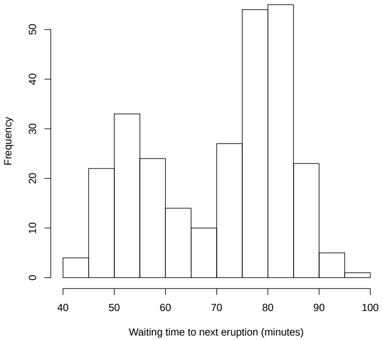
ExploRatory and statistical tools to investigate multimodality

Jose Ameijeiras–Alonso

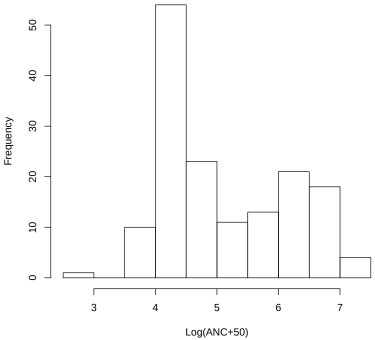
Department of Statistics and Operations Research
University of Santiago de Compostela

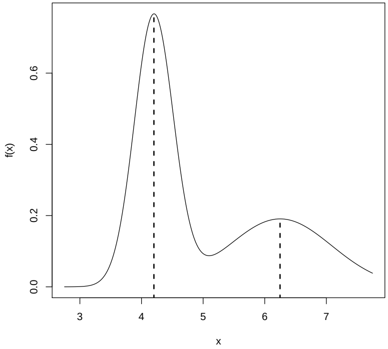


272 observations measuring the waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

`faithful {datasets}`
`hist {graphics}`

Azzalini, A. and Bowman, A. W. (1990)
A look at some data on the Old Faithful geyser.
Applied Statistics, **39**, 357–365.

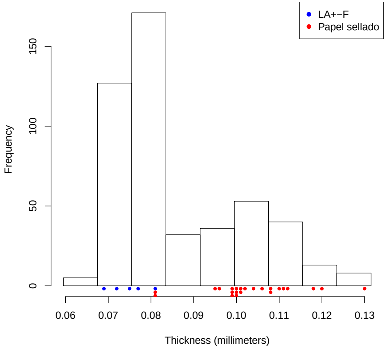


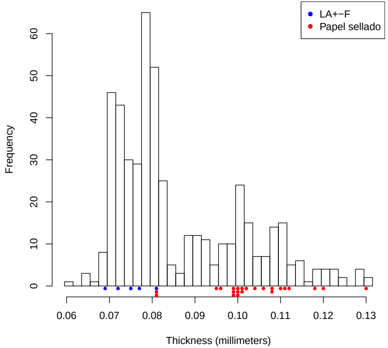


Acid–neutralizing capacity (ANC) measured in a sample of 155 lakes in North–Central Wisconsin.

`Acidity {mixAK}`
`norMix {nor1mix}`

Crawford, S. L. (1994)
An application of the Laplace method to finite mixture distributions.
Journal of the American Statistical Association, **89**, 259–267.



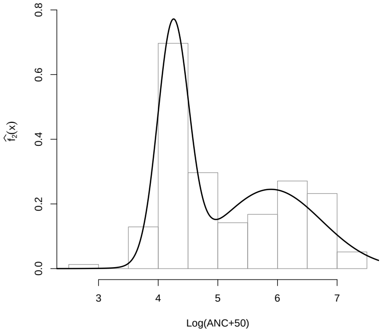


Thickness of 485 postal stamps, printed in Mexico, between 1872 and 1874 (The 1872 Hidalgo stamp issue of Mexico).

`stamp {bootstrap}`

Izenman, A. J. and Sommer, C. J. (1988)
Philatelic mixtures and multimodal densities.
Journal of the American Statistical Association, **83**, 941–953.

Mixture of M unimodal distributions, f_m:
$f_M(x) = \sum_{m=1}^{M} p_m f_m(x)$.



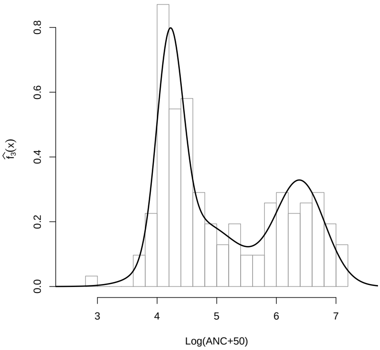
Estimation of the parameters of a univariate normal, $N(\mu_i, \sigma_i^2)$, mixture using the Likelihood Maximimization (fixing $M = 2$):
$\hat{\mu}_1 = 4.25; \hat{\mu}_2 = 5.89;$
$\hat{\sigma}_1 = 0.26; \hat{\sigma}_2 = 0.85;$
$\hat{p}_1 = 0.48; \hat{p}_2 = 0.52.$

`norMixEM {nor1mix}`

McLachlan, G. J. and Peel, D. (2000)
Finite Mixture Models.
John Wiley & Sons, Inc. New York.

