Minimum message length estimation of mixtures of multivariate Gaussian and von Mises-Fisher distributions*

Parthan Kasarapu, Lloyd Allison

Faculty of Information Technology, Monash University

parthan.kasarapu,lloyd.allison,(@monash.edu)

Abstract

Mixture modelling problem involves the inference of an optimal number of mixture components and their corresponding parameters. This paper discusses unsupervised learning of mixture models using the Bayesian inference paradigm of Minimum Message Length (MML). We propose a search method that is able to model the given data using a mixture of probability distributions by reliably balancing the trade-off between the mixture's complexity and its goodness-of-fit. The proposed inference method generalizes to mixture modelling problems involving many probability distributions, demonstrated here using the multivariate Gaussian and also the von Mises-Fisher (vMF) directional probability distribution. The effectiveness and practical utility is shown by applications in text clustering and mixture modelling of protein spatial orientation data.

1. Motivation

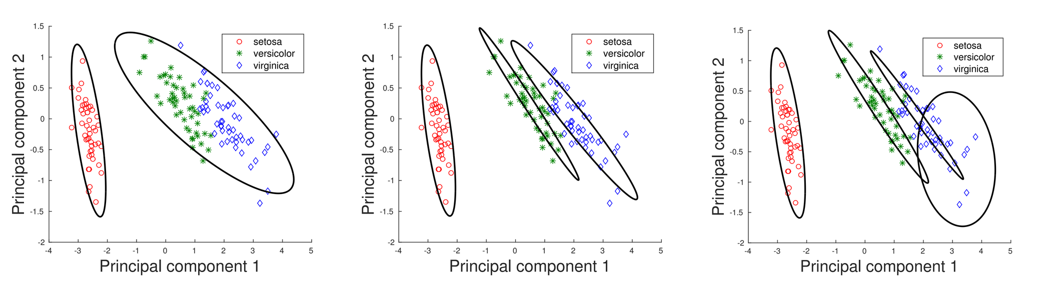


Figure 1: How many mixture components?

- Statistical model selection is important.
- Several competing models: which one to choose?
  - A criterion to compare models with the ability to compare models belonging to the same model class.
  - Based on the model's complexity and the goodness-of-fit

2. Minimum Message Length Framework

A Bayesian-information theoretic criterion to model data $\mathcal{D}$ using a hypothesis $\mathcal{H}$ (Wallace and Boulton, 1968)

- Bayes's theorem: $\Pr(\mathcal{H} \,\&\, \mathcal{D}) = \Pr(\mathcal{H}) \times \Pr(\mathcal{D}|\mathcal{H})$
- Shannon's observation: $I(\mathcal{H}) = -\log \Pr(\mathcal{H})$

$$\text{MML criterion: } \underbrace{I(\mathcal{H} \,\&\, \mathcal{D})}_{\text{Total}} = \overbrace{\underbrace{I(\mathcal{H})}_{\text{First part}}}^{\text{Complexity}} + \overbrace{\underbrace{I(\mathcal{D}|\mathcal{H})}_{\text{Second part}}}^{\text{Goodness-of-fit}}$$

$$\text{Optimal model: } \arg\min_{\mathcal{H}} I(\mathcal{H} \,\&\, \mathcal{D})$$

The total message length to describe $\mathcal{D}$ using a mixture with M component probability distributions:

1. *First part:* Encoding cost of the mixture weights and the parameters of the components.
2. *Second part:* Encoding cost of the data using the M-component mixture.

The MML framework is able to distinguish models belonging to the *same* model class. For example, all M-component mixtures have different first part message lengths depending on their constituent parameters.

3. Objectives

- MML-based estimation of the parameters of the multivariate Gaussian and von Mises-Fisher distributions.

