kaplan-Meier estimates.

## 6.3 NGG mixtures

### 6.3.1 NRMI construction

Yet another defining property of the DP is the construction as a normalized gamma process (Ferguson, 1973). The gamma process is a particular example of a much wider class of models known as completely random measures (CRM) (Kingman, 1993, chapter 8). Consider any non-intersecting measurable subsets $A_1, \ldots, A_k$ of the desired sample space $X$. The defining property of a CRM $\tilde{G}$ is that $\tilde{G}(A_j)$ be mutually independent. The gamma process is a CRM with $\tilde{G}(A_j) \sim \mathcal{G}(\alpha G_0(A), 1)$, for a probability measure $G_0$ and $\alpha > 0$. Normalizing $\tilde{G}$ by $G(A) = \tilde{G}(A)/\tilde{G}(X)$ defines a DP prior with base measure $\alpha G_0$.

**CRM and NGG.** Replacing the gamma process by any other CRM defines alternative BNP priors for random probability measures. Such priors are known as normalized random measures with independent increments (NRMI) and were first described in Regazzini et al. (2003) and include a large number of BNP priors. A recent review of NRMI's appears in Lijoi and Prünster (2010). Besides the DP prior, another interesting example is the normalized generalized gamma process (NGG), discussed in Lijoi et al. (2007). We write $G \sim \text{NGG}(\alpha, \kappa, \gamma, G_0)$. The NGG is indexed by a total mass parameter $\alpha > 0$, two more scalar parameters $\kappa \geq 0$ and $\gamma \in [0, 1)$ and a base probability measure $G_0$. In fact, the DP is a special case of the NGG with $\kappa = 1$ and $\gamma = 0$.

Like for any CRM, a realization from the generalized gamma process (before normalization) can be generated using the following constructive definition (Kingman, 1993; section 8.2). Assume we wish to generate a random measure $G$ on a measurable space $X$, for example $\Re^d$. We set up a Poisson process over $\Re^+ \times X$ with Poisson intensity $\nu(\tilde{\eta}, \mu)$. The choice of $\nu(\cdot)$ determines different CRM's. The arguments are already labeled in anticipation of the next step. For the generalized gamma process we use

$$\nu(\tilde{\eta}, \mu) = \rho(\tilde{\eta})\, \alpha G_0(\mu) \text{ with } \rho(\tilde{\eta}) = e^{-\kappa\tilde{\eta}} \tilde{\eta}^{-(1+\gamma)} / \Gamma(1-\gamma). \tag{6.14}$$

Let $(\tilde{\eta}_h, \mu_h)$, $h = 1, \ldots,$ denote a realization of this Poisson process. Then $\tilde{G} \propto \sum \tilde{\eta}_h \delta_{\mu_h}$ is a realization of the desired CRM. Here, $\tilde{G}$ is still the (un-normalized) CRM. The normalized measure $G$ rescales the weights $\tilde{\eta}_h$ to unit total mass.

**NGG mixture.** Barrios et al. (2013) discuss mixture models with an NGG prior on the mixing measure, similar to (6.9), but with an NGG prior replacing the DP prior.

$$y_i \mid \boldsymbol{\theta} \sim p(y_i \mid \theta_i),\ \theta_i \mid G \sim G,\ G \sim \text{NGG}(\alpha, \kappa, \gamma, G_0) \tag{6.15}$$

