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

Parthan Kasarapu & Lloyd Allison

Monash University, Australia

September 8, 2015

Presentation Outline

- Mixture modelling problem
- Minimum Message Length framework
- MML-based search method
- Evaluation of the proposed method
- von Mises-Fisher mixtures and applications.

Mixture models

$$\Pr(\mathbf{x}; \mathcal{M}) = \sum_{j=1}^{K} w_j f_j(\mathbf{x}; \mathbf{\Theta}_j)$$

Mixture models

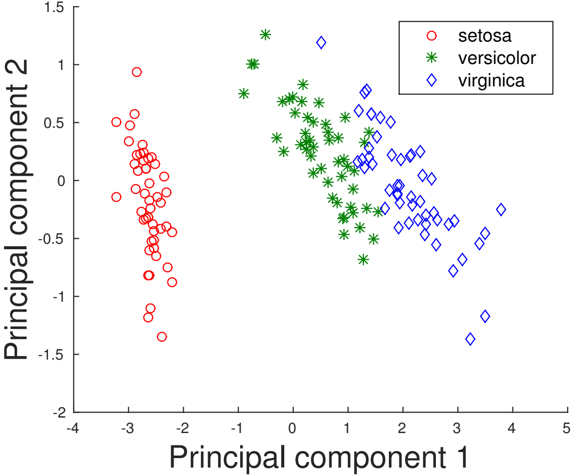
$$\Pr(\mathbf{x}; \mathcal{M}) = \sum_{j=1}^{K} w_j f_j(\mathbf{x}; \mathbf{\Theta}_j)$$

- Ubiquitously used
  - Modelling multi-modal data
- Component probability distributions of various kinds
  - Poisson, Exponential, Weibull, ...
  - multivariate Gaussian (Euclidean)
  - multivariate von Mises-Fisher (directional)

The Problem

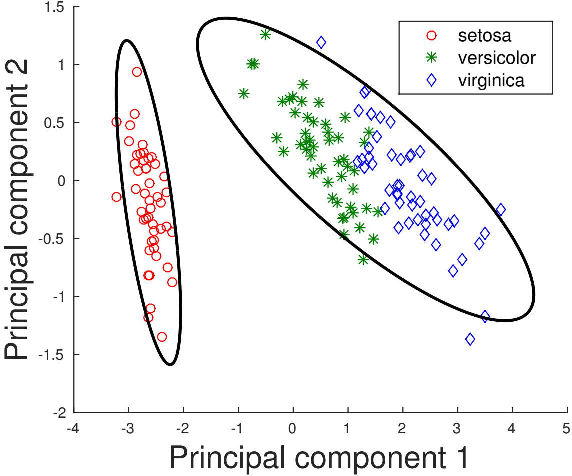
- Estimation of the parameters of the components.
  - Expectation-Maximization (EM) algorithm
- Determination of a suitable number of components
  - Objective function to compare two mixtures.

Motivation



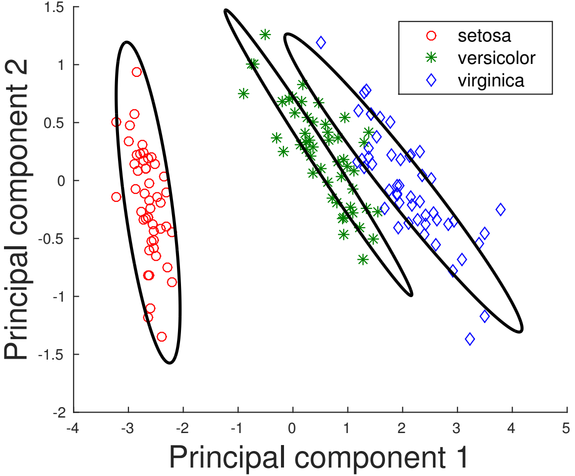
setosa
versicolor
virginica
Principal component 2
Principal component 1
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-4
-3
-2
-1
0
1
2
3
4
5

Motivation



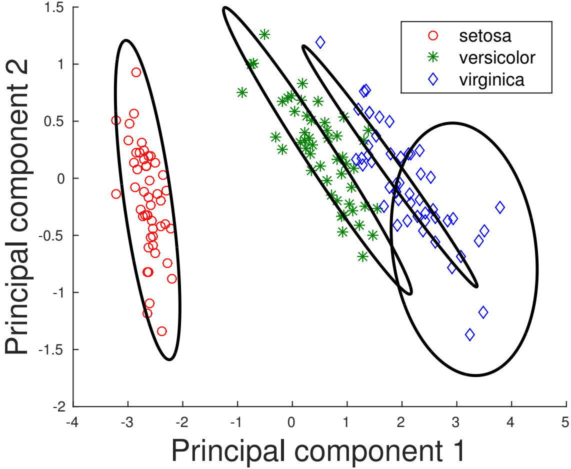
setosa
versicolor
virginica
Principal component 2
Principal component 1
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-4
-3
-2
-1
0
1
2
3
4
5

