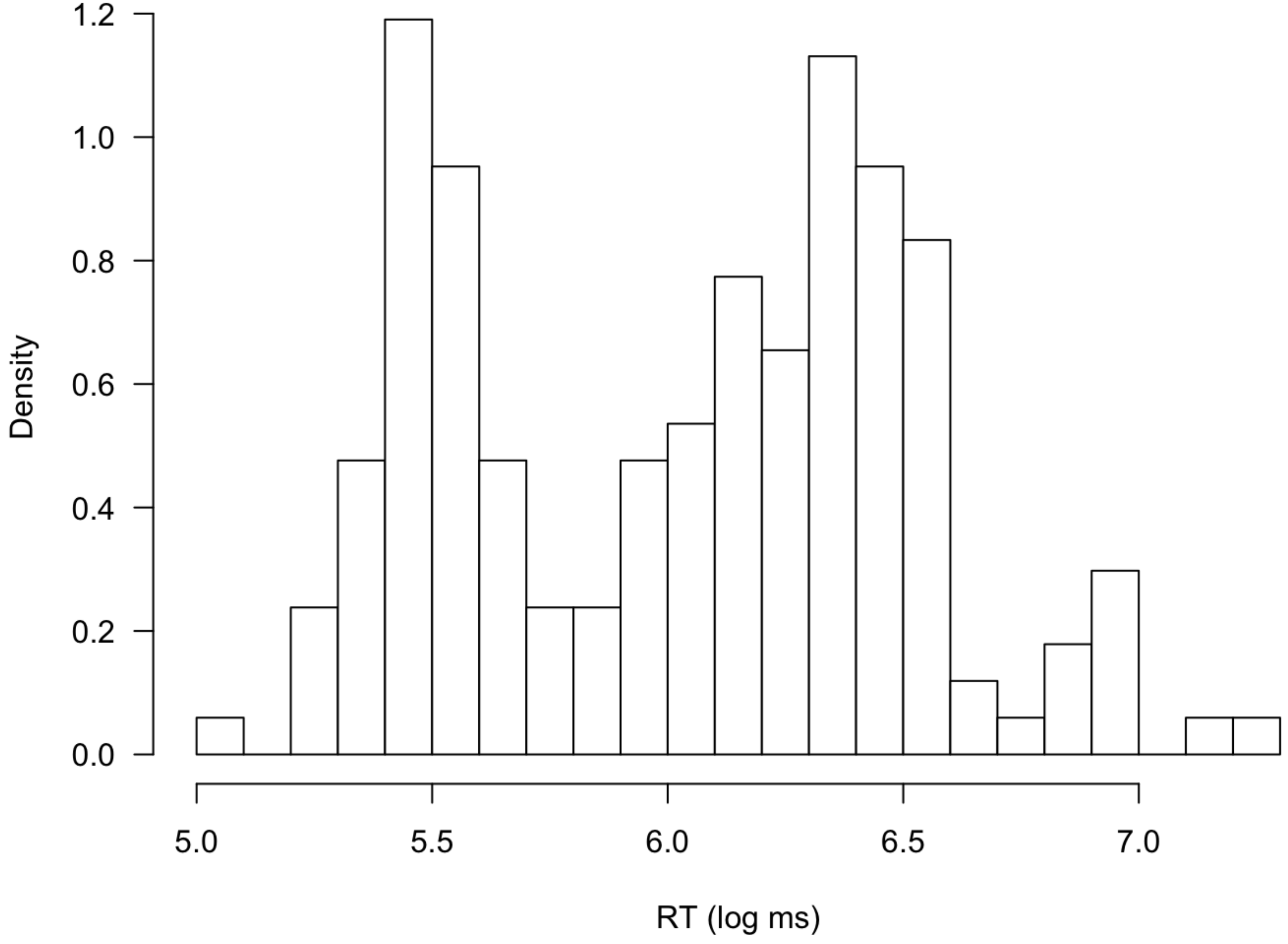
Mixture & latent class models

Ingmar Visser

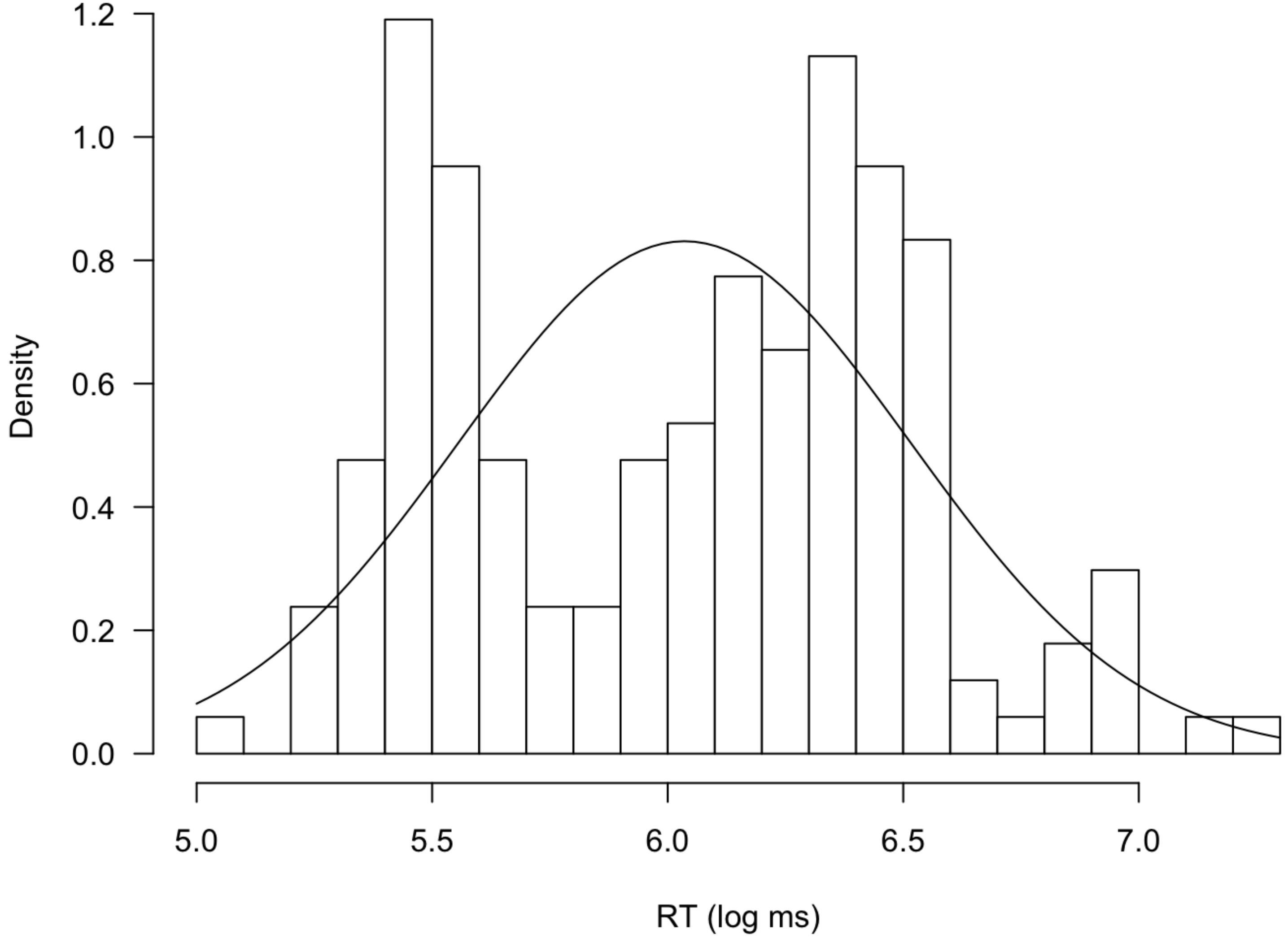
A motivating example, response times

The data below are response times (in log ms) from 168 trials of a lexical decision experiment.



