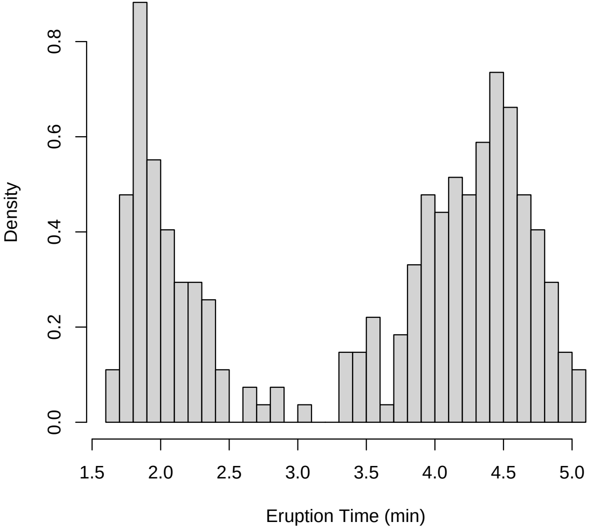
STATS 769

Density Estimation

Yong Wang
Department of Statistics
The University of Auckland

2021

Old Faithful Geyser: Eruption Time



Density Estimation

- For such data, it is quite clear that a simple distribution family, such as normal, cannot provide a good fit.
- We'd like to use a family of distributions that is flexible and adaptive to an arbitrary data set.
- They'd better be continuous, thus having density functions.
- Such problems are known as (nonparametric) density estimation.
- Unlike regression or classification problems, we don't have a response variable for density estimation.

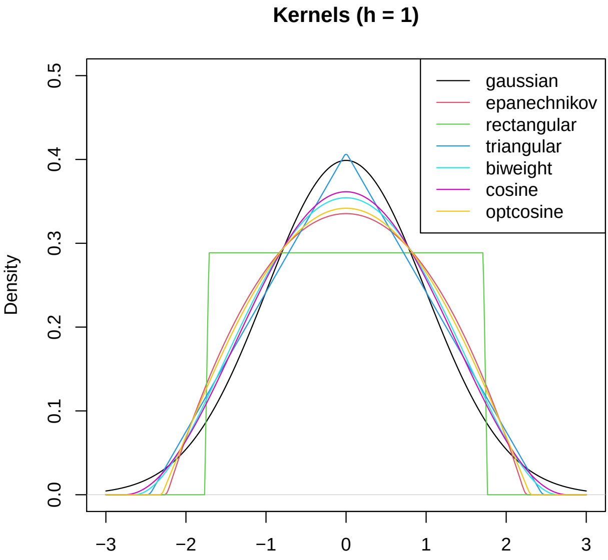
Kernels

- Usually, a kernel function $K(x)$ is a density function that is symmetric about 0 and has a unit variance.
  - For example, the standard normal (Gaussian) density.
- It plays the role of providing weights to observations in the neighbourhood of a point.
- The "size" of the neighbourhood is controlled by a smoothing parameter h, known as the bandwidth.
  - Often h is a scaling parameter, such as the standard deviation.
- Let's denote

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right).$$

- If $K(x)$ is the standard normal density, then $K_h(x)$ is the normal density with mean 0 and variance h^2.

Some Kernels



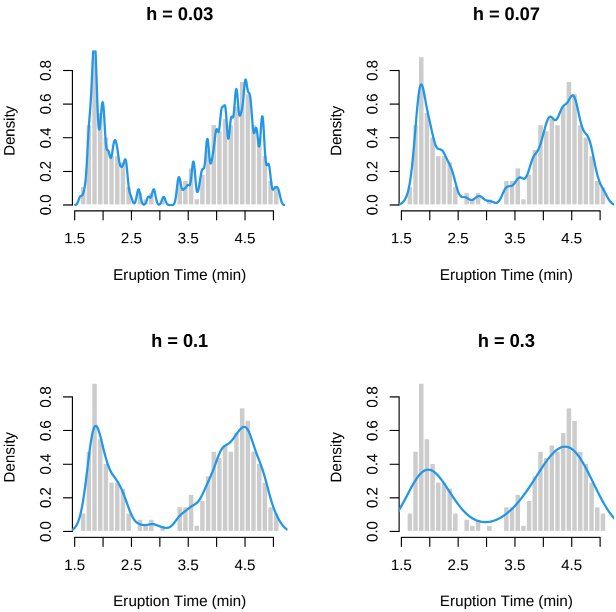
Kernel Density Estimation

- Denote the true density function by f.
- Given a random sample $x_1, \ldots, x_n$, the kernel density estimator (KDE) of f is defined as

$$\widehat{f}^{\mathrm{kde}}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i).$$

- The bandwidth h controls the smoothness of the estimate density estimate.
  - Too small a value for h gives an undersmoothed/overfitted density estimate.
  - Too large a value for h gives an oversmoothed/underfitted density estimate.
- It is critically important to choose an appropriate value for h.
- The choice of a kernel is considered less important, and the normal kernel is the most widely used.

Controlling the Smoothness



Prediction Error for Density Estimation

- The commonly-used prediction error for a density estimate $\widehat{f}$ is the integrated squared error:

$$\mathrm{ISE}(\widehat{f}; f) = \int [\widehat{f}(x) - f(x)]^2 \, \mathrm{d}x.$$

- For a random sample $x_1, \ldots, x_n$, it can be estimated by

$$\widehat{\mathrm{ISE}}(\widehat{f}) = \int [\widehat{f}(x)]^2 \, \mathrm{d}x - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}(x_i)$$

  (plus a constant).

