Supplemental Materials
for Bayesian Hierarchical Cross-Clustering

Dazhuo Li
Dept. of Comp. Eng. and Comp. Sci.
University of Louisville

Patrick Shafto
Dept. of Psychological and Brain Sciences
University of Louisville

1 Inductive proofs of predictive distributions

We begin with some preliminary definitions, then proceed to prove the consistency of the predictive distributions.

Definition 1.1. Define Nodes(T) as a function returning the identification numbers of all the nodes in tree T, i.e., if $T = (c, T_a, T_b, r_c)$, then

$$\text{Nodes}(T) = \begin{cases} \emptyset \text{ if } T = \emptyset, \\ \{c\} \cup \text{Nodes}(T_a) \cup \text{Nodes}(T_b) \text{ else.} \end{cases}$$

Definition 1.2. Define Parent(T, k) as a function returning identification number of the immediate parent of node k in a BHCC tree T, i.e.,

$$\text{Parent}(T, k) = \begin{cases} \emptyset \text{ if } k \text{ is root of } T, \\ \{c | T_c = (c, T_k, T_b, r_c) \vee T_c = (c, T_a, T_k, r_c)\} \text{ else.} \end{cases}$$

Definition 1.3. Define Path(T, k) as a function returning identification numbers of all the nodes in the path from the root to node k in a BHCC tree T, i.e.,

$$\text{Path}(T, k) = \begin{cases} \emptyset \text{ if } \{k\} \cap \text{Nodes}(T) = \emptyset, \\ \{k\} \cup \text{Path}(T, \text{Parent}(T, k)) \text{ else.} \end{cases}$$

Definition 1.4. Define $\omega(T, k)$ as the posterior probability that $\mathcal{L}(T_k)$ forms a view and merging other subtrees to T_k does not yield a view:

$$\omega(T, k) = r_k \prod_{c \in \text{Path}(T,k) - \{k\}} (1 - r_c).$$

Lemma 1.5. *The quantity defined in Equation 10 in the paper has a lower bound:*

$$\sum_{\boldsymbol{u} \in \text{Ptns}(T_c)} \prod_{v=1}^{\max(\boldsymbol{u})} \omega(T_c, k) p(\boldsymbol{x}_{\boldsymbol{u}=v} | X_{\cdot, \boldsymbol{u}=v}, \text{DPM}), \quad (1)$$

where the partitioning $\boldsymbol{u}$ *is represented in the form of a vector of indicator variables,* $\max(\boldsymbol{u})$ *is the number of clusters in partitioning* $\boldsymbol{u}$*, and* k *is the node in* T_c *such that the two vectors* $\boldsymbol{u} = v$ *and* $\mathcal{L}(T_k)$ *have the same set of dimensions.*

Proof. We show a proof by induction. If c is the leaf node, Ptns$(T_c) = (c)$, $\max(\boldsymbol{u}) = 1$, $k = c$ and $\omega(T_c, c) = r_c = 1$, thus Equation 1 becomes $p(\boldsymbol{x} | X_{\cdot, \mathcal{L}(T_c)}, \text{DPM})$ which is equal to the quantity in Equation 10 in the paper. Thus the lemma is true in the base case.

Our inductive hypothesis is that the lemma holds for the two subtrees T_a and T_b. That is,

$$p(\boldsymbol{x}_{\mathcal{L}(T_a)} | X_{\mathcal{L}(T_a)}, T_a) \geq \sum_{\boldsymbol{u}' \in \text{Ptns}(T_a)} \prod_{v'=1}^{\max(\boldsymbol{u}')} \omega(T_a, k') p(\boldsymbol{x}_{\boldsymbol{u}'=v'} | X_{\cdot, \boldsymbol{u}'=v'}, \text{DPM})$$

and same for $p(\boldsymbol{x}_{\mathcal{L}(T_b)} | X_{\mathcal{L}(T_b)}, T_b)$; Also note that

$$\omega(T_a, k') \geq (1 - r_c) \omega(T_a, k') = \omega(T_c, k'),$$

and same for $\omega(T_b, k'')$. Therefore

$$\begin{aligned} &(1 - r_c) p(\boldsymbol{x}_{\mathcal{L}(T_a)} | X_{\mathcal{L}(T_a)}, T_a) p(\boldsymbol{x}_{\mathcal{L}(T_b)} | X_{\mathcal{L}(T_b)}, T_b) \geq \\ &\sum_{\boldsymbol{u}' \in \text{Ptns}(T_a)} \prod_{v'=1}^{\max(\boldsymbol{u}')} \omega(T_c, k') p(\boldsymbol{x}_{\boldsymbol{u}'=v'} | X_{\cdot, \boldsymbol{u}'=v'}, \text{DPM}) \times \\ &\sum_{\boldsymbol{u}'' \in \text{Ptns}(T_b)} \prod_{v''=1}^{\max(\boldsymbol{u}'')} \omega(T_c, k'') p(\boldsymbol{x}_{\boldsymbol{u}''=v''} | X_{\cdot, \boldsymbol{u}''=v''}, \text{DPM}) \\ &= \sum_{\boldsymbol{u} \in \text{Ptns}(T_a) \times \text{Ptns}(T_b)} \prod_{v=1}^{\max(\boldsymbol{u})} \omega(T_c, k) p(\boldsymbol{x}_{\boldsymbol{u}=v} | X_{\cdot, \boldsymbol{u}=v}, \text{DPM}) \end{aligned} \quad (2)$$

