Case Study of Mixture Models

Rebecca C. Steorts

```r
set.seed(1234)
library(mixtools)
```

```
## mixtools package, version 2.0.0, Released 2022-12-04
## This package is based upon work supported by the Nationa
```

```r
knitr::opts_chunk$set(fig.width = 8,
                      fig.asp = 0.618,
                      fig.retina = 3,
                      out.width = "90%",
                      fig.align = "center")
```

Agenda

Case Study on Mixture Models

This has been adapted from `Advanced Data Analysis from an Elementary Point of View` Chapter 19 by Cosma Shalizi (publicly available online).

Data

Consider daily records of precipitation at Snoqualmie Falls, Washington from 1948 to the end of 1983. [1].

[1] The data set can be found at https://sites.stat.washington.edu/peter/stoch.mod.data.html

Data

- Each row of each data file is a different year; each column records, for that day of the year, the day's precipitation (rain or snow), in units of 1/100 inch.
- Due to leap-days, there are 366 columns, where the last column has an NA value for three out of four years.

Rainy days and such

Consider the distribution of the amount of precipitation on the wet days, such as rain, snow, hail, etc.

Goal: What is the distribution of the amount of precipitation on the wet days?

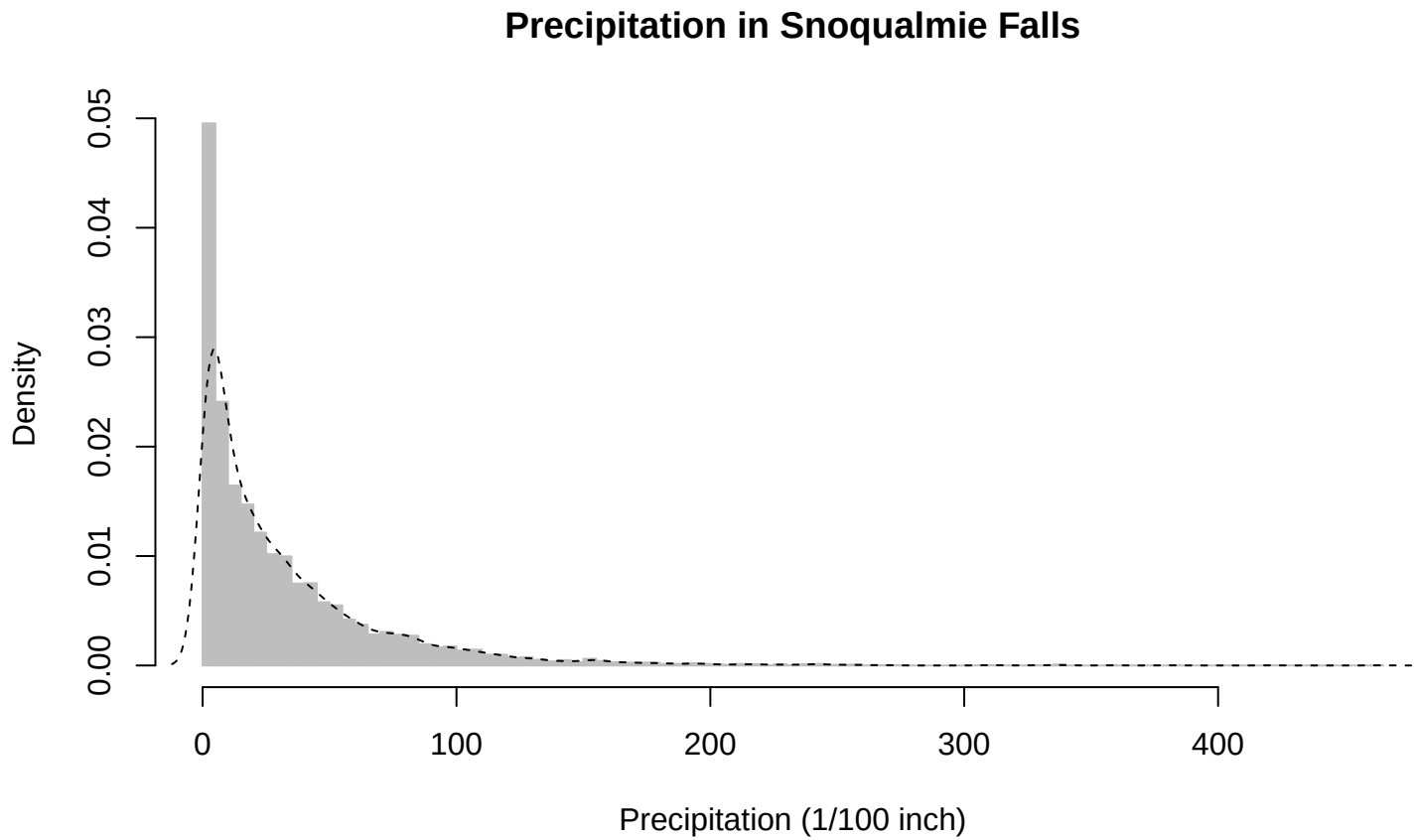
```
# skip the first line, a header
snoqualmie <- scan("http://www.stat.washington.edu/peter/bo
snoq <- snoqualmie[snoqualmie > 0]
```