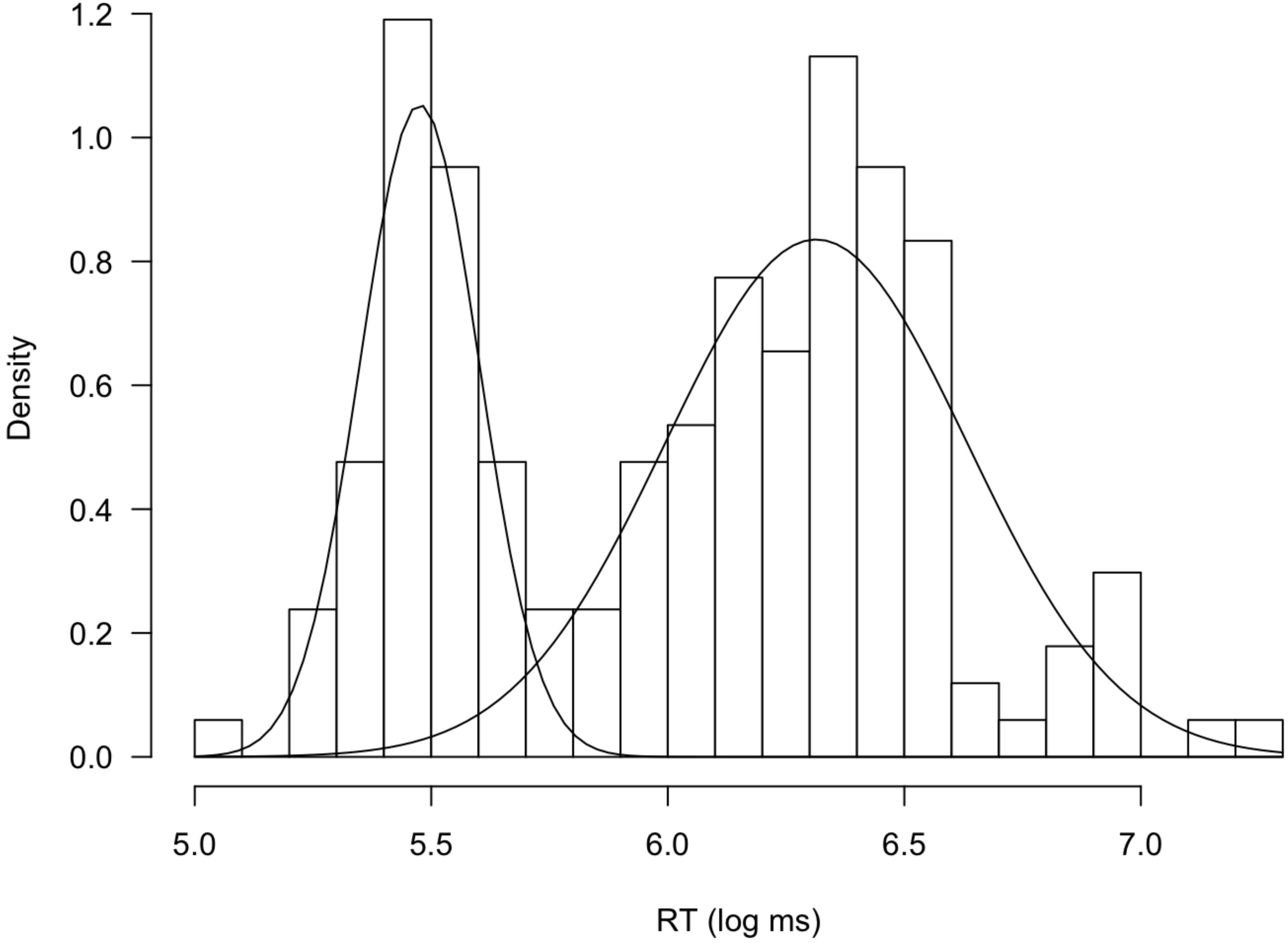
A motivating example, response times

A normal distribution does not seem very plausible…



A motivating example, response times

The bimodal nature of the data can be captured through two Normal distributions:



Mixture models

- each observation is assumed to be drawn from one of a number of distinct subpopulations (component distributions)
- which subpopulation an observation is drawn from is not directly observable (latent).
- *within* each subpopulation, observations are assumed to be relatively homogeneous, while there is more heterogeneity *between* subpopulations.

The term “subpopulation” should be understood in its statistical meaning as reflecting a probability distribution. For our

Mixture models: formal definition

A mixture distribution over observations $Y_t, t = 1, \ldots, T$, is defined as

$$p(Y_t = y) = \sum_{i=1}^{N} p(Y_t = y | S_t = i) P(S_t = i)$$

where

- $S_t \in \{1, \ldots, N\}$ denotes the latent state (a.k.a. “class”, “component”) of observation t
- $P(S_t = i)$ denotes the probability that the latent state at t equals i
- $p(Y_t = y | S_t = i)$ denotes the density of observation Y_t (evaluated at y), conditional upon the latent state

being $S_t = i$; i.e., it is the value of the i-th component density (evaluated at y).

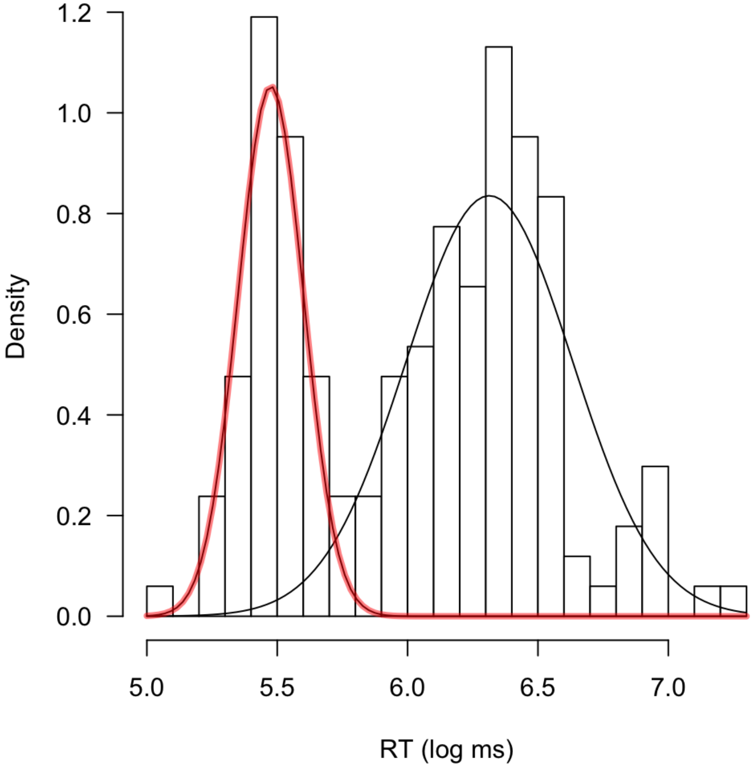
Mixture models: applied to RT data

Mixture distribution:

$$p(Y_t) = \sum_{i=1}^{2} p(Y_t | S_t = i) P(S_t = i)$$

Model components:

- $p(Y_t | S_t = 1) = N(5.48, 0.13)$



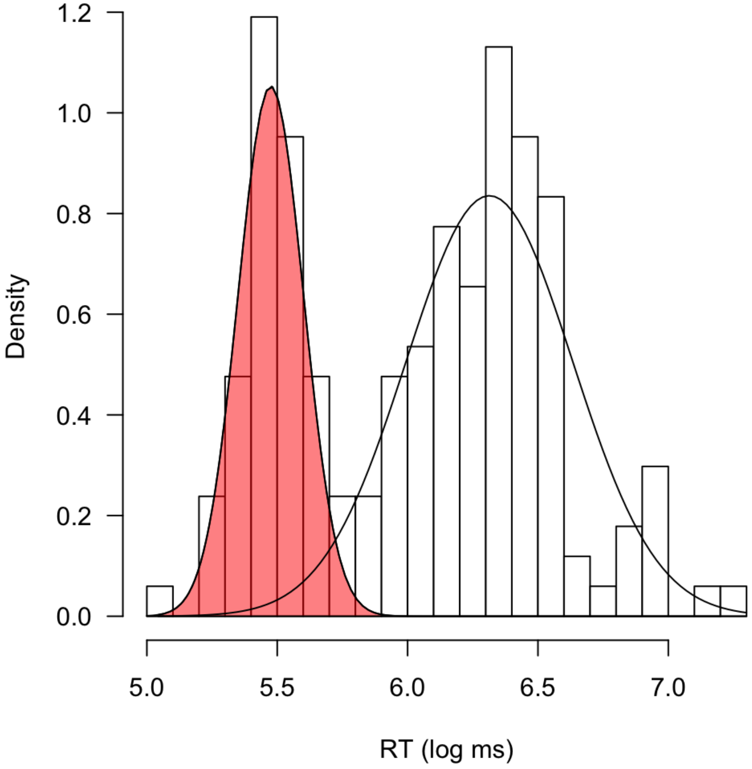
Mixture models: applied to RT data

Mixture distribution:

$$p(Y_t) = \sum_{i=1}^{N} p(Y_t | S_t = i) P(S_t = i)$$

Model components:

- $p(Y_t | S_t = 1) = N(5.48, 0.13)$
- $P(S_t = 1) = 0.33$



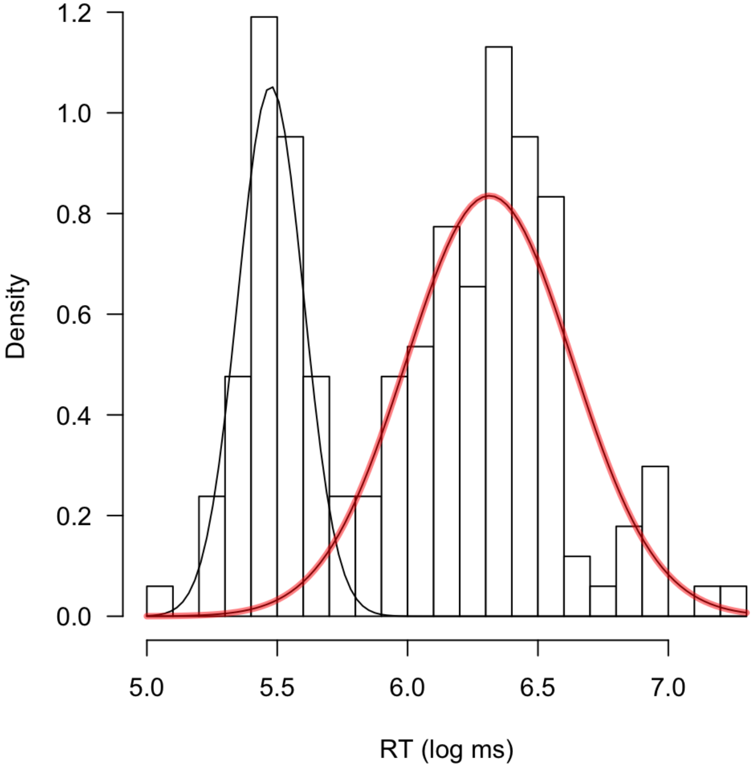
Mixture models: applied to RT data

Mixture distribution:

$$p(Y_t) = \sum_{i=1}^{N} p(Y_t | S_t = i) P(S_t = i)$$

Model components:

- $p(Y_t | S_t = 1) = N(5.48, 0.13)$
- $P(S_t = 1) = 0.33$
- $p(Y_t | S_t = 2) = N(6.31, 0.32)$