Motivation



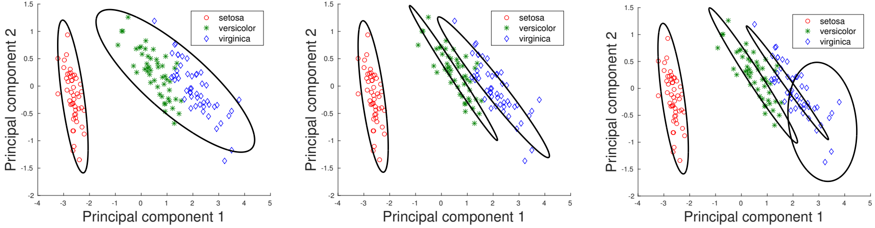
setosa
versicolor
virginica
Principal component 2
Principal component 1
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-4
-3
-2
-1
0
1
2
3
4
5

Motivation



setosa
versicolor
virginica
Principal component 2
Principal component 1
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-4
-3
-2
-1
0
1
2
3
4
5

Motivation



Statistical model selection is important.

Model selection and inference

- Several candidate models: which one to choose?
  - A criterion to compare models ...
  - Based on the model's complexity and the goodness-of-fit

Minimum Message Length (MML) Framework

Conceptualized by Wallace and Boulton (1968)

$$I(\mathcal{H}\&\mathcal{D}) = \underbrace{I(\mathcal{H})}_{\text{First part}} + \underbrace{I(\mathcal{D}|\mathcal{H})}_{\text{Second part}}$$

Minimum Message Length (MML) Framework

Conceptualized by Wallace and Boulton (1968)

$$I(\mathcal{H}\&\mathcal{D}) = \underbrace{I(\mathcal{H})}_{\text{First part}} + \underbrace{I(\mathcal{D}|\mathcal{H})}_{\text{Second part}}$$

- Two-part message:
  - $I(\mathcal{H})$: model complexity
  - $I(\mathcal{D}|\mathcal{H})$: goodness-of-fit

MML parameter estimation (Wallace and Freeman, 1987)

Single component ($\mathcal{H}$) with parameter $\boldsymbol{\Theta}_j$

$$I(\mathcal{H}\&\mathcal{D}) = I(\boldsymbol{\Theta}_j) + I(\mathcal{D}|\mathcal{H}) + \text{constant}$$

$$\text{where} \quad I(\boldsymbol{\Theta}_j) = -\log \frac{h(\boldsymbol{\Theta}_j)}{\sqrt{\mathcal{F}(\boldsymbol{\Theta}_j)}}$$

- Prior density $h(\boldsymbol{\Theta}_j)$
- *Expected* Fisher information $\mathcal{F}(\boldsymbol{\Theta}_j)$
- Negative log-likelihood $\approx I(\mathcal{D}|\mathcal{H})$

MML parameter estimation (Wallace and Freeman, 1987)

Mixture with K components ($\mathcal{H}$)

$$I(\mathcal{H}\&\mathcal{D}) = \underbrace{I(K) + I(\mathbf{w}) + \sum_{j=1}^{K} I(\mathbf{\Theta}_j)}_{\text{first part}} + I(\mathcal{D}|\mathcal{H}) + \text{constant}$$

MML parameter estimation (Wallace and Freeman, 1987)

Mixture with K components ($\mathcal{H}$)

$$I(\mathcal{H}\&\mathcal{D}) = \underbrace{I(K) + I(\mathbf{w}) + \sum_{j=1}^{K} I(\boldsymbol{\Theta}_j)}_{\text{first part}} + I(\mathcal{D}|\mathcal{H}) + \text{constant}$$

- An EM algorithm to estimate parameters ...
  - Component parameters are updated using their *MML* estimates!
- $I(\mathcal{H}\&\mathcal{D})$ is the scoring function.

Determining the number of components K

Several scoring functions ...

- AIC & BIC (Akaike, 1974; Schwarz et al., 1978)
- MDL (Rissanen, 1978)
- Approximated MML (Oliver et al., 1996; Roberts et al., 1998)
- ICL (Biernacki et al., 2000)
- MML-like (Figueiredo and Jain, 2002)

Determining the number of components K

Several scoring functions ...

