Introduction to `ungroup`

Marius D. Pascariu, Maciej J. Dańko, Jonas Schöley and Silvia Rizzi

September 1, 2018

1 Abstract

The `ungroup` 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, Gampe, and Eilers (2015).

2 Package Structure

The package has two top level functions `pclm` and `pclm2D`, two auxiliary functions (`control.pclm` and `control.pclm2D`), several generic functions (`plot`, `summary`, `fitted`, `residuals`). A dataset (`ungroup.data`) is provided as well for testing purposes.

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

```
# Load the package
library(ungroup)
```

3 Usage

3.1 Univariate Penalized Composite Link Model (PCLM)

The PCLM method (Eilers 2007) is based on the composite link model (Thompson and Baker 1981), 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 (Hastie and Tibshirani 1990).

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 (usually open)
nlast <- 26
# This results in the last group being [85, 110).
```

3.1.1 Fitting and ungrouping data using PCLM

The model can be fitted using `pclm` function:

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

3.1.2 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"              "deep"
## [5] "fitted"          "goodness.of.fit" "input"           "smoothPar"
```

3.1.3 Summary

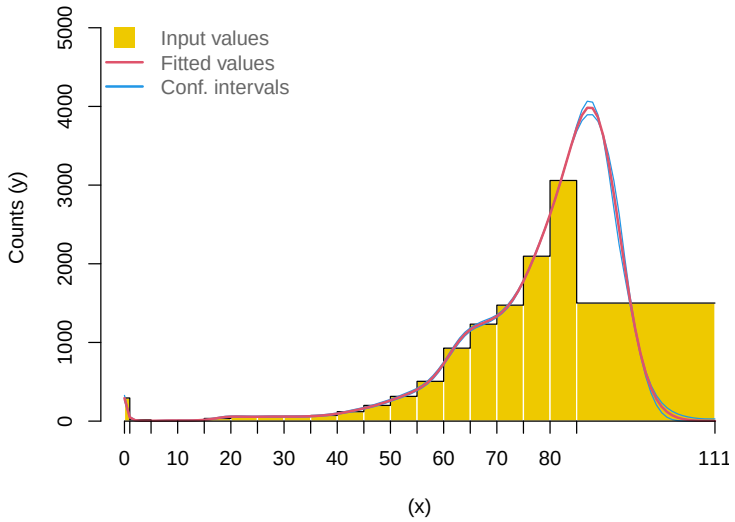
```
summary(M1)
##
## Penalized Composite Link Model (PCLM)
##
## Call:
## pclm(x = x, y = y, nlast = nlast)
##
## PCLM Type                    : Univariate
## Number of input groups       : 19
```

```
## Number of fitted values      : 111
## Length of estimate bins      : 1
## Smoothing parameter lambda   : 0
## B-splines intervals/knot (kr): 2
## B-splines degree (deg)       : 3
## AIC                          : 39.97
## BIC                          : 59.81
```

3.1.4 plot.pclm

Generic plot:

```
plot(M1)
```

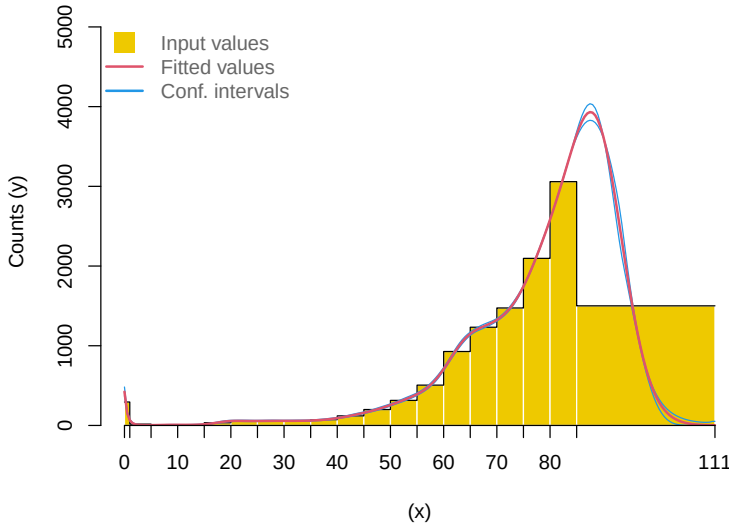



```
# Print first 6 fitted values
fitted(M1)[1:6]
##      [0,1)      [1,2)      [2,3)      [3,4)      [4,5)      [5,6)
## 292.254945  47.567040  12.031104   5.101512   3.653694   3.801854
```

