Graduation of mortality rates revisited

by

Takis Papaioannou and Athanasios Sachlas

takpap@unipi.gr asachlas@unipi.gr

University of Piraeus

1. Introduction

In this paper we critically review the existing methods of graduation or smoothing of mortality rates. We consider parametric methods such as methods based on mortality models [Gompertz, Makeham, Oppermann, Theile and Steffenson, Beard, Barnett, Heligman and Pollard et al.], generalized linear models, splines, smooth – junction interpolation. We also consider non – parametric methods of graduation such as the graphical method, weighted moving averages, Whittaker and Henderson where both the fit and the degree of smoothness are taken into account, the kernel method and graduation with reference to standard mortality rates (distribution). This last approach leads to methods of graduation using information theoretic ideas. Starting with Brockett’s idea to use the Kullback – Leibler divergence we explore the use of other divergence indices as for example the power divergence index with analogous linear and/or quadratic constraints.

2. Review of Methods of Graduation

One of the most important tasks in actuarial science is to describe the actual but unknown mortality pattern of a population. In order to achieve this, the actuary calculates from raw data the crude mortality rates, which usually form an irregular series. Because of this, it is common to revise the initial estimates with the aim of producing smoother estimates, with a procedure called graduation (London, 1985).

Graduation has two basic characteristics: smoothness and goodness of fit to the observed data. These two characteristics are in competition and in order to achieve one of them we have to sacrifice the other. Smoothness is usually measured by

\[ S = \sum_{x=1}^{n-1} (\Delta^3 v_x)^2 \]

while goodness of fit (fidelity) by

\[ F = \sum_{x=1}^{n} w_x (u_x - v_x)^2 \]

where \( v_x \) are
the graduated values, \( u_x \) are the initial values and \( w_x \) are weights. As weights it is usually used the reciprocal of the variance of \( U_x \), where \( U_x \) is the random variable that corresponds to the initial estimates \( u_x \).

In actuarial science, graduation might be done on the initial estimates of force of mortality \( \dot{\mu}_x \) or the death probabilities \( \dot{q}_x \) (Haberman, 1998). In order to graduate the force of mortality we assume that deaths are distributed as Poisson with parameter \( r_x^c \mu_x \), while when we want to graduate death probabilities we assume that deaths are modeled as Binomial distribution with parameters \( l_x \) and \( q_x \), where \( r_x^c, \mu_x, l_x \) and \( q_x \) are the “central exposure to risk”, the true but unknown force of mortality, the number of people at risk and the true but unknown death probabilities respectively.

There are lots of methods through which graduation can be obtained and they are basically classified into parametric and nonparametric ones. Through parametric methods one or more parametric models are being fitted to the initial estimates and so the graduated rates are being calculated. In nonparametric methods data are being combined at different values of the age and with appropriate techniques the graduated values are obtained.

The methods that fall under the parametric category are methods based on mortality models, generalized linear models, splines and smooth - junction interpolation. In graduation through mortality models we fit a model such as Gompertz, Makeham, Oppermann, Theile and Steffenson, Beard, Barnett, Heligman and Pollard et al. and we determine their parameters through a formal procedure such as maximum likelihood estimation. Some of the models refer to the force of mortality while others to death probabilities. In graduation through generalised linear models we fit the appropriate generalised linear model supposing that the number of deaths are modeled as independent realisations of Poisson random variables with parameter \( r_x^c \mu_x \), where \( \mu_x \) denotes that the force of mortality at age \( x + (1/2) \) and this supposes that it is constant in \((x, x+1)\). The same results are obtained for the graduation of force of mortality if we assume that the central exposures to risk are modeled as independent realisations of gamma random variables conditional to deaths \( d_x \), i.e.
\( \text{Gamma}(d_x, \mu, \nu) \). Under the assumption that the deaths are modeled as independent realisations of Binomial random variables with parameters \( l_x \) and \( q_x \), we graduate the death probabilities. In graduation through splines, cubic polynomials are being fitted over subranges of the data and special attention must be paid to the manner in which adjacent fitted function meets each other. The parameters of the spline function are being determined by the method of least squares. Smooth – junction interpolation is used when a limited number of initial estimates is known and a different interpolating arc is being fitted in each subrange of the data.

