Graduating Mortality Rates via Divergences

A. P. Sachlas and Takis Papaioannou
University of Piraeus

February 12, 2007

Abstract

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. There are several parametric and non-parametric methods to achieve this. In this paper initially we critically review the method of graduation using information theoretic ideas. Starting with Brockett's idea to use the Kullback - Leibler divergence (Brockett, 1991) we explore the use of a family of divergence indices - the family of power divergence statistics (Read and Cressie, 1998) - with analogous linear and/or quadratic constraints with the aim of finding the best divergence to use in order to obtain the "best" graduation. "Bestness" is defined as an acceptable level of both smoothness and goodness of fit. The results so far indicate that divergences with non-probability vectors in their arguments, as is the case with mortality rates, share, under some conditions, some of the properties of probabilistic or information theoretic divergences. The power divergence statistics also give results equivalent to those that other frequently used graduation methods give. A numerical investigation did not produce the best value or best values of λ, the power of the divergence statistic, for best graduation. The topic is under further investigation.

Keywords. Graduation, Information theory, Kullback-Leibler divergence, Cressie-Read divergence

1 Introduction

In order to describe the actual but unknown mortality pattern of a population, 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. Graduation has two basic characteristics: smoothness, which corresponds to the sum of the third derivatives of the

graduated values at $x = 1, 2, ..., n$, and goodness of fit to the observed data (London, 1985 and Benjamin and Pollard, 1992). 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 $S = \sum_{x=1}^{n-3} (\Delta^3 v_x)^2$, where v_x are the graduated values and u_x are the initial (crude) values of the "entity" to be graduated. Goodness of fit (fidelity) is measured by $F = \sum_{x=1}^{n} w_x (u_x - v_x)^2$ where w_x are weights. As weights the reciprocals of the variance of U_x's are usually used, where U_x is the random variable that corresponds to the initial estimates u_x, $x = 1, 2, ..., n$.

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 fit to the initial estimates and so the graduated rates are calculated. In nonparametric methods, data are 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. The existing nonparametric methods are the graphical ones, weighted moving averages, the Whittaker and Henderson method, the kernel method, graduation with reference to standard mortality rates and graduation using information theoretic ideas. 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 to use. However, there are some factors which can guide him to his decision. Among them are how smooth the graduated values should be, the range and form of actuarial data, the selection of parameters being displayed in the methods.

In this paper we explore the use of divergences as tools of graduation. In Section 2 we critically review the method of graduation using the Kullback - Leibler divergence. Since in graduation the divergence is between non - probability vectors, as a by product, we study the properties of the Kullback - Leibler divergence for non - probability vectors in the light of statistical information theory. In Section 3, we explore the use of the family of power divergence statistics (Read and Cressie, 1988) with the aim of finding the best divergence to use in order to obtain the "best" graduation. A numerical investigation is given in Section 4, while Section 5 contains concluding remarks.

2 Information Theoretic Graduation

2.1 Graduation via the Kullback - Leibler divergence

Zhang and Brockett (1987) tried to construct a smooth series of death probabilities $\{v_x\}$ which is as close 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 x, 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.

In order to obtain the graduated values, Zhang and Brockett (1987) minimize the Kullback - Leibler divergence between the crude death probabilities $\{u_x\}$ and the new death probabilities $\{v_x\}$, $D^{KL}(\mathbf{v}, \mathbf{u}) = \sum v_x \ln \frac{v_x}{u_x}$, subject to the constraints (i) - (v) by considering a dual problem of minimization. So instead of minimizing $D^{KL}(\mathbf{v}, \mathbf{u})$ subject to $\mathbf{v} \geq \mathbf{0}$ and $g_i(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T\mathbf{D}_i\mathbf{v} + \mathbf{b}_i^T\mathbf{v} + c_i \leq 0$, $i = 1, 2, ..., r$, 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:

maximize $-\mathbf{v}^T \exp\left\{\left[-\sum_{i=1}^{r} y_i(\mathbf{A}_i^T\mathbf{w}_i + b_i)\right]\right\} + \mathbf{c}^T\mathbf{y} - \frac{1}{2}\sum_{i=1}^{r} y_i \left\|\mathbf{w}_i\right\|^2$

subject to $\mathbf{v} \geq \mathbf{0}$ and $\mathbf{w}_i \in \mathbf{R}^{m_i}$.

