Supplementary material: Mixture models with a prior on the number of components

Jeffrey W. Miller
Department of Biostatistics, Harvard University
and
Matthew T. Harrison
Division of Applied Mathematics, Brown University

September 20, 2016

S1 Proofs

Proof of Theorem 3.1. Note that Equation S1 below is well-known (see, e.g., Green and Richardson (2001) or McCullagh and Yang (2008)); we derive it here for completeness. Letting $E_i = \{j : z_j = i\}$, and writing $\mathcal{C}(z)$ for the partition induced by $z = (z_1, \ldots, z_n)$, by Dirichlet-multinomial conjugacy we have

$$p(z|k) = \int p(z|\pi)p(\pi|k)d\pi = \frac{\Gamma(k\gamma)}{\Gamma(\gamma)^k} \frac{\prod_{i=1}^k \Gamma(|E_i| + \gamma)}{\Gamma(n + k\gamma)} = \frac{1}{(k\gamma)^{(n)}} \prod_{c \in \mathcal{C}(z)} \gamma^{(|c|)},$$

for $z \in [k]^n$, provided that $p_K(k) > 0$. Recall that $x^{(m)} = x(x+1)\cdots(x+m-1)$ and $x_{(m)} = x(x-1)\cdots(x-m+1)$, with $x^{(0)} = 1$ and $x_{(0)} = 1$ by convention; note that $x_{(m)} = 0$ when x is a nonnegative integer less than m. It follows that for any partition $\mathcal{C}$ of $[n]$,

$$\begin{aligned} p(\mathcal{C}|k) &= \sum_{z \in [k]^n : \mathcal{C}(z) = \mathcal{C}} p(z|k) \\ &= \#\Big\{z \in [k]^n : \mathcal{C}(z) = \mathcal{C}\Big\} \frac{1}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)} \\ &= \frac{k_{(t)}}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)}, \end{aligned} \tag{S1}$$

where $t = |\mathcal{C}|$, since $\#\{z \in [k]^n : \mathcal{C}(z) = \mathcal{C}\} = \binom{k}{t} t! = k_{(t)}$. Finally,

$$p(\mathcal{C}) = \sum_{k=1}^{\infty} p(\mathcal{C}|k) p_K(k) = \Big(\prod_{c \in \mathcal{C}} \gamma^{(|c|)}\Big) \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(n)}} p_K(k) = V_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)},$$

with $V_n(t)$ as in Equation 3.2. □

Proof of Equation 3.4. Theorem 3.1 shows that the distribution of $\mathcal{C}$ is as shown. Next, note that instead of sampling only $\theta_1, \ldots, \theta_k \overset{\text{iid}}{\sim} H$ given $K = k$, we could simply sample $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} H$ independently of K, and the distribution of $X_{1:n}$ would be the same. Now, $Z_{1:n}$ determines which subset of the i.i.d. variables $\theta_1, \theta_2, \ldots$ will actually be used, and the indices of this subset are independent of $\theta_1, \theta_2, \ldots$; hence, denoting these random indices $I_1 < \cdots < I_T$, we have that $\theta_{I_1}, \ldots, \theta_{I_T} | Z_{1:n}$ are i.i.d. from H. For $c \in \mathcal{C}$, let $\phi_c = \theta_{I_i}$ where i is such that $c = \{j : z_j = I_i\}$. This completes the proof. □

Proof of the properties in Section 3.1. Abbreviate $x = x_{1:n}$, $z = z_{1:n}$, and $\theta = \theta_{1:k}$, and assume $p(z, k) > 0$. Letting $E_i = \{j : z_j = i\}$, we have $p(x|\theta, z, k) = \prod_{i=1}^{k} \prod_{j \in E_i} f_{\theta_i}(x_j)$ and

$$\begin{aligned} p(x|z,k) &= \int_{\Theta^k} p(x|\theta, z, k) p(d\theta|k) = \prod_{i=1}^{k} \int_{\Theta} \Big[\prod_{j \in E_i} f_{\theta_i}(x_j)\Big] H(d\theta_i) \\ &= \prod_{i=1}^{k} m(x_{E_i}) = \prod_{c \in \mathcal{C}(z)} m(x_c). \end{aligned}$$

Since this last expression depends only on z, k through $\mathcal{C} = \mathcal{C}(z)$, we have $p(x|\mathcal{C}) = \prod_{c \in \mathcal{C}} m(x_c)$, establishing Equation 3.5. Next, recall that $p(\mathcal{C}|k) = \frac{k_{(t)}}{(\gamma k)^{(n)}} \prod_{c \in \mathcal{C}} \gamma^{(|c|)}$ (where $t = |\mathcal{C}|$) from Equation S1, and thus

$$p(t|k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(\mathcal{C}|k) = \frac{k_{(t)}}{(\gamma k)^{(n)}} \sum_{\mathcal{C}:|\mathcal{C}|=t} \prod_{c \in \mathcal{C}} \gamma^{(|c|)},$$