3.1.5 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)
## 211.751331  80.583500  32.679925  14.663349   7.463171   4.379768
# Number of fitted values
length(fitted(M2))
## [1] 222
```

3.1.6 control.pclm

For controlling the PCLM fitting process `control.pclm` 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 one can write:

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

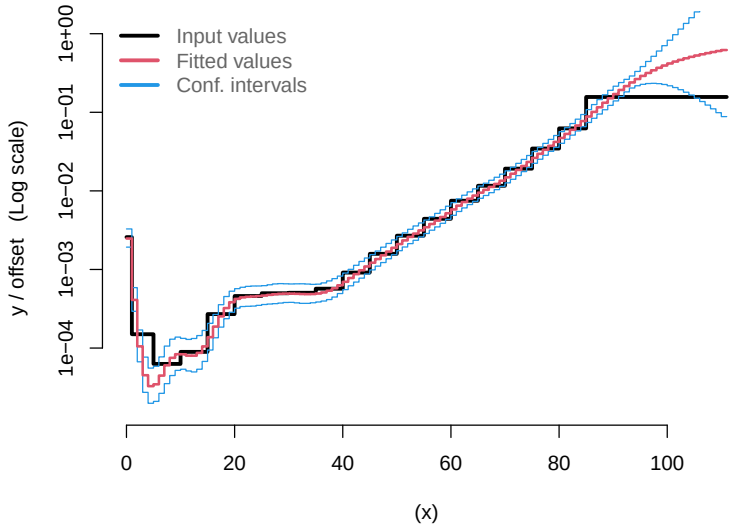
3.1.7 Offset

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")
```



3.2 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 the structure of the functions works 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.

The two-dimensional regression analysis combines two approaches: the PCLM for ungrouping in one dimension and two-dimensional smoothing with P-splines (Currie, Durban, and Eilers 2004). As an example we can ungroup age-specific distributions from the coarsely grouped data and smooth across adjacent calendar years to estimate both detailed age-at-death distributions and mortality time trends.

4 Acknowledgment

We thank Paul H.C. Eilers who provided insight and expertise that greatly supported the creation of this R package; and Catalina Torres and Tim Riffe for testing and offering feedback on the early versions of the software.

The authors are also grateful to the following institutions for their support:

- University of Southern Denmark;
- Max-Planck Institute for Demographic Research;
- SCOR Corporate Foundation for Science.

References

Currie, Iain D, Maria Durban, and Paul HC Eilers. 2004. "Smoothing and Forecasting Mortality Rates." *Statistical Modelling* 4 (4): 279–98.

Eilers, Paul HC. 2007. "Ill-Posed Problems with Counts, the Composite Link Model and Penalized Likelihood." *Statistical Modelling* 7 (3): 239–54. https://doi.org/10.1177/1471082X0700700302.

Hastie, Trevor J, and Robert J Tibshirani. 1990. "Generalized Additive Models." *Monographs on Statistics and Applied Probability* 43.

Rizzi, Silvia, Jutta Gampe, and Paul H. C. Eilers. 2015. "Efficient Estimation of Smooth Distributions from Coarsely Grouped Data." *American Journal of Epidemiology* 182 (2): 138–47. https://doi.org/10.1093/aje/kwv020.

Thompson, R, and RJ Baker. 1981. "Composite Link Functions in Generalized Linear Models." *Applied Statistics*, 125–31.