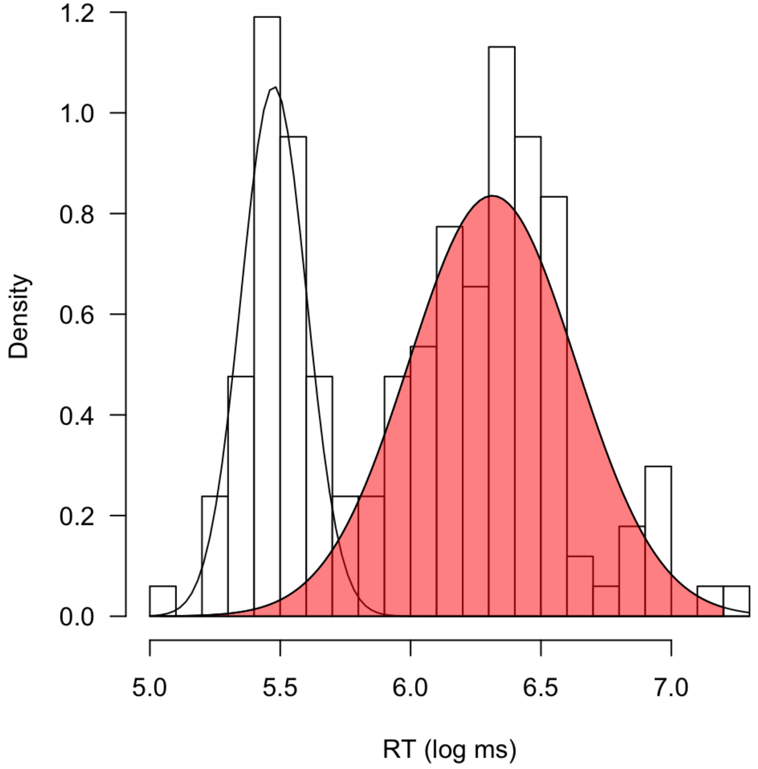
Mixture models: applied to RT data

Mixture distribution:

$$p(Y_t) = \sum_{i=1}^{N} p(Y_t | S_t = i) P(S_t = i)$$

Model components:

- $p(Y_t | S_t = 1) = N(5.48, 0.13)$
- $P(S_t = 1) = 0.33$
- $p(Y_t | S_t = 2) = N(6.31, 0.32)$
- $P(S_t = 2) = 0.67$



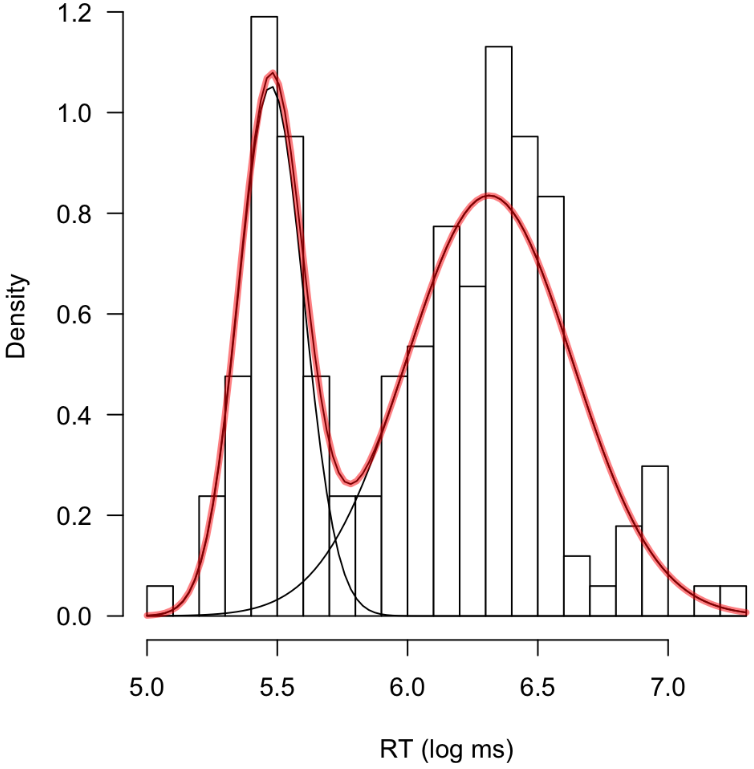
Mixture models: applied to RT data

Mixture distribution:

$$p(Y_t) = \sum_{i=1}^{N} p(Y_t | S_t = i) P(S_t = i)$$

Model components:

- $p(Y_t | S_t = 1) = N(5.48, 0.13)$
- $P(S_t = 1) = 0.33$
- $p(Y_t | S_t = 2) = N(6.31, 0.32)$
- $P(S_t = 2) = 0.67$
- $p(Y_t)$



Estimation

Estimation

Estimating a mixture model consists of estimating the parameters of each component distribution, and the component probabilities. Two main methods of estimation are

- Maximum likelihood estimation (MLE)
- Bayesian estimation

We will mainly focus on MLE. Maximum likelihood estimates have well-known desirable properties:

- Consistency
  (as sample size increases, the estimate converges in probability to the true value)
- Asymptotic normality
  (as sample size increases, the distribution of the estimate tends to a (multivariate) Normal distribution with covariance matrix equal to the inverse Fisher information matrix)
- Efficiency
  (no consistent estimator has lower asymptotic mean squared error)