(where the sum is over partitions $\mathcal{C}$ of $[n]$ such that $|\mathcal{C}| = t$) establishing Equation 3.6. Equation 3.7 follows, since

$$p(k|t) \propto p(t|k) p(k) \propto \frac{k_{(t)}}{(\gamma k)^{(n)}} p_K(k),$$

(provided $p(t) > 0$) and the normalizing constant is precisely $V_n(t)$. To see that $\mathcal{C} \perp K \mid T$ (Equation 3.8), note that if $t = |\mathcal{C}|$ then

$$p(\mathcal{C}|t, k) = \frac{p(\mathcal{C}, t|k)}{p(t|k)} = \frac{p(\mathcal{C}|k)}{p(t|k)},$$

(provided $p(t,k) > 0$) and due to the form of $p(\mathcal{C}|k)$ and $p(t|k)$ just above, this quantity does not depend on k; hence, $p(\mathcal{C}|t,k) = p(\mathcal{C}|t)$. To see that $X \perp K \mid T$ (Equation 3.9), note that $X \perp K \mid \mathcal{C}$; using this in addition to $\mathcal{C} \perp K \mid T$, we have

$$p(x|t,k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(x|\mathcal{C},t,k)p(\mathcal{C}|t,k) = \sum_{\mathcal{C}:|\mathcal{C}|=t} p(x|\mathcal{C},t)p(\mathcal{C}|t) = p(x|t).$$

□

Proof of Theorem 4.1. Let $\mathcal{C}_\infty$ be the random partition of $\mathbb{Z}_{>0}$ as in Section 3.3, and for $n \in \{1,2,\dots\}$, let $\mathcal{C}_n$ be the partition of $[n]$ induced by $\mathcal{C}_\infty$. Then

$$p(\mathcal{C}_n|\mathcal{C}_{n-1},\dots,\mathcal{C}_1) = p(\mathcal{C}_n|\mathcal{C}_{n-1}) \propto q_n(\mathcal{C}_n)\, I(\mathcal{C}_n \setminus n = \mathcal{C}_{n-1}),$$

where $\mathcal{C}\setminus n$ denotes $\mathcal{C}$ with element n removed, and $I(\cdot)$ is the indicator function ($I(E) = 1$ if E is true, and $I(E) = 0$ otherwise). Recalling that $q_n(\mathcal{C}_n) = V_n(|\mathcal{C}_n|)\prod_{c\in\mathcal{C}_n}\gamma^{(|c|)}$ (Equation 3.1), we have, letting $t = |\mathcal{C}_{n-1}|$,

$$p(\mathcal{C}_n|\mathcal{C}_{n-1}) \propto \begin{cases} V_n(t+1)\gamma & \text{if } n \text{ is a singleton in } \mathcal{C}_n, \text{ i.e., } \{n\} \in \mathcal{C}_n \\ V_n(t)(\gamma + |c|) & \text{if } c \in \mathcal{C}_{n-1} \text{ and } c \cup \{n\} \in \mathcal{C}_n, \end{cases}$$

for $\mathcal{C}_n$ such that $\mathcal{C}_n \setminus n = \mathcal{C}_{n-1}$ (and $p(\mathcal{C}_n|\mathcal{C}_{n-1}) = 0$ otherwise). With probability 1, $q_{n-1}(\mathcal{C}_{n-1}) > 0$, thus $V_{n-1}(t) > 0$ and hence also $V_n(t) > 0$, so we can divide through by $V_n(t)$ to get the result. □

Proof of Theorem 4.2. Let $G \sim \mathcal{M}(p_K, \gamma, H)$ and let $\beta_1,\dots,\beta_n \overset{\text{iid}}{\sim} G$, given G. Then the joint distribution of $(\beta_1,\dots,\beta_n)$ (with G marginalized out) is the same as $(\theta_{Z_1},\dots,\theta_{Z_n})$ in the original model (Equation 2.1). Let $\mathcal{C}_n$ denote the partition induced by $Z_1,\dots,Z_n$ as usual, and for $c \in \mathcal{C}_n$, define $\phi_c = \theta_I$ where I is such that $c = \{j : Z_j = I\}$; then, as in the proof of Equation 3.4, $(\phi_c : c \in \mathcal{C}_n)$ are i.i.d. from H, given $\mathcal{C}_n$.