The discussion in Barrios et al. (2013) is more general, allowing for any other NRMI, but the NGG is a sufficiently rich model for most purposes. However, in comparison with the DP prior the additional flexibility of the NGG is important for modelling. This is extensively discussed in De Blasi et al. (2014) and in Barrios et al. (2013). For example, consider two clusters $k$ and $\ell$ with cluster sizes $n_k > n_\ell$. As before let $z_i$ denote a latent cluster membership indicator in an equivalent hierarchical model version of (6.15), let $\boldsymbol{z}_{-i} = (z_j,\ j \neq i)$ and define $n_k^- = \sum_{j \neq i} I(z_j = k)$ to be cluster sizes without the $i$-th unit. Then a priori $p(z_i = k \mid \boldsymbol{z}_{-i})/p(z_i = \ell \mid \boldsymbol{z}_{-i}) = (n_k^- - \gamma)/(n_\ell^- - \gamma)$. The implication for data analysis is that cluster sizes under NGG priors with $\gamma > 0$ tend to be more concentrated, with few large clusters including most experimental units. Perhaps more importantly the implied prior on the number of clusters, $K$, is more flexible under the NGG prior, in the sense that for matching prior means, hyperprior parameters can be chosen to allow for substantially more prior variance for $K$. This allows the number of clusters to be a posteriori adjusted as needed for the data. Under the DP prior, $p(K)$ is centered around approximately $\alpha \log(n)$. That is, prior centering determines $\alpha$, leaving no more flexibility to inflate prior variance.

### 6.3.2 Posterior simulation for NGG mixtures

Most importantly, posterior inference under (6.15) is still easily implemented. Barrios et al. (2013) as well as Favaro and Teh (2013) and Argiento et al. (2010) outline specific MCMC algorithms. Both are based on a representation of the posterior distribution for NRMI's discussed in James et al. (2009), under independent sampling, $\theta_i \sim G$ as in (6.15) and an NRMI prior for $G$. Details of the general result are not needed for the upcoming algorithm for NGG mixtures. We only outline the setup, and give specific details for the NGG mixture. The representation involves a model augmentation of the posterior $p(\boldsymbol{\theta}, \tilde{G} \mid \boldsymbol{y})$ under (6.15) with a latent variable, $u$, using

$$p(u \mid \boldsymbol{\theta}, \tilde{G}, \boldsymbol{y}) = p(u \mid K) \propto u^{n-1} (u + \kappa)^{K\gamma - n} e^{-\alpha(u+\kappa)^\gamma / \gamma}. \tag{6.16}$$

For the following description of the algorithm it is convenient to distinguish atoms of $G$ that are matched with currently imputed $\theta_i$ versus unmatched atoms. Also posterior inference is easiest discussed for the random measure $\tilde{G}$, before normalization. As before let $\{\theta_k^\star,\ k = 1, \ldots, K\}$ denote the unique $\theta_i$. Then

$$\tilde{G} = \sum_{k=1}^{K} \eta_k^\star \delta_{\theta_k^\star} + \tilde{G}_C \text{ with } \tilde{G}_C = \sum_{h=1}^{\infty} \tilde{\eta}_h \delta_{\mu_h}. \tag{6.17}$$

Note that the split of $\tilde{G}$ in (6.17) implicitely is a function of $\boldsymbol{\theta}$ and can only be used when conditioning on $\boldsymbol{\theta}$. In the MCMC implementation we approximate $\tilde{G}_C$ by using the $H$ terms with largest $\tilde{\eta}_h$ only. This is possible since the algorithm for generating $\tilde{G}_C$ samples the $\tilde{\eta}_h$

in decreasing order. We can therefore assume that the weights are indexed by decreasing order, $\tilde{\eta}_h \geq \tilde{\eta}_{h+1}$. Let $\tilde{\boldsymbol{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_H)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_H)$. Finally, let $\boldsymbol{\eta}^\star = (\eta_1^\star, \ldots, \eta_K^\star)$ and $\boldsymbol{m} = (n_1, \ldots, n_K)$.

**An MCMC scheme for NGG mixture models.** We describe the particular Gibbs sampling implementation that is proposed in Barrios et al. (2013). The following steps define transition probabilities of an MCMC scheme for the posterior distribution $p(\boldsymbol{\theta}, \tilde{G}, u \mid \boldsymbol{y})$ under (6.15), augmented with (6.16). The algorithm includes five transition probabilities. Let $[a \mid b, c]$ indicate sampling from the conditional distribution of parameter $a$ given $b, c$. All distributions are complete conditional posterior distributions, with the absence of any variables in the conditioning set indicating conditional indpendence. In some cases dependence on $\boldsymbol{\theta}$ is only indirectly through $\boldsymbol{z}$, or even just $K$ or $n_j$. The five steps are (i) $[u \mid K]$, (ii) $[\eta_k^\star \mid u, n_k]$; (iii) $[\theta_k^\star \mid \boldsymbol{z}, \boldsymbol{y}]$; (iv) $[\boldsymbol{\eta}, \boldsymbol{\mu} \mid u]$; (v) $[\theta_i \mid \tilde{G}]$.

