Robust Statistical Methods for Image Processing

Kalpesh Patil
Electrical Engineering
IIT Bombay
Email: kalpeshpatil@iitb.ac.in

Guide: Prof. Suyash Awate
Computer Science Engineering
IIT Bombay
Email: suyash@cse.iitb.ac.in

Co-guide: Prof. S. N. Merchant
Electrical Engineering
IIT Bombay
Email: merchent@ee.iitb.ac.in

***Abstract*—Usually performance of a statistical technique is highly affected by the presence of outliers in data. Unsupervised clustering methods like GMM are found to be very susceptible to such outliers. Hence we studied robust technique like SMM and compared their results with GMM. We also studied Variational Bayesian counterparts of these techniques (VBGMM and VB-SMM respectively), which will be used for Bayesian Inferencing. We wish to apply these robust techniques on fMRI clustering in future.**

INTRODUCTION

Presence of outliers is a practical issue faced by many of the statistical models. Especially presence of outliers is prevalent in Medical Imaging. Functional Magnetic Resonance Imaging (fMRI) is a technique used in neuroimaging to analyze temporal aspects of activations in various parts of the brain. Unsupervised Clustering is used to learn patterns in those activations.

Gaussian Mixture Models (GMM) are widely used for clustering. But they are very much susceptible to large outliers. We wanted to incorporate robustness into our models. Hence SMM (Student-t Mixture Models) are used to model data containing outliers. The basic idea is to use a heavy-tailed distribution (heavier than Gaussian) to model data containing outliers. Sometimes we are also interested in 'inferencing' using the model, which requires computation of some intractable integrals. Variational Bayesian is one of the approaches to solve this. Hence we also studied VBGMM and VBSMM which are counterparts of GMM and SMM in Bayesian framework.

This report will first go through major prior literature in this domain, then we will elaborate some results obtained from the implementation of those papers, with few additional experiments of our own on a toy dataset. Later we will briefly go through potential applications of these techniques in fMRI domain and further plan for research.

LITERATURE REVIEW

GMM

GMM is one of the most widely used model for clustering. [1, Chapter 9] has studied GMM in depth. Each data point is modeled as weighted sum of Gaussians. $\{\ \mu_k,\ \Sigma_k,\ \pi_k\ \}$ are the parameters of the model (mean, covariance matrices and weights for gaussians). Cluster labels are treated as hidden variables in this model. The parameters are optimized using EM algorithm. EM algorithm consists of an *E-step* and and an *M-step*, which are performed alternatively. *E-step* computes expectation of log-likelihood using current estimates of parameters. *M-step* maximizes expectation found in *E-step*. This provides updates for the parameters of the model. Please refer [1, Chapter 9] for detailed analysis of GMM using EM algorithm.

SMM

While using GMM, an inherent assumption that datapoints from a particular cluster come from Gaussian distribution, is made. Although this is justifiable in the case of pure data or data corrupted with Gaussian noise, it fails to model large outliers. Gaussian distribution puts very less probability for points far away from the center (outliers), hence a heavy tailed distribution is required. McLachlan et. al [2] have studied finite mixture of Student-t distributions for unsupervised clustering. A student-t distribution can be looked upon as a combination of infinitely many Gaussian distributions with scaled covariance matrices. Lets define u as a scaling parameter for Σ for each data point. Assume u is generated from $Gamma(\frac{\nu}{2}, \frac{\nu}{2})$ distribution. Consider combination of such infinite scaled gaussians.

$$\int \phi(y_j; \mu, \frac{\Sigma}{u}) du$$

Above integral results in student-t distribution given below

$$f(y_j; \mu, \Sigma, \nu) = \frac{\Gamma(\frac{\nu+p}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})(1+\delta(y_j;\mu,\Sigma)/\nu))^{\frac{\nu+p}{2}}}$$
$$\delta(y_j; \mu, \Sigma) = (y_j - \mu)^T \Sigma^{-1} (y_j - \mu)$$

Here p is dimensionality of the data points and ν is the parameter for degree of freedom. It controls the heavy-tailedness of the distribution. Lower DoF implies heavier tails. At $\nu = \infty$ it approaches Gaussian distribution. In SMM we have two sets hidden variables, class labels and scaling factors. Labels (z_{ij}) is 1 if y_j belongs to class i, else zero. According to the definition of the 'scaling parameter' u_i, we can define following random variables.

$$Y_j|(u_j, z_{ij} = 1) = N(\mu_i, \frac{\Sigma_i}{u_j})$$
$$U_j|(z_{ij} = 1) = gamma(\frac{v_i}{2}, \frac{v_i}{2})$$

Iterative update scheme is developed using EM algorithm as before. Note that update for degree of freedom (ν_i) doesn't have closed form solution. Hence a complicated equation

involving digamma function needs to be solved to get the updates. Further Henning et. al [3] have shown that outliers have to be much larger for the breakdown of an SMM compared to that of GMM. This was also observed during our experiments.

