Statistical inference problems with applications to structural biology

Parthan Kasarapu

Monash University

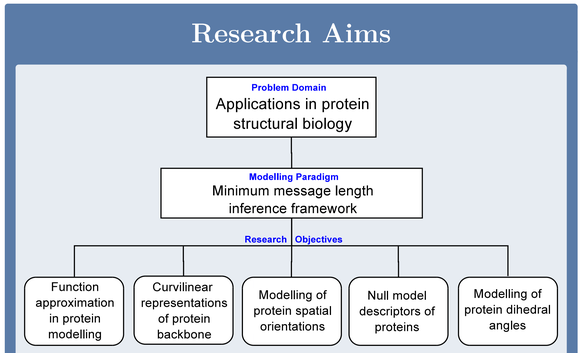
MONASH University Information Technology

NICTA

Abstract

My thesis explores the problem of model selection and inference based on the Bayesian information-theoretic principle of minimum message length (MML). The inference framework has enabled the selection of optimal models by using the constituent parameters to better balance the trade-off between the *model's complexity* and its *goodness-of-fit* to the data. This is demonstrated in the context of mixture modelling of probability distributions by developing a generic search method to determine the optimal number of mixture components that describe the given data. This modelling paradigm is explored in detail on a variety of real-world data, specifically on spatial orientation data of protein three-dimensional structures. Furthermore, the framework has been used for concise representations of protein folding patterns using a combination of non-linear parametric curves. Results of this work have a wide-variety of uses including direct applications in protein structural biology.

Research Aims



- MML-based parameter estimation of some relevant directional probability distributions amongst others.
- Design of a *general* mixture modelling apparatus to enable the optimal selection of the number of mixture components.
- *Specific* applications in protein structure modelling.

Motivation

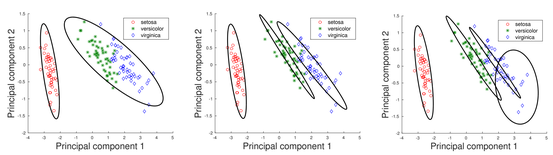


Figure 1: Into how many classes would you classify the data?

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

Minimum Message Length Framework

Inference methodology to model data $\mathcal{D}$ using a hypothesis $\mathcal{H}$

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

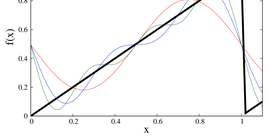
$$\text{MML criterion: } I(\mathcal{H}\,\&\,\mathcal{D}) = \underbrace{I(\mathcal{H})}_{\text{First part}} + \underbrace{I(\mathcal{D}|\mathcal{H})}_{\text{Second part}}$$

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

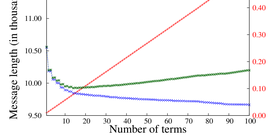
Function approximation

Increasing the number of terms decreases the error of fit at the expense of an overly complex model.

$$f(x) = \begin{cases} \dfrac{x}{T} & \text{if } x \in [0, T) \\ f(x-T) & \text{if } x \geq T \\ f(x+T) & \text{if } x < 0 \end{cases} \qquad \underbrace{f(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{2n\pi x}{T} + b_n \sin\frac{2n\pi x}{T}\right)}_{\text{Fourier decomposition}}$$



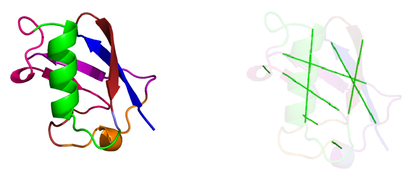
(a) Fourier approximations



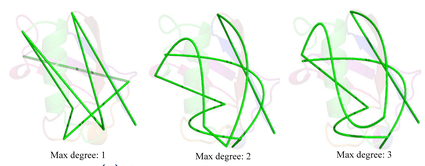
(b) Variation of the message length

Figure 2: Approximating the Fourier expansion of a *Sawtooth* function

Abstracting protein folding patterns



(a) Secondary structure representation (b) A lossy representation



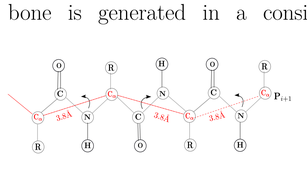
(c) Competing Bézier segmentations

An optimal segmentation achieves to maximize the economy of description and minimize the loss of structural information by minimizing the *two-part* message length

1. *First part:* Explain the segmentation
2. *Second part:* Explain the protein coordinates using the segmentation

Modelling the protein directional data

The directional data corresponding to the protein backbone is generated in a consistent canonical orientation.