The mixture likelihood

Let $\theta = (\theta_{resp}, \theta_{prior})$ denote a general vector of parameters with elements θ_{resp} for the component distributions and elements θ_{prior} for the component probabilities. The likelihood function of the model parameters θ of general mixture model can be written as

$$\begin{aligned} L(\theta|y_{1:T}) &= p(y_{1:T}|\theta) \\ &= \prod_{t=1}^{T} \sum_{i=1}^{N} P(S_t = i|\theta_{prior}) p(y_t|S_t = i, \theta_{resp}) \end{aligned}$$

where e.g., for the previous example

- $\theta_{resp} = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ contains the means and standard deviations of two mixture components
- $\theta_{prior} = (P(S = 1), P(S = 2))$ contains the two mixture probabilities.

Maximising the mixture likelihood

The maximum likelihood estimates $\hat{\theta}$ (MLEs) of the model parameters θ are defined as

$$\hat{\theta} = \arg\max_{\theta} L(\theta | y_{1:T})$$

i.e. as the values of the parameters at the maximum of the likelihood function.

There are generally no analytical expressions to obtain the MLEs. Two popular methods for MLE are numerical optimization and the Expectation-Maximization (EM) algorithm.

Numerical optimization of the likelihood

Numerical optimization routines are iterative procedures which, from provided starting values, change parameter values in the direction of a (local) minimum.

Expectation-Maximization (EM)

If we knew the value of the states S_t, MLE would be easy. The EM algorithm can be viewed as a method which iteratively (1) imputes expected values for the unknown states to (2) computes MLEs, after which latent states are imputed with new values, etc.

Formally, the EM algorithm works with the joint or *complete-data* log likelihood

$$\log p(y_{1:T}, S_{1:T}|\theta) = \sum_{t=1}^{T} \log P(S_t|\theta_{\text{prior}}) + \sum_{t=1}^{T} \log p(y_t|S_t, \theta_{\text{resp}})$$

Expectation-Maximization (EM)

EM uses the *expected value* of the complete-data log-likelihood under initial parameters θ':

$$
\begin{aligned}
Q(\theta, \theta') &= E_{\theta'}\left[\sum_{t=1}^{T} \log p(S_t | \theta_{prior}) + \sum_{t=1}^{T} \log p(y_t | S_t, \theta_{resp})\right] \\
&= \sum_{t=1}^{T} \sum_{i=1}^{N} \gamma_t(i) \log P(S_t = i | \theta_{prior}) + \sum_{t=1}^{T} \sum_{i=1}^{N} \gamma_t(i) \log P(y_t | S_t = i, \theta_{resp})
\end{aligned}
$$

where $\gamma_t(i)$ are the posterior state probabilities under θ'.

Expectation-Maximization (EM)

The EM algorithm starts with an initial guess for the parameters, $\theta^{(0)}$. On each iteration $k = 1, \ldots$, an improved parameter vector $\theta^{(k)}$ is then determined as

$$\theta^{(k)} = \arg\max_{\theta} Q(\theta, \theta^{(k-1)})$$

The algorithm is stopped when either the difference between $\theta^{(k)}$ and $\theta^{(k-1)}$, or the difference between $Q(\theta^{(k)}, \theta^{(k-1)})$ and $Q(\theta^{(k-1)}, \theta^{(k-2)})$, is sufficiently small.

Numerical optimization or EM?

- Both are guaranteed to arrive at a *local* maximum of the likelihood. This means that the results are dependent on the starting values. It is good practice to try a range of starting values.
- EM can be more robust, but needs analytical expressions for the conditional MLEs
  - The *generalized* EM algorithm only requires conditional estimates which increase the likelihood
- EM is difficult (or impossible) when parameters are constrained. Constraining parameters is much easier in numerical optimization.
- Generally, EM makes larger jumps at the start, while numerical optimization converges quicker near the (local) maximum.
  - The methods can be combined, starting with EM and then using numerical optimization when EM iterations make small changes

Inference

Inference

Inference in mixture models:

- Determining the value of a particular parameter
- Determining the number of mixture components (model selection)

While the first inference problem is relatively straightforward, the second is tricky.

Testing parameters

Testing whether a parameter in a mixture model has a particular value is generally done through a likelihood ratio test. Let θ_u denote the parameter vector and let θ_r denote the parameter vector with the respective elements fixed to the test values. Then

$$\begin{aligned} LR &= -2 \log \frac{L(\theta_u | y_{1:T})}{L(\theta_r | y_{1:T})} \\ &= -2(\log L(\theta_u | y_{1:T}) - \log L(\theta_r | y_{1:T})) \end{aligned}$$

asymptotically follows a $\chi^2(\nu)$-distribution with degrees of freedom $\nu = \dim(\theta_u) - \dim(\theta_r)$, provided that * θ_r is an interior point of the parameter space of θ_u; e.g., the test values should not lie on the bounds of the parameter space.

