Review of **Multiple Clusterings with Pouch Latent Tree Models**

*Executive Summary*

This paper is either a reject or an accept with major revision. It is possible that there is a nugget of something interesting in the paper, but the story is told in an unnecessarily and persistently confusing way. Additionally, there is a chance that the “results” in the paper derive from merely from having stipulated a model with more parameters, though this is difficult to tell from the paper itself. If this opinion is completely skewed with respect to the other reviewers, then, at a minimum, a very large number of grammatical errors (at least one per paragraph) must be fixed in order for the paper to make it up to the minimum standard of publication for any English language journal. This review should be considered a “first pass” review, if the paper is to be seriously considered for publication I would like a chance to review it again. Although, theoretically, it might have been possible to verify more of the technical claims of the paper, it is the opinion of this reviewer that far too much of the onus was placed on the reader; major details of estimation, model specification, and experimental design all being missing.

*Summary*

This paper is about Gaussian mixture model clustering where the class identifiers are not univariate discrete variables but are instead vector-valued discrete variables imbued with a first-order, Markov-structured “prior”. There is a complicated and potentially interesting interplay between the latent “class” variables and the structure of the class conditional means and covariance matrices. Individual or groups of *dimensions* of continuous, vector-valued observations are either singly or in groups (“pouches”) given their own class conditional means and covariances. The resulting Gaussian mixture model (with full-dimensional class conditional means and covariances) consists of a potentially very large number of “classes” (indexed by the vector-valued class variable) that index and assign observations to class conditional multivariate normal densities whose means and covariances are the product of all “pouch” Gaussian densities indexed by the elements of the latent class vectors. A message passing and greedy search algorithm are developed for estimating the model. Experimental results are given that suggest the pouch model is a superior clustering model, particularly when the data can simultaneously be clustered in more than one, potentially “conflicting” way.

*Review*

In this reader’s opinion this paper is far more confusing than it needs to be. While this reader appreciates the authors’ attempts to make the paper readable by staying high level in most technical sections, ultimately the paper suffers from having too few details where they might be helpful to the reader and too little intuition where it would really be helpful. To be honest, I’ve read the paper approximately five times and *still* remain uncomfortable with my level of understanding about what the claims are and what really has been accomplished. Part of this may stem from the pervasive, distracting grammatical errors and poor writing. A complete listing of the grammatical errors would take almost as many pages as the paper itself (there being one per paragraph nearly) so this reviewer merely suggests a thorough review for grammaticality.

*Criticisms*

I think that the title of the paper is somewhat misleading. “Latent tree” implies that there is a latent tree that is given a prior and then either estimated from data or averaged over. The tree in the PLTM is not a latent variable in any normal sense of the word. Additionally, it took this reader 3 readings of the paper to get used to the word “pouch.” Perhaps something more directly indicative of the model characteristics would be better.

Variable selection usually refers to choosing a subset of variables to retain in the model (as in L1 penalized regression). In this case, something different is going on and it doesn’t become (relatively) clear until much later in the paper. Describing the process of identifying different subsets of the observation vectors to cluster differently is not variable selection per se. I think “model selection” is probably a better way to describe what is going on, however, this perspective is troubling because the model space being explored is not penalized in complexity in any way other than via AIC, a penalty known to be quite loose.

Interestingness of clusters is not a good metric for a machine learning journal unless extensive human judgement is included in the experiments section (which it is not). Held out predictive accuracy (not included, but should be) and correctness of labeling data with known class structure (included, but not fair comparison in my opinion) are more appropriate metrics to include. More comments on the experiments section follow.

In section two, variables are used before their definitions are given, i.e. RV vs. value, vector vs. scalar, etc. (not provided until 3 pages later).

In section two you refer to a Gaussian mixture model as having a single scalar latent variable and being, therefore, limited to a single clustering. There are many, many Bayesian approaches to Gaussian mixture modeling which allow for a *distribution* over clusterings. None of these methods are considered, a major flaw in the paper as far as I’m concerned.

