Comparing non-parametric methods for ungrouping coarsely aggregated age-specific distributions

Rizzi, Silvia; Thinggaard, Mikael; Vaupel, James W. ; Jacobsen, Rune

Comparing non-parametric methods for ungrouping coarsely aggregated age-specific distributions

Silvia Rizzi[1,2], Mikael Thinggaard[1,2], James W. Vaupel[1,2,3] and Rune Lindahl-Jacobsen[1,2]

[1]Max Planck Odense Center on the Biodemography of Aging, Odense, Denmark
[2]University of Southern Denmark, Unit of Epidemiology, Biostatistics and Biodemography
[3]Max Planck Institute for Demographic Research, Rostock, Germany

1 Background

In demography grouped data are common. Examples are abridged life tables or abridged fertility data. In this study we focus on aggregated age-specific death counts, collected usually by 5-years of age with an open-ended interval starting at age 85 [1, 2]. Such coarsely aggregated data constitute an obstacle for detailed analysis. It is useful to estimate age-specific distributions on a detailed grid of ages, e.g. by single year of age, particularly when information about nonagenarians and centenarians is needed.

Parametric and non-parametric methods have been suggested in the demographic literature [3] to split grouped data but little attention has been given to the last open-ended age intervals. Our aim here is to compare and evaluate different non-parametric approaches for ungrouping demographic data. We study non-parametric methods because of their flexibility in modelling age-specific patterns that follow very different trajectories. Popular techniques that estimate detailed distributions from coarsely grouped data are spline interpolation methods. In the Human Mortality Database [4], for example, cubic splines are fitted to the cumulative number of deaths to split aggregated death counts in single year age steps [5]. The interpolation algorithm used in the Human Fertility Database [6] is the Hermite cubic spline interpolation [7,8]. An alternative is the cubic spline with Hyman filter [9, 10]. Another non-parametric model that has been recently proven to efficiently ungroup aggregated data is the penalized composite link model [11, 12].

We examine the performance of these different ungrouping methods in an empirical application. To do so we compare original NORDCAN data by single-year of age with the estimated distributions resulting from the models. We show that the penalized composite link model outperforms spline interpolation methods in presence of wide open-ended intervals.

2 Methods and Data

For our comparison study we chose the novel penalized composite link model for ungrouping [11] and commonly used spline interpolation methods that always result in positive estimates, i.e. the cubic spline with Hyman filter [9, 10] and the Hermite cubic spline interpolation [7, 8]. All three methods are ready to use in the statistical software R. For the penalized

composite link model, R code is provided in Appendix 2 of the paper. The spline interpolation with Hyman filter is implemented in the demography R package [13] under *cm.spline* function; while the piecewise cubic Hermite interpolating polynomial can be found in the signal R package [14] under *interp1* function with *"pchip"* method.

We test the models against age-specific deaths from colorectal cancer for Denmark in years 1980, 1990, 2000 and 2010 combined. Data were obtained from the Danish Cancer Society [15]. All deaths are collected by single-year of age from age 0 up to the last age of recorded events. They can therefore serve as a "golden standard" for comparison with the estimates of the different ungrouping methods. To compare the performance of the selected methods, we artificially grouped the death counts according to two grouping schemes: First into 5-year age classes and then into 5-year age classes with an open-ended interval starting at age 85. While in theory the tail area could be unlimited, in age-at-death applications there is a maximum number beyond which no observation is expected. We set 105 as the maximum age and we complete the histogram with an age group from 105 to 115 with 0 counts. This allows to efficiently estimate age-specific distributions with wide, open-ended age intervals. We apply the different models to the artificially grouped data and provide graphical representation of the different ungrouped estimates against the empirical counts. For additional comparison we propose the integrated squared error (ISE).

3 Results

Figure 1 reports the estimated distributions together with the empirical data.

