1 Is Equation 9 correct?

Recall Equation 9 in Heller and Ghahramani (2005)[1] is

$$p(\mathbf{x} \mid \mathcal{D}) = \sum_{k \in \mathcal{N}} \omega_k p(\mathbf{x} \mid \mathcal{D}_k)$$

where $\mathcal{N}$ is the set of all nodes in the tree, $\omega_k \stackrel{\text{def}}{=} r_k \prod_{i \in \mathcal{N}_k} (1 - r_i)$ is the weight on cluster k, and $\mathcal{N}_k$ is the set of the nodes on the path from the root node to the parent node k.

We're beginning to suspect this is wrong, and that ω_k is not correctly defined. To sample a point from the posterior predictive distribution, since BHC is a mixture model where each node is a component, there must be some process of sampling a node of BHC before sampling from that node. This would result in the weighted sum in Equation 9 above and is implied by both ω_k and one of Heller's presentations.[2]

Specifically, the paper describes this process as "recursing through the tree starting at the root node." So if r_k is our probability of stopping at node k during this recursion, ω_k represents the probability of ending up at node/component k if we do *not* stop at any of the nodes along the path from the root of the tree to node k's parent $(1 - r_i)$, and we *do* stop at node k (r_k).

However, ω_k doesn't seem to include any notion of choosing between left and right children. It's calculated based on probabilities assuming a fixed path from the root to the node, $\mathcal{N}_k$. Thus, we think ω_k is an overestimate of the true value, and the implementation confirms this: the sum of ω_k as defined in Equation 9 is consistently > 1 for BHC trees, while ω_k *must* sum to 1 (see Section 3).

To be accurate, we think ω_k needs to include the probability of recursing from the root node to node k and *not* descending into the other subtrees along the path.

2 Proposed fix

Stuart Sale's pyBHC package[3] doesn't include functionality for evaluating the likelihood of a new point according to the posterior predictive distribution, but *does* have functionality for sampling from the posterior predictive (Heller's BHC code[4] doesn't). In his procedure, he starts at the root node and chooses between the left and right children with probability proportional to the ratio of the sizes of the left and right subtrees. Specifically, if n_k is the number of leaves under node k, at node k he descends into the left child with probability $n_{\text{left}}/(n_k)$ and the right child with 1 minus that probability. So in this case,

[1] `http://www2.stat.duke.edu/~kheller/bhcnew.pdf`
[2] Last slide of `http://www.gatsby.ucl.ac.uk/~heller/present/bhc/Ranbhc.ppt`
[3] `https://github.com/stuartsale/pyBHC/blob/master/pyBHC/bhc.py#L181-L187`
[4] `http://www.gatsby.ucl.ac.uk/~heller/code/bhc/`

$$\omega_k \stackrel{\text{def}}{=} r_k \prod_{i \in \mathcal{N}_k} (1 - r_i) R_{\mathcal{N}_k}(i)$$

where the existing symbols are defined as before, and

$$R_{\mathcal{N}_k}(i) = \begin{cases} n_{\text{Left}(i)}/n_i & \text{if Left}(i) \in \mathcal{N}_k \text{ or Left}(i) = k \\ n_{\text{Right}(i)}/n_i & \text{otherwise} \end{cases},$$

i.e. $R_{\mathcal{N}_k}(i)$ is the ratio of the sizes of the children of a node in $\mathcal{N}_k$, where the ratio is dependent on which child is also in the path $\mathcal{N}_k$.

To show this is true, we would ideally like to do the following three things:

1. verify that the posterior predictive distribution with ω_k computed as above makes sensible predictions,
2. derive this equation by "rearranging the sum over all tree-consistent partitionings into a sum over all clusters in the tree", and
3. verify with one of the people who originally worked with BHC (e.g. Yang Xu or Katherine Heller) that our intuitions are correct and that Equation 9 is wrong.

3 ω_k must sum to 1

If $p(x)$ is a valid probability distribution, then $\int p(x)\, dx = 1$. Therefore,

$$\begin{aligned} 1 &= \int p(\mathbf{x} \mid \mathcal{D})\, d\mathbf{x} \\ &= \int \sum_{k \in \mathcal{N}} \omega_k p(\mathbf{x} \mid \mathcal{D}_k)\, d\mathbf{x} \\ &= \sum_{k \in \mathcal{N}} \omega_k \int p(\mathbf{x} \mid \mathcal{D}_k)\, d\mathbf{x} \\ &= \sum_{k \in \mathcal{N}} \omega_k. \end{aligned}$$

So in BHC (and *any* mixture model), the weights must sum to 1.