Determining the number of states

While a 2-component mixture is formally nested under a 3-component mixture, the nesting relation is not unique (the 2-component mixture can be derived from the 3 component mixture by fixing a component probability to 0, or by fixing the parameters of two components to be identical to each other.)

In addition, the restriction involves fixing parameters on the bound of the space, e.g., setting $P(S_t = i) = 0$.

This means that the likelihood-ratio statistic LR does **not** asymptotically follow a χ^2-distribution.

Solutions:

- “Empirically” determine the actual distribution (e.g., parametric bootstrap)
- Use model selection criteria (e.g., AIC, BIC)

The parametric bootstrap

Rather than approximating the distribution of the LR statistic with a known distribution, we can use sampling techniques to obtain an "empirical" estimate of the distribution.

The parametric bootstrap LR test to compare a k-component mixture to a $(k + l)$-component mixture (with $l \geq 1$) is as follows:

1. Fit the k and $(k + l)$-component mixture models to obtain maximum likelihood estimates $\hat{\theta_k}$ and $\hat{\theta_{k+l}}$ and compute the likelihood ratio LR_{obs}.
2. For $i = 1, \ldots, M$:

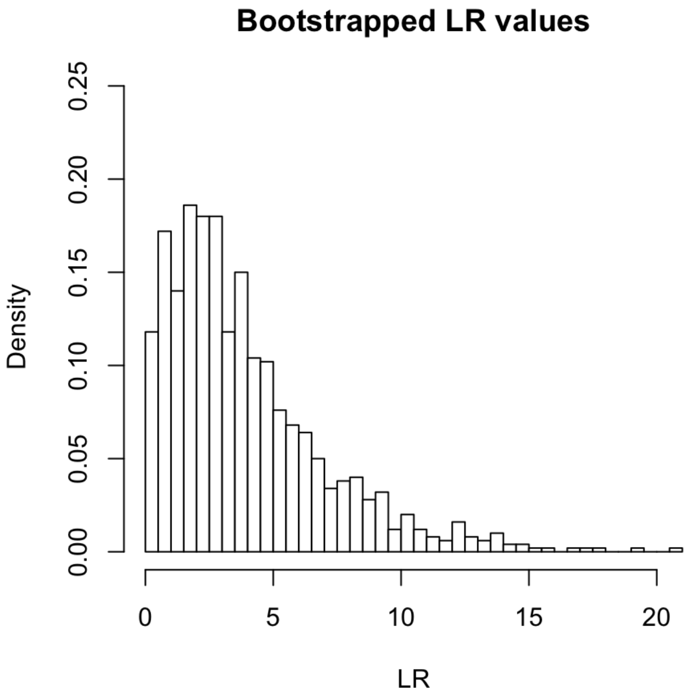
- Use the fitted k-component model to simulate a bootstrap sample $y^{(i)}_{1:T} \sim p(\cdot|\hat{\theta_k})$
- Fit the k and $(k + l)$-component mixture models to the bootstrap data $y^{(i)}_{1:T}$ and calculate the likelihood ratio $LR^{(i)}$ for these models

3. The estimated exceedence probability $P(LR \geq LR_{obs})$ is the proportion of bootstrap values $LR^{(i)}$ which are larger.

Parametric bootstrap example

Compare a 2-component to a 3-component mixture.

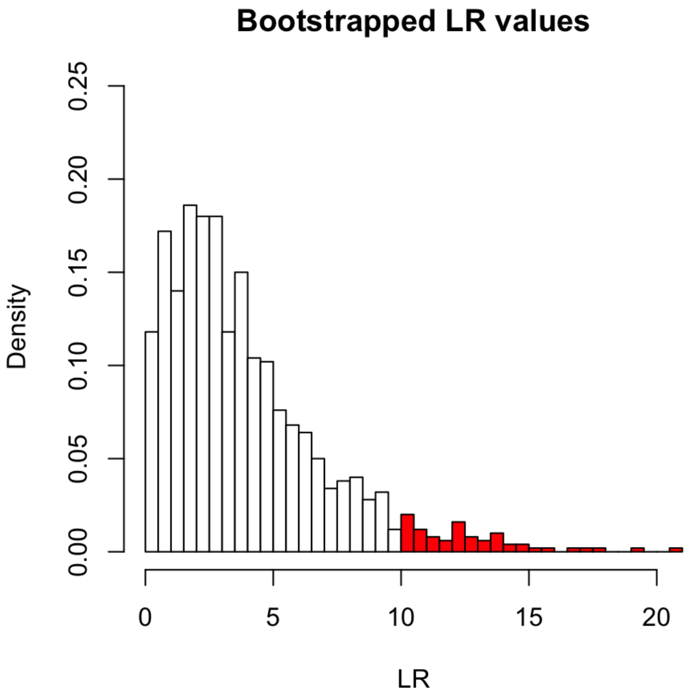
- Observed $\mathrm{LR}_{\mathrm{obs}} = 10.062$



Parametric bootstrap example

Compare a 2-component to a 3-component mixture.

