AN R SNIPPET FOR ADAPTIVE BETA KERNEL GRADUATION. AN APPLICATION TO ITALIAN MORTALITY DATA

Anna Maria Altavilla, Angelo Mazza, Antonio Punzo

1. Introduction

Mortality rates are age-specific indicators commonly used in demography. They are also widely adopted by actuaries, in the form of mortality tables, to calculate life insurance premiums, annuities, reserves, and so on. Producing these tables from a suitable set of crude (or raw) mortality rates is called graduation.

To be specific, the d_x deaths at age x can be seen as arising from a population, initially exposed to the risk of death, of size e_x. This is commonly summarized via the model $d_x \sim Bin(e_x, q_x)$, where $q_x$represents the true, but unknown, probabiliy of dying at age x.

The crude rate $\dot{q}_x$ is the observed counterpart of q_x. Graduation is necessary because crude data usually presents abrupt changes, which do not agree to the dependence structure supposedly characterizing the true rates (London, 1985). In fact, a common prior opinion about their form is that each true mortality rate is closely related to its neighbors. So, the logical step is to graduate the crude rates to produce smooth estimates, $\hat{q}_x$, of the true rates. This is done by systematically revising the crude rates, in order to remove any random fluctuations.

In Altavilla, Mazza and Punzo (2011) was introduced a function for the R statistical environment (R Core Team, 2013) allowing for nonparametric graduation using the discrete beta kernel estimator proposed by Mazza and Punzo (2011). Kernel smoothing is one of the most popular statistical methods for nonparametric graduation. The genesis of this model starts with the consideration that, although age X is in principle a continuous variable, it is typically truncated in some way, such as age at last birthday, so that it takes values on the discrete set $\mathcal{X} = \{0,1,...,\omega\}$, ω being the highest age of interest. In this model, discrete beta distributions are adopted as kernel functions, in order to overcome the problem of boundary bias commonly arising from the use of symmetric kernels. The support $\mathcal{X}$ of the discrete beta, which can be asymmetric, in fact matches the age range and this, when smoothing is made near the boundaries, allows avoiding the allocation of weight outside the support (e.g. negative or unrealistically high ages). Mazza and Punzo (2013a) propose an adaptive bandwidth discrete beta kernel estimator, in which the bandwidth

is allowed to vary at each age, according to the reliability of the data as expressed by the e_x, while in Mazza and Punzo (2013b) a different approach, in which reliability is measured via the reciprocal of the *variation coefficient* (VC), was proposed.

Here we will present an R code snippet that allows for both the adaptive bandwidth discrete beta kernel estimator based on the VC. An application to mortality data relative to the Italian male population for the year 2009 is presented. The code snippet and the data used in the paper may be requested to the authors.

2. The adaptive discrete beta kernel estimator

Given the crude rates $\dot{q}_y$, y $\in \mathcal{X}$, the Nadaraya-Watson kernel estimator of the true but unknown mortality rates q_x the evaluation age x is

$$\hat{q}_x = \sum_{y\in\mathcal{X}} \frac{k_h(y;m=x)}{\sum_{j\in\mathcal{X}} k_h(j;m=x)} \dot{q}_y = \sum_{y\in\mathcal{X}} K_h(y;m=x)\dot{q}_y, \quad x \in \mathcal{X}, \tag{1}$$

where $k_h(\cdot;m)$ is the discrete kernel function (hereafter simply named kernel), $m \in \mathcal{X}$ is the single mode of the kernel, $h > 0$ is the so-called bandwidth governing the bias-variance trade-off, and $K_h(\cdot;m)$ is the normalized kernel. As kernels, in (1) we use

$$k_h(x;m) = \left(x + \frac{1}{2}\right)^{\frac{m+\frac{1}{2}}{h(\omega+1)}} \left(\omega + \frac{1}{2} - x\right)^{\frac{\omega+\frac{1}{2}-m}{h(\omega+1)}}. \tag{2}$$

In the normalized version, $K_h(\cdot;m)$ corresponds to the discrete beta probability mass function of Punzo and Zini (2012), parameterized according to the mode m and another parameter h that is closely related to the distribution variability. Substituting (2) in (1), we obtain the discrete beta kernel estimator that was introduced in Mazza and Punzo (2011).