Mixture of M unimodal distributions, f_m:

$$f_M(x) = \sum_{m=1}^{M} p_m f_m(x).$$



- $H_0 : M = M_0$ vs $H_a : M = M_1$ (for some $M_1 > M_0$).
- $2(\log L(\hat{f}_{M_1}) - \log L(\hat{f}_{M_0}))$.
- Bootstrap samples are generated from $\hat{f}_{M_0}$.
- P–values $(B = 100)$: $0.00(M_0 = 1)$, $0.04(M_0 = 2)$, $0.34(M_0 = 3)$.

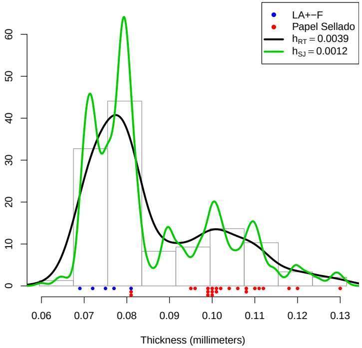
`boot.comp {mixtools}`

McLachlan, G. J. and Peel, D. (2000)
Finite Mixture Models.
John Wiley & Sons, Inc. New York.

Given a random sample $(X_1, \ldots, X_n)$ from some unknown density f, the KDE is given by:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

- K is a unimodal kernel function, e. g., $N(0,1)$.
- $h > 0$ is the smoothing parameter.



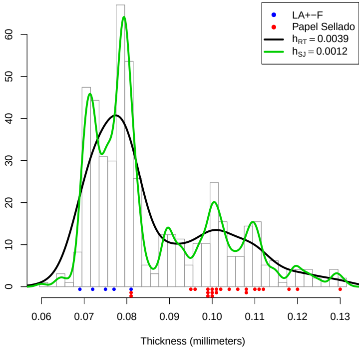

```
density {stats}
bw.SJ {stats}
```

Wand M. P. and Jones M. C. (1995)
Kernel Smoothing.
Chapman and Hall. London.

Given a random sample $(X_1, \ldots, X_n)$ from some unknown density f, the KDE is given by:

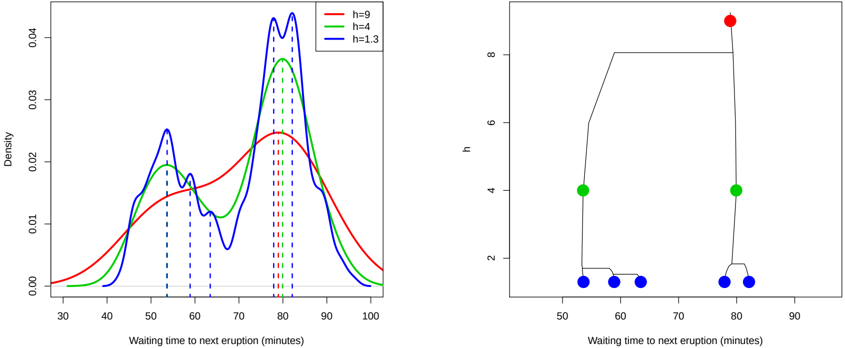
$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

- K is a unimodal kernel function, e. g., $N(0, 1)$.
- $h > 0$ is the smoothing parameter.



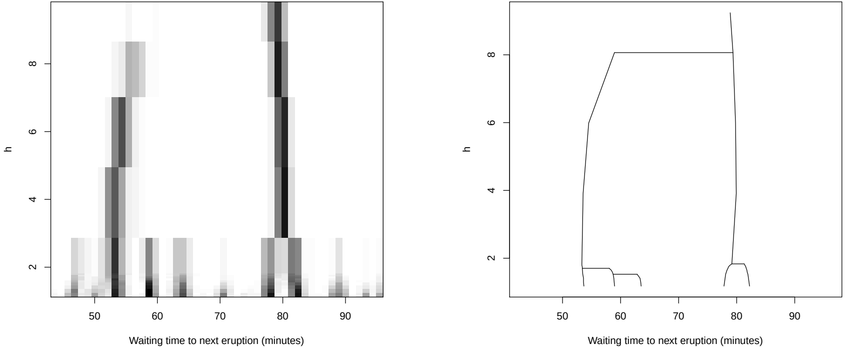

```
density {stats}
bw.SJ {stats}
```

Wand M. P. and Jones M. C. (1995)
Kernel Smoothing.
Chapman and Hall. London.



