NIPS 2009
Neural Information Processing Systems (NIPS) Conference
December 7-10, 2009
Hyatt Regency, Vancouver, B.C.,
Canada

Reviews For Paper

Paper ID 615

Title An Alternative Prior process for Nonparametric Bayesian Clustering

---

Masked Reviewer ID: Reviewer 1

Review:

Question	
	This paper analyzes properties of the less-utilized uniform process (UN) as compared to the more common Dirichlet process (DP) and Pitman-Yor (PY) process, and then demonstrates that the uniform process provides a better description for the distribution over clusters present in document clustering applications. The UN avoids the "rich-get-richer" assumption of the DP/PY. Although the PY has a second parameter that allows for faster growth in the number of clusters with the size of the dataset relative to the DP, this process still assumes a general trend of fewer clusters with large numbers of observations and more clusters with small numbers of observations. The UP on the other hand, provides what the authors conjecture as an asymptotically uniform distribution on cluster size (demonstrated through a simulated study presented in this paper). The paper also provides an asymptotic result on the expected number of clusters generated.
Comments to author(s). First provide a summary of the paper, and then address the following criteria: Quality, clarity, originality and significance. (For detailed reviewing guidelines, see http://nips.cc /PaperInformation /ReviewerInstructions)	Overall, the paper is well-written and the results presented on a document clustering dataset clearly demonstrate the utility of considering a uniform process prior over a DP prior. Other comments: * It would obviously be nice to have an analytic result for Eq. (9), or at least bounds. Even loose bounds would be nice. A sentence describing why this is challenging for this process would help since analytic results were presented for the DP and PY. * Emphasize in Sec. 3.4 that it is the size of the clusters that are of concern. The authors allude to this end of the 2nd paragraph (still relying on the reader to make the jump that this is a good feature for the applications of interest), but it would help to highlight it more. In a first reading one thinks, "Okay, you can get similar growth rates in # of clusters, so who cares about the uniform process?" * Figure 1 and especially Figure 2 are hard to read (and are massive in file size!) The results from Figure 2 are key in substantiating the claim made in Eq. (9), so it would be nice to be able to read them more easily. Also, be clear what the black o's are. The text clearly states what the red lines are, and that the estimates are averaged over 1000 independent partitions, but the black o's are not clearly connected to this idea. * It took me an extremely long time to parse out what the terms "Between-Partition" vs. "Between-Ordering" referred to. The most confusing part is that the authors speak of computing the st.dev. of log P(C) for L different partitions c, but that sentence also says "to different orderings". The next

	sentence says for a "fixed ordering". Since "ordering" versus "partitioning" is what is being analyzed, it would be helpful to make this clear from the get-go. Maybe add a picture? * In both the intro and conclusion, the authors describe the "rich-get-richer" phenomenon as if it something that is always undesirable and often overlooked. However, in many applications, this property is what makes the DP and PY so useful and is an appropriate description of the data clustering. Perhaps just emphasis that, in some applications, this property is not desirable yet the DP and PY are still used. * In the supplemental material, C(\theta,\epsilon) is not defined. Also, I think the Borel-Cantelli lemma relies on a limiting assumption that is not present. Finally, it would be nice to mention the monotone convergence lemma in justifying the first step of Sec. 2 here. (Also, there's a typo of "n" instead of "N" in the displayed equation above Eq. (6).) There are other little things, too. Overall, a cleaner version of the supplemental material would make the derivations easier to follow.
Please summarize your review in 1-2 sentences	This paper analyzes asymptotic properties of the uniform process as compared to the more common Dirichlet and Pitman-Yor processes. The paper appears to be theoretically sound and provides a demonstration that the uniform process provides a better description for the distribution over clusters present in document clustering applications.
Numerical score	7: Good paper, accept
Confidence	4: Reviewer is confident but not absolutely certain

Masked Reviewer ID: Reviewer 2

Review:

Question	
Comments to author(s). First provide a summary of the paper, and then address the following criteria: Quality, clarity, originality and significance. (For detailed reviewing guidelines, see http://nips.cc/PaperInformation/ReviewerInstructions)	This paper proposes use of a relatively new class of nonparametric prior, addressing some issues associated with DP and PY priors ("relatively" new because it was proposed and used elsewhere). Properties of the uniform prior are discussed, including that it is not an exchangeable prior, and example results are shown for a toy example and for documents (a set of patents). The paper is generally well written, but I do not like the way the DP and PY processes were discussed. Specifically, it is stated that the goal is to place {X_1,\dots,X_N} into clusters. Then, it is stated that if X_{N+1}=\tilde{X}_k, then X_{N+1} goes into cluster k. Hence, data X go in the same cluster if they are *exactly* the same. This doesn't make sense. The data in general are never the same. However, the *parameters* of the model may be the same. Hence, the clustering should be performed on the parameters of some model f(\beta), with X_k\sim f(\beta_k), and \beta_k\sim G with G\sim DP(\theta G_0), where \theta is (your) scalar parameter from the DP, and G_0 is the base measure. In this way the model parameters \beta are the same when data come from the same cluster, but the data itself is not all the same within a given cluster. The same type of thing is true of the PY and the uniform process. When you do this, there needs to be a discussion of the base measure. Then, it becomes clear that the statement on p. 2 is wrong: "The Dirichlet process prior has a single parameter \theta." This is not true: it has a positive scalar \theta *and* a base measure G_0. The same is true of the PY process.