Discrete beta kernels possess two peculiar characteristics. Firstly, their shape, fixed h, automatically changes according to the value of m. Secondly, the support of the kernels matches the age range $\mathcal{X}$, so that no weight is assigned outside the data support; this means that the order of magnitude of the bias does not increase near the boundaries. Further details are in Mazza and Punzo (2011). In (1), a low value of h puts more emphasis on fit than on smoothness. Rather than restricting h to a fixed value, a more flexible approach is to allow the bandwidth to vary according to the reliability of the data measured in a convenient way. Thus, for ages in which the reliability is relatively larger, a low value for h results in an estimate that more closely reflects the crude rates. For ages in which the reliability is smaller,

such as at old ages, a higher value for *h* allows the estimate of the true rates of mortality to progress more smoothly; this means that at older ages we are calculating local averages over a greater number of observations. This technique is referred to as a variable or *adaptive kernel estimator* because it is characterized by an adaptive bandwidth $h_x(s)$ which depends on the reliability l_x and is function of a further sensitive parameter *s*.

The reliability l_x can be inserted into the basic model (1) in a number of ways (Gavin *et al.*, 1995); here we adopt a natural formulation according to which

$$h_x(s) = hl_x^s, \quad x \in \mathcal{X}$$

where *h* is the global bandwidth and $s \in [0,1]$. Reliability decides the shape of the local factors, while *s* is necessary to dampen the possible extreme variations that can arise between young and old ages. If $s = 0$, we have the fixed bandwidth estimator.

Using (2) we are calculating a different bandwidth for each age, leading model (1) to become

$$\hat{q}_x = \sum_{y\in\mathcal{X}} \frac{k_{h_x}(y;m=x)}{\sum_{j\in\mathcal{X}} k_{h_x}(j;m=x)} \dot{q}_y = \sum_{y\in\mathcal{X}} K_h(y;m=x)\dot{q}_y, \quad x \in \mathcal{X}, \tag{3}$$

where the notation $h_x$is used to abbreviate $h_x(s)$. Thus, for each evaluation age *x*, the $\omega + 1$ discrete beta distributions $K_{h_x}(\cdot; m = x)$ vary for the placement of the mode as well as for their variability.

According to the model $d_x \sim Bin(e_x, \dot{q}_x)$, where $\dot{q}_x$ is the maximum likelihood estimate of q_x, a natural index of reliability is represented by the reciprocal of a relative measure of variability. As relative measure of variability, Mazza and Punzo (2013b) adopt the variation coefficient, which can be computed as

$$VC_x = \frac{\sqrt{e_x\dot{q}_x(1-\dot{q}_x)}}{e_x\dot{q}_x}, \quad x \in \mathcal{X},$$

and it is normalized, so that $l_x^s \in [0,1]$.

In (3) two parameters need to be selected: sensitivity, *s*, and global bandwidth, *h*. Although both parameters could be selected by cross-validation, we prefer to choose *s* subjectively, as in Gavin *et al.* (1995, see also Mazza and Punzo, 2013b). Once *s* has been chosen, cross-validation can be still used to select *h;* instead of the standard residual sum of squares, Mazza and Punzo (2011) suggest the use of the sum of the squares of the proportional differences

$$CV(h|s) = \sum_{x\in\mathcal{X}} \left(\frac{\dot{q}_x}{\hat{q}_x^{(-x)}} - 1\right)^2; \tag{4}$$

this is a commonly used divergence measure in the graduation literature because, since the high differences in mortality rates among ages, we want the mean relative square error to be low (see Heligman and Pollard, 1980).

3. Discrete beta kernel graduation using the R statistical environment

This section introduces the essential elements needed for doing the adaptive discrete beta kernel graduation using the code snippet that we developed for the `R` statistical environment (R Core Team, 2013). The main function, `adpDbkGrad`, does the adaptive beta kernel graduation. Its arguments are:

- `obsqx`, a numeric vector, containing the observed mortality rates;
- `exposures`, a numeric vector of the same size of `obsqx`, providing the exposed to the risk. It is required if the bandwidth is adaptive;
- `bwtype`, a string, allowing the user to select the type of bandwidth. It may be `"FX"` for a fixed bandwidth and `"VC"` for an adaptive bandwidth based on the variation coefficient;
- `h` and `s`, scalars, providing values for the two smoothing parameters;
- `cv`, a logical. If it is `TRUE` then `h` is computed by means of cross-validation;
- `cvres`, a string. If it is `"propres"` then cross-validation selects `h` by minimizing the proportional sum of squares in (4), while if it is `"res"` the standard sum of square residuals is minimized. Default value is `"propres"`;
- `logit`, a logical. If it is `TRUE` then a logit transformation is applied to the data before graduating, and then data are back-transformed to obtain the estimate of the true rates; its default value is `FALSE`;
- `omega`, a scalar, setting the upper age limit. Its default value is the length of the vector `obsqx` minus one.