Variational Bayesian Approximation

In order to find optimal model required for the given data based on evidence of data given model structure, we need to evaluate following integral. [1, Chapter 10]

$$Pr(X|H_M) = \int_\theta Pr(X|\theta, H_M)Pr(\theta|H_M)d\theta$$

Here $X = \{x_n\}_{n=1}^{n=N}$ is the observed data, H_M is model structure used for modeling that data and θ is the set of parameters. Usually computing $Pr(\theta|H_M)$ leads to intractable integrals and hence approximations are used. Sampling techniques like MCMC (Markov Chain Monte Carlo) are some of the numerical simulations based methods to approximate these integrals. One must have to ensure that they have converged properly. Other approach is to use Variational Bayesian techniques. Here we assume that joint posterior over latent variables and parameters factorizes into components. One of the most important use of variational based approach is the ability to compute total likelihood (variational lower bound of evidence) of the data. This allows one to compare various models based on their likelihoods.

Probabilistic graphical models are used to denote relationship (dependence) amongst random variables. *Bayesian Networks* aka *Directed Graphical Models* are one of the kinds of graphical models used to visualize joint distribution of a set random variables. A node denotes a random variable. Conditional distribution of each node is given by the edges connecting it to the parent node. A shaded node implies given random variable is observed and a box around random variables/s denotes set of multiple (N) IID random variables. This kind of visualization helps in building intuition for the conditional dependence of random variables, hence highly useful in Bayesian framework.

In Bayesian framework parameters like mean, covariances etc. will also be treated like random variables having specific priors. We use Dirichlet prior for component weights (π_k) and Gaussian-Wishart prior for mean and precision of each of the Gaussian components. Note that we have used conjugate priors in each of the cases i.e. if posterior distribution also belongs to the same family of distribution of prior, then the prior is called conjugate prior. Detailed treatment of Variational Bayesian approach for GMM (VBGMM) is given in [1, Chapter 10].

Archambeau et. al [4] have done similar analysis for Variational Bayesian approach for mixture of Student-t distributions (VBSMM). Following appropriate conjugate priors are chosen for given parameters.

- Mixture proportions: Dirichlet prior $D(\pi|\kappa_0)$ is used as prior for $\{\pi_m\}$.
- Means and precisions: Gaussian-Wishart prior $NW(\mu_m, \Lambda_m|\theta_{NW_0})$ is chosen. Gaussian Wishart is product of Gaussian and Wishart function $N(\mu_m|m_0, \eta_0\Lambda_m)W(\Lambda_m|\gamma_0, S_0)$
- Degree of Freedom: No prior is assumed over $\{\nu_m\}$.

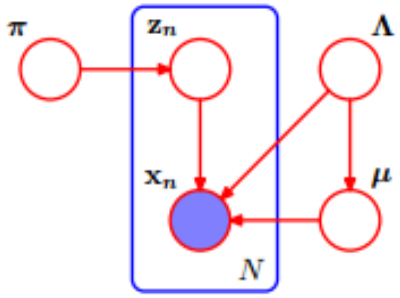

Fig. 1. Graphical model for VBGMM. Image credits: [1]

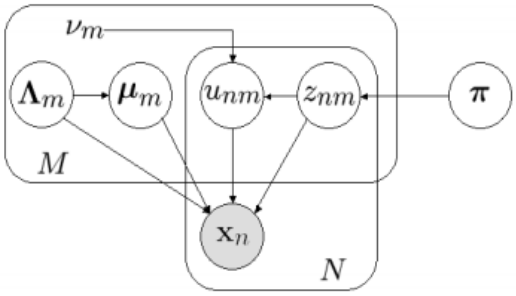

Fig. 2. Graphical model for VBSMM. Image credits: [4]

Thus joint posterior can be factorized as follows.

$$Pr(\theta_s|H_M) = D(\pi|\kappa_0)\prod_{m=1}^{M} NW(\mu_m, \Lambda_m|\theta_{NW_0}). \quad (1)$$

VBEM algorithm is used to maximize the lower bound. The crux of the VBEM algorithm is the use of EM algorithm in variational bayesian setting to tackle with circular dependencies of latent variables while taking expectation. It consists of two steps; VBE and VBM step. Details of the algorithm can be found in [5, Chapter 2], [1, Chapter 10]. Application of VBEM to the problem of clustering using mixture of Student-t distributions can be found in [4].

