Density Estimation
Bandwidth Choice by Leave-one-out Maximum Likelihood

Azzarito Domenico, Daniel Reverter, Alexis Vendrix

29 September, 2025

Histogram

1.

We want to find a similar relationship between the histogram estimator of the density function $\hat{f}_{hist}(x)$ and its leave-one-out version, $\hat{f}_{hist,(-i)}(x)$, when both are evaluated at x_i.

Let the histogram be defined by bins B_j of common width b and let $j(x)$ be the index function indicating the interval containing x. The density estimator using all n observations is given by:

$$\hat{f}_{\text{hist}}(x_i) = \frac{N_{j(x_i)}}{nb}$$

where $N_{j(x_i)}$ is the count of data points in bin $B_{j(x_i)}$.

The leave-one-out estimator, built using $n-1$ data points, evaluates at x_i as:

$$\hat{f}_{\text{hist},(-i)}(x_i) = \frac{N_{j(x_i)} - 1}{(n-1)b}$$

By substituting $N_{j(x)} = nb \cdot \hat{f}_{\text{hist}}(x_i)$, we derive the relationship:

$$\hat{f}_{\text{hist},(-i)}(x_i) = \frac{n}{n-1}\hat{f}_{\text{hist}}(x_i) - \frac{1}{(n-1)b}$$

This is the desired relationship between the full histogram estimator and the leave-one-out version, both evaluated at an observation x_i.

2.

Read the CD rate data set and call x the first column.

```
cdrate.df <-read.table("data/cdrate.dat")
x <- cdrate.df[,1]
```

Then define

$$A < -min(x) - 0.05 * diff(range(x))Z < -max(x) + 0.05 * diff(range(x))nbr < -7$$

```
# Define the range for the histogram
A <- min(x) - 0.05 * diff(range(x))
Z <- max(x) + 0.05 * diff(range(x))
nbr <- 7

cat("A =", A, "\n")
```
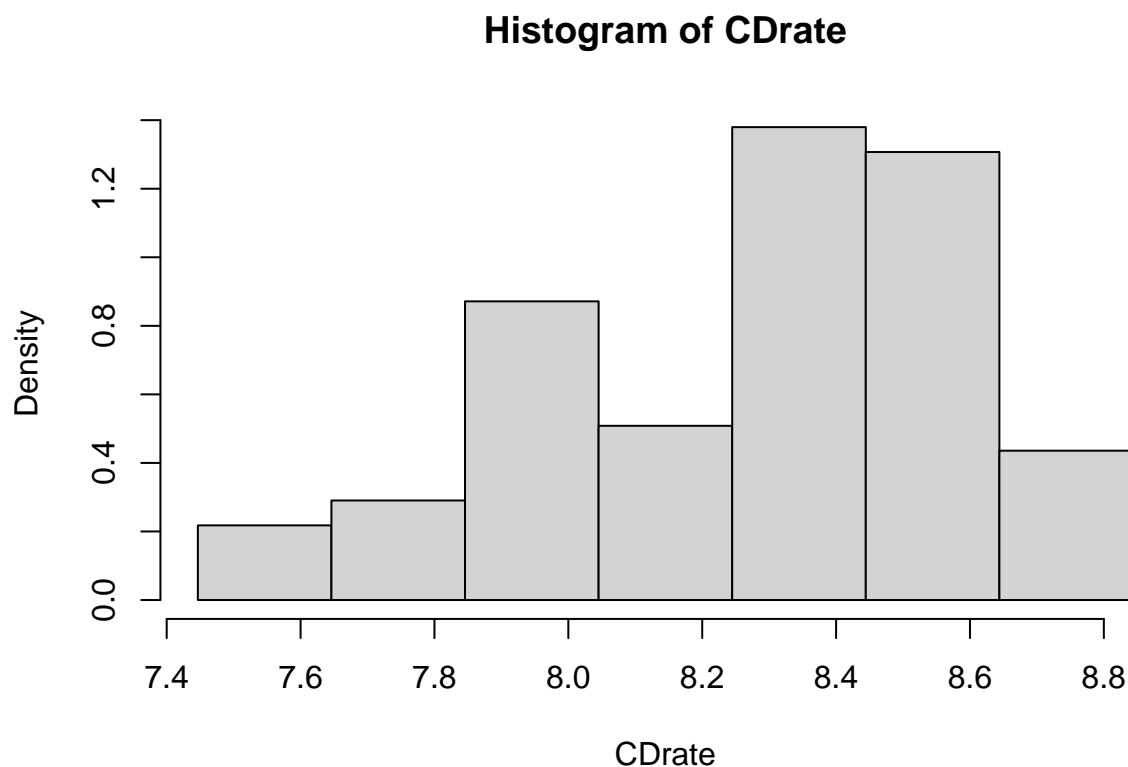
```
## A = 7.4465
```

```
cat("Z =", Z, "\n")
```

```
## Z = 8.8435
```

and plot the histogram of x as

$$hx < -hist(x, breaks = seq(A, Z, length = nbr + 1), freq = F)$$

```
# Plot the histogram
hx <- hist(x, breaks = seq(A, Z, length = nbr + 1), freq = FALSE,
           main = "Histogram of CDrate", xlab = "CDrate")
```


Histogram of CDrate

The following sentence converts this histogram into a function that can be evaluated at any point of $\mathbb{R}$, or at a vector of real numbers:

```
hx_f <- stepfun(hx$breaks, c(0, hx$density, 0))
```

Use `hx_f` to evaluate the histogram at the vector of observed data x. Then add the points $(x_i, \hat{f}_{\text{hist}}(x_i))$, $i = 1, \ldots, n$, to the histogram you have plotted before.

```
# Evaluate the histogram estimator at each data point
f_hat <- hx_f(x)
```

```
# Add the points to the plot
plot(hx, freq = FALSE, main = "Histogram of CDrate", xlab = "CDrate")
points(x, f_hat, col = "blue", pch = 19, cex = 0.5)
legend(x = 7.5, y = 1.3, "f_hat(x_i)", col = "blue", pch = 19)
```