Explore the data

Let's perform an exploratory data analysis of the data to understand it a bit, where we consider a histogram and then consider a kernel density estimator.

Kernel density estimation is a statistical tool that attempts to estimate the true shape of a distribution by smoothing out the existing data. Let's look at an example for this application.

EDA


Precipitation in Snoqualmie Falls
Density
0.00
0.01
0.02
0.03
0.04
0.05
0
100
200
300
400
Precipitation (1/100 inch)

KDE

What is problematic regarding the kernel density estimator?

Hint: Look at the x-axis very carefully.

Mixture models

Could we consider a two-component mixture model? Explain why or why not.

Two component mixture model

```
snoq.k2 <- normalmixEM(snoq,k=2,maxit=100,epsilon=0.01)
```

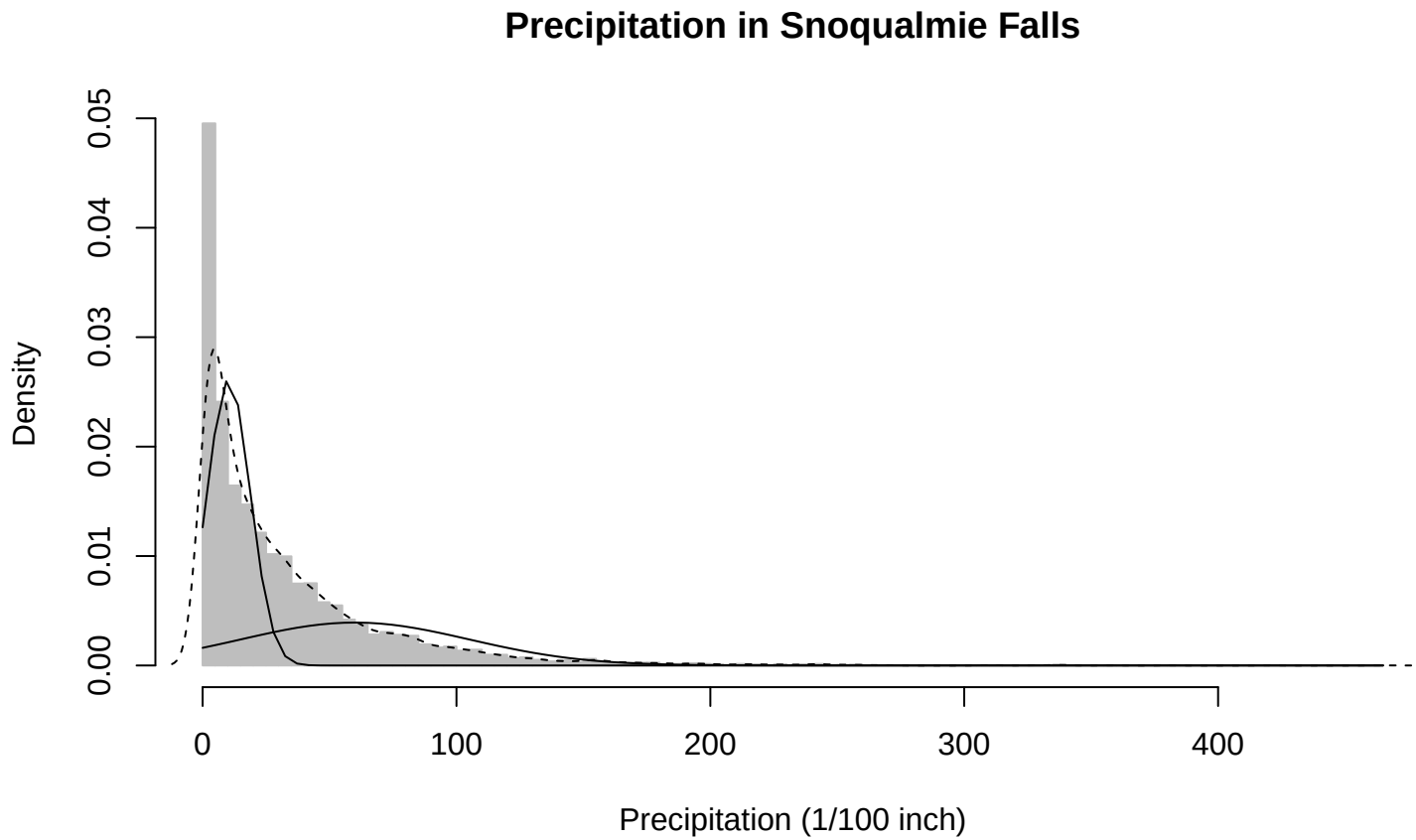
```
## number of iterations= 81
```

```
summary(snoq.k2)
```

```
## summary of normalmixEM object:
##           comp 1    comp 2
## lambda  0.557535  0.442465
## mu     10.266508 60.010330
## sigma   8.510558 44.997478
## loglik at estimate:  -32681.21
```

Interpret the output of the package.

Two component mixture model



Two component mixture model

Visually, how is the fit?

Can we be more rigorous?