Experiments and Results

GMM and SMM

To demonstrate better performance of SMM w.r.t. GMM in the presence of outliers, we implemented and ran an experiment similar to [2].

Experimental setup: We create a 2D toy dataset for clustering as follows. The datasets consist of 3 components of Gaussian distribution having means

$$\mu_1 = [-2, 0], \qquad \mu_2 = [0, 0], \qquad \mu_3 = [2, 0]$$

and common covariance matrix as given below

$$\Sigma_i = \begin{bmatrix} 0.2 & 0 \\ 0 & 2 \end{bmatrix}$$

In the first experiment, we sampled 200 data points from each of three distributions and certain number of outliers which we will vary. Outliers are sampled from the following set uniformly,

$$\big\{(x_1, x_2) \in [-5, 5] \times [-8, 8] \text{ such that } (x < -4 \text{ or } x > 4) \text{ and } (y < -5 \text{ or } y > 5)\big\}$$

We try to fit GMM as well as SMM with three components to the data. Means are initialized using k-means clustering of data and covariance matrices are initialized with the value of covariance of the data. Means and covariances are estimated according to [2]. Fig. 3 displays performance of GMM and SMM respectively on a data containing no outliers i.e. 600 total points with 200 coming from each of the original components. Equiprobable contours are plotted. Red implies higher probability and blue implies lower probability. It can be observed that both GMM and SMM performed equally well in absence of outliers and estimated means and covariance matrix were close to original. Now we add 20 outliers to the data distributed as mentioned above. Similar to fig 3, we plot equiprobable contours for the data containing outliers in fig. 4. In fig. 4 bunch of points at the corners, away from the cluster centers are the outliers. As we can observe that GMM has failed miserably in modeling actual data, while SMM has performed really well. The means and covariance matrices of SMM components were close to actual means and covariances. GMM means and covariances are not close to the original values.

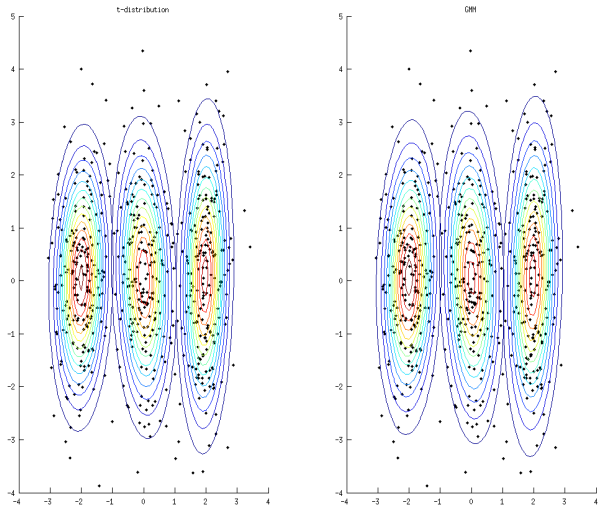

Fig. 3. Performance of GMM and SMM on data with containing no outliers. Left image corresponds to SMM while right image corresponds to GMM.

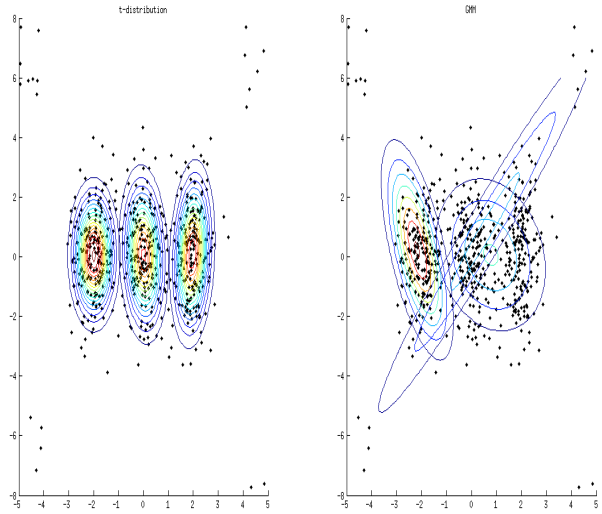

Fig. 4. Performance of GMM and SMM on data with outliers. Left image corresponds to SMM while right image corresponds to GMM.