- AIC & BIC (Akaike, 1974; Schwarz et al., 1978)
- MDL (Rissanen, 1978)
- Approximated MML (Oliver et al., 1996; Roberts et al., 1998)
- ICL (Biernacki et al., 2000)
- MML-like (Figueiredo and Jain, 2002)

We propose a comprehensive MML formulation with no assumptions.

Determining the number of components K

Search method: existing approaches ...

- Choose the K that has the best EM outcome.
- Figueiredo and Jain (2002) propose an improved method.
  - Begin with a large number of components.
  - Iteratively eliminate the redundant ones.
- MML-based Snob (Wallace and Boulton, 1968) ...
  - Perturb the current mixture.
  - Assumes independent assumption on the attributes.

Proposed search method

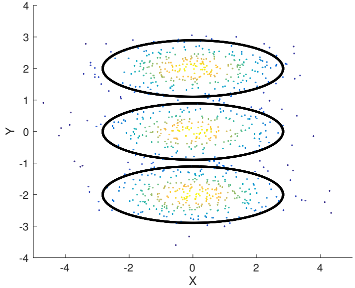
Basic idea

Perturb a K-component mixture through a series of operations so that the mixture escapes a presumably sub-optimal state to an improved state.

Operations include ...

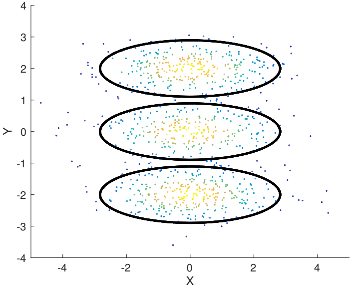
- *Split*
- *Delete*
- *Merge*

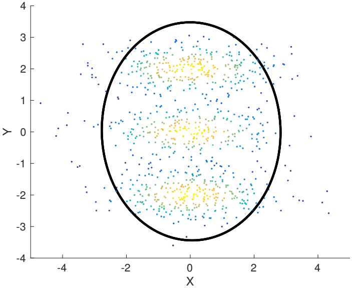
Illustrative example of the search method



Original mixture with three components.

Illustrative example of the search method

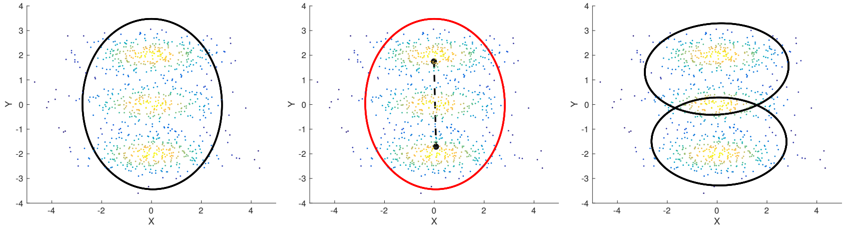




Original mixture with three components.

Begin with a one-component mixture.

Illustrative example of the search method

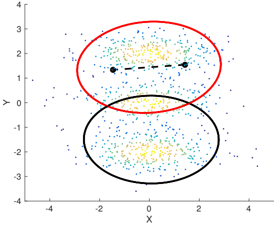


(a) $I = 22793$ bits (b) Splitting (c) $\mathbf{I = 22673}$ bits

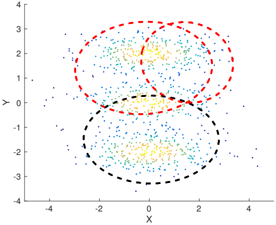
Split operation

A parent component is split to find locally optimal children leading to a $(K+1)$-component mixture.

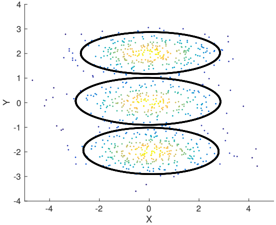
Illustrative example of the search method