Histogram of CDrate
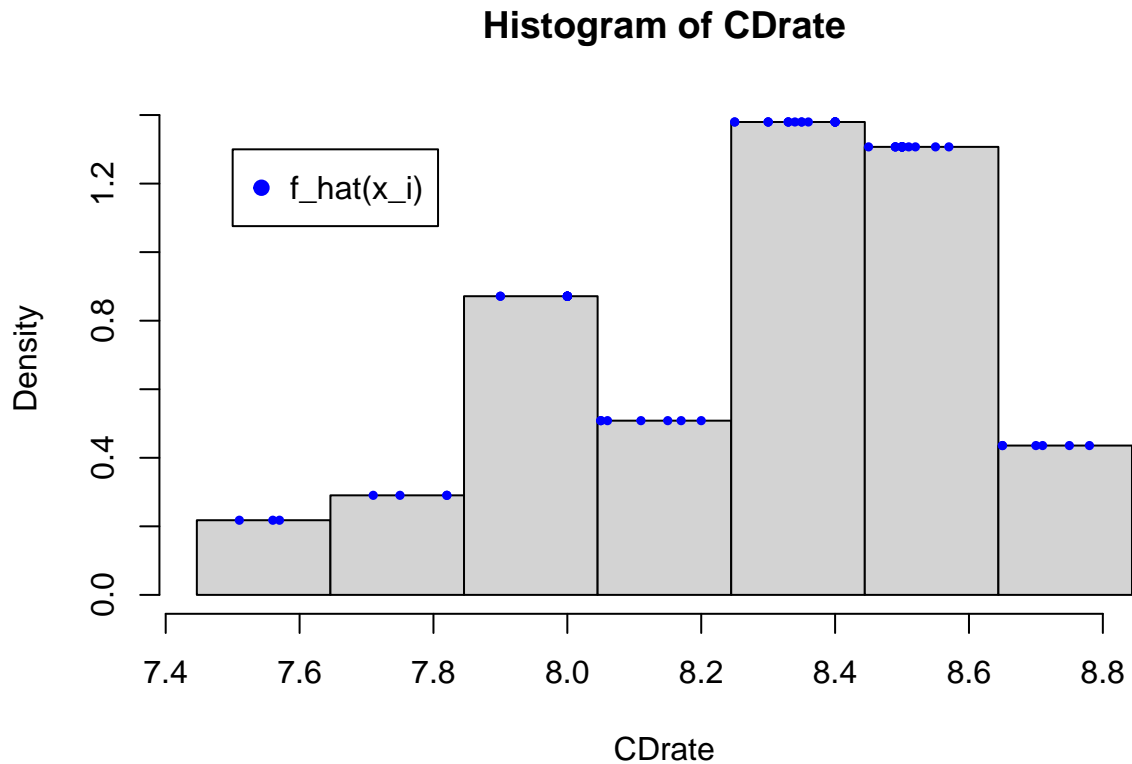
3.

Use the formula you have found before relating $\hat{f}_{\text{hist}}(x_i)$ and $\hat{f}_{\text{hist},(-i)}(x_i)$ to compute $\hat{f}_{\text{hist},(-i)}(x_i)$, $i = 1, \ldots, n$. Then add the points $(x_i, \hat{f}_{\text{hist},(-i)}(x_i))$, $i = 1, \ldots, n$, to the previous plot.

```
# Calculate bin width
b <- (Z - A) / nbr
n <- length(x)

# Calculate the leave-one-out estimates
f_loo <- (n / (n - 1)) * f_hat - 1 / ((n - 1) * b)

# Add the points to the plot
plot(hx, freq = FALSE, main = "Histogram with Full and LOO Densities", xlab = "CDrate")
points(x, f_hat, col = "blue", pch = 19, cex = 0.5)
points(x, f_loo, col = "red", pch = 4, cex = 0.5)
legend(x = 7.5, y = 1.3, c("f_hat(x_i)", "f_loo(x_i)"), col = c("blue", "red"), pch = c(19, 4))
```

Histogram with Full and LOO Densities
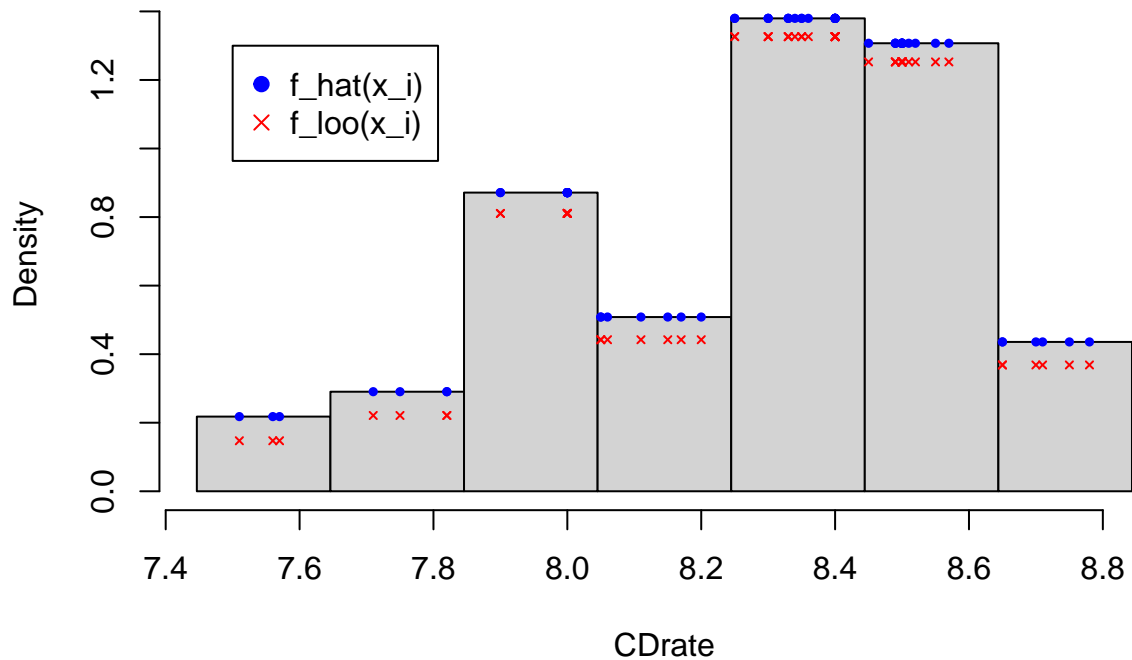
The blue dots show the histogram estimator of the density function for each data point using the full dataset. As you can see, all dots within the same bin are at the same height, being the top of that bin's bar.