Constraints (i) - (v) may easily be written in the form of $g_i(\mathbf{v})$; we see that we have $r = 32$ constraints. Solving the above dual problem, we can easily find the graduated values v_x^* by using the equality $\ln(v_x^*/u_x^*) = \zeta_x^*$, $x = 1, 2, ..., n$ provided that $\zeta^* = -\sum_{i=1}^{r} y_i^*(\mathbf{A}_i^T\mathbf{w}_i^* + b_i)$. y_i^* and $\mathbf{w}_i^*$ are the solution of the dual problem. In this way we obtain the minimum divergence estimator.

2.2 Kullback - Leibler directed divergence involving non - probability vectors

In the discrete case Kullback - Leibler measures of information based on $(\mathbf{p}^*, \mathbf{q}^*)$, is defined by $I^{KL}(\mathbf{p}^*, \mathbf{q}^*) = \sum_i p_i^* \ln \frac{p_i^*}{q_i^*}$ (Kullback, 1951) and it may be considered as directed divergence. The Kullback - Leibler directed divergence, is defined for probability vectors and shares some properties that all information measures share. Papaioannou (1985, 2001) presents in detail the properties of information measures. Namely these are: nonnegativity, additivity - subadditivity, conditional inequality, maximal information, invariance under sufficient transformations, convexity, loss of information, sufficiency in experiments, appearance in Cramer - Rao inequalities, invariance under parametric transformations, nuisance parameter inequality, order preserving property and asymptotic behavior.

However, when we use Kullback - Leibler directed divergence in graduation, we have mortality rates $\mathbf{u}$ and $\mathbf{v}$ which are not probability distributions since we have $\sum_{x=1}^{n} u_x > 1$ and $\sum_{x=1}^{n} v_x > 1$. Brockett (1991) states "that $D^{KL}(\mathbf{v}, \mathbf{u}) = \sum_{x=1}^{n} v_x \ln \frac{v_x}{u_x}$ is still a measure of fit even in the non - probability situation because the mortality rates are non - negative and because of the assumed constraints" .

In the sequel we investigate whether the Kullback - Leibler directed divergence between two non - probability vectors can be considered as a measure of information. This is done by examining its properties in the light of general properties of measures of information and divergence.

Definition 1. *The Kullback - Leibler directed divergence between two* $n \times 1$ *non - probability vectors* $\mathbf{p}$ *and* $\mathbf{q}$*, is defined by*

$$D^{KL}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

where $\sum_{i=1}^{n} p_i \neq 1$ *and* $\sum_{i=1}^{n} q_i \neq 1$.

Lemma 1. *For the Kullback - Leibler directed divergence with non - probability vectors, it holds that*

$$D^{KL}(\mathbf{p}, \mathbf{q}) = \left(\sum_{i=1}^{n} p_i \right) \left[I^{KL}(\mathbf{p}^*, \mathbf{q}^*) + \ln k \right],$$

where $k = \sum_{i=1}^{n} p_i / \sum_{i=1}^{n} q_i$*, and* $I^{KL}(\mathbf{p}^*, \mathbf{q}^*)$ *is the Kullback - Leibler measure involving probability vectors* $\mathbf{p}^*$ *and* $\mathbf{q}^*$*, where the elements of* $\mathbf{p}^*$ *and* $\mathbf{q}^*$ *are the normalized elements of* $\mathbf{p}$ *and* $\mathbf{q}$*, i.e.* $p_i^* = p_i / \sum_{i=1}^{n} p_i$ *and* $q_i^* = q_i / \sum_{i=1}^{n} q_i$*,* $i = 1, ..., n$.