(d) Initial means

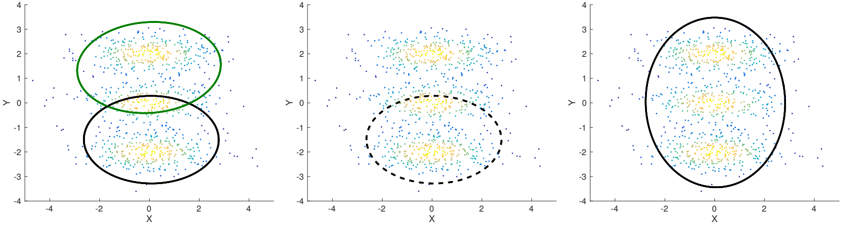


(e) $I = 22691$ bits



(f) $\mathbf{I} = \mathbf{22460}$ bits

Illustrative example of the search method

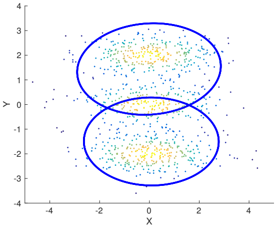


(g) Deleting (h) $I = 25599$ bits (i) $I = 22793$ bits

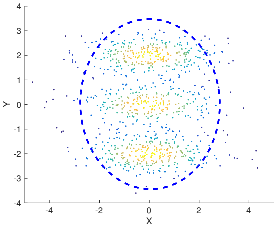
Delete operation

A component is deleted to find an optimal $(K-1)$-component mixture.

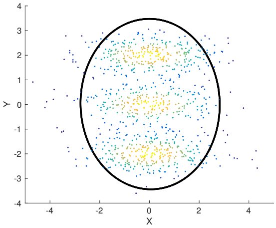
Illustrative example of the search method



(j) Merging



(k) Initialization



(l) $I = 22793$ bits

Merge operation

A pair of *close* components are merged to find an optimal $(K-1)$-component mixture.

Evolution of the mixture model

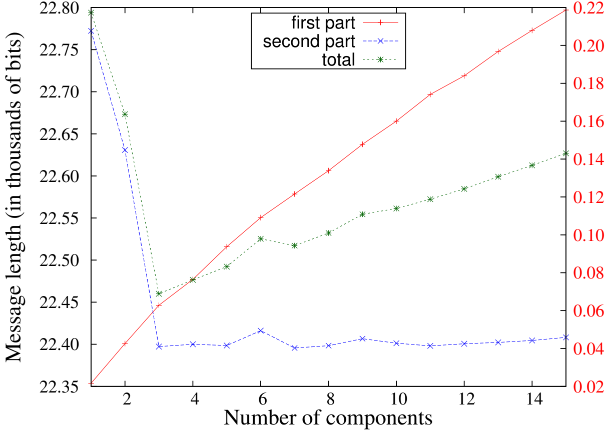
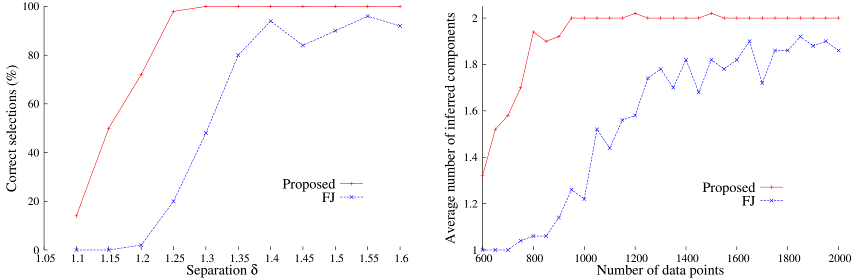


Figure: Variation of the individual parts of the total message length with increasing components.

Performance of the proposed method

Comparison with the search method of Figueiredo and Jain (2002)



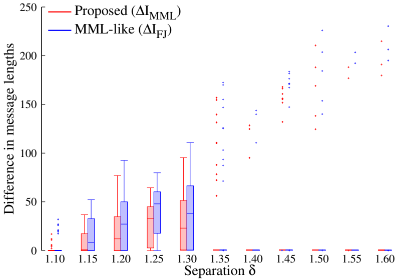
(a) (b)

Figure: 10-dimensional Gaussian mixture simulations (a) Percentage of **correct selections** with varying δ for a fixed sample size of $N = 800$ (b) **Average number** of inferred mixture components with different sample sizes and $\delta = 1.20$ between component means.

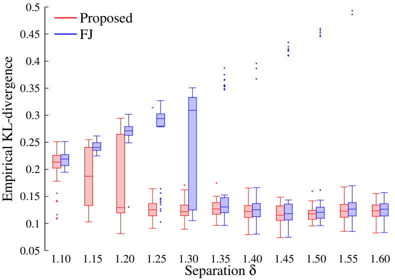
Performance of the proposed method

Comparison methodology

$$\Delta I_{MML} = I_{MML}(\mathcal{M}^{FJ}) - I_{MML}(\mathcal{M}^*) \quad \text{and} \quad \Delta I_{FJ} = I_{FJ}(\mathcal{M}^{FJ}) - I_{FJ}(\mathcal{M}^*)$$



(a)