The existing nonparametric methods are the graphical ones, weighted moving averages, Whittaker and Henderson method, the kernel method, graduation with reference to standard mortality rates and graduation using information theoretic ideas. In the graphical method we try to fit by hand a smooth curve that passes as closer as possible to the initial estimates. In the graduation through weighted moving averages each graduated value is produced as a weighted average of \( 2m + 1 \) initial values while in the Whittaker and Henderson method we minimize the function

\[
M = F + hS = \sum_{x=1}^{n} w_x (u_x - v_x)^2 + h \sum_{x=1}^{n} \Delta^3 v_x^2
\]

with respect to the graduated values \( v_x \).

In the kernel method we use the kernel estimators in order to graduate the death probabilities while in the method with reference to standard mortality rates we assume that the graduated values exhibit a pattern similar to that of the standard mortality rates. Finally, Brockett (1991) minimize the Kullback – Leibler divergence subject to mathematical and actuarial constraints in order to obtain a series of values – the graduated ones – that are the least indistinguishable from the initial estimates.

A question that arises is which is the best method for graduation of actuarial data? There is no an explicit answer in bibliography. It is in the actuary’s ease which method he will use. However, there are some factors that can guide him to his decision. The first factor is how smooth the graduated values he wants to be. If we want to give emphasis in smoothness the graphic method is not recommended. Also the parametric methods are assumed to give smooth results and the degree of smoothness can only change if we use another model. Nonparametric methods enable the actuary to vary the amount of smoothness by varying the value of some parameters.
Another important factor is the range and form of the actuarial data. For example, if the rates are grouped and we want graduated values for all the values of age a parametric method should be used. Also, if we have a few initial estimates the graphical method and graduation with reference to standard mortality rates are appropriate while the weighted moving averages method is not recommended because it does not give graduated values for the m first and m last initial values.

A third factor is the selection of the parameters being displayed in the methods. In some methods, the easy interpretation of the parameters and their effect can guide the researcher to the selection of their values, as in graduation through mortality models, while in other methods he has to use different values of the parameters and then to compare the results, as it happens with the smoothing parameter in the Whittaker – Henderson method and Brockett’s method.

The factor of the amount of numerical computations is not nowadays important because it is relatively easy to apply all the methods using appropriate computer programming.

3. Information Theoretic Graduation

3.1 Kullback - Leibler Statistic

Zhang and Brockett (1987) tried to construct a smooth series of death probabilities \( \{v_x\} \) which is as closer as possible to the observed series \( \{u_x\} \) and in addition they assumed that the true but unknown underlying mortality pattern is (i) smooth, (ii) increasing with age, i.e. monotone, (iii) more steeply increasing in higher ages, i.e. convex. They also assumed that (iv) the number of deaths in the graduated data equals the number of deaths in the observed data, and (v) the total age of death in the graduated data equals the total age of death in the observed data. By the term total age of death we mean the sum of the product of the number of deaths at every age by the corresponding age.

The constraints (i) – (v) may be written in matrix notation as:

\[
(i) \ (Av)^T (Av) = v^T A^T Av \leq M
\]
where $A = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 3 & -3 & 1 & \cdots & 0 \\ 0 & 0 & -1 & 3 & -3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$, \textit{smoothness}

(ii) $Bv \geq 0$

where $B = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$, \textit{increasing with age}

(iii) $Cv \geq 0$

where $C = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$, \textit{convexity}

(iv) $I^T v = I^T u$ \textit{equality in numbers of deaths}

where $I = (l_x, l_{x+1}, \ldots, l_{x+n})^T$ and

(v) $m^T v = m^T u$ \textit{equality in total age}

where $m = (xl_x, (x+1)l_{x+1}, \ldots, (x+n)l_{x+n})^T$ and $^T$ denotes the transposition of the matrix or the vector.

In order to obtain the graduated values, Zhang and Brockett (1987) minimize the Kullback–Leibler divergence $I^{KL}(v, u) = \sum_{x=1}^{a} v_x \ln \frac{v_x}{u_x}$ subject to the constraints (i) – (v) by considering a dual problem of minimization. So instead of minimizing
\[ I^{KL}(\mathbf{v},\mathbf{u}) \text{ subject to } \mathbf{v} \geq \mathbf{0} \text{ and } g_i(\mathbf{v}) = \frac{1}{2} \mathbf{v}^T \mathbf{D}_i \mathbf{v} + \mathbf{b}_i^T \mathbf{v} + c_i \leq 0, \quad i = 1,2,...,r, \]

where \( \mathbf{D}_i \) is a positive semidefinite matrix for each \( i \) and \( \mathbf{b}_i, c_i \) are constants, they maximize the dual problem which is:

\[
\begin{aligned}
\text{maximize} & \quad -\mathbf{v}^T \exp \left[ -\sum_{i=1}^{r} y_i (\mathbf{A}_i^T \mathbf{w}_i + \mathbf{b}_i) \right] + \mathbf{c}^T \mathbf{y} - \frac{1}{2} \sum_{i=1}^{r} y_i |\mathbf{w}_i|^2 \\
\text{subject to} & \quad \mathbf{y} \geq \mathbf{0} \quad \text{and} \quad \mathbf{w}_i \in \mathbb{R}^n.
\end{aligned}
\]

It is easy to see that the constraints (i) – (v) may be written in the form \( g_i(\mathbf{v}) \).