In the cross-validation routine, minimization is performed using the Levenberg-Marquardt nonlinear least-squares algorithm, as implemented in the package `minpack.lm` (Elzhov et al., 2010); this package has to be installed before running our code.

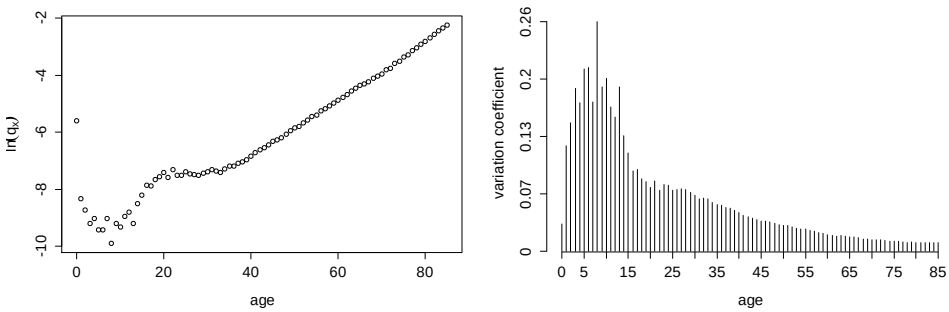
4. An application to the 2009 Italian males.

The `adpDbkGrad` function will be applied to the 2009 Italian males' probabilities of dying and January 1st population, for the age range 0 to 85. Data come from the Human Mortality Database (2013).

To begin the analysis, program and data have to be loaded; if we assume that within the `R` working directory there are a text file named "adpDbkGrad.R" containing the code in appendix and a file named "ItalyM2009.RData" containing the two variables `obsqx` and `exposure`, this may be done with the commands

```
R> source("adpDbkGrad.R")
R> load("ItalyM2009")
```

Figure 1 – *Observed probabilities of dying (on the left), and variation coefficient (on the right), for the 2009 Italian males over the age range.*



The plot in Figure 1 shows the observed probabilities of dying and the estimated variation coefficient over the age range. As usual in the graduation literature, a logarithmic scale is used for the mortality rates and this makes variations at lower orders of magnitude more visible. It may be noted easily that whenever the estimated variation coefficient is lower, such as at older ages, observed data appear smoother, while where there is a greater variability, such as for ages around 8, data exhibit a bumpier behavior. The differences in variability over the age range points out to the usefulness of an adaptive approach. The command used to perform the graduation, using a variation coefficient based adaptive bandwidth is:

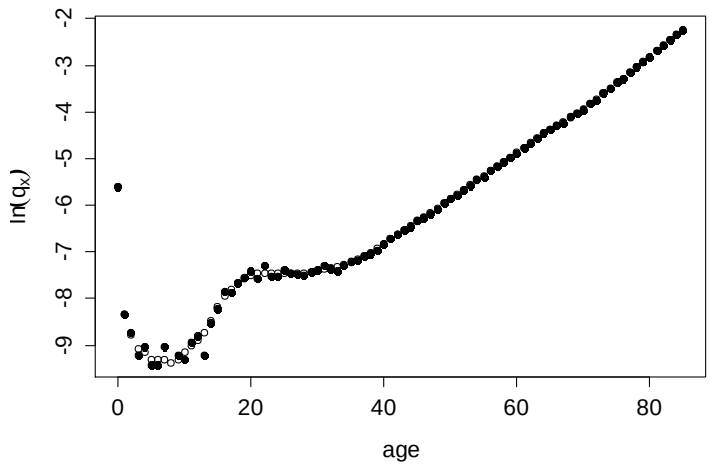
```
R> qxest <- adpDbkGrad (obsqx=obsqx, exposure=exposure,
 +          bwtype="VC")
```

Default values have been used for all the omitted arguments, so `s` is equal to 0.2, cross-validation has been used to select `h`, by minimizing the proportional sum of squares (`cvres="propres"`), `omega` has been set at the highest observed age (`length(obsqx)-1`), and the logit transformation has not been applied (`logit=FALSE`).

