STATS 606 Peer Review –
A Novel Method for Sufficient Dimension Reduction on Gaussian Mixture Models

Mason Ferlic & Easton Huch

April 2022

1 Summary

The authors propose a novel method for sufficient dimension reduction where they model the joint density of the features and labels as a Gaussian mixture model. They propose four algorithms to learn a sufficient statistics which maximizes the mutual information. The problem has a closed form solution for $p = 1$ but requires constrained optimization for the more general case. For SDR of a Gaussian mixture model the joint distribution (X, Y) is modeled by a GMM. The Expectation of the mutual information can be computed by the EM algorithm. The authors propose three methods for the constrained optimization problem: (1) the natural gradient of the objective function. (2) Cayley transformation; which constructs a descent curve in a different space. (3) Projected gradient descent. The authors also propose a method for choosing the optimal p based on the ratio of mutual information of the SDR and the MI of the full joint distribution. Lastly, the authors present pseudocode overview of the 3 methods and simulations with the GMM on synthetic and real datasets. The GMM SDR is also compared to other well-know dimension reduction methods. The authors show their method works as well or better than most alternatives.

Overall, this paper was fascinating and we found the use of a GMM and the mutual information metric to be elegant. It was interesting to see the closed-form analytical solution for $p = 1$ and helps validate the theory when moving to higher-dimensions. Some suggestions for the authors would be to remove projected gradient descent if not implemented. If possible we would like to see some performance metrics from the different algorithms and if the Cayley implementation offers any improvements. In the discussion section maybe note if there were limitations or implementation issues you found.

2 Comments/Questions

- Since f is not a convex function and there are multiple minimizers how do you ensure the solution is the global min or unique? Do different starting values impact this?

- It might be nice to have a bit more explanation as to how this method differs from competing methods.
- Do you implement the EM algorithm for (2)?
- In the natural gradient methods is there an analytical solution for $\nabla f(A)$ or G?
- For method 2, are we finding the minimum at each iteration along the decent curve? Or just take a fixed step?
- Can you explain what the constraints are in the optimization problem? Wasn't very clear to me. $A^T A = I_p$?
- Suggestion: Don't introduce method 3 if you don't use it.
- The CAT names are funny but the acronyms are not totally clear tbh
- How long does the search for optimal p take to run? Also, would be cool to see some run time metrics for each algo.
- For visualizing the transformation did you set $p = 5$ or use an adaptive approach?
- What is it about Figure 2 that shows that your method and SAVE are working well? Is it because you have a small number of large values and a large number of small values?
- In Table 1 it looks like your method is performing poorly on certain datasets (corr < 0.1) compared to the other methods. Do you have some ideas why? Is the algo not converging in high-dim? I think a better comparison would be to use MSPE since you split test and train sets.
- Hows does knowing/estimating the GMM parameters a priori impact the estimation of A? Should estimation of the GMM and maximization of information proceed jointly? If not maybe explain why.
- The paper is quite dense in terms of its technical material. You might consider moving more of the technical details and algorithms to the appendix. Doing so would allow more room for explanation, comparison, and synthesis.

3 Corrections

- Ann Arbor is misspelled in the title