Therefore, we have the following equivalent construction for $(\beta_1,\dots,\beta_n)$:

$$\begin{aligned} &\mathcal{C}_n \sim q_n, \text{ with } q_n \text{ as in Section 3.3} \\ &\phi_c \overset{\text{iid}}{\sim} H \text{ for } c \in \mathcal{C}_n, \text{ given } \mathcal{C}_n \\ &\beta_j = \phi_c \text{ for } j \in c,\ c \in \mathcal{C}_n, \text{ given } \mathcal{C}_n, \phi. \end{aligned}$$

Due to the self-consistency property of $q_1, q_2, \ldots$ (Proposition 3.3), we can sample $\mathcal{C}_n, (\phi_c : c \in \mathcal{C}_n), \beta_{1:n}$ sequentially for $n = 1, 2, \ldots$ by sampling from the restaurant process for $\mathcal{C}_n | \mathcal{C}_{n-1}$, sampling $\phi_{\{n\}}$ from H if n is placed in a cluster by itself (or setting $\phi_{c \cup \{n\}} = \phi_c$ if n is added to $c \in \mathcal{C}_{n-1}$), and setting β_n accordingly.

In particular, if the base measure H is continuous, then the ϕ's are distinct with probability 1, so conditioning on $\beta_{1:n-1}$ is the same as conditioning on $\mathcal{C}_{n-1}, (\phi_c : c \in \mathcal{C}_{n-1}), \beta_{1:n-1}$, and hence we can sample $\beta_n | \beta_{1:n-1}$ in the same way as was just described. In view of the form of the restaurant process (Theorem 4.1), the result follows. □

We use the following elementary result in the proof of Theorem 5.1; it is a special case of the dominated convergence theorem.

Proposition S1.1. *For $j = 1, 2, \ldots,$ let $a_{1j} \geq a_{2j} \geq \cdots \geq 0$ such that $a_{ij} \to 0$ as $i \to \infty$. If $\sum_{j=1}^{\infty} a_{1j} < \infty$ then $\sum_{j=1}^{\infty} a_{ij} \to 0$ as $i \to \infty$.*

Proof of Theorem 5.1. For any $x > 0$, writing $x^{(n)}/n! = \Gamma(x+n)/(n!\,\Gamma(x))$ and using Stirling's approximation, we have

$$\frac{x^{(n)}}{n!} \sim \frac{n^{x-1}}{\Gamma(x)}$$

as $n \to \infty$. Therefore, the $k = t$ term of $V_n(t)$ (Equation 3.2) is

$$\frac{t_{(t)}}{(\gamma t)^{(n)}}\, p_K(t) \sim \frac{t!}{n!}\, \frac{\Gamma(\gamma t)}{n^{\gamma t - 1}}\, p_K(t).$$

The first $t-1$ terms of $V_n(t)$ are 0, so to prove the result, we need to show that the rest of the series, divided by the $k = t$ term, goes to 0. (Recall that we have assumed $p_K(t) > 0$.) To this end, let

$$b_{nk} = (\gamma t)^{(n)} \frac{k_{(t)}}{(\gamma k)^{(n)}}\, p_K(k).$$

We must show that $\sum_{k=t+1}^{\infty} b_{nk} \to 0$ as $n \to \infty$. We apply Proposition S1.1 with $a_{ij} = b_{t+i,t+j}$. For any $k > t$, $b_{1k} \geq b_{2k} \geq \cdots \geq 0$. Further, for any $k > t$,

$$\frac{(\gamma t)^{(n)}}{(\gamma k)^{(n)}} \sim \frac{n^{\gamma t - 1}}{\Gamma(\gamma t)}\, \frac{\Gamma(\gamma k)}{n^{\gamma k - 1}} \longrightarrow 0$$

as $n \to \infty$, hence, $b_{nk} \to 0$ as $n \to \infty$ (for any $k > t$). Finally, observe that $\sum_{k=t+1}^{\infty} b_{nk} \leq (\gamma t)^{(n)} V_n(t) < \infty$ for any $n \geq t$. Therefore, by Proposition S1.1, $\sum_{k=t+1}^{\infty} b_{nk} \to 0$ as $n \to \infty$. This proves the result. □

