Bayesian Nonparametric Mixture Modeling with Unimodal Kernels
Carlos E. Rodriguez & Stephen G. Walker
School of Mathematics, Statistics and Actuarial Science, University of Kent, UK.

Abstract

Within the context of mixture modeling, the normal distribution is typically used as the components distribution. However, if a cluster is skewed or heavy tailed, then the normal distribution will be inefficient and many may be needed to model a single cluster. Here, we present an attempt to solve this problem. We define a cluster, in the absence of further information, to be a group of data which can be modeled by a unimodal density function. Hence, our intention is to use a family of distribution functions, to replace the normal, for which the only constraint is unimodality: one cluster modeled by a unimodal density function.

The difficult aspect of the Bayesian model is to construct a suitable MCMC algorithm to sample from the correct posterior distribution. The key will be the introduction of strategic latent variables and the use of the Product Space view of Reversible Jump methodology.

Keywords: Cluster analysis, Dirichlet Process, Mixture model, Slice sampler, Product space and reversible jump.

Normal Distribution

If just a density estimate is needed the use of the normal distribution is justified. However, for the modeling of clusters, it does have some serious issues: if a cluster is skewed or heavy tailed, then the normal will be inefficient and many may be needed to model a single cluster. To motivate our proposal we can cite two important works in Bayesian mixture modeling:

“The underlying assumption is that each galactic cluster is a normal component. If the distribution of a galactic cluster is skewed or has a very light or heavy tail, then we may use two or more normal components to fit one galactic cluster component.”

Escobar and West (1995)

“In each case, the high overall number of components can be related in part to the skewness of the data, two or three normals being sometimes needed to fit one skewed component”

Richardson and Green (1997)

Ideas

Our objective is to perform density estimation and clustering:

• Can not do with normal kernels: one cluster sometimes being modeled by two or more normals
• Use a components distribution for which the only constraint is unimodality: one cluster modeled by a unimodal distribution
• By necessity such distribution should be modeled nonparametrically!
• Model k explicitly

$$f_k(y) = \sum_{j=1}^{k} w_j f_j(y)$$

Unimodal Density

We use the Mixture of Dirichlet Process Model, Lo (1984) and Ferguson (1973):

$$f_k(y) = \int_k k(\theta, \lambda, \mu) \, d\theta$$

In our case:

$$k(\theta, \lambda, \mu) = U(\theta, \lambda, \mu)$$

then (2) defines a unimodal distribution which we will use as the components distribution.

Prior Predictive

Setting $f_0$ and $f_0(\theta)$ as a $\Gamma(\lambda, \alpha, \beta)$ distribution and density function (parametrized such that $E(\theta) = \alpha/\beta$), the prior predictive or prior guess can be calculated integrating out $G$ to yield

$$f(y|\mu, \lambda, \alpha, \beta) = \frac{\exp(\lambda)}{2(\alpha - 1)} \left(1 - \frac{\exp(\lambda)}{2(\alpha - 1)}\right)$$

The important point here is to understand how $\mu, \lambda, \alpha$, and $\beta$ influence the prior predictive.

Mixture of Unimodal Kernels

Then for $k$ unimodal kernels, we define

$$f_k(y) = \sum_{j=1}^{k} w_j f_j(y)$$

• There are no assumptions about the shape of the components
• $k$ has a proper meaning: the number of clusters modeled by a unimodal density

We include the joint discrete latent variables ($z_i, d_i$),

$$f(z_i, d_i|\theta_i, \phi_i) = U(\theta_i, \phi_i, \theta_i, \phi_i, \mu_i)$$

then given ($z_i, d_i$), $y_i$ is drawn from its respective uniform distribution.

Examples

To demonstrate our model, we analyzed two data sets. The Galaxy data: that consists of the velocities (in 1000 km/sec) of 82 distant galaxies diverging from our own, from six well separated conic sections of the Corona Borealis region, and the enzyme data: that concerns the distribution of enzymatic activity in the blood of an enzyme involved in the metabolism of a carcinogenic substances, among a group of 245 unrelated individuals.

References