### 3.2 Power Divergence Statistic

Starting with Brockett’s idea of minimizing the Kullback – Leibler divergence in order to find the best fitting series of graduated values \( \{v_i\} \) subject to the constraints (i) to (v), in this paper we explore the use of the power divergence index.

Cressie and Read (1984) introduced the family of power divergence statistics, defined as

\[ 2n I^\lambda(p,q) = \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^{q} p_i \left[ \left( \frac{p_i}{q_i} \right)^{\lambda} - 1 \right], \quad (3.2.1) \]

where \( p, q \) are two probability measures and \( \lambda \) is a real valued parameter, chosen by the user. For the values \( \lambda = 0 \) and \( \lambda = -1 \) the statistic is defined as the limit of \( 2n I^\lambda(p,q) \) as \( \lambda \to 0 \) and \( \lambda \to -1 \), respectively.

It can be easily seen (Cressie and Read, 1988) that for \( \lambda = 1 \) statistic (3.2.1) is equal to \( X^2 \) statistic, for \( \lambda = 0 \) is equal to \( G^2 \) statistic, for \( \lambda = -1 \) is equal to the Kullback – Leibler divergence, for \( \lambda = -(1/2) \) equals the Freeman – Tukey statistic \( F^2 \) and for \( \lambda = -2 \) equals the Neyman – modified \( X^2 \). In order to take the desirable properties of both \( G^2 \) and \( X^2 \) statistics, Cressie and Read (1984) proposed a statistic which lies between of them taking \( \lambda = 2/3 \), i.e. statistic (3.2.1) becomes

\[ 2n I^{3/2}(p,q) = \frac{9}{5} \sum_{i=1}^{q} p_i \left[ \left( \frac{p_i}{q_i} \right)^{3/2} - 1 \right]. \quad (3.2.2) \]

The power divergence has the properties of other measures of divergence such as nonnegativity, symmetry, continuity, nonadditivity and strong nonadditivity.
3.3 Mortality Data Graduation Using the Power Divergence

The problem is to find the best fitting values $v_x$, which satisfy the mathematical and actuarial constraints (i) to (v) and are the least distinguishable from the initial estimates $u_x$. The above constrained problem can be easily solved by using any of the readily available non-linear programming codes.

For the illustration of the method we will use the data displayed in Table 1 that are taken from London (1985). In Table 1 are shown the age, the number of exposed to risk, the number of deaths and the initial (ungraduated) estimates of the mortality rates at each value of age.

<table>
<thead>
<tr>
<th>Age $x$</th>
<th>Exposed to risk $l_x$</th>
<th>Number of deaths $d_x$</th>
<th>Initial estimates $u_x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>135</td>
<td>6</td>
<td>0.044</td>
</tr>
<tr>
<td>71</td>
<td>143</td>
<td>12</td>
<td>0.084</td>
</tr>
<tr>
<td>72</td>
<td>140</td>
<td>10</td>
<td>0.071</td>
</tr>
<tr>
<td>73</td>
<td>144</td>
<td>11</td>
<td>0.076</td>
</tr>
<tr>
<td>74</td>
<td>149</td>
<td>6</td>
<td>0.040</td>
</tr>
<tr>
<td>75</td>
<td>154</td>
<td>16</td>
<td>0.104</td>
</tr>
<tr>
<td>76</td>
<td>150</td>
<td>24</td>
<td>0.160</td>
</tr>
<tr>
<td>77</td>
<td>139</td>
<td>8</td>
<td>0.058</td>
</tr>
<tr>
<td>78</td>
<td>145</td>
<td>16</td>
<td>0.110</td>
</tr>
<tr>
<td>79</td>
<td>140</td>
<td>13</td>
<td>0.093</td>
</tr>
<tr>
<td>80</td>
<td>137</td>
<td>19</td>
<td>0.139</td>
</tr>
<tr>
<td>81</td>
<td>136</td>
<td>21</td>
<td>0.154</td>
</tr>
<tr>
<td>82</td>
<td>126</td>
<td>23</td>
<td>0.183</td>
</tr>
<tr>
<td>83</td>
<td>126</td>
<td>26</td>
<td>0.206</td>
</tr>
<tr>
<td>84</td>
<td>109</td>
<td>26</td>
<td>0.239</td>
</tr>
</tbody>
</table>