In step (i) we generate $u$ from (6.16). Favaro and Teh (2013) recommend to instead sample $v = \log(u)$, as the complete conditional distribution $p(v \mid K)$ turns out to be log concave, allowing for easier r.v. generation.

In step (ii) we update $\eta_k^\star$, $k = 1, \ldots, K$, by generating from the complete conditional posterior which under the NGG simplifies to

$$p(\eta_k^\star \mid u, n_j) = \mathcal{G}(n_j - \gamma, \kappa + u),$$

In step (iii) we draw from the complete conditional posterior distribution for $\theta_k^\star$. This step is identical to (6.10).

Step (iv) updates $\tilde{G}_C$. James et al. (2009) show that conditional on $u$ the random $\tilde{G}_C$ is again a CRM with Poisson intensity $\nu(\tilde{\eta}, \mu)$ replaced by an updated version. In the case of the NGG this simpifies to

$$\tilde{G}_C \sim \text{NGG}(\alpha, \kappa + u, \gamma, G_0).$$

We can use the following easy algorithm to generate $\tilde{G}_C = \sum \tilde{\eta}_h \delta_{\mu_h}$. Ferguson and Klass (1972) introduce a clever scheme to generate the weights $\tilde{\eta}_h$ in decreasing order. The construction requires a function $N(v) = \int_v^\infty \rho(\tilde{\eta}) d\tilde{\eta}$, using the factor $\rho(\tilde{\eta})$ from definition (6.14), with $\kappa^\star = \kappa + u$ in place of $\kappa$. That is, $N(v) = \frac{\alpha}{\Gamma(1-\gamma)} \int_v^\infty e^{-(\kappa+u)\tilde{\eta}} \tilde{\eta}^{-(1+\gamma)} d\tilde{\eta}$. Next let $\xi_1, \xi_2, \ldots$ denote a realization from a unit rate Poisson process, that is $\xi_h - \xi_{h-1} \sim \mathcal{E}(1)$ are i.i.d. exponential draws (starting with $\xi_0 = 0$). Then

$$\tilde{\eta}_h = N^{-1}(\xi_h).$$

In words, plot the function $N(v)$ against $v \geq 0$, mark the $\xi_h$ on the vertical axis, and then use $N^{-1}(\cdot)$ to map $\xi_1, \xi_2, \ldots,$ to the horizontal $v$-axis. The construction delivers $\tilde{\eta}_h$, already ordered by decreasing size. The construction of $\tilde{G}_C$ is completed by generating the locations $\mu_h \sim G_0$, i.i.d.

Finally, in step (v) we resample $\boldsymbol{\theta}$ by generating $p(\theta_i \mid \tilde{G}, y_i) \propto \tilde{G}(\theta_i) p(y_i \mid \theta_i)$. Write $\tilde{G} = \sum_{\ell=1}^\infty \tilde{w}_\ell \delta_{m_\ell}$, using a single running index $\ell$ for all terms in (6.17). Then $p(\theta_i = m_\ell) \propto \tilde{w}_\ell \, p(y_i \mid m_\ell)$.

The described MCMC is implemented in the R package `BNPdensity`, which is available in the CRAN package repository (`http://cran.r-project.org/`).

An alternative MCMC scheme for posterior inference under model (6.15) is descrbied in Favaro and Teh (2013) and also in Argiento et al. (2010). Favaro and Teh (2013) describe what can be characterized as a modified version of the Polya urn. Recall that the Polya urn defines the marginal distribution of $(\theta_1, \ldots, \theta_n)$ under the DP prior, after marginalizing with respect to $G$. Similarly, Favaro and Teh (2013) describe a method for sampling $p(\theta_1, \ldots, \theta_n \mid u, \boldsymbol{y})$, marginalizing with respect to $G$. Generating $u \mid \boldsymbol{\theta}$ proceeds as in step (i), above. Additionally, they describe the complete conditional posterior distributions for the NGG hyperparameters. This allows to augment model (6.15) with a hyperprior on the NGG parameters.