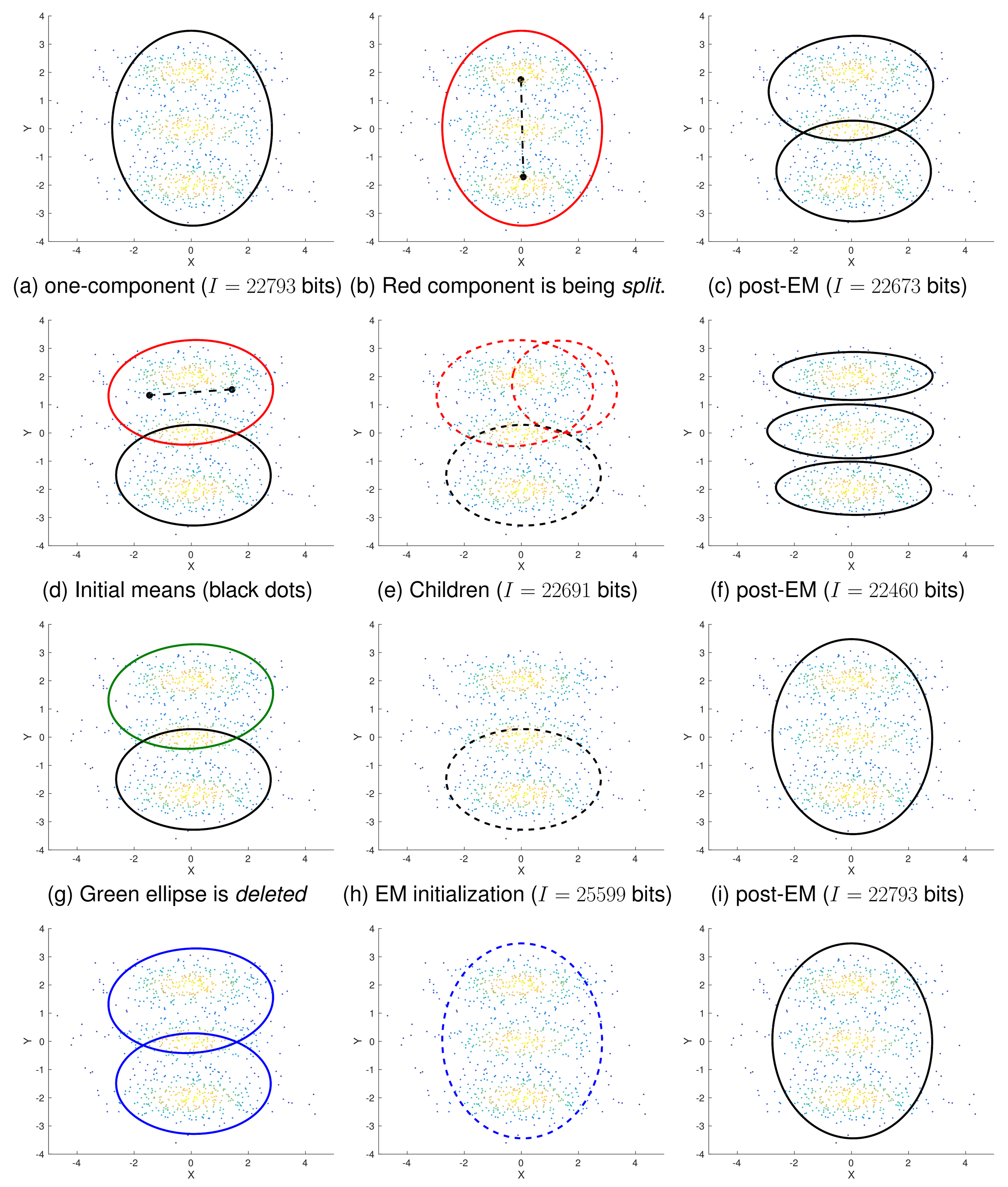
  As compared to the maximum likelihood estimators, the **derived MML estimates** have *lower* bias, mean-squared error, and Kullback-Leibler (KL) divergence.

- A generalized MML-based search heuristic to infer the optimal number of mixture components that best explain the observed data. The search implements a generic approach to mixture modelling and allows, in this instance, the use of d-dimensional Gaussian and vMF distributions.

  The proposed methodology:
  - Includes an *accurate MML formulation* unlike the MML-like approximation of Figueiredo and Jain (2002).
  - Makes *no assumptions* pertaining to the form of the component distribution.

Machine Learning: volume 100, issue 2 (2015), pp. 333-378

4. Proposed Search Method



(a) one-component ($I = 22793$ bits) (b) Red component is being *split*. (c) post-EM ($I = 22673$ bits)

(d) Initial means (black dots) (e) Children ($I = 22691$ bits) (f) post-EM ($I = 22460$ bits)

(g) Green ellipse is *deleted* (h) EM initialization ($I = 25599$ bits) (i) post-EM ($I = 22793$ bits)

(j) Blue ellipses are *merged* (k) EM initialization (l) mixture post-EM ($I = 22793$ bits)

Figure 2: Progression of the search for the optimal mixture.

- Begin with a one-component mixture.
- At any intermediate stage, the components of a M-component mixture are perturbed using *split*, *delete*, and *merge* operations.
  - *Split:* A parent component is split to find locally optimal children leading to a $(M+1)$-component mixture.
  - *Delete:* A component is deleted to find an optimal $(M-1)$-component mixture.
  - *Merge:* A pair of *close* components are merged to find an optimal $(M-1)$-component mixture.