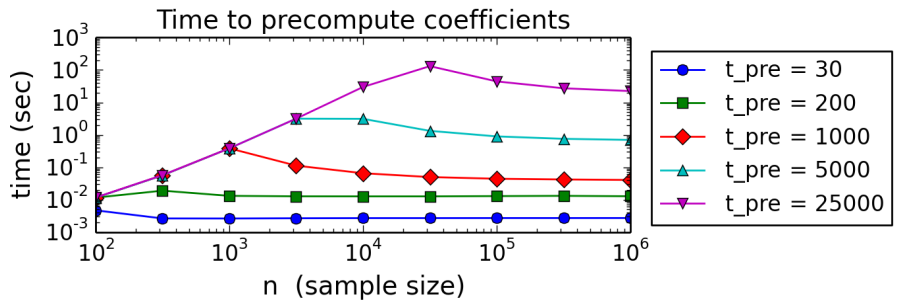

Figure S1: Amount of time required to precompute the MFM coefficients $V_n(1), \ldots, V_n(t_{\text{pre}})$ for various values of t_{pre}, for increasing n.

Proof of Theorem 5.2. For any $t \in \{1, \ldots, k\}$,

$$p_n(K = t \mid T = t) = \frac{1}{V_n(t)} \frac{t_{(t)}}{(\gamma t)^{(n)}} \, p_K(t) \longrightarrow 1 \tag{S2}$$

as $n \to \infty$ (where p_n denotes the MFM distribution with n samples), by Equation 3.7 and Theorem 5.1. For any $n \geq k$,

$$p(K = k \mid x_{1:n}) = \sum_{t=1}^{k} p(K = k \mid T = t, x_{1:n}) \, p(T = t \mid x_{1:n}),$$

and note that by Equations 3.9 and S2, $p(K = k \mid T = t, x_{1:n}) = p_n(K = k \mid T = t) \longrightarrow I(k = t)$ for $t \leq k$. The result follows. □

S2 Precomputation time for the MFM coefficients

In all of the empirical demonstrations in this paper, the largest value of t visited by the sampler was less than 30. Thus, in each case it was sufficient to precompute $V_n(t)$ for $t = 1, \ldots, 30$, and reuse these values throughout MCMC sampling.

To see how long this precomputation would take if the sample size n and/or the number of clusters were much larger, Figure S1 shows the amount of time required to compute $V_n(t)$ for $t = 1, \ldots, t_{\text{pre}}$, for each $t_{\text{pre}} \in \{30, 200, 1000, 5000, 25000\}$, for increasing values of n, when $K \sim \text{Geometric}(0.1)$ and $\gamma = 1$. For $t_{\text{pre}} = 30$ it only takes around 0.001 seconds for any n. For much larger values of t_{pre} it takes longer, but the time required relative to MCMC sampling would still be negligible. The reason why the computation time decreases

as n grows past t_{pre} is that, as discussed in Section 3.2, the infinite series for $V_n(t)$ (Equation 3.2) converges more rapidly when n is much bigger than t.

S3 Formulas for some posterior quantities

Posterior on the number of components k

The posterior on $t = |\mathcal{C}|$ is easily estimated from posterior samples of $\mathcal{C}$. To compute the MFM posterior on k, note that

$$p(k|x_{1:n}) = \sum_{t=1}^{\infty} p(k|t, x_{1:n})p(t|x_{1:n}) = \sum_{t=1}^{n} p(k|t)p(t|x_{1:n}),$$

by Equation 3.9 and the fact that t cannot exceed n. Using this and the formula for $p(k|t)$ given by Equation 3.7, it is simple to transform our estimate of the posterior on t into an estimate of the posterior on k. For the DPM, the posterior on the number of components k is always trivially a point mass at infinity.

Density estimates

Using the restaurant process (Theorem 4.1), it is straightforward to show that if $\mathcal{C}$ is a partition of $[n]$ and $\phi = (\phi_c : c \in \mathcal{C})$ then

$$p(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) \propto \frac{V_{n+1}(t+1)}{V_{n+1}(t)} \gamma\, m(x_{n+1}) + \sum_{c \in \mathcal{C}} (|c| + \gamma) f_{\phi_c}(x_{n+1}) \tag{S1}$$

where $t = |\mathcal{C}|$, and, using the recursion for $V_n(t)$ (Equation 3.10), this is normalized when multiplied by $V_{n+1}(t)/V_n(t)$. Further,

$$p(x_{n+1} \mid \mathcal{C}, x_{1:n}) \propto \frac{V_{n+1}(t+1)}{V_{n+1}(t)} \gamma\, m(x_{n+1}) + \sum_{c \in \mathcal{C}} (|c| + \gamma) \frac{m(x_{c \cup \{n+1\}})}{m(x_c)}, \tag{S2}$$

with the same normalization constant. Therefore, when $m(x_c)$ can be easily computed, Equation S2 can be used to estimate the posterior predictive density $p(x_{n+1}|x_{1:n})$ based on samples from $\mathcal{C} \mid x_{1:n}$. When $m(x_c)$ cannot be easily computed, Equation S1 can be used to estimate $p(x_{n+1}|x_{1:n})$ based on samples from $\mathcal{C}, \phi \mid x_{1:n}$, along with samples $\theta_1, \ldots, \theta_N \overset{\text{iid}}{\sim} H$ to approximate $m(x_{n+1}) \approx \frac{1}{N} \sum_{i=1}^{N} f_{\theta_i}(x_{n+1})$.