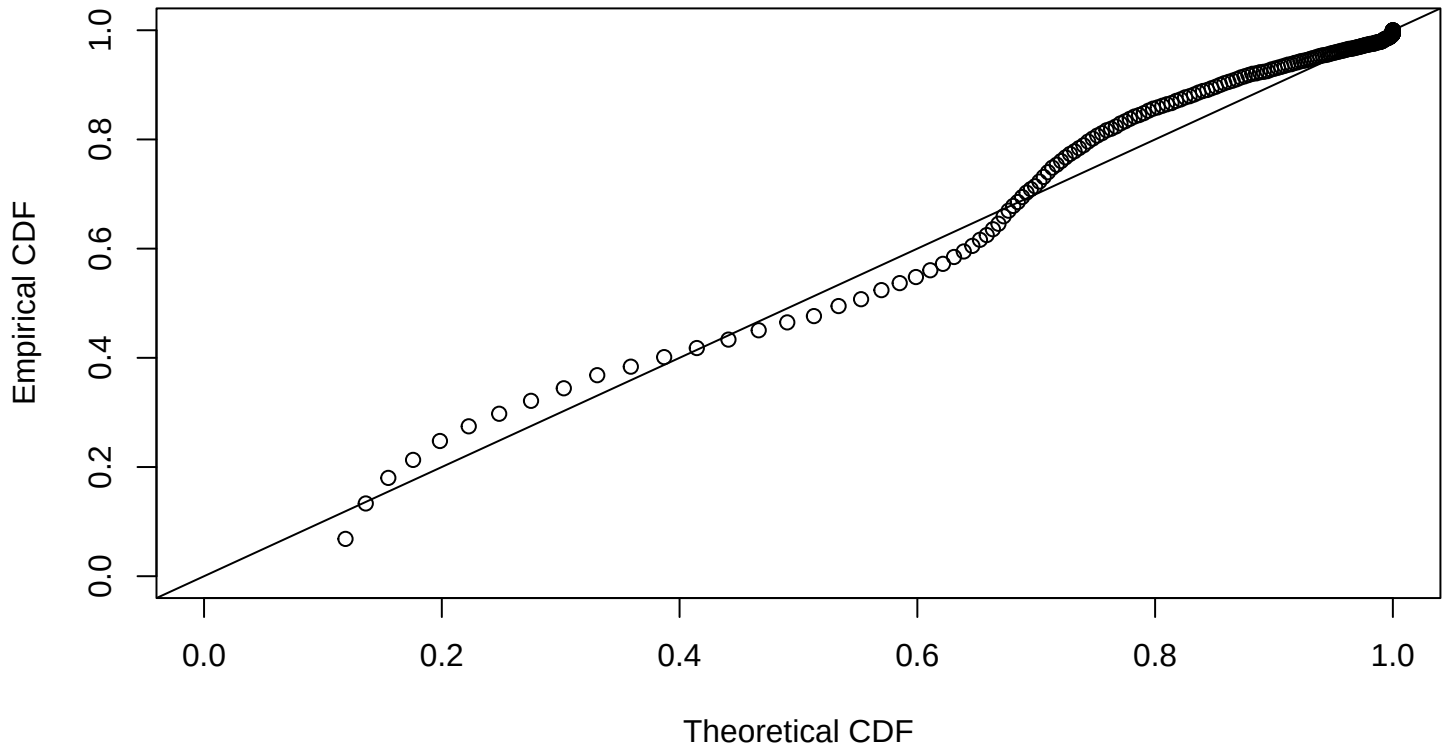
Let's assess the two-component mixture model more rigorously using a calibration plot.

Plots the empirical CDF versus theoretical CDF.

Specifically, for each distinct value of precipitation x, we plot the fraction of days predicted by the mixture model to $\leq x$ precipitation on the horizontal axis, versus the actual fraction of days $\leq x$.[2]

[2] If you'd like to see more mathematical rigor behind why this works, see Chapter 19!

Calibration Plot



We would like to see that our observations are typically pretty well centered on the straight line.

Instead, they are pretty noisy, which matches our visual intuition as well.

What about more clusters?

- We could use more clusters? Thoughts on what we could try next?

Next steps

We will try looking at more clusters and selecting these using cross validation.

Cross Validation

Cross validation (CV) evaluates the model performance on unseen data.

We split given data into folds or subsets, where we:

1. use one of the folds as a validation set and the remaining folds to train our model.
2. We repeat the process many times, each time using a different fold as the validation set.
3. We then average the results from each validation set to produce a more robust estimate of the model's performance.

We do this to try and avoid over-fitting. For more information about CV, see this post (https://www.geeksforgeeks.org/cross-validation-machine-learning/) for different types of CV.