Both the observed and graduated data are depicted in Figure 1. The plot shows how graduated data retain the important aspects coming from the changes of the mortality rates and, at the same time, leave out random noise. It is possible to note a small but prominent hump, peaking around 19 years of age. This excess mortality is known in literature as "accidental hump"; risk-taking and surplus mortality are signatures of the male human's early adult years. Although the statistical influence of the accidental hump on survival and life expectancy is small, on a logarithmic-scale the hump is visible relative to the low mortality typical of the final stages of male puberty. Since $\dot{q}_x$ from age 25 to about 80 grows at an approximately exponential rate, with the logarithmic scale data points in that region almost lay over a straight line. Plots were obtained using the standard plot function, as:

```
R> plot(0:(length(qxest)-1), log(qxest), type="p",
+       xlab="age", ylab=expression(ln(q[x])))
R> points(0:(length(qxest)-1),log(obsqx), pch=19)
```

Figure 2 – *Observed (◦) and graduated (•) male Italian mortality rates, in logarithmic scale for the year 2009.*



5. Conclusions

In this paper we have discussed an R function specifically conceived for adaptive nonparametric graduation of discrete finite functions, such are age-dependent mortality data, proposed by Mazza and Punzo (2013a, 2013b). Over other graduation techniques, this one has the advantage that kernel functions are chosen from a family of conveniently discretized and re-parameterized beta densities; since their support matches the age range boundaries, the estimates are free of boundary bias. The adaptive bandwidth features estimates that more closely reflect the crude rates at ages in which reliability is larger, and smoother estimates at ages in which reliability is smaller, such as at old ages. Reliability may be based on exposure or on the variation coefficient.

An application to 2009 mortality data for Italian males has been proposed. Since the area considered is relatively small, random variations made raw data slightly noisy. Graduated data, on the other side, were smooth and consistent.

References

ALTAVILLA AM, MAZZA A, PUNZO A. 2012. Beta kernel graduation of mortality data in R. An application to the Enna province. *Rivista Italiana di Economia, Demografia e Statistica*, vol. LXVI – N. 3/4, p. 15-22, ISSN: 0035-6832.

ELZHOV TV, MULLEN KM, BOLKER B 2010. minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK. *R package version 1.1-6.*

GAVIN, J., HABERMAN S., AND VERRALL R. 1995. Graduation by kernel and adaptive kernel methods with a boundary correction. *Transactions of the Society of Actuaries*, 47, pp.173–209.

HELIGMAN L., POLLARD J. 1980. The age pattern of mortality. *Journal of the Institute of Actuaries*, 107, pp.49–80.

HUMAN MORTALITY DATABASE 2013. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org (data downloaded on March 2013).

MAZZA A., PUNZO A. 2011. Discrete beta kernel graduation of age-specific demographic indicators. In INGRASSIA S., ROCCI R., VICHI M., (Eds), *New Perspectives in Statistical Modeling and Data Analysis*, Springer, pp. 127–134.

MAZZA A., PUNZO A. 2013a. Graduation by Adaptive Discrete Beta Kernels. In Giusti A, Ritter G, Vichi M (Eds.), *Classification and Data Mining*, Springer, pp. 243-250.

MAZZA A., PUNZO A. 2013b. Using the Variation Coefficient for Adaptive Discrete Beta Kernel Graduation. In Giudici P, Ingrassia S, Vichi M (Eds.), *Statistical Models for Data Analysis*, Springer International Publishing, pp. 225-232.

PUNZO A., ZINI A (2012). Discrete Approximations of Continuous and Mixed Measures on a Compact Interval. *Statistical Papers*, 53(3), 563-575.

R Core Team (2013). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria.

SUMMARY

Various approaches have been proposed in literature for the kernel graduation of mortality rates. Among them, this paper considers an adaptive bandwidth discrete beta kernel estimator, a recent proposal conceived to intrinsically reduce boundary bias and in which age is pragmatically considered as a discrete variable. Furthermore, the adaptive bandwidth features estimates that more closely reflect the crude rates at ages in which reliability is larger, and smoother estimates at ages in which reliability is smaller, such as at old ages. Reliability is based on the variation coefficient. In this paper, we present an implementation of this estimator for the R statistical environment. An application to 2009 male Italian mortality data is also presented.

Anna Maria ALTAVILLA, Professor, University of Catania, Department of Economics and Business, altavil@unict.it

Angelo MAZZA, Assistant Professor, University of Catania, Department of Economics and Business, a.mazza@unict.it

Antonio PUNZO, Assistant Professor, University of Catania, Department of Economics and Business, antonio.punzo@unict.it