

Comments on "Statistics of Optical Colors of KBOs and Centaur's"

Zhengyuan Zhu

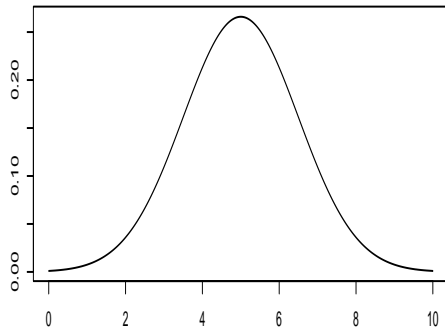
Department of Statistics and Operations
Reserach

University of North Carolina at Chapel Hill

June 2006

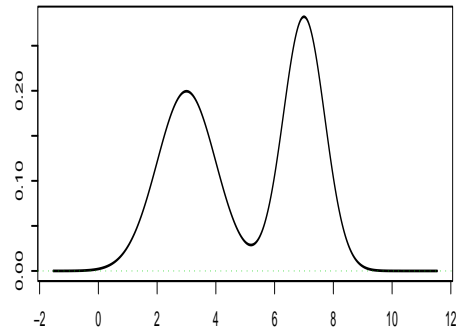
What's the dip test? (Hartigan and Hartigan (1985))

- Probability density function (PDF) $f(x)$, Cumulative distribution function (CDF) $F(x)$, and empirical CDF (ECDF) $F_n(x)$
- Unimodal CDF: convex in $(-\infty, m)$, concave in $[m, \infty)$.
- Bimodal cdf: one bump
- Let $G^* = \arg \min \sup_x |F_n(x) - G(x)|$, where $G(x)$ is a unimode CDF.
- Dip statistic: $d = \sup_x |F_n(x) - G^*(x)|$