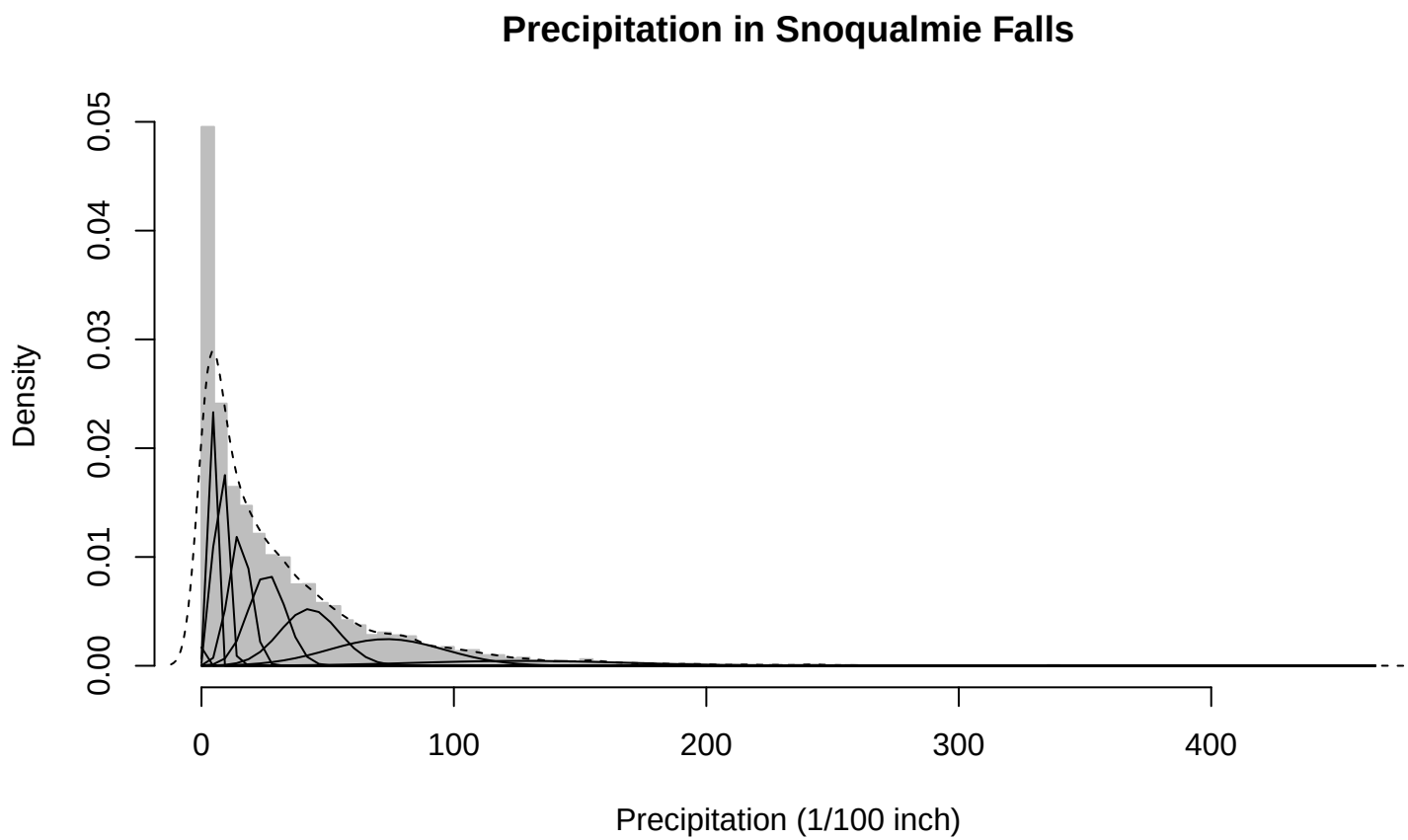
Selecting the Number of Clusters by Cross-Validation

For illustrative purposes, we will do a random split of the data for cross validation.