The posterior predictive density is, perhaps, the most natural estimate of the density. However, following Green and Richardson (2001), a simpler way to obtain a natural estimate is by assuming that element $n+1$ is added to an existing cluster; this will be very similar to the posterior predictive density when n is sufficiently large. To this end, we define $p_*(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) = p(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}, |\mathcal{C}_{n+1}| = |\mathcal{C}|)$, where $\mathcal{C}_{n+1}$ is the partition of $[n+1]$, and observe that

$$p_*(x_{n+1} \mid \mathcal{C}, \phi, x_{1:n}) = \sum_{c \in \mathcal{C}} \frac{|c| + \gamma}{n + \gamma t} f_{\phi_c}(x_{n+1})$$

where $t = |\mathcal{C}|$ (Green and Richardson, 2001). Using this, we can estimate the density by

$$\frac{1}{N} \sum_{i=1}^{N} p_*(x_{n+1} \mid \mathcal{C}^{(i)}, \phi^{(i)}, x_{1:n}), \tag{S3}$$

where $(\mathcal{C}^{(1)}, \phi^{(1)}), \ldots, (\mathcal{C}^{(N)}, \phi^{(N)})$ are samples from $\mathcal{C}, \phi \mid x_{1:n}$. The corresponding expressions for the DPM are all very similar, using its restaurant process instead. The density estimates shown in this paper are obtained using this approach.

These formulas are conditional on additional parameters such as γ for the MFM, and α for the DPM. If priors are placed on such parameters and they are sampled along with $\mathcal{C}$ and ϕ given $x_{1:n}$, then the posterior predictive density can be estimated using the same formulas as above, but also using the posterior samples of these additional parameters.

S4 Small components

In Section 7.3, we noted that the DPM tends to introduce one or two tiny extra components, and it is natural to wonder whether the DPM would fare better if the data were actually drawn from a mixture with an additional one or two small components. To see, we modify the data distribution from Section 7.3 to be a four-component mixture in which w_1 is reduced from 0.45 to 0.44, and the fourth component has weight $w_4 = 0.01$, mean $\mu_4 = \left(\begin{smallmatrix} 8 \\ 11 \end{smallmatrix}\right)$, and covariance $C_4 = \left(\begin{smallmatrix} 0.1 & 0 \\ 0 & 0.1 \end{smallmatrix}\right)$. We use exactly the same model and inference parameters as in Section 7.3.

Figure S2 shows that the MFM still more accurately infers the number of clusters than the DPM. We expect that in order to have a situation where the DPM performs

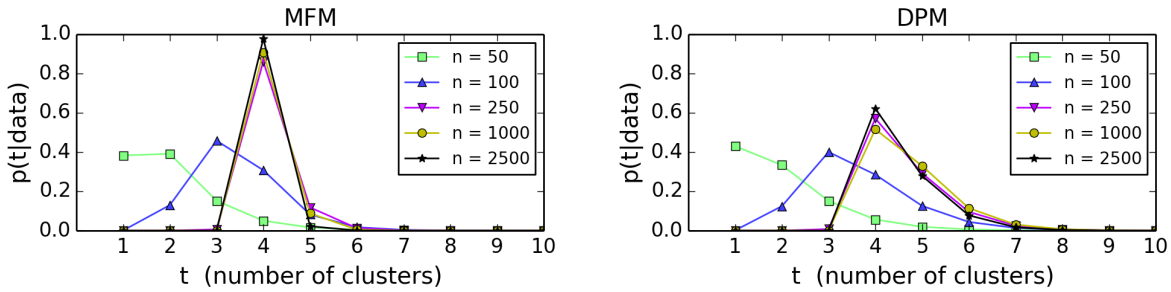

Figure S2: Posterior on the number of clusters t for the MFM and DPM on data from the modified bivariate example with a small fourth component.

more favorably in terms of clustering and inferring the number of clusters, the number of components would have to be sufficiently large relative to the sample size, or infinite.

References

P. J. Green and S. Richardson. Modeling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, June 2001.

P. McCullagh and J. Yang. How many clusters? *Bayesian Analysis*, 3(1):101–120, 2008.