Meanwhile, for the trivial partitioning $(\mathcal{L}(T_c))$ (recalling that $(\mathcal{L}(T_c))$ represents the partitioning where all dimensions in $\mathcal{L}(T_c)$ are assigned to the same cluster), we have $\max(\boldsymbol{u}) = 1$, $k = c$ and $\omega(T_c, c) = r_c$. Thus for $\boldsymbol{u} = (\mathcal{L}(T_c))$

$$\begin{aligned} &\prod_{v=1}^{\max(\boldsymbol{u})} \omega(T_c, k) p(\boldsymbol{x}_{\boldsymbol{u}=v} | X_{\cdot, \boldsymbol{u}=v}, \text{DPM}) \\ &= r_c p(\boldsymbol{x}_{\mathcal{L}(T_c)} | X_{\cdot, \mathcal{L}(T_c)}, \text{DPM}) \end{aligned} \quad (3)$$

By definition, $\mathrm{Ptns}(T_c) = (\mathcal{L}(T_c) \cup \mathrm{Ptns}(T_a) \times \mathrm{Ptns}(T_b)$. Therefore combining the results from Equation 2 and 3, we see the lemma is true. □

Lemma 1.6. *The quantity defined in Equation 11 in the paper is equal to the quantity:*

$$\sum_{k \in \mathrm{Nodes}(T_c)} \omega(T_c, k) p(\boldsymbol{y} | X_{\cdot, \mathcal{L}(T_k)}, \mathrm{DPM}) \tag{4}$$

which sums over the prediction w.r.t. all the nodes in T_c weighted by the posterior of the nodes.

Proof. We show a proof by induction. If c is a leaf node, then $T_a = T_b = \emptyset$ and $r_c = 1$. By definition, $p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_c)}, T_c) = p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_c)}, \mathrm{DPM})$; Meanwhile, $\mathrm{Nodes}(T_c) = \{c\}$, $\omega(T_c, c) = r_c = 1$, $\mathcal{L}(T_c) = \{c\}$, thus Equation 4 becomes $p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_c)}, \mathrm{DPM})$. Thus the lemma is true in the base case.

Our inductive hypothesis is that Equation 4 holds for the two subtrees T_a and T_b, i.e.,

$$p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_a)}, T_a) = \sum_{k \in \mathrm{Nodes}(T_a)} \omega(T_a, k) p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_k)}, \mathrm{DPM})$$

and same for T_b. Meanwhile, by definition, $(1 - r_c)\omega(T_a, k) = \omega(T_c, k)$ and same for T_b; and $r_c = \omega(T_c, c)$. Therefore,

$$\begin{aligned} p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_c)}, T_c) &= \omega(T_c, c) p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_c)}, \mathrm{DPM}) + \\ &\sum_{k \in \mathrm{Nodes}(T_a)} \omega(T_c, k) p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_k)}, \mathrm{DPM}) + \\ &\sum_{k \in \mathrm{Nodes}(T_b)} \omega(T_c, k) p(\boldsymbol{y}|X_{\cdot,\mathcal{L}(T_k)}, \mathrm{DPM}) \end{aligned}$$

Further notice $\mathrm{Nodes}(T_c) = \{c\} \cup \mathrm{Nodes}(T_a) \cup \mathrm{Nodes}(T_b)$, thus the lemma is true. □

Lemma 1.7. *The quantity defined in Equation 13 in the paper is equal to the quantity:*

$$\sum_{k \in \mathrm{Path}(T_c, j)} \omega(T_c, k) p(X_{i,j} | X_{\cdot, \mathcal{L}(T_k)}, \mathrm{DPM}) \tag{5}$$

which sums over the prediction w.r.t. the nodes on the path from root to dimension j, weighted by the posterior of the nodes.

Proof. We show a proof by induction. If c is a leaf node, $\mathrm{Path}(T_c, j) = \{c\} \cap \{j\}$, $\omega(T_c, c) = r_c = 1$, $\mathcal{L}(T_c) = \{c\}$, thus Equation 5 becomes $p(X_{i,j}|X_{\cdot,\mathcal{L}(T_c)}, \mathrm{DPM}) \times |\{c\} \cap \{j\}|$ which is also the case in Equation 13 in the paper ($T_a = T_b = \emptyset$ and $r_c = 1$). Thus the lemma is true in the base case.

Our inductive hypothesis is that Equation 5 holds for the two subtrees T_a and T_b, i.e.,

$$p(X_{i,j}|X_{\cdot,\mathcal{L}(T_a)}, T_a) = \sum_{k \in \mathrm{Path}(T_a, j)} \omega(T_a, k) p(X_{i,j}|X_{\cdot,\mathcal{L}(T_k)}, \mathrm{DPM})$$

and same for T_b. Meanwhile, by definition, $(1 - r_c)\omega(T_a, k) = \omega(T_c, k)$ and same for T_b; and $r_c = \omega(T_c, c)$. Therefore,

$$\begin{aligned} &p(X_{i,j}|X_{\cdot,\mathcal{L}(T_c)}, T_c) = \\ &\omega(T_c, c) p(X_{i,j}|X_{\cdot,\mathcal{L}(T_c)}, \mathrm{DPM}) \times |\{j\} \cap \mathcal{L}(T_c)| + \\ &\sum_{k \in \mathrm{Path}(T_a, j)} \omega(T_c, k) p(X_{i,j}|X_{\cdot,\mathcal{L}(T_k)}, \mathrm{DPM}) + \\ &\sum_{k \in \mathrm{Path}(T_b, j)} \omega(T_c, k) p(X_{i,j}|X_{\cdot,\mathcal{L}(T_k)}, \mathrm{DPM}) \end{aligned}$$

Assume j is a leaf node of subtree T_a, then $\mathrm{Path}(T_b, j) = \emptyset$, $\mathrm{Path}(T_c, j) = \{c\} \cup \mathrm{Path}(T_a, j)$ and $|\{j\} \cap \mathcal{L}(T_c)| = 1$, thus the lemma is true. □