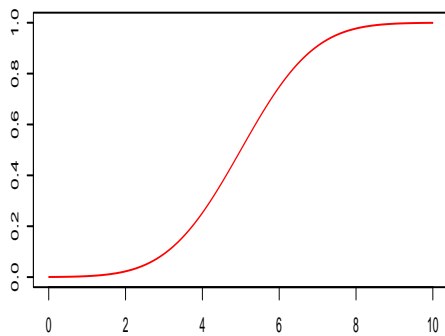


Unimode PDF

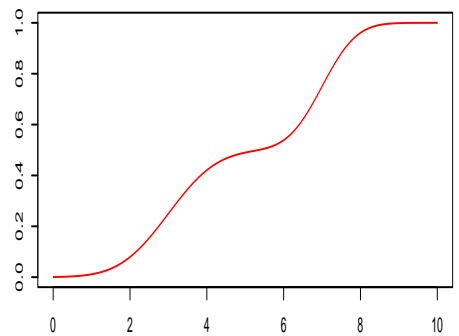
NM2.37_10.5



bimode PDF

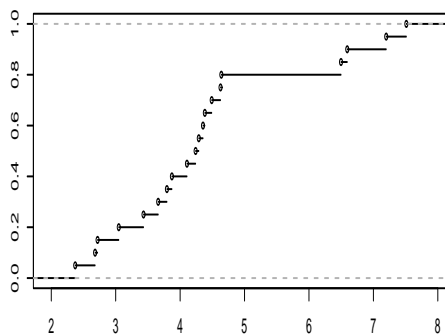


Unimode CDF



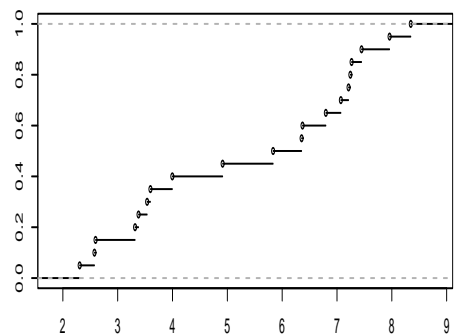
bimode CDF

`ecdf(rnorm(20, 5, 1.5))`



Unimode ECDF

`ecdf(rnorMix(20, t1))`



bimode ECDF

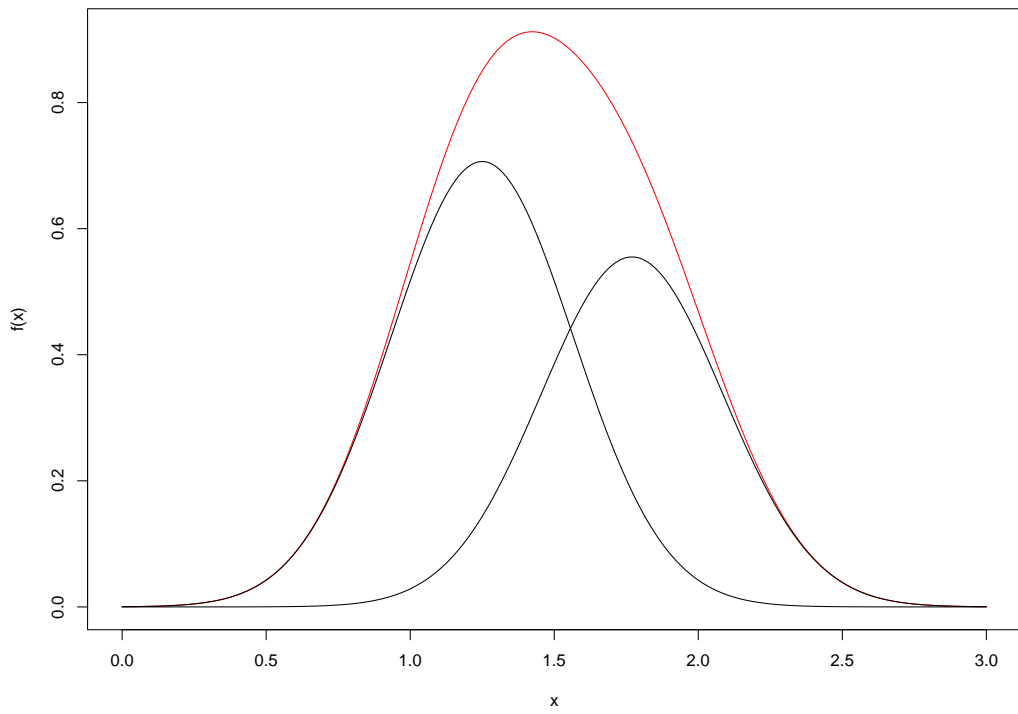
Compute G^* and distribution of d

- The Greatest Convex Minorant (GCM): imagining an elastic string bounding the function from below.
- The Least Concave Majorant (LCM): elastic string bounding the function from above.
- Theorem: G^* is a GCM on $(-\infty, x_L)$, LCM on (x_U, ∞) , and maximum constant slope on $[x_L, x_U]$
- Distribution of d : by simulation assuming F uniform, which gives the largest d among a wide class of unimodal distributions.

Bimodality or two clusters?

Dip test is for bimodality. However, mixture of two distributions does not necessarily result in a bimodal distribution.

Example: Suppose type I objects have color distribution $N(1.25, 0.1)$, type II objects have color distribution $N(1.77, 0.1)$, and among the observations 56% are type I and 44% type II. Next page shows a plot of the mixture distribution.



Facts: (Behboodian (1970)) A mixture of two normal distributions is either unimodal or bimodal. A sufficient condition to have unique mode:

$$|\mu_1 - \mu_2| \leq 2\min(\sigma_1, \sigma_2).$$

If we want to know whether data can be classified into two (or more) meaningful classes of objects, we may need to go beyond dip test.

Finite mixture models and Model-based clustering

- Assume what we observed is a mixture of several component probability distributions:

$$f(x) = \sum_{i=1}^k p_i f_i(x; \theta_i), \quad \sum p_i = 1, p_i > 0.$$

- Each component represents a cluster.
- Statistical problem: determine the number of clusters k , and estimate p_i and θ_i .

How to Compute?

- For fixed k , estimate parameters using maximum likelihood method, usually via EM algorithm.
- Choosing k can be formulated as a model selection problem using approximate Bayes factor such as BIC or other criteria.
- Trade-off between number of components and complexity of component models. Initial values.

Available software?

Lots of choices. In R, can use package `mclust` (Fraley and Raftery JASA 2002).

Here are the clustering results for color distributions of Plutinos:

```
library(mclust)
plu.clust <- Mclust(pluAll, 1,3)
plu.clust$BIC
      E      V
1 -23.66207 -23.66207
2 -19.56507 -22.56174
3 -24.88548 -29.57214
```

```
> plu.clust$sigma2
[1] 0.02045387
> plu.clust$mu
      1      2
1.251093 1.770691
$pro
[1] 0.5594534 0.4405466
```

Not shown here: probability that each object belongs to which class.

Could we use more than one variables to classify KBOs?

Clustering of multivariate data: great progress in recent years, many software available.

Early reference on finite mixture models: Pearson (1894)

books:

Titterington, Smith, and Makov, Statistical analysis of finite mixture distributions (1985)

McLachlan and Peel, Finite Mixture Models, (2002)