- With the PE available, one can use cross-validation, say, to find an appropriate value for the bandwidth h.

Mixture Models

- If one simple distribution can not fit well to the data, maybe we can consider their combination.
- A mixture distribution/model is a (special) linear combination of distributions.
- A mixture density with m components has the form

$$f(x) = \pi_1 f_1(x) + \cdots + \pi_K f_K(x),$$

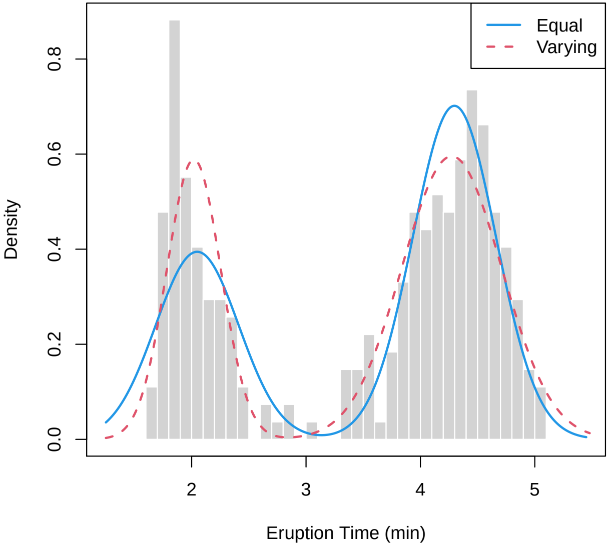
  where each f_j is a component density and is associated with a mixing proportion π_j, subject to $\pi_j > 0$ and $\sum_{j=1}^K \pi_j = 1$.
- Each f_j is usually a simple density function, e.g., a normal density as given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

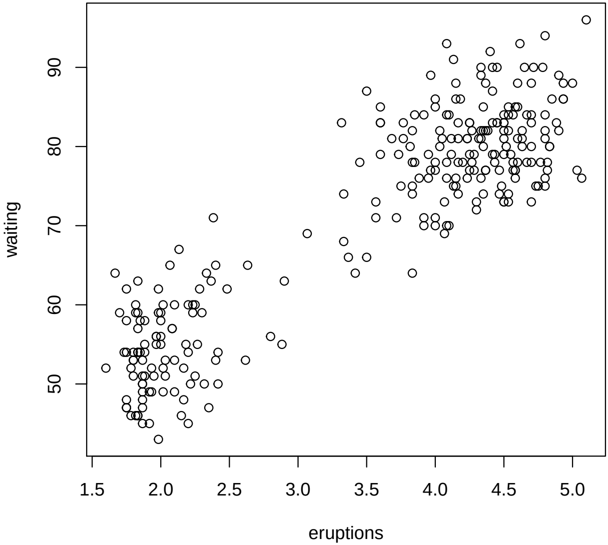
Types of Normal Mixtures

- A normal mixture has normal distributions as its components.
- There are two main subfamilies of normal mixtures:
  - Equal variances (homoscadestic): All normal components share an identical variance σ^2 (but may have different means μ_j).
  - Varying variances (heteroscadestic): Each normal component has its own variance σ_j^2 (and mean μ_j).
- A normal mixture can be fitted by maximum likelihood, typically via the Expectation-Maximisation (EM) algorithm.

Two-component Mixture: Equal *vs.* Varying Variances



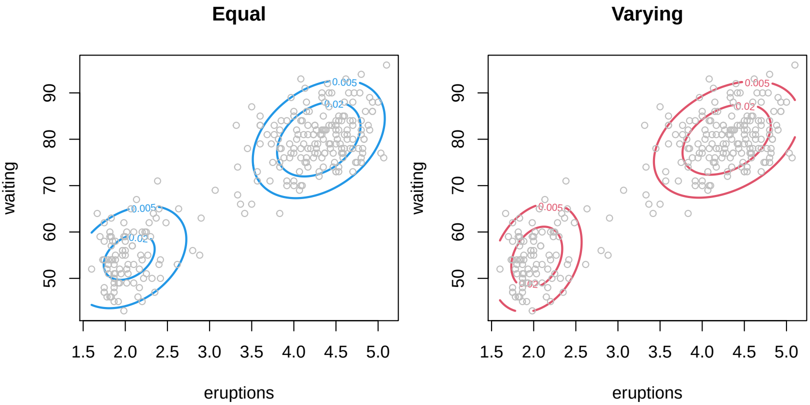
Old Faithful Geyser: Bivariate Data



Multivariate Normal Mixtures

- For multivariate data in $\mathbb{R}^d$, we can consider using a multivariate mixture that has multivariate normal components.
- There are two main subfamilies of multivariate normal mixtures:
  - Equal variances (homoscadestic): All normal component distributions share an identical variance-covariance matrix Σ (but may have different means μ_j).
  - Varying variances (heteroscadestic): Each normal component distribution has its own variance-covariance matrix Σ_j (and mean μ_j).
- There are other subfamilies for various restrictions on the $d \times d$ variance-covariance matrix, by decomposing it into several factors.
- The EM algorithm is almost the only computational tool for maximum likelihood estimation of a multivariate normal mixture.

Equal *vs.* Varying Σ



Recommended Readings

ISLv2 (basics):

- (No relevant section found)

ESL (advanced):

- Sections 6.6.1, 6.8, 8.5.1

Other Books:

- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.
- McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. John Wiley & Sons.