pclm User Manual

Marius D. Pascariu, Silvia Rizzi and Maciej J. Danko

December 8, 2017

1 Abstract

The `pclm` R package introduces a versatile method for ungrouping histograms (binned count data) assuming that counts are Poisson distributed and that the underlying sequence on a fine grid to be estimated is smooth. The method is based on the composite link model and estimation is achieved by maximizing a penalized likelihood. Smooth detailed sequences of counts and rates are so estimated from the binned counts. Ungrouping binned data can be desirable for many reasons: Bins can be too coarse to allow for accurate analysis; comparisons can be hindered when different grouping approaches are used in different histograms; and the last interval is often wide and open-ended and, thus, covers a lot of information in the tail area. Age-at-death distributions grouped in age classes and abridged life tables are examples of binned data. Because of modest assumptions, the approach is suitable for many demographic and epidemiological applications. For a detailed description of the method and applications see Rizzi et al. (2015).

2 Package Structure

The package has 2 top level functions `pclm` and `pclm2D` an auxiliary function `pclm.control`, several generic functions (`plot`, `summary`, `fitted`) and a test dataset (`pclm.data`).

All functions are documented in the standard way, which means that once you load the package using `library(pclm)` you can just type for example `?pclm` to see the help file.

```
# Load the package
library(pclm)
##
## R Package  : pclm
## Name       : Penalized Composite Link Model for Efficient Estimation
##               of Smooth Distributions from Coarsely Binned Data
## Authors    : M.D. Pascariu, S. Rizzi, and M.J. Danko
## Last Update: December 15, 2017
```

3 Univariate Penalized Composite Link Model (PCLM)

The PCLM method is based on the composite link model, which extends standard generalized linear models. It implements the idea that the observed counts, interpreted as realizations from Poisson distributions, are indirect observations of a finer (ungrouped) but latent sequence. This latent sequence represents the distribution of expected means on a fine resolution and has to be estimated from the aggregated data. Estimates are obtained by maximizing a penalized

likelihood. This maximization is performed efficiently by a version of the iteratively reweighted least-squares algorithm. Optimal values of the smoothing parameter are chosen by minimizing Bayesian or Akaike's Information Criterion.

This is an example of estimation of the smooth age at death distributions from grouped death counts. First we have to define some grouped data:

```
# Input data
# x: Age groups
x <- c(0, 1, seq(5, 85, by = 5))
x
##  [1]  0  1  5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85
# y: Death counts in the age group
y <- c(294, 66, 32, 44, 170, 284, 287, 293, 361, 600, 998,
       1572, 2529, 4637, 6161, 7369, 10481, 15293, 39016)
# offset: Population exposed to risk in the age group
offset <- c(114, 440, 509, 492, 628, 618, 576, 580, 634, 657,
            631, 584, 573, 619, 530, 384, 303, 245, 249) * 1000
# nlast: the size of the last age interval (usualy open)
nlast <- 26
# This results in the last group being [85, 110).
```

3.1 Fitting and ungrouping data using PCLM

The model can be fitted using `pclm` function:

```
M1 <- pclm(x, y, nlast)
```

3.2 Output

It generates different types of output stored in the created object. See `pclm` help page for detailed information about the output list (`?pclm`).

```
ls(M1)
## [1] "bin.definition"  "call"           "ci"             "fitted"
## [5] "goodness.of.fit" "input"          "smoothPar"
```

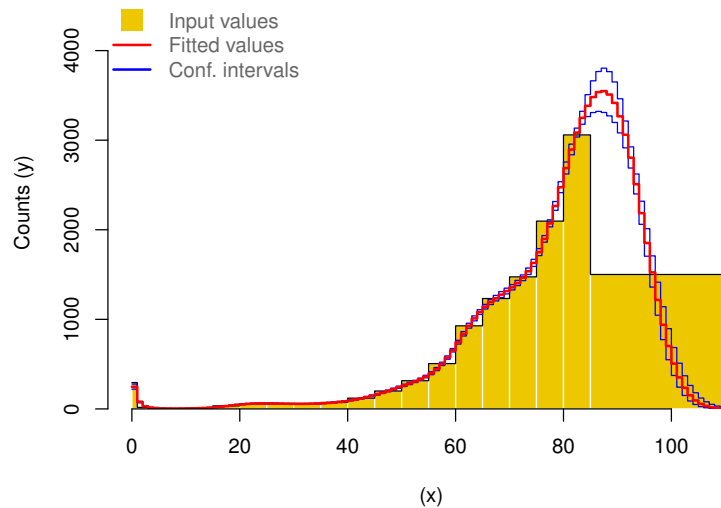## 3.3 Summary

```
summary(M1)
##
## Penalized Composite Link Model (PCLM)
##
## Call:
## pclm(x = x, y = y, nlast = nlast)
```

```
## 
## PCLM Type                 : Univariate
## Number of input groups    : 19
## Number of fitted values   : 111
## Length of estimate bins   : 1
## Smoothing parameter lambda : 1
## B-splines intervals/knot  : 6
## B-splines degree          : 3
## AIC                       : 82.75
## BIC                       : 97.24
```