The red crosses show the histogram estimator of the density function for each data point if that point had been excluded from the calculation.

The red crosses are always slightly lower than the blue dots. Because when we "leave out" a data point x_i, the count of points in its bin N_k decreases by one. Since the density is calculated as (count / total), reducing the count naturally leads to a lower density estimate for that bin.

4.

Compute the leave-one-out log-likelihood function corresponding to the previous histogram, at which `nbr=7` has been used.

```r
# We only take the log of positive values. If f_loo is 0, log(f_loo) is -Inf.
# This happens when a point is the only one in its bin.
looCV_log_lik_7 <- sum(log(f_loo[f_loo > 0]))
cat("Leave-one-out log-likelihood for nbr=7:", looCV_log_lik_7)
```

```
## Leave-one-out log-likelihood for nbr=7: -16.58432
```

5.

Consider now the set `seq(1,15)` as possible values for `nbr`, the number of intervals of the histogram. For each of them compute the leave-one-out log-likelihood function (`looCV_log_lik`) for the corresponding histogram.

```r
n <- length(x)
nbr_values <- 1:15
```

4

```r
looCV_log_lik_nbr <- numeric(length(nbr_values))

for (i in seq_along(nbr_values)) {
  current_nbr <- nbr_values[i]
  b <- (Z - A) / current_nbr

  # Create histogram object
  hx <- hist(x, breaks = seq(A, Z, length = current_nbr + 1), plot = FALSE)

  # Create step function
  hx_f <- stepfun(hx$breaks, c(0, hx$density, 0))

  # Calculate f_hat and f_loo
  f_hat <- hx_f(x)
  f_loo <- (n / (n - 1)) * f_hat - 1 / ((n - 1) * b)

  # Calculate and store looCV log-likelihood
  looCV_log_lik_nbr[i] <- sum(log(f_loo[f_loo > 0]))
}

# Plot the results
plot(nbr_values, looCV_log_lik_nbr, type = "b", pch = 19,
     xlab = "Number of Bins (nbr)", ylab = "LOO Log-Likelihood",
     main = "LOO Cross-Validation for Number of Bins")

# Find the optimal nbr
optimal_nbr <- nbr_values[which.max(looCV_log_lik_nbr)]
abline(v = optimal_nbr, col = "red", lty = 2)
legend("bottomleft", legend = paste("Optimal nbr =", optimal_nbr), col = "red", lty = 2)
```
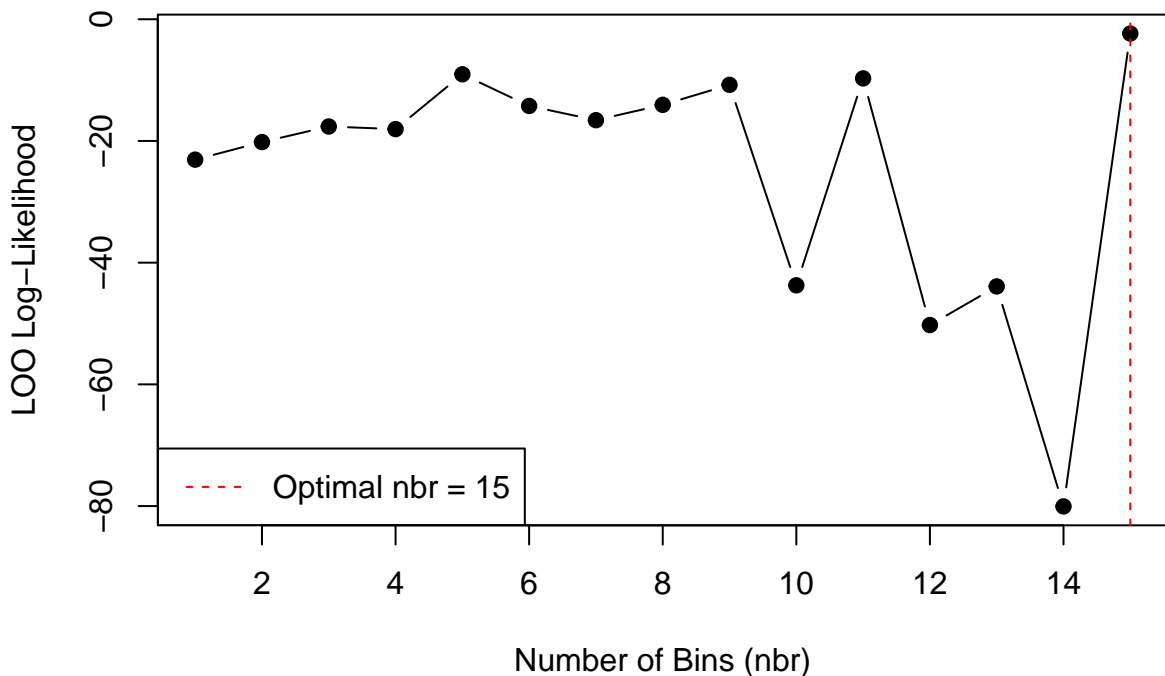
LOO Cross–Validation for Number of Bins

The plot shows the optimal value of nbr as that at which looCV_log_lik takes its maximum is nbr=15, which is therefore considered the optimal choice for this dataset. One could argue that the Log-likelihood steadily increases from nbr=1 to nbr=9, and then from nbr=10 onward the Log-likelihood changes erratically, probably due to the high variance nature of choosing a high bandwidth, making nbr=15 the maximum due to randomness. It could be wise to choose nbr=9 as a good tradeoff, but basing our choice purely on the maximization of the Log-likelihood we would have to choose nbr=15.