Quantitative analysis: To quantitatively measure improvement in performance of SMM over GMM we used few metrics described below.

- Average Euclidean distance between cluster centers($d_{cluster-centers}$):

$$d_{cluster-centers} = \frac{1}{K}\sum_{k=0}^{K} ||c_k^{est} - c_k^{orig}||_2$$

Here c_k is the center of the k^{th} mixture component. c_k^{est} and c_k^{orig} are estimated and original cluster centers respectively. $d_{cluster-centers}$ gives an idea about how far away estimated means are from the actual means.

- Mean cosine similarity between eigenvectors of covariance matrices($s_{eigen-cos}$): We calculate sum of absolute cosine similarities between each eigenvector of estimated covariance matrix and actual covariance matrix and average it over number of components.

$$s_{eigen-cos} = \frac{1}{K}\sum_{k=0}^{K}\sum_{j=0}^{j=d} abs\left(\frac{v_{jk}^{est}.v_{jk}^{orig}}{||v_{jk}^{est}||_2||v_{jk}^{orig}||_2}\right)$$

Here v_{jk} is j^{th} eigenvector of k^{th} cluster component, d is number of eigenvectors (equals to dimensionality in most cases) and K is total number of mixture components. Eigenvectors of a covariance matrix give idea about direction of equiprobable ellipsoidal contours. They represent major and minor axis of the ellipse. The maximum value for $s_{eigen-cos}$ is d (when eigenvectors of all mixture components align exactly). Note that absolute value is chosen because v^{est} and $-v^{est}$ imply the same axis. Hence similarity of v_{orig} from both should be treated equally.

TABLE I
Performance of GMM and SMM with and without outliers

	without outliers		with outliers	
Quantity	GMM	SMM	GMM	SMM
$d_{cluster-centers}$	0.0820	0.0782	1.1148	0.0883
$s_{eigen-cos}$	1.9987	1.9987	1.7896	1.9987

For better models $d_{cluster-centers}$ should be low (close to zero) and $s_{eigen-cos}$ high (close to $d = 2$). Table I shows that both (GMM and SMM) perform equally well when no outliers are present. But in presence of outliers, only SMM was able to perform well.

As we had mentioned earlier degree of freedom in Student-t distribution is the parameter which allows model to have heavy-tailed distribution. In our experiment we observe that optimal degree of freedom was very high for the case of no outliers for all the components, while it was low for the case of data corrupted with outliers. This is expected, because at $DoF = \infty$, student-t tends to Gaussian distribution.

VBGMM and VBSMM

We also carried out experiment in Byesian framework using Variational Bayesian approach. Experimental setup was same as before (original means, covariances and number of mixture components). The major benefit of VB approach is that it allows one to claculate likelihood (lower variational bound) of data given the model structure only i.e. parameters are also treated like random variables and likelihood is computed by taking their prior distributions into account.

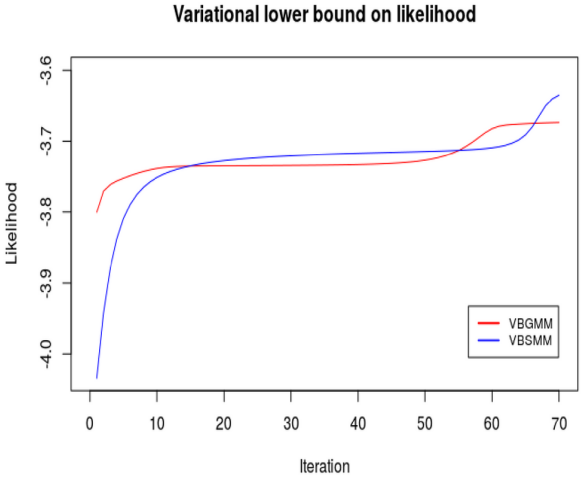

Fig. 5. Variational lower bound on likelihood with iterations of VBEM algorithm when no outliers are present.