	Also, once one understands the relationship of the data to the model parameters, one recognizes that the model appears in the likelihood. The likelihood and the characteristics of the data will impact the number of clusters, somewhat mitigating the "rich-get-richer" criticism of DP/PY. If the data are very different, the likelihood will impose that they belong in different clusters, regardless of how many samples are already in a given cluster. The asymptotic behavior of the DP and PY processes is well known. While something is stated very briefly about the uniform process, there is no proof in the paper itself. This may have been the most interesting thing in the paper, but it is excluded from the formal paper. A claim is made that the lack of exchangeability of the proposed process is not important in practice, and one example is shown. It seems hard to justify this statement based upon a single example.
Please summarize your review in 1-2 sentences	A uniform process is proposed to replace/complement the DP/PY processes. There are some problems with the model presentation, and the statements about exchangeability are not sufficiently established based upon a single real example.
Numerical score	3: A clear rejection
Confidence	4: Reviewer is confident but not absolutely certain

Masked Reviewer ID: Reviewer 3

Review:

Question	
Comments to author(s). First provide a summary of the paper, and then address the following criteria: Quality, clarity, originality and significance. (For detailed reviewing guidelines, see http://nips.cc/PaperInformation/ReviewerInstructions)	The authors present new asymptotic properties for the clustering properties of the uniform process, and compare it with the Dirichlet and Pitman-Yor process. Simulations supporting the theoretical analysis are presented, and an application of document clustering is presented. Overall, I found this to be an interesting paper. The concepts are presented in a clear and intuitive way. I have the following comments: The authors state that the uniform process (UP) was originally introduced in [12], but upon checking I did not see any formulation of the UP in the above mentioned reference. Could the authors clarify this point? Could the authors give an example of a posterior computation for the UP? Also, are there conjugacy issues with the UP? How one draws from a UP? Could the authors provide a reference when stating (4)? I think the authors mean "UP" instead of "UN" in (8) and (9). In Section 5, the authors mention that the UP is not exchangeable over observations, but (10) clearly shows exchangeability. Could the authors clarify this point? How about computational complexity with inference using UP? Is it possible to obtain a variational Bayesian (VB) solution for the UP?
Please summarize your review in 1-2 sentences	An interesting, well balanced, and clearly articulated paper.
Numerical score	7: Good paper, accept
Confidence	5: Reviewer is absolutely certain

Masked Reviewer ID: Reviewer 4

Review:

Question	
Comments to author(s). First provide a summary of the paper, and then address the following criteria: Quality, clarity, originality and significance. (For detailed reviewing guidelines, see http://nips.cc/PaperInformation/ReviewerInstructions)	The paper attempts to describe a new nonparametric prior (the so called uniform process) for clustering. They do this by trying to define a species sampling model. I think that the goal is worthwhile (trying to find nonparametric models that a priori favor uniform cluster sizes). However, the paper describes a model that is not exchangeable, which is in my opinion a more serious flaw than the author(s) admit: 1) Lack of exchangeability means, among other things, that the joint distribution of all model parameters depends on how the observations are ordered. The authors argue that the results are robust such ordering. However the "robustness" is not the main issue: More importantly, the lack of exchangeability implies that (in the notation of Section 6) we cannot write Y_i \sim X_i X_i \sim G G \sim P(G|\theta, G_0) BECAUSE SUCH P(G|\theta, G_0) DOES NOT EXIST (if it did, the joint distribution for the X_i's would be exchangeable, that is De Finetti's theorem!!!). Therefore, the model description at the bottom of page 6 and top of page 7 is wrong. 2) I would be happy to see the model used in cases where there is a natural ordering for the data (i.e., a time series model), or if the observations, in spite of being exchangeable a priori, arrive sequentially. But I feel that the current application of the model is inappropriate. 3) More generally, the authors should frame the uniform process in the context of species sampling models, which include the Dirichlet process and the Poisson-Dirichlet process among others. There is extensive literature on species sampling models that the authors should check and reference, for example, exact formulas for the expected number of clusters (the size of the clusters) on species sampling models have been obtained in Lijoi, Prunster and Mena (2007), Biometrika.
Please summarize your review in 1-2 sentences	I would not accept the paper in its current form. I do not believe that a non-exhangeable model is appropriate for exchangeable data.
Numerical score	3: A clear rejection
Confidence	4: Reviewer is confident but not absolutely certain