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

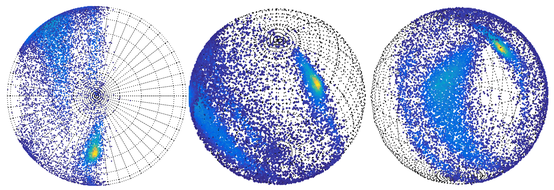


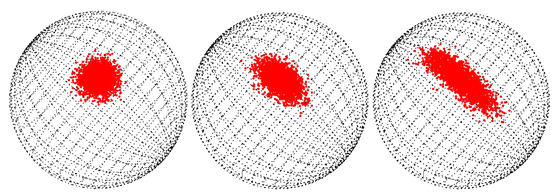
Figure 3: Empirical distribution (different orientations of the sphere).

Directional distributions defined on the surface of the sphere: von Mises-Fisher (vMF) and Kent distributions.

- vMF can model symmetrically distributed directional data.
- Kent distribution is suitable to model asymmetrical data as it has an *eccentricity* parameter controlled by β.

$$\text{Kent density} \propto \exp\{\underbrace{\kappa \boldsymbol{\gamma}_1^T \mathbf{x}}_{\text{linear term}} + \underbrace{\beta(\boldsymbol{\gamma}_2^T \mathbf{x})^2 - \beta(\boldsymbol{\gamma}_3^T \mathbf{x})^2}_{\text{non-linear term}}\}$$

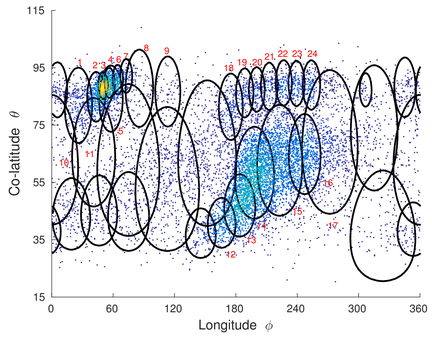
- Kent is a generalization of the vMF distribution ($\kappa > 0$ and $\beta = 0$). In comparison, for a *uniform distribution* on the sphere, $\kappa = \beta = 0$.
- Shown below are example illustrations of Kent distributions with eccentricities 0 (corresponding to a vMF), 0.5 and 0.9 respectively.



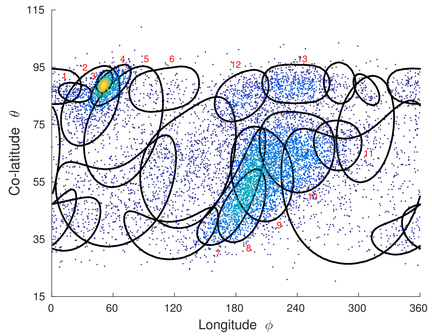
Modelling performance

Modelling distribution	Message length (in bits)	Bits per residue
Uniform	6.895×10^6	27.434
vMF mixture	6.449×10^6	25.656
Kent mixture	$\mathbf{6.442 \times 10^6}$	**25.630**

The *Kent* mixture serves as a superior null model that provides a benchmark in terms of the amount of compression to describe a database of protein structures.



(a) vMF mixture (37 components)

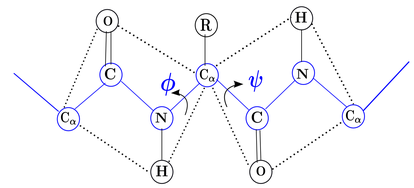


(b) Kent mixture (23 components)

Figure 4: Contours of vMF and Kent mixture components (θ and ϕ in degrees)

Modelling protein dihedral angles

The protein backbone dihedral angle pairs (ϕ, ψ) are different from the previously considered directional data.



- $\phi, \psi \in [0, 2\pi)$, and hence, *cannot* be modelled using Kent distributions.
- The dihedral angles, are therefore, modelled using mixtures of bivariate von Mises distributions defined on the surface of a torus.

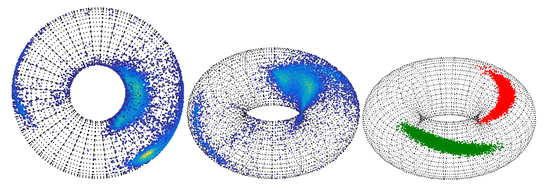


Figure 5: Shown above is the empirical distribution of the dihedral angles. An example realization of the bivariate von Mises (on the right) demonstrates the suitability to model the data using mixtures of toroidal distributions.