Finally, plot the histogram of x using the optimal value of `nbr`.

```
hist(x, breaks = seq(A, Z, length = optimal_nbr + 1), freq = FALSE,
     main = paste("Optimal Histogram (nbr =", optimal_nbr, ")"),
     xlab = "CDrate")
```

Optimal Histogram (nbr = 15)
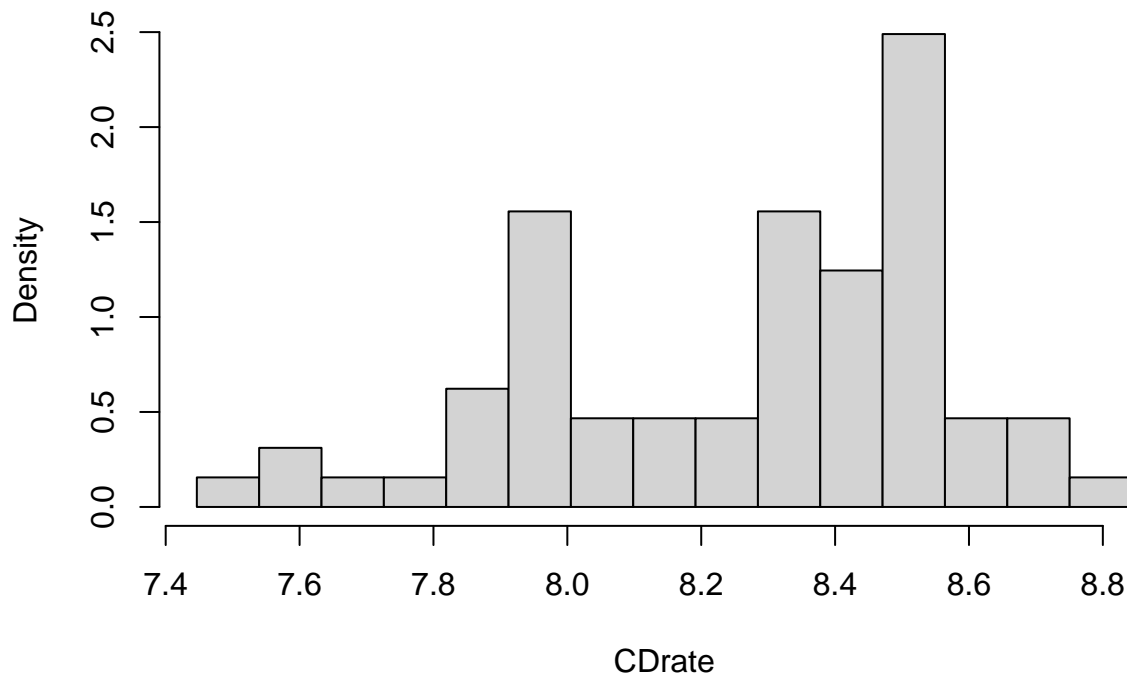
6.