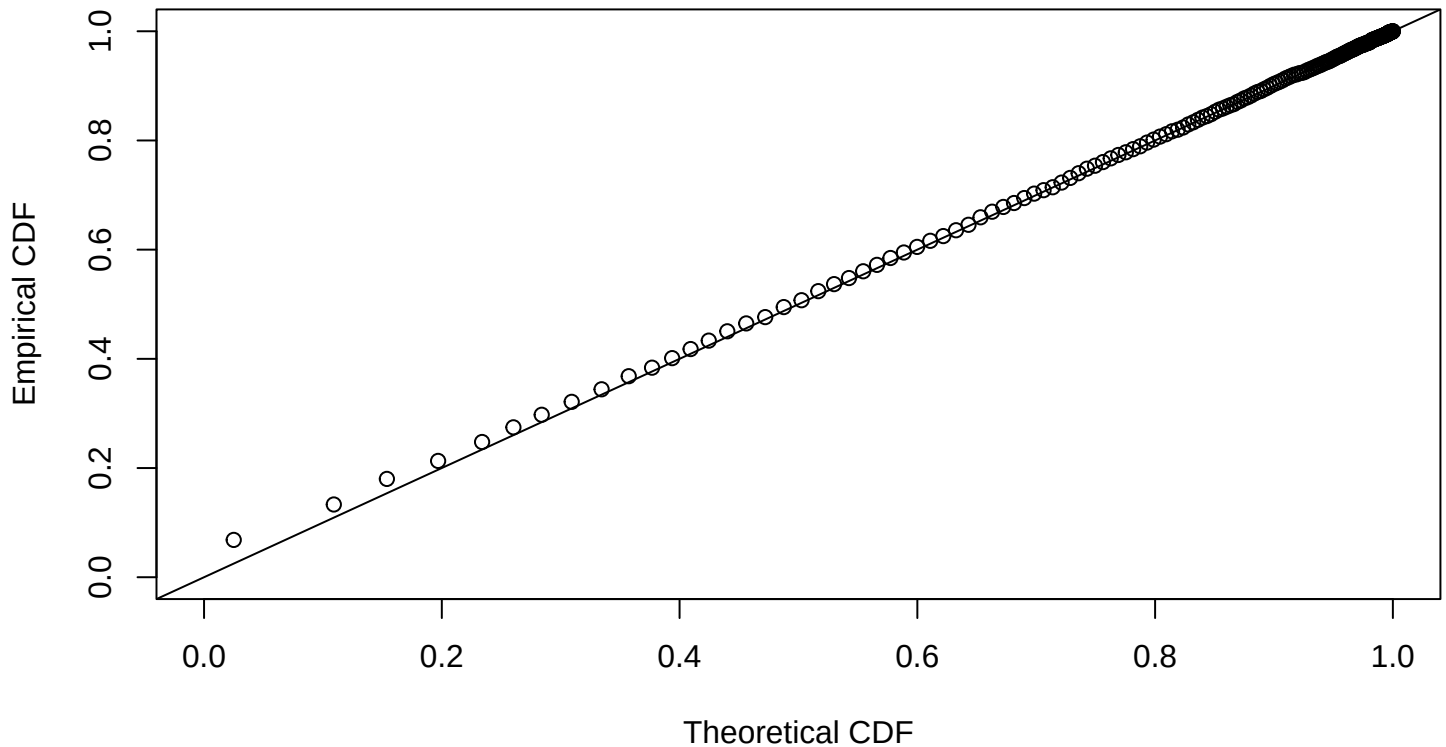
```
## number of iterations= 46
## number of iterations= 16
## number of iterations= 94
## number of iterations= 85
## number of iterations= 343
## number of iterations= 188
## number of iterations= 75
## number of iterations= 105
## One of the variances is going to zero;  trying new start
## number of iterations= 128
```