The mode tree plot relates the locations of modes in density estimates with the bandwidths used for their construction.

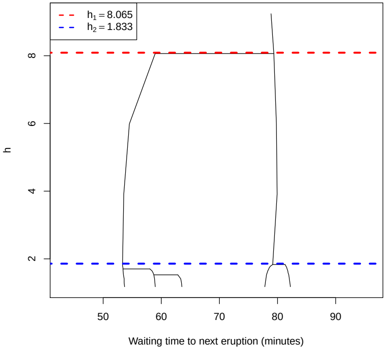
Minnotte, M. C. and Scott, D. W. (1993)
The mode tree: A tool for visualization of nonparametric density features.
Journal of Computational and Graphical Statistics, **2**, 51–68.



The mode forest looks simultaneously at a large collection of mode trees generated from the empirical distribution of the original data.

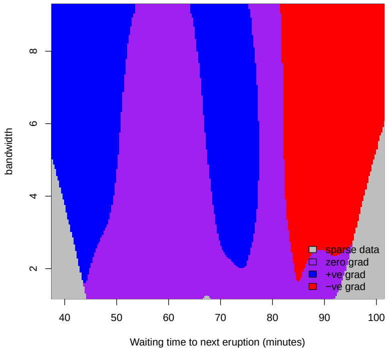
Minnotte M. C., Marchette D. J. and Wegman, E. J. (1998).
The Bumpy Road to the Mode Forest.
Journal of Computational and Graphical Statistics, **7**, 239–251.

- $H_0 : j \leq k$ vs $H_a : j > k$, where j is the real number of modes.
- $h_k = \min\{h : \hat{f}_h$ has at most k modes$\}$.
- Bootstrap samples are generated from $\hat{f}_{h_k}$.
- Reject if $h_k < Q_\alpha(h_k^*)$.
- P–values $(B = 500)$: $0.006(k = 1)$; $0.820(k = 2)$.



Silverman, B. W. (1981).
Using kernel density estimates to investigate multimodality.
Journal of the Royal Statistical Society. Series B, **43**, 97–99.

For each pair (x, h), with $h > 0$, the SiZer computes the confidence interval for $\hat{f}'_h$ (with $\alpha = 0.05$).



If the interval:

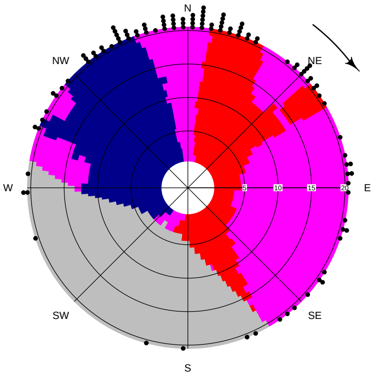
- is above zero, the smoothed curve is significantly increasing (blue).
- is below zero, the smoothed curve is significantly decreasing (red).
- contains zero, the derivative is not significantly different from zero (purple).

If there is not enough data (gray).

`SiZer {feature}`

Chaudhuri, P. and Marron, J. S. (1999).
SiZer for Exploration of Structures in Curves.
Journal of the American Statistical Association, **94**, 807–823.

For each pair (x, h), with $h > 0$, the SiZer computes the confidence interval for $\hat{f}_h'$ (with $\alpha = 0.05$).



If the interval:

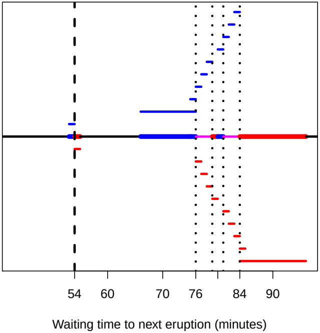
- is above zero, the smoothed curve is significantly increasing (blue).
- is below zero, the smoothed curve is significantly decreasing (red).
- contains zero, the derivative is not significantly different from zero (purple).

If there is not enough data (gray).

`circsizer.density {NPCirc}`

Oliveira, M., Crujeiras, R.M. and Rodríguez–Casal (2014).
CircSiZer: an exploratory tool for circular data.
Environmental and Ecological Statistics,**21**, 143–159.

Simultaneous confidence statements for the existence and location of local increases and decreases of a density f.