Let `b` be the common width of the bins of a histogram. Consider the set `seq((Z-A)/15,(Z-A)/1,length=30)` as possible values for `b`. Select the value of `b` maximizing the leave-one-out log-likelihood function, and plot the corresponding histogram.

```
b_values <- seq((Z - A) / 15, (Z - A) / 1, length = 30)
looCV_log_lik_b <- numeric(length(b_values))

for (i in seq_along(b_values)) {
  current_b <- b_values[i]

  # Create histogram object with specified bin width
  hx <- hist(x, breaks = seq(A, Z + current_b, by = current_b), plot = FALSE)

  # Create step function
  hx_f <- stepfun(hx$breaks, c(0, hx$density, 0))
```

```
  # Calculate f_hat and f_loo
  f_hat <- hx_f(x)
  f_loo <- (n / (n - 1)) * f_hat - 1 / ((n - 1) * current_b)

  # Calculate and store looCV log-likelihood
  looCV_log_lik_b[i] <- sum(log(f_loo[f_loo > 0]))
}

# Plot the results
plot(b_values, looCV_log_lik_b, type = "b", pch = 19,
     xlab = "Bin Width (b)", ylab = "LOO Log-Likelihood",
     main = "LOO Cross-Validation for Bin Width")

# Find the optimal b
optimal_b <- b_values[which.max(looCV_log_lik_b)]
abline(v = optimal_b, col = "red", lty = 2)
legend("topright", legend = paste("Optimal b =", round(optimal_b, 3)), col = "red", lty = 2)
```
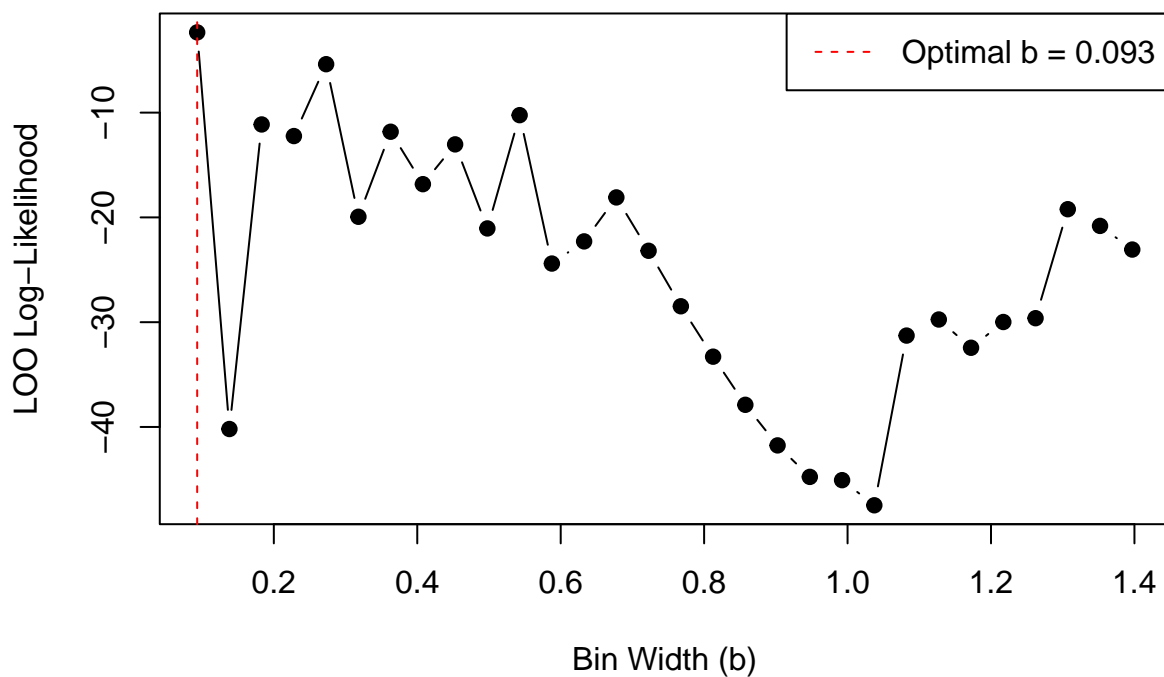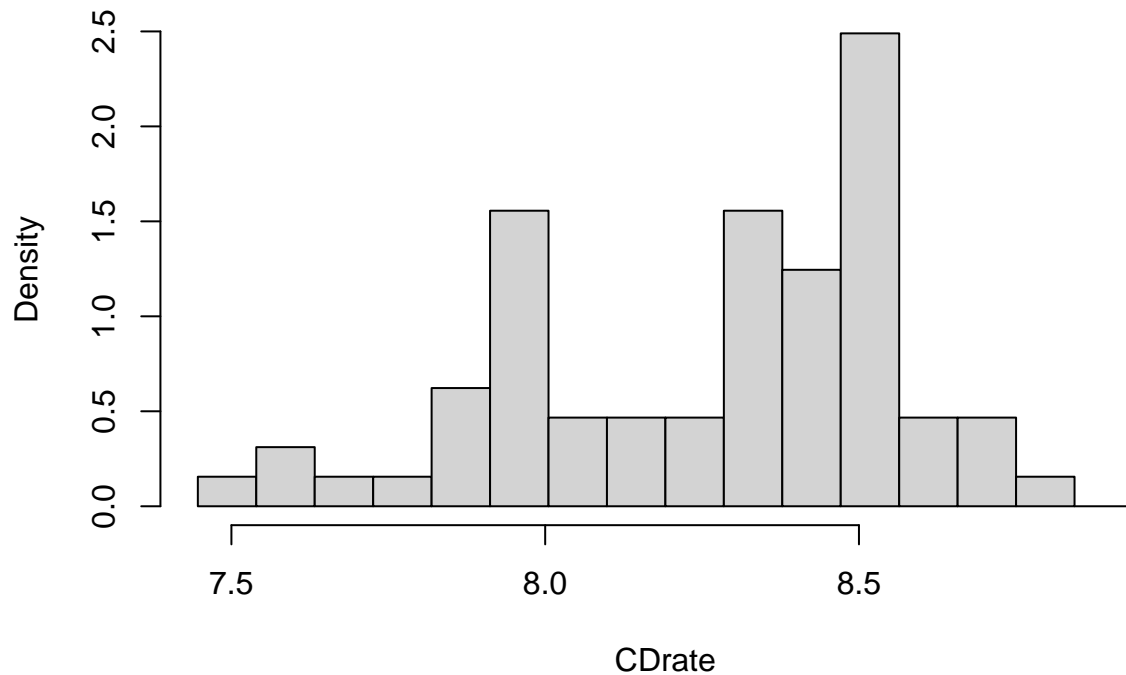
LOO Cross−Validation for Bin Width

Plot the corresponding histogram.

```
hx_optimal_b <- hist(x, breaks = seq(A, Z + optimal_b, by = optimal_b), plot = FALSE)
plot(hx_optimal_b, freq = FALSE,
     main = paste("Optimal Histogram (b =", round(optimal_b, 3), ")"),
     xlab = "CDrate")
```

Optimal Histogram (b = 0.093)

As we can see, through both methods, maximizing by choosing the number of bins and maximizing by choosing the bin size, we find the same optimal histogram with the highest log-likelihood (LLO).

7.

Recycle the functions `graph.mixt` and `sim.mixt` to generate $n = 100$ data from

$$f(x) = (3/4)N(x; m = 0, s = 1) + (1/4)N(x; m = 3/2, s = 1/3)$$

```r
# Generate 100 observations from f(x) = (3/4)N(x; m = 0, s = 1) +(1/4) N(x; m = 3/2, s = 1/3)
set.seed(123)
n <- 100
mu <- c(0,3/2)
sigma <- c(1,1/3)
alpha <- c(3/4,1/4)
x <- sim.mixt(n=n, k=2, mu=mu, sigma=sigma, alpha=alpha)
f_x <- graph.mixt(k=2, mu=mu, sigma=sigma, alpha=alpha, graphic = F)
```