The perturbations provide the best chance for the intermediate mixture to escape a local optimum.

- The perturbed mixture with the greatest improvement to the *two-part* message length is retained. The procedure is repeated until there is no improvement.

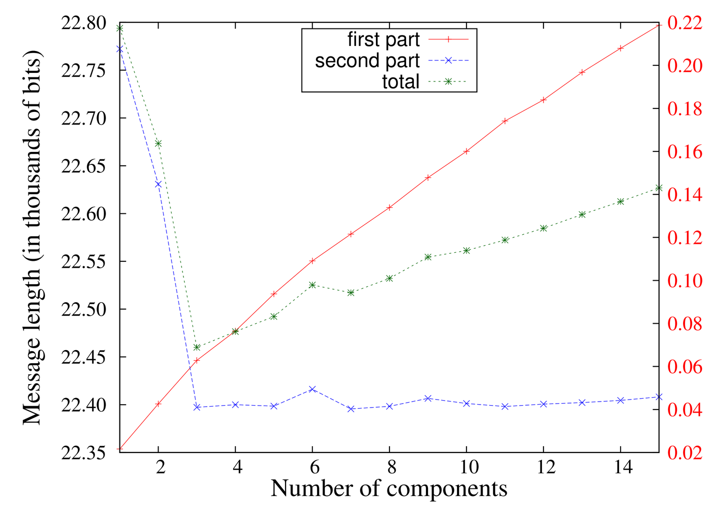


Figure 3: Variation of the individual parts of the total message length with increasing components.

5. Performance of the search method

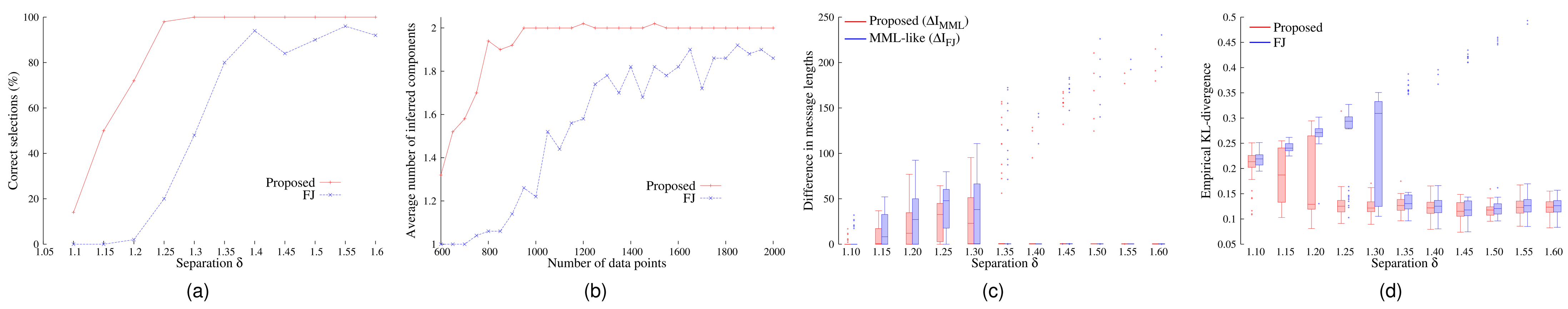
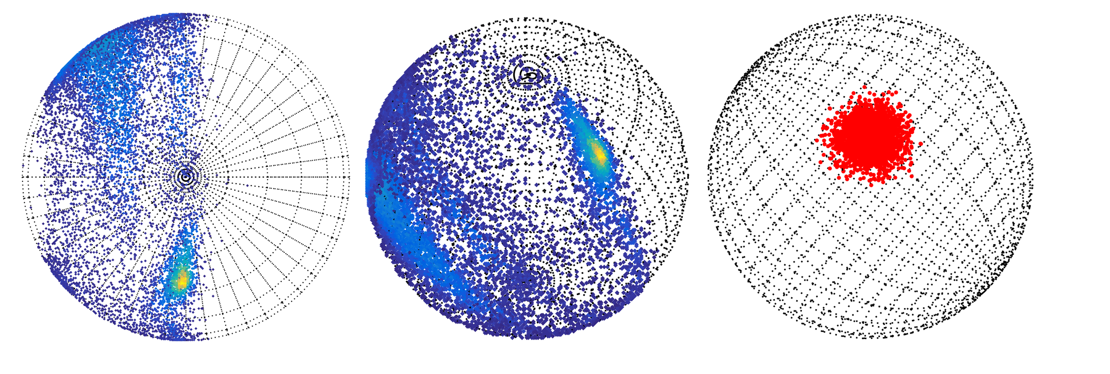