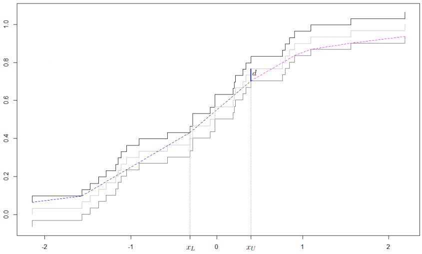


- It only depends on the ordered sample.
- At least one of the extrema of the interval must be known (and finite).
- It does not allow repeated data.

`modeHunting {modehunt}`

Dümbgen, L. and Walther, G. (2008).
Multiscale Inference about a density.
The Annals of Statistics, **36**, 1758–1785.

The dip test measures multimodality in a sample by the maximum difference, over all sample points, between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference.

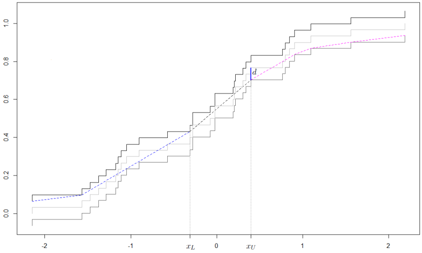


- Under the assumption that the distribution is unimodal, it generates a modal interval (x_L, x_U).
- In the example of waiting time, the modal interval is $(73, 86)$.

`dip {diptest}`
`dip.test {diptest}`

Hartigan, J. A. and Hartigan, P. M. (1985).
The Dip Test of Unimodality.
Journal of the American Statistical Association,**86**, 738–746.

The dip test measures multimodality in a sample by the maximum difference, over all sample points, between the empirical distribution function, and the unimodal distribution function that minimizes that maximum difference.



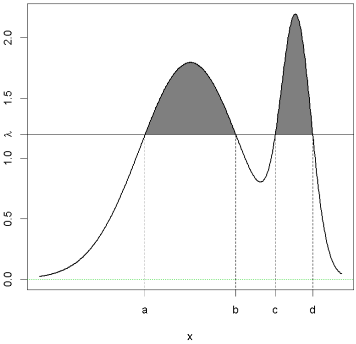
- Test: Resamples are generated from the uniform distribution.
- Reject if $d(\mathcal{X}) < Q_\alpha(d(\mathcal{U}^*))$.
- In the example of waiting time, the p–value $(B = 2000)$ is 0.001.

```
dip {diptest}
dip.test {diptest}
```

Hartigan, J. A. and Hartigan, P. M. (1985).
The Dip Test of Unimodality.
Journal of the American Statistical Association, **86**, 738–746.

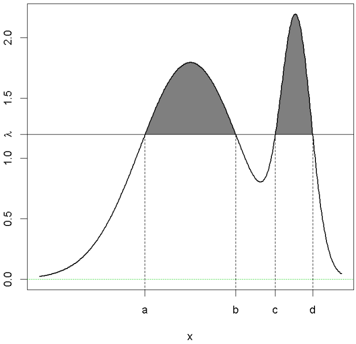
Under the assumption that f has (at most) k modes, excess mass can be empirically estimated by

$$E_{n,k}(P_n,\lambda) = \sup_{C_1,\ldots,C_k}\left\{\sum_{l=1}^{k} P_n(C_l) - \lambda||C_l||\right\},$$



- $D_{n,k}(\lambda) = E_{n,k+1}(P_n,\lambda) - E_{n,k}(P_n,\lambda)$.
- $\Delta_{n,k} = \max_{\lambda}\{D_{n,k}(\lambda)\}$.
- Our proposal: Resamples generated from a modified $\hat{f}_{h_k}$.
- In the example of waiting time, the p–values ($B = 500$) are: $0(k=1)$, $0.214(k=2)$.

Müller, D. W. and Sawitzki, G. (1991)
Excess mass estimates and tests for multimodality
The Annals of Statistics, **13**, 70–84.



- $D_{n,k}(\lambda) = E_{n,k+1}(P_n, \lambda) - E_{n,k}(P_n, \lambda)$.
- $\Delta_{n,k} = \max_{\lambda}\{D_{n,k}(\lambda)\}$.
- Our proposal: Resamples generated from a modified $\hat{f}_{h_k}$.
- In the example of waiting time, the p–values ($B = 500$) are: $0 (k = 1)$, $0.214 (k = 2)$.

Müller, D. W. and Sawitzki, G. (1991)
Excess mass estimates and tests for multimodality
The Annals of Statistics, **13**, 70–84.

ExploRatory and statistical tools to investigate multimodality

Jose Ameijeiras–Alonso

Department of Statistics and Operations Research
University of Santiago de Compostela