Let b be the bin width of a histogram estimator of $f(x)$ using the generated data. Select the value of b maximizing the leave-one-out log-likelihood function, and plot the corresponding histogram.

```r
#We will use the same method as before
A <- min(x) - 0.05 * diff(range(x))
Z <- max(x) + 0.05 * diff(range(x))

b_values <- seq((Z - A) / 15, (Z - A) / 1, length = 30)
looCV_log_lik_b <- numeric(length(b_values))

for (i in seq_along(b_values)) {
```

```r
  current_b <- b_values[i]

  # Create histogram object with specified bin width
  hx <- hist(x, breaks = seq(A, Z + current_b, by = current_b), plot = FALSE)

  # Create step function
  hx_f <- stepfun(hx$breaks, c(0, hx$density, 0))

  # Calculate f_hat and f_loo
  f_hat <- hx_f(x)
  f_loo <- (n / (n - 1)) * f_hat - 1 / ((n - 1) * current_b)

  # Calculate and store looCV log-likelihood
  looCV_log_lik_b[i] <- sum(log(f_loo[f_loo > 0]))
}

# Plot the results
plot(b_values, looCV_log_lik_b, type = "b", pch = 19,
     xlab = "Bin Width (b)", ylab = "LOO Log-Likelihood",
     main = "LOO Cross-Validation for Bin Width")

# Find the optimal b
optimal_b <- b_values[which.max(looCV_log_lik_b)]
abline(v = optimal_b, col = "red", lty = 2)
legend("topright", legend = paste("Optimal b =", round(optimal_b, 3)), col = "red", lty = 2)
```
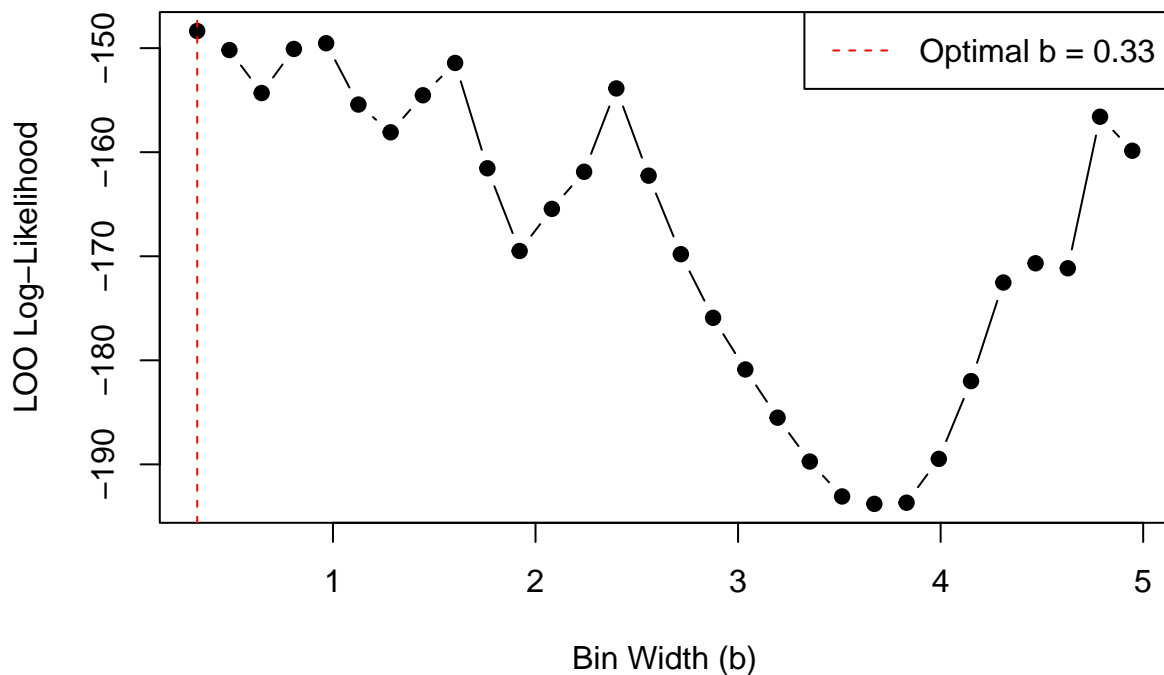
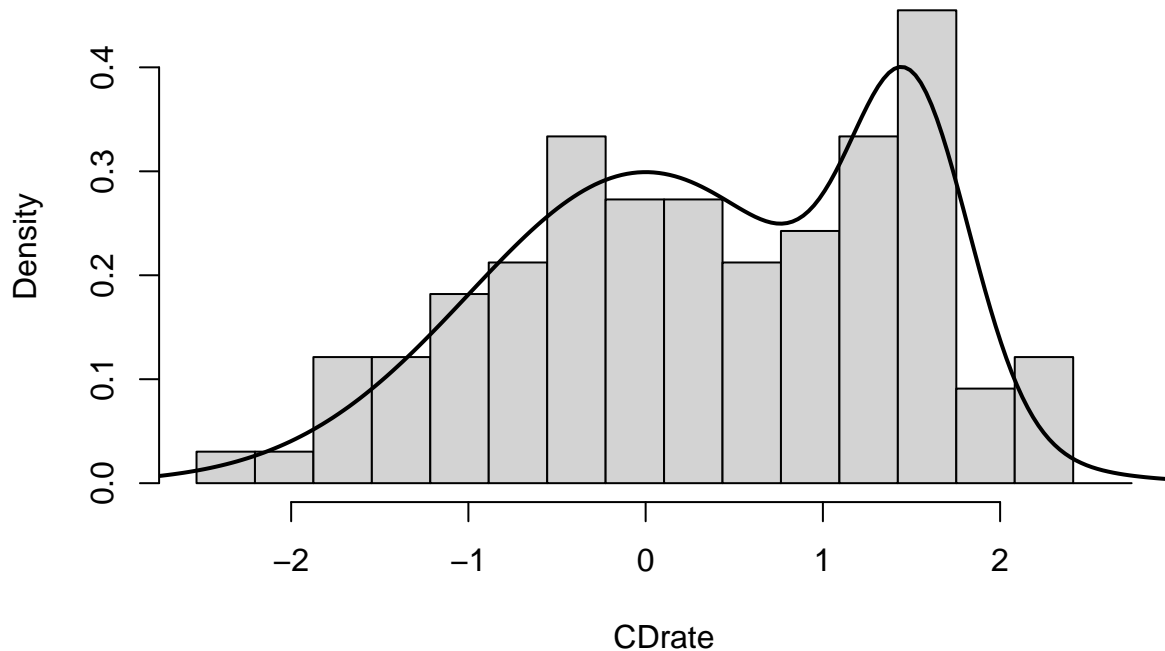## LOO Cross–Validation for Bin Width


```r
hx_optimal_b <- hist(x, breaks = seq(A, Z + optimal_b, by = optimal_b), plot = FALSE)
plot(hx_optimal_b, freq = FALSE,
     main = paste("Optimal Histogram (b =", round(optimal_b, 3), ")"),
```

```
      xlab = "CDrate")
lines(f_x$x, f_x$fx, lwd = 2)
```

Optimal Histogram (b = 0.33)