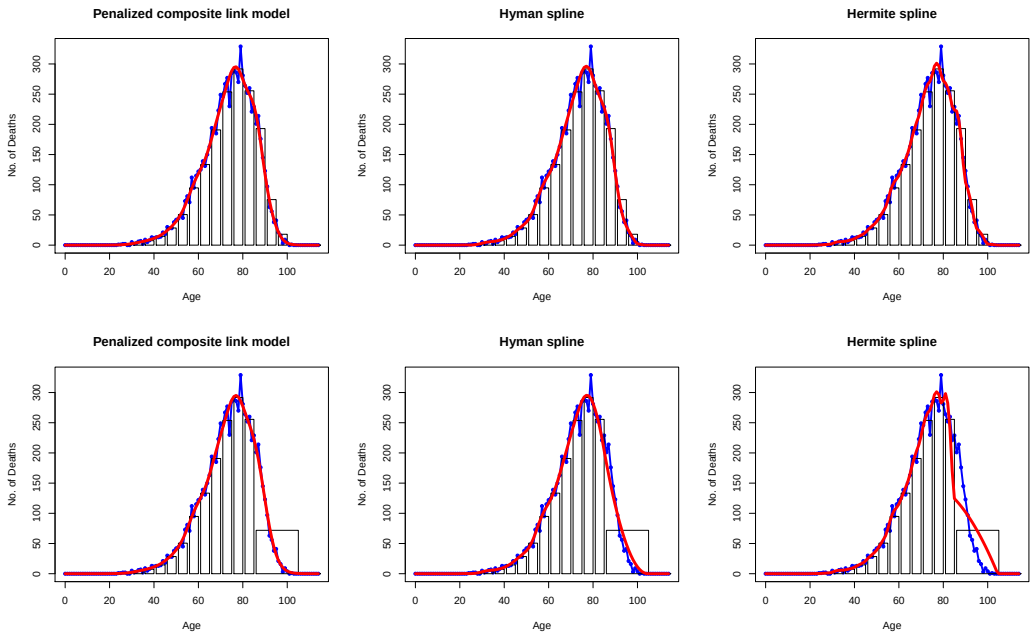

Figure 1: Age-at-death for colorectal cancers in Denmark for 1980, 1990, 2000 and 2010. Empirical data (blue line with overplotted points). Top: Models' estimates from 5-years age groups (red smooth lines). Bottom: Models' estimates from 5-years age groups with open-ended age interval 85+ (red smooth lines).

For equal interval length of 5 years the three methods perform well and similarly. In the setting of open-ended interval, the Hermite spline does not accurately redistribute the observations into the tail area. The spline with Hyman filter also looses goodness of fit. The penalized composite link model seems on the contrary not affected by the wide open-ended age interval.

To better highlight the comparison between the models we report the integrated squared error (ISE) in the table below. The penalized composite link model (pclm) clearly outperforms the spline interpolation methods in case of the open-ended interval.

Method	pclm	hyman spline	hermite spline
Bins of 5-years	0.000139	0.000131	0.000144
Bins of 5-years with 85+	0.000136	0.000219	0.000924

4 Conclusion

We have compared different methods to split age-specific aggregated data into a fine grid of single-year of ages. The strength of these non-parametric methods is that they can model a wide range of distributions. Indeed age-at-death from various causes of deaths or age-at-onset of different diseases follow disparate age-specific patterns, e.g. bimodal, skewed to the right or to the left. Other relevant examples in demography are uncompleted cohort fertility patterns by age-groups, age-at-first marriage with open-intervals at the left and right hand side of the distribution or onset of infectious diseases grouped in days or weeks. All methods can work with input data grouped in intervals of unequal width and can cope with groups of 0 counts. They can also be extended to estimate detailed age-specific rates. We found that for age groups of 5-years age length, the selected methods show similar results. However for wider groupings such as open age intervals, the penalized link model performs best.

References

[1] World Health Organization, Media center Fact sheets. Available at `http://www.who.int/whosis/mort/download/en/index.htm`. Last access May 4, 2015.

[2] European Commission, Eurostat). Statistics Database, Luxembourg. Available at `http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database`. Last access May 4, 2015.

[3] Kostaki A. Panousis V. Expanding an abridged life table. *Demographic Research*, 5:1–22, 2001.

[4] Human Mortality Database (HMD). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). `http://www.mortality.org/`. Last access May 4, 2015.

[5] Wilmoth J.R. Andreev K. Jdanov D. Glei D.A. Methods protocol for the human mortality database. http://www.mortality.org/public/docs/methodsprotocol.pdf. Last revised May 31, 2007.

[6] Human Fertility Database (HFD).Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). `http://www.humanfertility.org`. Last access August 10, 2015.

[7] Jasilioniene A. Jdanov D. A. Sobotka T. Andreev E. M. Zeman K. Shkolnikov V. M. Methods protocol for the human fertility database. http://www.humanfertility.org/docs/methods.pdf. Last revised July 30, 2015.

[8] Fritsch F.N. Carlson R.E. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17:238–246, 1980.

[9] Hyman J.M. Accurate monotonicity preserving cubic interpolation. *Journal on Scientific and Statistical Computing*, 4:645–654, 1983.

[10] Smith L. Hydman R.J. Wood S.N. Spline interpolation for demographic variables: The monotonicity problem. *Journal of Population Research*, 21:95–98, 2004.

[11] Rizzi S. Gampe J. Eilers P.H.C. Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*, 182:138–147, 2015.

[12] Eilers P.H.C. Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7:239–254, 2007.

[13] Rob J. Hyndman. *demography*. R package version 1.18 `https://cran.r-project.org/web/packages/demography/index.html`, downloaded on August 10, 2015.

[14] Uwe Ligges et al. *signal*. R package version 0.7-6 `https://cran.r-project.org/web/packages/signal/index.html`, downloaded on August 10, 2015.

[15] Engholm G. Ferlay J. Christensen N. Kejs A.M.T. Johannesen T.B. Khan S. Milter M.C. Ólafsdóttir E. Petersen T. Pukkala E. Stenz F. Storm H.H. NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 7.0 (17.12.2014). Association of the Nordic Cancer Registries. Danish Cancer Society. Available from http://www.ancr.nu. Last access October 10, 2015.