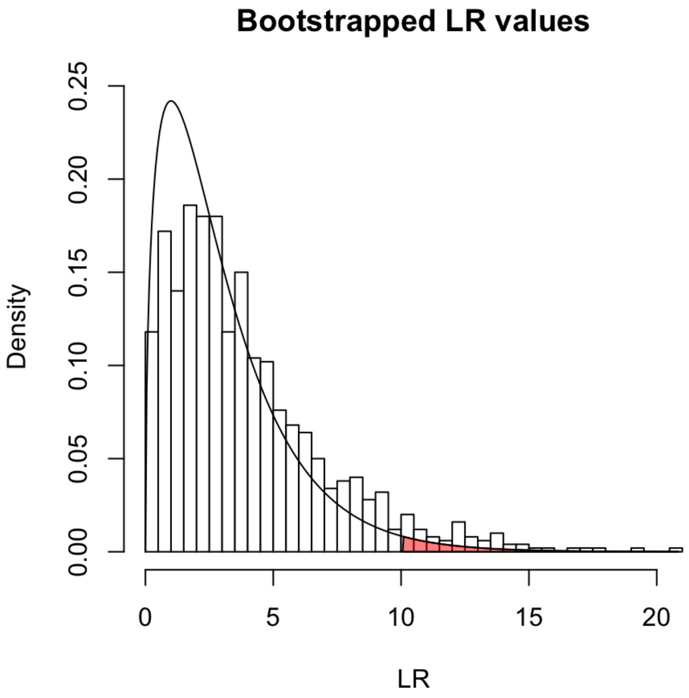
- Observed $\mathrm{LR}_{\mathrm{obs}} = 10.062$
- Estimated $\mathrm{P}(\mathrm{LR} \geq 10.062) = 0.053$



Parametric bootstrap example

Compare a 2-component to a 3-component mixture.

- Observed $\mathrm{LR}_{\mathrm{obs}} = 10.062$
- Estimated $P(\mathrm{LR} \geq 10.062) = 0.053$
- Using a $\chi^2(3)$ distribution, the p-value would be $P(\mathrm{LR} \geq 10.062) = 0.018$



The 2-component mixture is rejected when assuming a χ^2 distribution, but not when using the parametric bootstrap!

Akaike Information Criterion (AIC)

A good model M minimizes the Kullback-Leibler discrepancy between the “true” model $p^*(Y)$ and the approximating model $M = p(Y|\theta_M)$

$$\begin{aligned} D_{KL}(p^*||M) &= \int p^*(y) \log \frac{p^*(y)}{p(Y|\theta_M, M)} dy \\ &= E_{p*}[\log p^*(y)] - E_{p*}[\log p(Y|\theta_M, M)] \end{aligned}$$

$E_{p*}[\log p^*(y)]$ is a constant that is identical for each model. The Akaike Information criterion (Akaike, 1973) is an asymptotic approximation of (twice) the important part of D_{KL}:

$$\begin{aligned} \text{AIC} &= -2 \log L(\hat{\theta_M}|y_{1:T}) + 2k \\ &\approx -2E_{p*}[\log p(Y|\theta_M, M)] \end{aligned}$$

where k is the number of freely estimated parameters in the ML estimate $\hat{\theta_M}$

Bayesian Information Criterion (BIC)

Main idea: a good model obtains the maximum posterior probability

$$p(M|y) \propto p(y|M)p(M)$$

The key quantity is the marginal likelihood

$$p(y|M) = \int_{\theta} p(y|\theta_M, M)p(\theta_M|M)d\theta$$

The Bayesian Information Criterion (Schwarz, 1978) is an asymptotic approximation to the -2 log marginal likelihood

$$\begin{aligned} \text{BIC} &= -2\log L(\hat{\theta_M}|y_{1:T}) + k\log T \\ &\approx -2\log p(y_{1:T}|M) \end{aligned}$$

where k is the number of freely estimated parameters in the ML estimate $\hat{\theta_M}$ and T is the total number of observations.

Which to use?

Both the AIC and BIC are widely used and neither the AIC nor the BIC require that the true model be in the set of models under consideration.

- BIC is **consistent**: if the true model is in the set, it will pick this model with probability 1 as $T \rightarrow \infty$. The AIC is not consistent.
- The AIC is **asymptotically efficient**: if the true model has infinite parameters, the AIC will asymptotically select the model which minimizes the mean squared error of prediction. The BIC is not asymptotically efficient.
- The empirical bootstrap works well, but is computationally expensive.

Nylund, Asparouhov, & Muthén (2007): BIC works best out of information measures, the parametric bootstrap works best overall.

State classification

After estimating the parameters of a mixture model, we are often interested in determining which component each observation belongs to. This is usually done with the posterior probabilities. Each observation is assigned to the component with the maximum posterior probability:

$$
\begin{aligned}
\hat{s_t} &= \arg\max_s P(S_t = s | y_t, \hat{\theta}) \\
&= \arg\max_s P(S_t = s | \hat{\theta}_{\text{prior}}) p(y_t | S_t = s, \hat{\theta}_{\text{resp}})
\end{aligned}
$$