The description of the model would be aided tremendously by specifying a full generative model so that readers know what model inference is being performed in. As it stands one can only guess about the regularization provided by AIC and the limitations of the greed search through model structures. A reader has no way of knowing whether or not the estimation procedure is good because, among other things, it isn’t clear what the model is that’s being estimated. A number of figures showing what data from the generative model underlying the PLTM looks like would help as well, either with a full description of the generative model or as the paper stands.

Everything before section 4 could (almost) be removed (or shortened significantly – sometimes less is more). Section 4 would be aided by a clearer distinction between vector-valued and scalar-valued RV’s and values. In particular P(x\_i|y) = N(x\_i|\mu\_y, \Sigma\_y) is very confusing because it isn’t clear whether y is a vector, a scalar, or the last element of a vector. The Example should be expanded and made more clear. The last paragraph of the example is extremely confusing as the dimensionality of \Sigma\_{\pi{x\_i}} is unclear. Is the dimensionality the dimensionality of the number of variables in the pouch or the dimensionality of the entire observation vector? My assumption is that it is the latter (because of the multiplication in eqn. 2 but then the description of 1 on the diagonal and .5 otherwise doesn’t make sense). The way the paper reads right now the example is essential to understanding what’s going on, but it is not well integrated with the surrounding text. In particular it reads as if it is describing a general PLTM but is in fact describing a single PLTM whose parameterization is arbitrary. A picture of data drawn from this PLTM (PCA projections would be one way to go here) would be tremendously helpful. Additionally, actually showing the structure of both the individual covariances (in matrix form) and the resulting joint covariance (in matrix form) would be invaluable.

Equations one and two which describe the relationship between the PLTM and a regular Gaussian mixture model are the clearest and most helpful in the paper, however, they also are the cause for the most concern. These equations show the equivalence of the PLTM to a Gaussian mixture model with a large number of components, each with a richly structured mean and covariance. This is a huge cause for concern in my opinion because, for the rest of the paper, the PLTM is compared to GMM’s with lower complexity/number of components. It will come as no surprise to any reader that, when properly regularized and estimated, a model with more parameters is more expressive that one with fewer. Nothing in the paper provides evidence that anything more than this is happening. To be more clear, in the experiments you choose the best possible clustering amongst all the dimensions of the latent vector Y. What would happen if you fit K different Gaussian mixture models on randomly chosen subsets of the observation dimensions (such that the total number of free parameters in all K GMM’s is equal to the number of free parameters in the PLTM) and then chose the “best” clustering? I bet that one of the random subset Gaussian mixtures would be competitive with the best PLTM.

Section 5 should be expanded to include all details necessary to implement the estimation algorithm. No reader could implement the estimation algorithm from the details in section 5. At a minimum pseudocode for all operations should be included even if the mathematical details are excluded.

I am generally not a fan of greedy search algorithms, especially when the maxima of the objective isn’t clearly defined in the sense that it is not clear whether or not, given data generated from a fixed PLTM, the true model is the maxima of the estimation objective equation on page 13. I suspect very strongly that it is not and that the estimator is not consistent in the statistical sense. Furthermore, no guidance, theoretical or practical, is given to the reader to indicate whether or not the greedy procedure is likely to find the maxima of the objective, whether or not the objective is a consistent estimator.

Section 8.2 is the most confusing section of the paper because the equation on page 15 which explains how the PLTM is evaluated (NMI(C;Y)) is not explained clearly. Exactly how are I(C;Y), H(C) and H(Y) evaluated? In particular, P(C|d\_k) (the probability distribution of the true class given a data vector) is used without definition). If the class label is given then isn’t this distribution degenerate?

I disagree on principle with the evaluation metric of choosing to evaluate the “best” clustering. It is possible to compute both the mutual information and the individual entropies of a random vector. Why not compute the true NMI from the vector (if NMI remains the metric. I would prefer almost any other metric from Meila’s “Comparing Clusterings”)?

I anticipate either a reject or a major revision. I will withhold additional comments until a new version of the paper is submitted, if it is to be so. In general, I believe that the paper requires a large amount of additional effort to bring it to publishable quality.