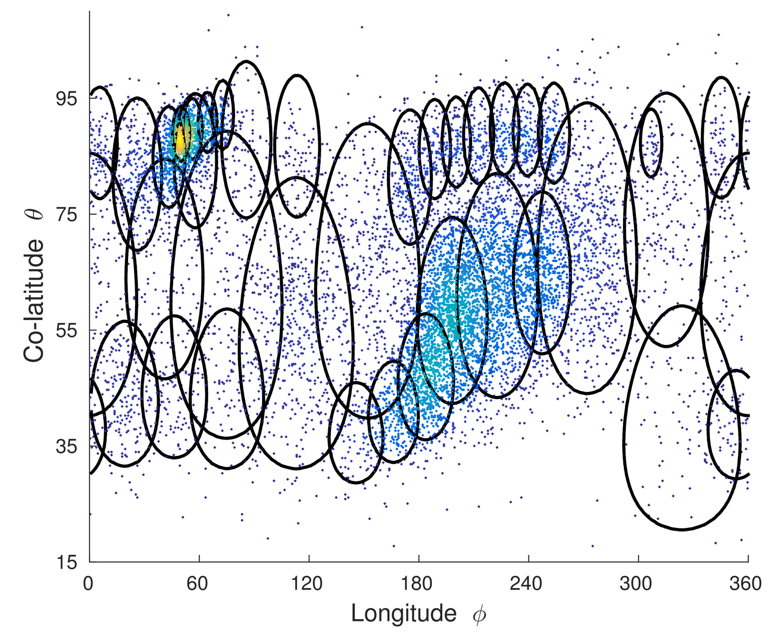
Figure 4: Results of the 10-dimensional Gaussian mixture simulations compared to that of Figueiredo and Jain (2002) (a) Percentage of correct selections with varying δ (separation between the component means) for a fixed sample size of $N = 800$ (b) Average number of inferred mixture components with different sample sizes and $\delta = 1.20$. (c) Difference in message lengths of inferred mixtures (d) Box-whisker plot of KL-divergence of inferred mixtures.

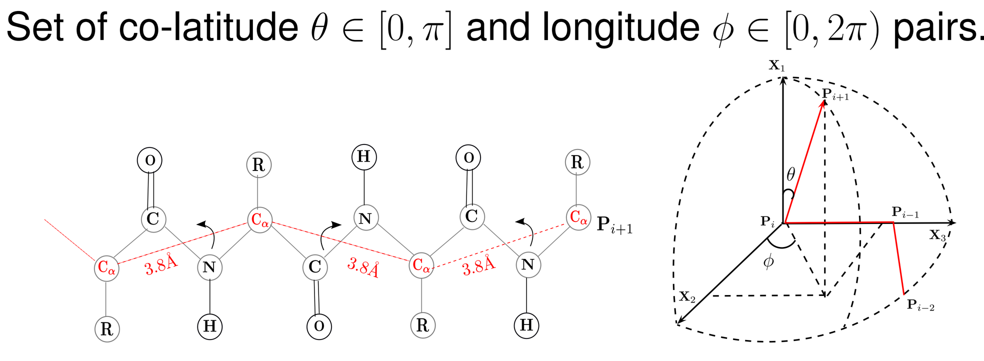
6. Mixture modelling using von Mises-Fisher distributions

Mixture modelling of protein directional data

- Data corresponds to unit vectors on the sphere.
- Set of co-latitude $\theta \in [0, \pi]$ and longitude $\phi \in [0, 2\pi)$ pairs.







Text clustering

- Data corresponds to the *normalized vector representations* of text documents (Banerjee et al., 2005).
- The vMF *directional* probability distribution is used to model unit vectors on the surface of a unit hypersphere.

Clusters		Evaluation metric	Methods of vMF parameter estimation				
True	Inferred		Banerjee	Tanabe	Sra	Song	MML
3	3	Message length	100678069	100677085	100677087	100677080	**100676891**
		Avg. F-measure	0.9644	0.9758	0.9758	**0.9780**	0.9761
		Mutual Information	0.944	0.975	0.975	**0.982**	0.976
20	21	Message length	728497453	728498076	728432625	728374429	**728273820**
		Mutual Information	1.313	1.229	**1.396**	1.377	1.375

Table 1: Clustering performance on the two datasets: (a) Classic3 (b) CMU_Newsgroup. The MML mixtures *consistently* have lower message lengths.

References

A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005.

M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

MONASH University Information Technology NICTA

ECML-PKDD, 7 - 11 September 2015, Porto, Portugal