Table 1: Raw mortality data

In this paper, in order to minimize the statistic (3.2.1) we used an on line version of the Lingo 8 program. We minimized subject to the constraints (i) to (v), the statistic with parameter $\lambda = 1$, i.e. the $X^2$ statistic, the power divergence with $\lambda = 2/3$ and the power divergence with $\lambda = -1$, i.e. the Kullback–Leibler divergence in order to confirm Brockett’s results (Brockett and Zhang, 1986). The results are shown in Table 2. We see that the power divergence with $\lambda = -1$ does not give the same results as Brockett’s especially for the first and last age. The power divergence for $\lambda = 1$ and $\lambda = 2/3$ give better results. However all graduations give almost the same value for
the measure of smoothness \( S = \sum_{x=1}^{n-3} (\Delta^x v_x)^2 \). The graphical representation of the graduated values is given in Figure 1.

<table>
<thead>
<tr>
<th>Age (Brockett’s solution)</th>
<th>MDI</th>
<th>Power divergence with ( \lambda = -1 ) (Kullback - Leibler)</th>
<th>Power divergence with ( \lambda = 1 ) (( X^2 ) statistic)</th>
<th>Power divergence with ( \lambda = 2/3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>0.068</td>
<td>0.056</td>
<td>0.065</td>
<td>0.063</td>
</tr>
<tr>
<td>71</td>
<td>0.068</td>
<td>0.062</td>
<td>0.068</td>
<td>0.066</td>
</tr>
<tr>
<td>72</td>
<td>0.072</td>
<td>0.068</td>
<td>0.071</td>
<td>0.070</td>
</tr>
<tr>
<td>73</td>
<td>0.076</td>
<td>0.074</td>
<td>0.073</td>
<td>0.074</td>
</tr>
<tr>
<td>74</td>
<td>0.080</td>
<td>0.080</td>
<td>0.076</td>
<td>0.077</td>
</tr>
<tr>
<td>75</td>
<td>0.084</td>
<td>0.088</td>
<td>0.083</td>
<td>0.084</td>
</tr>
<tr>
<td>76</td>
<td>0.088</td>
<td>0.096</td>
<td>0.090</td>
<td>0.091</td>
</tr>
<tr>
<td>77</td>
<td>0.093</td>
<td>0.103</td>
<td>0.097</td>
<td>0.099</td>
</tr>
<tr>
<td>78</td>
<td>0.103</td>
<td>0.111</td>
<td>0.107</td>
<td>0.108</td>
</tr>
<tr>
<td>79</td>
<td>0.118</td>
<td>0.120</td>
<td>0.118</td>
<td>0.118</td>
</tr>
<tr>
<td>80</td>
<td>0.134</td>
<td>0.139</td>
<td>0.139</td>
<td>0.139</td>
</tr>
<tr>
<td>81</td>
<td>0.155</td>
<td>0.160</td>
<td>0.161</td>
<td>0.161</td>
</tr>
<tr>
<td>82</td>
<td>0.181</td>
<td>0.181</td>
<td>0.184</td>
<td>0.183</td>
</tr>
<tr>
<td>83</td>
<td>0.207</td>
<td>0.201</td>
<td>0.206</td>
<td>0.205</td>
</tr>
<tr>
<td>84</td>
<td>0.244</td>
<td>0.221</td>
<td>0.229</td>
<td>0.227</td>
</tr>
<tr>
<td>( S )</td>
<td>0.000211</td>
<td>0.000164</td>
<td>0.000213</td>
<td>0.000245</td>
</tr>
</tbody>
</table>

Table 2: Graduated values
4. Conclusions

After a theoretical and numerical evaluation of the existing methods through which graduation can be obtained, we have concluded that none of them can be thought as better or more correct, as they give almost the same results. So it depends on the actuary which method he will use. It also depends on the environment (setting) of the problem, its constraints and the purpose of graduation.

Furthermore, other divergence measures, such as the power divergence index, can be used for graduation. In the numerical illustration, minimization of the power divergence statistic for various values of $\lambda$ gave almost the same results as the results obtained by Brockett and Zhang (1986). However it did not give the same graduated values for $\lambda = -1$, i.e. the Kullback–Leibler divergence, as in the Brockett’s paper apparently due to differences in the optimization codes.

References