3.4   plot.pclm

Generic plot:

```
plot(M1)
```
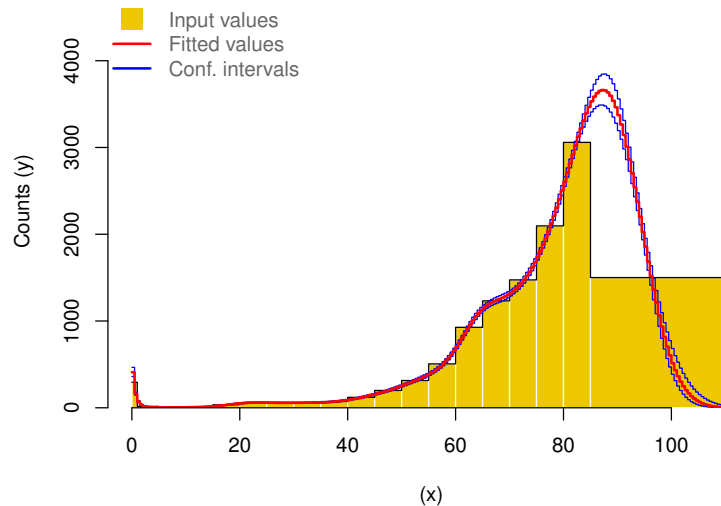


```
# Print first 6 fitted values
fitted(M1)[1:6]
##      [0,1)      [1,2)      [2,3)      [3,4)      [4,5)      [5,6)
## 247.566572  77.077259  29.441542  13.643942   7.585935   5.003955
```

3.5   out.step

By default `pclm` ungroups data in intervals of length `1`. If higher granularity is required `out.step` argument can be used to specify this. For example obtaining groups 222 groups of length 0.5 one can try:

```
M2 <- pclm(x, y, nlast, out.step = 0.5)
plot(M2)
```

```
# Print first 6 fitted values
fitted(M2)[1:6]
##    [0,0.5)    [0.5,1)     [1,1.5)    [1.5,2)     [2,2.5)     [2.5,3)
## 204.838340  80.316232  35.344804  17.410947   9.574988   5.862967
# Number of fitted values
length(fitted(M2))
## [1] 222
```

3.6   pclm.control

For controlling the PCLM fitting process `pclm.control` provides several options. The list of
arguments needs to be specified using the `control` argument. For example, if we want to optimize
the smoothing parameters in order to obtain a fit characterized by the small `AIC` level we can
write:

```
# Optimise smoothing parameters: lambda, kr and deg
M3 <- pclm(x, y, nlast, control = list(lambda = NA,
                                       kr = NA,
                                       deg = NA,
                                       opt.method = 'AIC'))
```

If we want to optimize only the degree of the splines needed to create equally-spaced B-splines
basis we can set `deg = NA` and assign a fixed value to the other smoothing parameters (`lambda`
and `kr`).

```
M4 <- pclm(x, y, nlast, control = list(lambda = 5,
                                       kr = 5,
                                       deg = NA,
                                       opt.method = 'AIC'))
```
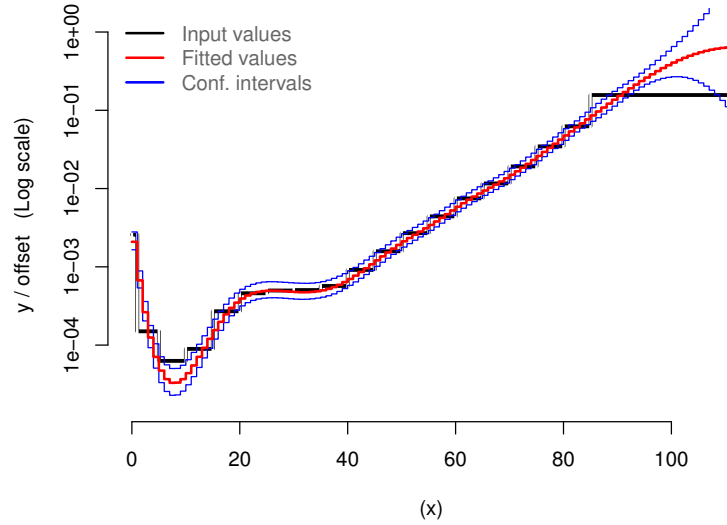
4

3.7 Offset

The `offset` argument can be used to estimate smooth death rates. `offset` must be a vector of the same length as `y`.

```
M5 <- pclm(x, y, nlast, offset)
```

Generic plot:

```
plot(M5, type = "s")
```


4 Two-dimensional Penalized Composite Link Model

The PCLM can be extended to a two-dimensional regression problem. The two-dimensional Penalized Composite Link Model to ungroup simultaneously coarse distributions for adjacent years can be fitted using `pclm2D` function, and it works in exactly the same way as `pclm`. See the examples provided in the help page. Note that `pclm2D` might be slower, depending on the data and model specification provided in the functions.