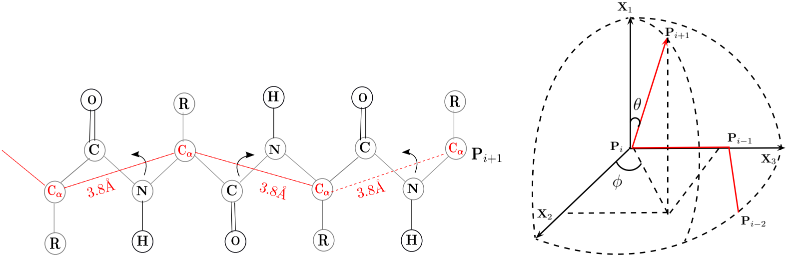
(b)

Figure: (a) **Difference in message lengths** of inferred mixtures (b) Box-whisker plot of **KL-divergence** of inferred mixtures.

Mixtures of von Mises-Fisher (vMF) distributions

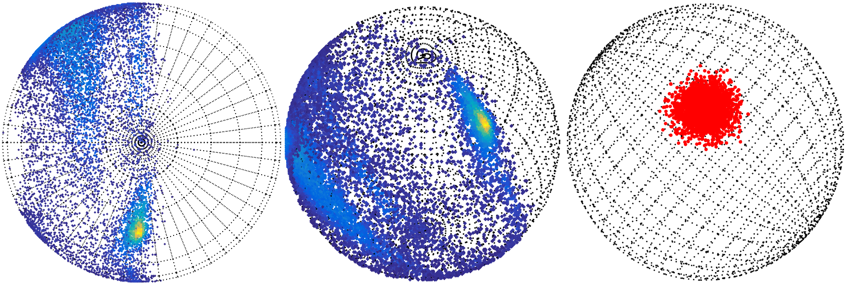
- vMF is analogous to a *symmetric* Gaussian wrapped on the hypersphere.
- Suitable for modelling directional data.
- Mixtures of vMF distributions inferred for ...
  - Describing protein data.
  - High-dimensional text clustering.

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.

Mixture modelling of protein directional data



Optimal number of vMF mixture components

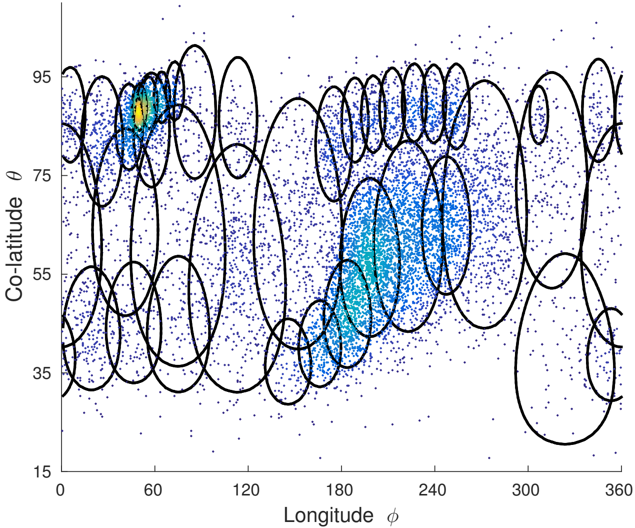


Figure: 37-component mixture

Improved descriptors of protein data

Null model	Total message length (millions of bits)	Bits per residue
Uniform	6.895	27.434
vMF mixture	**6.449**	**25.656**

Text clustering

Data corresponds to the *normalized vector representations* of text documents (Banerjee et al., 2005).

Text clustering

Data corresponds to the *normalized vector representations* of text documents (Banerjee et al., 2005).

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: Clustering performance on the two datasets: (a) Classic3 ($d = 4358$)(b) CMU_Newsgroup ($d = 6448$).

The MML mixtures *consistently* have lower message lengths.

Summary

- MML-based parameter estimation of ...
  - Multivariate Gaussian and vMF distributions
- Design of the mixture modelling apparatus ...
  - Selection of the optimal number of components.
  - Applications to modelling protein directional data and text clustering.

P. Kasarapu, L. Allison, Minimum message length estimation of mixtures of multivariate Gaussian and von Mises-Fisher distributions, *Machine Learning*, 100(2-3):333-378, 2015.

Thank you.

References I

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Dec 1974.

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.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

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.

J. J. Oliver, R. A. Baxter, and C. S. Wallace. Unsupervised learning using MML. In *Machine Learning: Proceedings of the 13th International Conference*, pages 364–372, 1996.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

S. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11): 1133–1142, Nov 1998.

G. Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

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

C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(3):240–265, 1987.