Fig. 5 shows how variational lower bound on likelihood increases in each iteration of VBEM algorithm for the case of no outliers. In absence of outliers we can observe that both VBGMM and VBSMM predicts approximately same likelihood of the data. Now we add 40 outliers sampled from

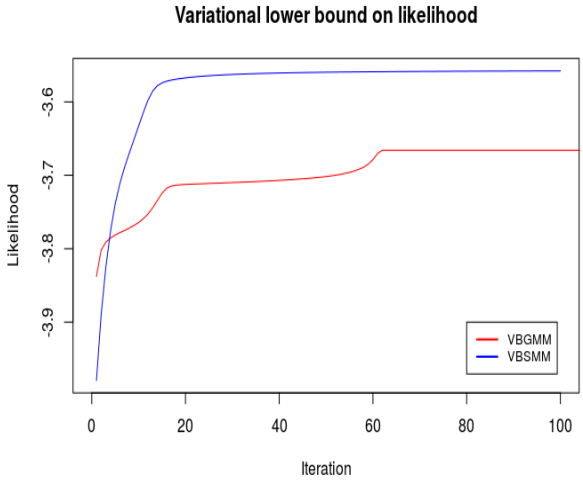

Fig. 6. Variational lower bound on likelihood with iterations of VBEM algorithm when outliers are present.

distribution mentioned in previous section and plot variational lower bound on likelihood in Fig. 6. We can observe that there is significant amount of difference between likelihoods of VBGMM and VBSMM, when outliers are added to the data. This implies that SMM is a better model than GMM in presence of outliers.

Applications in Image Processing: fMRI

Functional Magnetic Resonance Imaging (fMRI) is a technique used in neuroimaging to observe patterns in activations of various parts of brain. fMRI is basically a 4 dimensional object i.e. 3 dimensional voxel data collected at various time instants. Clustering of this time series data gives an idea about which parts of the brain are stimulated synchronously. Various approaches like k-means [6], GMM [7], Ward's hierarchical clustering [8], Spectral clustering [8],c-means [9] ICA [10] etc. have been studied before. We believe that SMM with some regularization in the form of spatial smoothening prior would perform better due to its robustness against outliers in fMRI data.

Conclusion and Future Work

It has been shown that Student-t Mixture Models perform better than Gaussian Mixture Models when there are outliers present in data. Some metrics are defined to quantify closeness of estimated values to the actual values. The methods are also explored in Bayesian framework and it is shown that variational lower bound on likelihood is better for VBSMM than VBGMM in presence of outliers. Some of techniques used for fMRI clustering are studied. In future we would like to implement SMM or VBSMM on fMRI data. Also one needs to tackle curse of dimensionality while analyzing a very high dimensional data like fMRI. Further we also want to modify these algorithms to incorporate spatial and temporal priors while performing clustering on actual fMRI data.

REFERENCES

[1] C. Bishop, "Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn," *Springer, New York*, 2007.

[2] G. J. McLachlan, S.-K. Ng, and R. Bean, "Robust cluster analysis via mixture models," *Austrian Journal of Statistics*, vol. 35, no. 2, pp. 157–174, 2006.

[3] C. Hennig, "Breakdown points for maximum likelihood estimators of location-scale mixtures," *Annals of Statistics*, pp. 1313–1340, 2004.

[4] C. Archambeau and M. Verleysen, "Robust bayesian clustering," *Neural Networks*, vol. 20, no. 1, pp. 129–138, 2007.

[5] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University of London United Kingdom, 2003.

[6] A. Venkataraman, K. R. Van Dijk, R. L. Buckner, and P. Golland, "Exploring functional connectivity in fmri via clustering," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 441–444.

[7] G. Garg, G. Prasad, L. Garg, and D. Coyle, "Gaussian mixture models for brain activation detection from fmri data," *International Journal of Bioelectromagnetism*, vol. 13, no. 4, pp. 255–260, 2011.

[8] B. Thirion, G. Varoquaux, E. Dohmatob, and J.-B. Poline, "Which fmri clustering gives good brain parcellations?" *Frontiers in neuroscience*, vol. 8, p. 167, 2014.

[9] M. H. Lee, C. D. Hacker, A. Z. Snyder, M. Corbetta, D. Zhang, E. C. Leuthardt, and J. S. Shimony, "Clustering of resting state networks," *PloS one*, vol. 7, no. 7, p. e40370, 2012.

[10] C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith, "Investigations into resting-state connectivity using independent component analysis," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 360, no. 1457, pp. 1001–1013, 2005.