Compare with the results obtained using the Scott's formula:

$$b_{Scott} = 3.49 St.Dev(X)n^{-1/3}$$

```
scott_b <- 3.49 * sd(x) * (length(x)^(-1/3))
hx_scott_b <- hist(x, breaks = seq(A, Z + scott_b, by = scott_b), plot = FALSE)
ymax <- max(c(hx_scott_b$y, f_x$fx))
plot(hx_scott_b, freq = FALSE, ylim = c(0, ymax),
     main = paste("Optimal Histogram (b_scott =", round(scott_b, 3), ")"),
     xlab = "CDrate")
lines(f_x$x, f_x$fx, lwd = 2)
```

Optimal Histogram (b_scott = 0.831)

CDrate
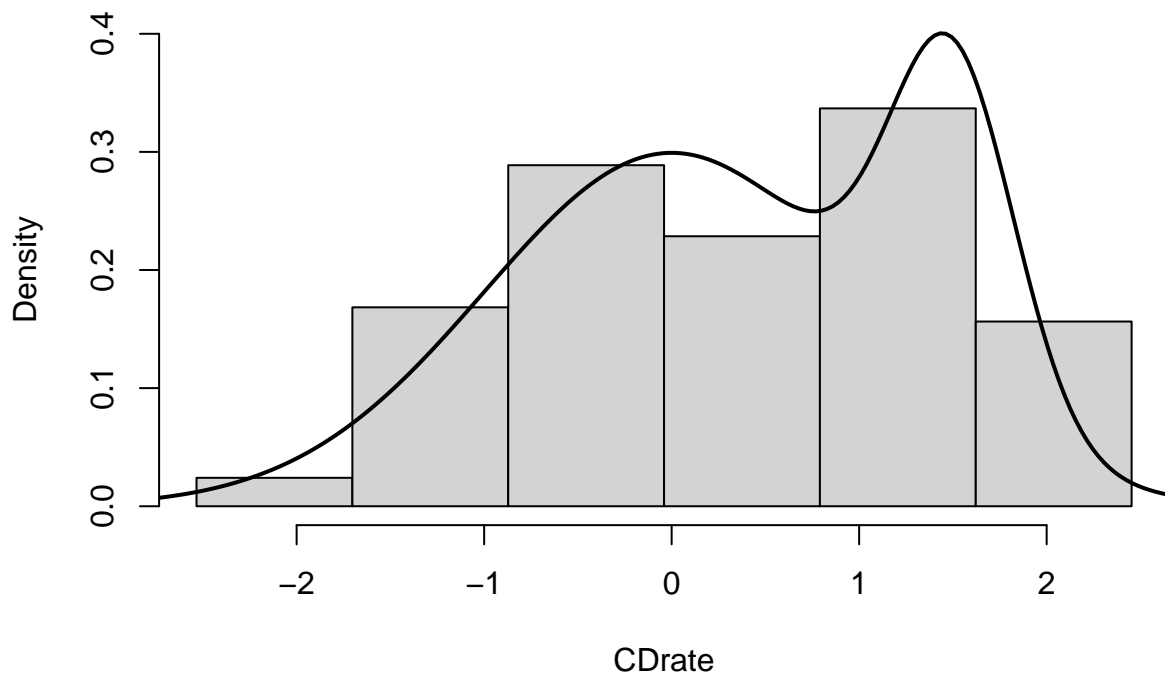
In this case, the b that maximized the leave-one-out log-likelihood provided a better fit than Scott's formula for choosing b. This can be proved by calculating the MSE of the estimators.

```r
# Compute histogram densities as step functions
hx_scott_f <- stepfun(hx_scott_b$breaks, c(0, hx_scott_b$density, 0))
hx_optimal_f <- stepfun(hx_optimal_b$breaks, c(0, hx_optimal_b$density, 0))

# Evaluate histogram estimates at the grid of f_x
f_hat_scott <- hx_scott_f(f_x$x)
f_hat_optimal <- hx_optimal_f(f_x$x)

# Compute MSE for Scott's rule and Optimal b
mse_scott <- mean((f_hat_scott - f_x$fx)^2)
mse_optimal <- mean((f_hat_optimal - f_x$fx)^2)

# Print results
cat("MSE (Scott's rule):", mse_scott, "\n")
```

```
## MSE (Scott's rule): 0.002943868
```

```r
cat("MSE (Optimal b):", mse_optimal, "\n")
```

```
## MSE (Optimal b): 0.002224444
```

Kernel Density Estimator

8.

Consider the vector x of data you have generated before from the mixture of two normal. Use the relationship

$$\hat{f}_{h,(-i)}(x_i) = \frac{n}{n-1}\left(\hat{f}_h(x_i) - \frac{K(0)}{nh}\right)$$