## 6.4 BNP mixtures with random partitions

Recall that we started the discussion by observing that a mixture model (6.1) can naturally be thought of as a mixture with respect to a mixing measure, as in (6.2), and we proceeded by assuming BNP priors on the mixing measure. When the BNP prior generates discrete probability measures with infinitely many atoms, such as the DP and the NRMI prior, this construction leads to an infinite mixture.

There is another feature of hierarchical models like (6.2) with a discrete BNP prior $p(G)$ that naturally leads to a mixture model. Consider the posterior predictive distribution $F_{n+1}(y_{n+1}) = p(y_{n+1} \mid \boldsymbol{y})$. With an argument similar to (6.12) for $i = n+1$, but without conditioning on $y_{n+1}$ and with instead an additional convolution with $p(y_{n+1} \mid z_{n+1} = k, \boldsymbol{y})$, we find

$$F_{n+1}(y_{n+1}) \equiv p(y_{n+1} \mid \boldsymbol{y}) \propto \sum_{k=1}^K n_k p(y_{n+1} \mid \boldsymbol{y}_k^\star) + \alpha h_0(y_{n+1}). \tag{6.18}$$

Let $F(y) = \int p(y \mid \theta) \, dG(\theta)$, as earlier and let $\overline{F} = E(F \mid \boldsymbol{y})$ denote the posterior expectation. Then $F_{n+1} = \overline{F}$. That is, the posterior predictive (6.18) coincides with the posterior expectation on the random probability measure. This is easily seen by considering $p(y_{n+1} \leq c \mid \boldsymbol{y}) = E\{p(y_{n+1} \leq c \mid F, \boldsymbol{y}) \mid \boldsymbol{y}\} = E\{F(c) \mid \boldsymbol{y}\}$. Here we overload notation to let $F(c)$ indicate the c.d.f. under the probabilty measure $F$.

In the outlined construction the nature of $F_{n+1}$ as a mixture model arises from the implied random partition $p(\rho_n)$ under i.i.d. sampling $\theta_i \sim G$ from the discrete random probability measure $G = \sum \eta_h \delta_{\mu_h}$. In that case the mixture model $F_{n+1}$ is just another manifestation of the assumed mixture model $F(y) = \sum \eta_h p(y \mid \theta_h)$, and does not introduce fundamentally new structure. In fact, any exchangeable random partition $p(\rho_n)$ can be argued to arise from such a construction. See, for example, Lee et al. (2013b) for a review. The attraction of exchangeable random partitions is coherence and mathematical tractability. However, if the inference goal is a posterior predictive in the form of a mixture model, as in (6.18), then the same form can be achieved with any underlying random partition model $p(\rho_n)$, including possibly non-exchangeable random partitions.

**Locally weighted mixtures.** An attractive general framework for random partitions are the product partition models (PPMs; Hartigan, 1990) which take the form

$$p(\rho_n = \{S_1, \ldots, S_K\}) \propto \prod_{j=1}^K c(S_j). \tag{6.19}$$

for some functions $c(S_j)$, which are known as the *cohesion* functions. The cohesion functions are restricted to be nonnegative functions of $S_j$, but in principle, any such function is valid. If $c(S)$ is only a function of the size of $S$, then the resulting model for $\rho_n$ is invariant under permutations of the indices. If additionally $p(\rho_n) = \sum_{z_{n+1}} p(\rho_{n+1})$, then we are back to exchangeable partitions. For example, with $c(S) = M \times (|S| - 1)!$ the PPM reduces to the Polya urn (6.6). More general, it can be shown that the family of all PPM models with cohesion function $c(S) = c(|S|)$ that depend on $S$ only indirectly through the cardinality and that define exchangeable random partitions coincides with the family of so-called Gibbs-type