Mixture Model

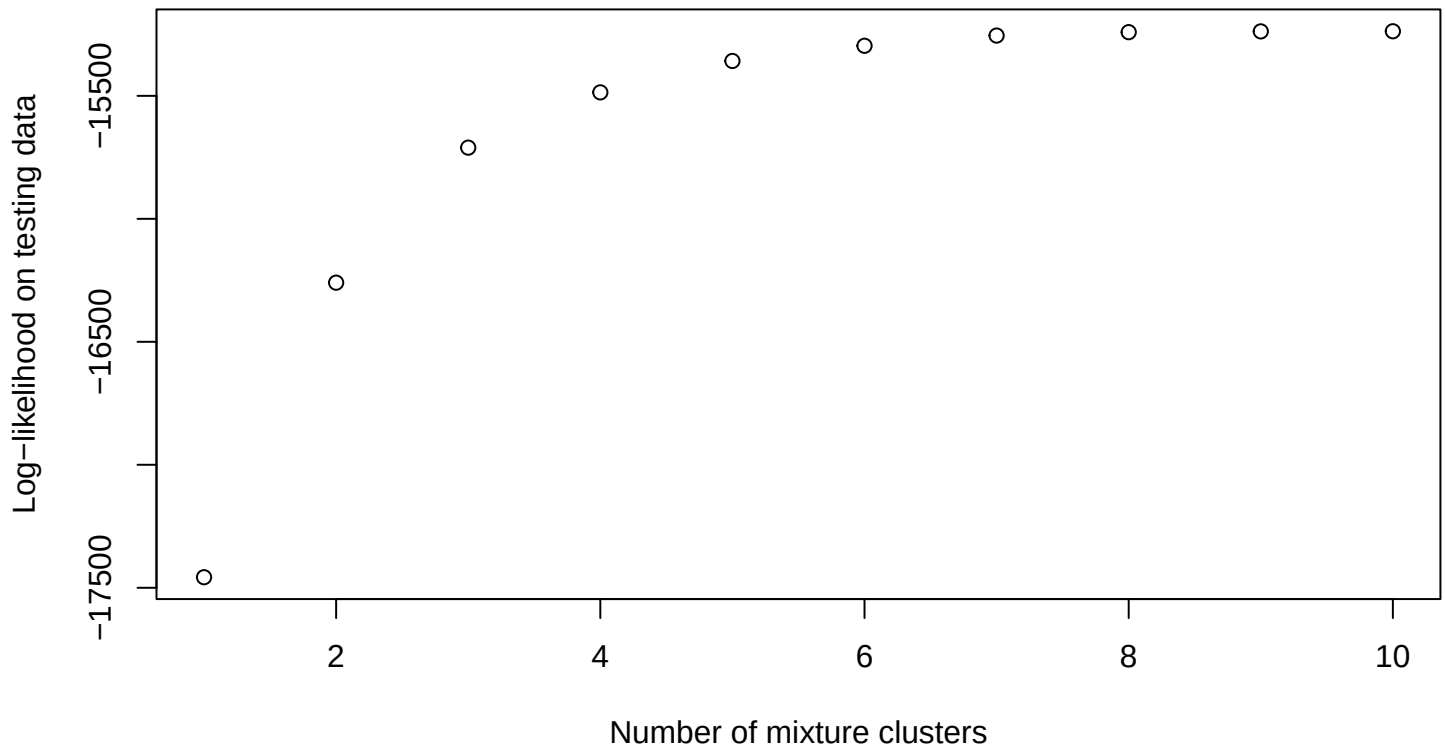
```
## number of iterations= 192
```



Calibration Plot



Log-Likelihood



Digging Deeper

Are there really nine types of rainy or wet days?

Two ways forward:

1. Statistical.
2. Substantive: Checking, digging into the data at hand or relying on what we know about weather.

Statistical

Suppose we care about only caring about describing the distribution of the data and predicting future precipitation.

If this is **all** we care about, then it doesn't matter whether the nine-cluster mixture is true.

Cross validation did not choose the model because there are truely nine types of wet days. It picked this model based on the best trade-off between estimated bias and variance.

Substantive

For this particular problem, nine types of wet days seems strange, but perhaps one could check with a weather expert if this was crucial to their research.

Digging into the data doesn't lead to anything revealing or helpful.

Alternatives

Instead of cross validation, we could select the number of clusters using a hypothesis test (or a likelihood ratio test).

The LRT has an approximate chi-squared distribution. Given that we are using the EM algorithm, this is not ideal.

One alternative is using a parametric bootstrap, which simulates data and estimates the sampling distribution of the parameter estimates.

This uses the `boot.comp` function in the `mixtools` package.

Alternatives

Consider the type-II generalized Pareto distribution, where

$$p(x) \propto (1 + x/\sigma)^{-\theta-1}.$$

Specifically, this can be written as a two-step process as follows:

1. $X \mid Z \mid Exp(\sigma/Z)$, $p(X \mid Z) = \sigma e^{-\sigma x}$.
2. $Z \mid \Gamma(shape = \theta, rate = 1)$.

See Arnold (1983), Macquire et al. (1952).

Assignment

Investigate a mixture of exponentials for the Snoqualmie Falls data set and report your findings.