to select the value of h maximizing the leave-one-out log-likelihood function, and plot the corresponding kernel density estimator.

```r
n <- length(x)
K0 <- dnorm(0)

kx0 <- density(x)
base_bw <- kx0$bw
h_values <- seq(base_bw/5, base_bw*3, length.out = 50)

loo_loglik_h <- numeric(length(h_values))

for (i in seq_along(h_values)) {
  h <- h_values[i]
  kx <- density(x, bw = h, kernel = "gaussian", n = 1024,
                from = min(x) - 3*h, to = max(x) + 3*h)
  kx_f <- approxfun(x = kx$x, y = kx$y, rule = 2)
  f_hat <- kx_f(x)                        # approx f_hat at each xi
  f_loo <- (n / (n - 1)) * (f_hat - K0 / (n * h))
  pos <- f_loo > 0
  if (any(pos)) {
    loo_loglik_h[i] <- sum(log(f_loo[pos]))
  } else {
    loo_loglik_h[i] <- -Inf
  }
}

optimal_h_approx <- h_values[which.max(loo_loglik_h)]
cat("Optimal h (approx) =", optimal_h_approx, "\n")
```
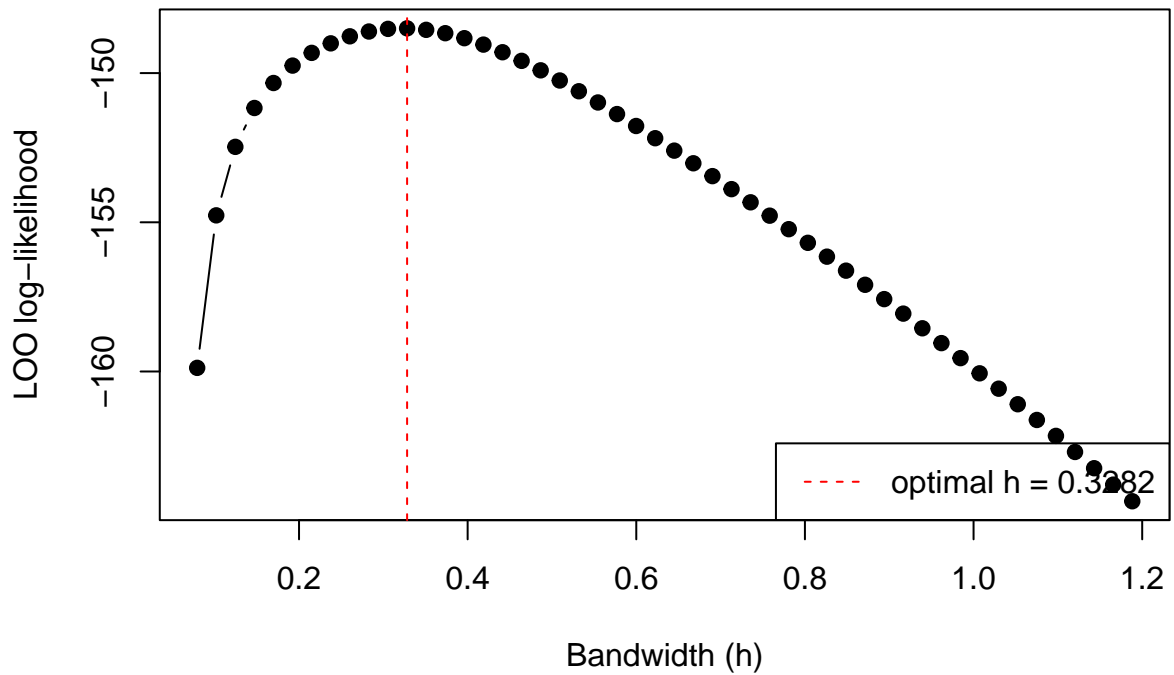
```
## Optimal h (approx) = 0.3281848
```

```r
plot(h_values, loo_loglik_h, type = "b", pch = 19,
     xlab = "Bandwidth (h)", ylab = "LOO log-likelihood",
     main = "LOO CV for KDE bandwidth (approx)")
abline(v = optimal_h_approx, col = "red", lty = 2)
legend("bottomright", legend = paste("optimal h =", round(optimal_h_approx, 4)),
       col = "red", lty = 2)
```

LOO CV for KDE bandwidth (approx)

- - - optimal h = 0.3282

LOO log-likelihood (y-axis)

Bandwidth (h) (x-axis)

```r
kx_opt <- density(x, bw = optimal_h_approx, kernel = "gaussian")

# Common ylim so both functions fit inside the plot
ymax <- max(c(kx_opt$y, f_x$fx))


plot(kx_opt,
     ylim = c(0, ymax),
     main = paste("KDE (h =", round(optimal_h_approx, 4), ")"),
     col = "blue",
     lwd = 2)

lines(f_x$x, f_x$fx, col = "black", lwd = 2)

legend("topright",
       legend = c("KDE", "f(x)"),
       col = c("blue", "black"),
       lwd = 2,
       bty = "n")
```
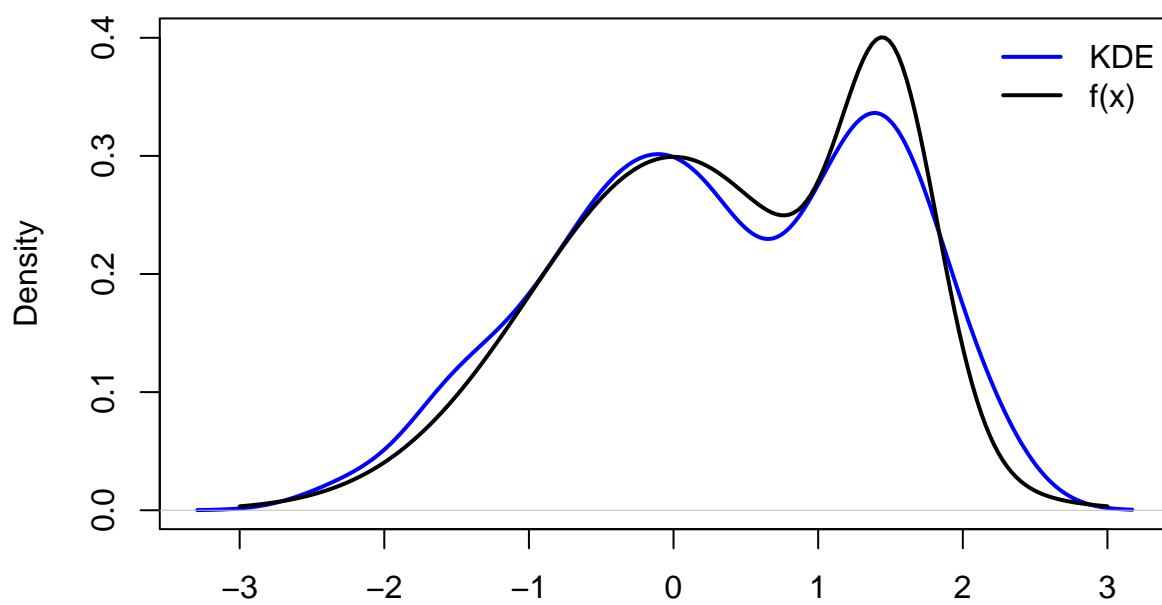
KDE (h = 0.3282)

N = 100   Bandwidth = 0.3282