**Meta-Reviews:**

|  |  |
| --- | --- |
| **Question** |  |
| Detailed Comments | This paper kernelizes the recent method of moments tensor factorization algorithm for learning latent variable models, which permits estimating a much richer set of mixture distributions compared to before. Experiments could be strengthened. The paper could also better clarify the delta over previous work. |

|  |  |
| --- | --- |
| **Masked Reviewer ID:** | Assigned\_Reviewer\_1 |
| **Review:** |  |

|  |  |
| --- | --- |
| **Question** |  |
| Overall Rating | Weak reject |
| Reviewer confidence | Reviewer is knowledgeable |
| Detailed comments for the authors | Kernel embedding is a well-known technique to map the distributions into an infinite dimensional RKHS. It is also well known that this mapping is injective. In this paper, authors use this embedding trick to estimate the parameters of a multi-view latent variable model. Authors present a spectral algorithm (based on the spectral decompositions of the gram matrices of different views) to estimate these parameters.   Since authors use the well-known kernel trick in addition to an existing algorithm to estimate the parameters, the paper is limited in technical novelty. The proposed algorithm, as also mentioned in the paper, is merely a Kernelized version of an existing algorithm. However, it is interesting to see how this algorithm provides a unified framework and subsumes many of the existing algorithms.   Results presented in this paper are not convincing enough. It is not clear, in the real dataset, how authors came up with splitting those 5 dimensions into three different views. Although it is clear in the formulation, it is not clear in this experiment that what a view means. In general, a view can be created in two ways – one is, you can divide the dataset along the dimensions to create different views (as you are doing), and second, you can group examples themselves, each group representing one view. It seems that in this experiment, you are using both kind of views. If true, it is important to see why one needs to split the dataset along the feature dimesnsion to create these different views, and how one can come up with such split.   The baselines used in this paper are rather inferior, one of them is EM, which is generative. Anywhere where the underlying generative assumption is wrong, the method is bound to perform worse. Similarly, the other baseline which is not only generative but also restrictive in the sense that the center should like on a k-dimensional span. It would be interesting to see the performance of the proposed approach in the light of other parametric methods.   Authors claims to have experimented on 30 datasets, the figure only shows 24 datasets. |
| Summary of evaluation / Paper's main strengths and weaknesses | . |
| I have read and considered the authors' response. (To be answered only after the response period.) | No |

|  |  |
| --- | --- |
| **Masked Reviewer ID:** | Assigned\_Reviewer\_3 |
| **Review:** |  |

|  |  |
| --- | --- |
| **Question** |  |
| Overall Rating | Weak reject |
| Reviewer confidence | Reviewer is knowledgeable |
| Detailed comments for the authors | The paper combines spectral methods with kernel mean embeddings to infer latent variable models.  I'm not an expert for spectral methods, though, there is an important point on the embedding side that is worth raising. The strength of embeddings lies in the calculation of integrals and expectations not in density estimation. Since the task in the paper is to learn latent variable models the authors seem forced to move away from expectations towards densities.   This puts them in a setting that is not natural for the embeddings and my feeling is that the authors are well aware of this since the necessary move from expectations to densities is pushed aside as much as possible. The main description of how they do this is L340-L343 that is they set  p(x|h) = E\_(y~h) k(x,y)   where E\_(y~h) is the inferred mean embedding for the case that the latent variable is h. Now this is not a great way to estimate a density. First of all you need a kernel which concentrates around x; think of a delta-distribution. So rule out polynomial kernels and, I guess, kernels on structures like graphs.   It is possible, and I think this is the only reasonable setting, to use kernels which concentrate around x and have a bandwidth parameter that shrinks towards zero with sample size. Though, you are getting very close to kernel density estimation here and one wonders what you gain from the embedding approach. In particular, why not do directly a kernel density estimation for the conditionals instead of estimating the densities through the embeddings?   In my opinion this is a really crucial point and at the very least a proper discussion is needed why either kernel density estimation cannot be performed (I don't see why one shouldn't be able to use it) or, if it can, why there is no comparison to kernel density estimation. In particular, arguments are needed why embeddings are worth the effort and experimental results to support the claims.   I also find the current experimental discussion about how the density extraction from the embeddings plays in insufficient. In 7.1 probabilities are directly estimated, in 7.2 the maximum posterior -- which I guess means probabilities for the h's are estimated and then the highest probability is chosen? The bandwidth will here be an important factor to concentrate in the right rate around x. It is said some form of cross validation is performed in 7.2 using the log-likelihood. Some more details would be appreciated. |
| Summary of evaluation / Paper's main strengths and weaknesses | The paper is mostly presented well. However, I have considerable concerns about the extraction of densities from mean embeddings. This is rather unnatural and I miss a clear argument why this is something one wants to do and why other non-parametric methods like kernel density estimation are not suitable. |
| I have read and considered the authors' response. (To be answered only after the response period.) | No |

|  |  |
| --- | --- |
| **Masked Reviewer ID:** | Assigned\_Reviewer\_4 |
| **Review:** |  |

|  |  |
| --- | --- |
| **Question** |  |
| Overall Rating | Weak accept |
| Reviewer confidence | Reviewer is knowledgeable |
| Detailed comments for the authors | This paper brings together two emerging themes: the first pertains to the resurgence of the method of moments for estimating latent variable models using approximate tensor decomposition schemes (Anandkumar, Hsu, Kakade et al, 2012), and the second theme pertains to RKHS embeddings of probability distributions for extending spectral techniques to non-parametric estimation settings (Smola et al, 2007 and several recent papers). The focus in this paper is on "multi-view" latent variable models where observed variables are conditionally independent given a discrete hidden variable. Moment tensors in the kernelized settings are replaced by higher order covariance operators, whose tensor decomposition after appropriate whitening steps yields the mean embeddings of the underlying conditional distributions. The algorithm can be implemented on Gram matrices given a set of observations. The paper extends previous sample complexity results for this kernelized setting and some promising empirical results are reported.  Overall, the paper is well written and the contributions follow naturally from previous related literature.  Do other models such as ICA or LDA also admit such a kernelization?  Section 5.2 should be rewritten with pure kernelized notation (no feature spaces - only Gram matrices). Is Algorithm 1 only a partial description - e.g., I do not see the tensor power method invoked anywhere there. As I understand it, the power method is run on deflated versions of the tensor for i=1-to-k. Can these steps be entirely implemented with Gram matrices? If so please clarify, as such a full description will help appreciate the full computation complexity of the proposed algorithm. For the case of p views, are these Gram matrices of size pm x pm - if so scalability becomes a concern rather quickly?  Last line of the paper: "more robustness" -> "more robust"  It may be useful to provide a short proof sketch for Theorem 2, or atleast how the theorem relates to previous results. What is being added to the proof technique in this specific kernelized context? This will help make the paper a bit more self-contained.  In Figure 4, the benefits of the kernel spectral approach over EM-GMM are not consistent. How was the claim that these are datasets where the multiview assumption is "heavily violated" verified? As in Figure 2, is Hsu et al not applicable on this problem for some reason? It would have made the paper stronger to provide more empirical support on real datasets. |
| Summary of evaluation / Paper's main strengths and weaknesses | Nice paper combining RKHS embedding ideas with tensor decomposition techniques for latent variable estimation - though in some respects the algorithmic extensions are straightforward. The proposed approach is backed with empirical results and theoretical analyses. The empirical section could be strengthened. The paper can be improved by making some sections more self-contained and clarifying precisely how some of the analysis builds on prior results. |
| I have read and considered the authors' response. (To be answered only after the response period.) | No |