Paper ID	615
Title	An Alternative Prior process for Nonparametric Bayesian Clustering

	Question	Response
1	Author(s)' feedback: Please focus your responses on factual clarifications. Text is limited to 4000 characters.	**Review 1:** **An analytic result or bounds for eq 9 would be a nice addition -- we consider it an important direction of our future research in this area.** **Review 2:** **The UP formulation is given on p438 (2nd para., 2nd col.) of [12]. For eq 4, a recent reference is eq 10 of "Dirichlet Processes" by Teh (Encyclopedia of Machine Learning; submitted), though it also appears in "Nonparametric hierarchical Bayes via sequential imputations" by Liu (1996).** **To draw the d^th data point (x_d) from a UP mixture model:** **c_d | c_{< d} ~ eq 3** **n_c ~ G_0** **x_d ~ F(n_{c_d})** **where c_d is the cluster assignment for x_d, n_c are the params for cluster c, G_0 is the UP base measure and F is the data distribution.** **Eq 10 is conditioned on a particular ordering of the observations, so it does not assume exchangeability over orderings.** **Gibbs sampling for the UP has slightly greater computational complexity than for the DP, because calculating the conditional posterior of c_d given c_{\d} is more complicated (see eq 14). Also, for exchangeable data one must (approx.) marginalize over orderings of the data.** **It should be possible to derive a VB algorithm.** **Review 3:** **The formalism on p6-7 is not correct for the UP -- this was an oversight. We will correct it as follows: "The model assumes the following generative process: the tokens w_d that comprise each document, indexed by d, are drawn from a document-specific distribution over words phi_d, which is drawn from a document-specific Dirichlet distribution with base measure n_d and concentration parameter beta. The base measure is obtained using the uniform process by selecting a cluster assignment from the uniform process, using eq 3. If an existing cluster is selected then n_d is set to to the cluster-specific distribution over words for that cluster. If a new cluster is selected, then a new cluster-specific distribution over words is drawn from base measure G_0 and n_d is set to this distribution. Finally G_0 is chosen to be a hierarchical Dirichlet distribution.** **c_d | c_{< d} ~ equation 3**

n_c ~ G_0
phi_d ~ Dir(phi_d | n_{c_d}, beta)
w_d ~ Mult(phi_d)

where c_d is the cluster assignment for the d^th document."

Indeed, there is a large literature on species sampling models. Our work was informed by several such papers, including "Prediction rules for exchangeable sequences related to species sampling" by Hansen & Pitman (2000). For space reasons, we only included references that we felt were essential for our exposition.

Applied researchers in machine learning and stats are routinely forced to make assumptions when modeling real data. Even though the use of exchangeable priors can provide many practical advantages for clustering tasks, exchangeability itself is just one particular modeling assumption. In reality, most data generating processes are not actually exchangeable -- a good example would be news stories, which are published at different times and therefore have an associated temporal ordering. If one is willing to make an exchangeability assumption, than the DP prior is a natural choice, though it comes with additional assumptions about the size distribution of clusters. These assumptions will be reasonable in certain situations, but unreasonable in others. We have chosen to explore alternatives that remove the exchangeability assumption in order to make better prior assumptions about cluster sizes. We demonstrate empirically the benefits of this choice despite the additional complications that result from using a non-exchangeable prior. We do not feel that it is always necessary to restrict ourselves to exchangeable models, which can impose other poor assumptions, when alternatives do exist. The idea that exchangeable models must be used to model "exchangeable" data is not an uncommonly held viewpoint and serves to illustrate the importance of this paper -- exchangeability is a modeling choice that can sometimes have a high price in practice.