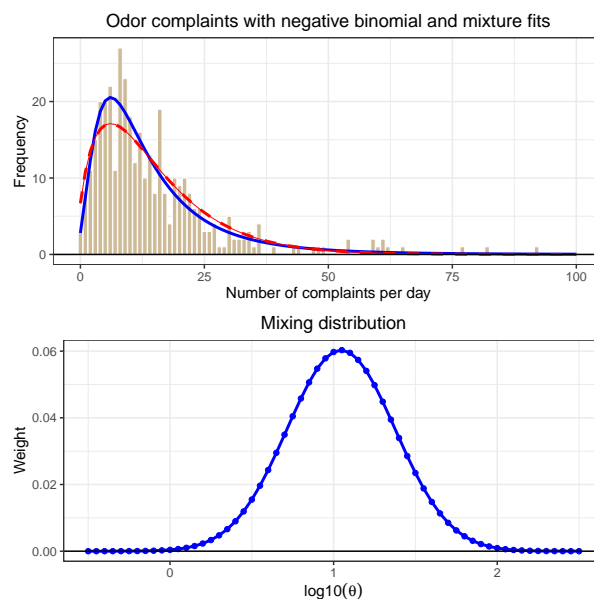


Penalized CLM and negative binomial estimation (Complaints data)



Top: distribution of daily number of complaints about odour annoyance (bars), with the estimated negative binomial distribution (broken red line) and the fit of the PCLM model (solid blue line). Bottom: the estimated mixing distributions. The dots indicate the grid points used for the model. R code in `f-complaints2.R`

```
# Penalized CLM and negative binomial estimation (Complaints data)
# A graph in the book 'Practical Smoothing. The Joys of P-splines'
# Paul Eilers and Brian Marx, 2019

library(ggplot2)
library(gridExtra)
library(JOPS)
library(MASS)

# Negative binomial probabilities
negbin = function(m, k, p) {
  f = rep(0, m + 1)
  f[1] = (1 - p) ^ k
  for (i in 1:m) f[i + 1] = f[i] * p * (k + i - 1) / i
  return(f)
}

# Deviance of an NB fit to a histogram (counts in y, mids in x)
nbdev = function(k, x, y) {
  g = sum(x * y) / sum(y)
  p = g / (g + k)
  m = length(x)
  f = sum(y) * negbin(m - 1, k, p)
  dev = sum(y * log((y + 1e-05) / f[1:m]))
  return(dev)
}

# Get the data
data(Complaints)
x = Complaints$freq
y = Complaints$count

# Optimize NB parameter
```

```

op = optimize(nbdev, interval = c(1, 2), x = x, y = y)
k = op$minimum

# Fit negative binomial distribution
g = sum(x * y) / sum(y)
p = g / (g + k)
m = length(x)
f = sum(y) * negbin(m - 1, k, p)
Fnb = data.frame(x = x, y = f)

# Prepare for PCLM
lla <- seq(-0.5, 2.5, by = 0.05)
lap <- 10 ^ lla
C <- outer(x, lap, dpois)
n <- ncol(C)

# Prepare B-spline matrix and penalty matrix
B <- bbase(lla, nseg = 10, bdeg = 3)
fit = pclm(y, C, B, lambda = 1e+06, pord = 3)
Ffit = data.frame(x = x, y = fit$mu)
gam = fit$gamma
gam = gam / sum(gam)

dev2 = sum(y * log((y + 1e-05) / fit$mu))
aic = op$objective + 4
aic2 = dev2 + 2 * fit$ed
cat("NB parameter, AIC1, AIC2:", k, aic, aic2, "\n")

Data = data.frame(x = x, y = y)

plt1 = ggplot(Data, aes(x = x, y = y, col = I("white"), fill = I("wheat3"))) +
  geom_bar(stat = "identity", size = 0.03) +
  geom_line(data = Ffit, col = "blue", size = 1) +
  geom_line(data = Fnb, col = "red", size = 1, linetype = 2) +
  geom_line(data = Fnb, col = "red", size = 0.3, linetype = 1) +
  geom_hline(yintercept = 0) +
  xlab("Number of complaints per day") + ylab("Frequency") +
  ggtitle('Odor complaints with negative binomial and mixture fits') +
  JOPS_theme()

plt2 = qplot(lla, gam, color = I("blue")) +
  geom_line(color = "blue", size = 1) +
  xlab(expression(paste("log10"(theta)))) + ylab("Weight") +
  geom_hline(yintercept = 0) +
  ggtitle('Mixing distribution') +
  JOPS_theme()

# Make and save graph
grid.arrange(plt1, plt2, nrow = 2, ncol = 1)

```