Proposition 1. *(The nonnegativity property)*

$$D^{KL}(\mathbf{p}, \mathbf{q}) \geq 0, \tag{1}$$

if one of the following conditions holds:

$$(i) \sum_{i=1}^{n} p_i \geq \sum_{i=1}^{n} q_i, \ (ii) \ \sum_{i=1}^{n} p_i < \sum_{i=1}^{n} q_i \ and \ \ln k > -I^{KL}(\mathbf{p}^*, \mathbf{q}^*).$$

Equality in (1) holds if $\mathbf{p} = \mathbf{q}$ *or* $\ln k = -I^{KL}(\mathbf{p}^*, \mathbf{q}^*)$*. Moreover if* $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i$ *then* $D^{KL}(\mathbf{p}, \mathbf{q}) \geq 0$ *if and only if* $\mathbf{p} = \mathbf{q}$.

Note that $D^{KL}(\mathbf{p}, \mathbf{q}) = 0$ does not necessarily imply $\mathbf{p} = \mathbf{q}$ unless $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i$. The (directed) divergence $D^{KL}(\mathbf{p}, \mathbf{q})$ used by Brockett is not a proper divergence as it was explained before.

Definition 1 has obvious extensions to the bivariate and multivariate case. We present the related definitions for the bivariate case.

Definition 2. *(Bivariate Divergence) Let* $p_i(x, y)$*,* $i = 1, 2$*, be two bivariate measures (non - probability functions) associated with two discrete variables* X, Y *in* R^2 *for which*

it holds $\sum_x \sum_y p_i(x,y) > 1$. *We define the Kullback - Leibler directed divergence between two bivariate non - probability functions* p_1, p_2 *as*

$$D^{KL}_{X,Y}(p_1,p_2) = \sum_x \sum_y p_1(x,y) \ln \frac{p_1(x,y)}{p_2(x,y)}.$$

Definition 3. *(Conditional Divergence) For the discrete variables* X, Y *and the bivariate non - probability functions* $p_i(x,y)$, $i = 1,2$, *as given above let* $f_i(x) = \sum_y p_i(x,y)$, $h_i(y|x) = \frac{p_i(x,y)}{f_i(x)}$, $g_i(y) = \sum_x p_i(x,y)$, *and* $r_i(x|y) = \frac{p_i(x,y)}{g_i(y)}$, $i = 1,2$. *We set*

$$D^{KL}_{Y|X=x}(h_1,h_2) = \sum_y h_1(y|x) \ln \frac{h_1(y|x)}{h_2(y|x)}, D^{KL}_{X|Y=y}(r_1,r_2) = \sum_x r_1(x|y) \ln \frac{r_1(x|y)}{r_2(x|y)}$$

and define

$$D^{KL}_{Y|X}(h_1,h_2) = E_X\left[D^{KL}_{Y|X=x}(h_1,h_2)\right] = \sum_x f_1(x) \sum_y h_1(y|x) \ln \frac{h_1(y|x)}{h_2(y|x)},$$

$$D^{KL}_{X|Y}(r_1,r_2) = E_Y\left[D^{KL}_{X|Y=y}(r_1,r_2)\right] = \sum_y g_1(y) \sum_x r_1(x|y) \ln \frac{r_1(x|y)}{r_2(x|y)}.$$

Proposition 2. *(Strong Additivity) Let* p_1, p_2 *be two bivariate non - probability functions associated with two discrete variables* X, Y *in* R^2 *as in Definition 2. Then*

$$D^{KL}_{X,Y}(p_1,p_2) = D^{KL}_X(f_1,f_2) + D^{KL}_{Y|X}(h_1,h_2) = D^{KL}_Y(g_1,g_2) + D^{KL}_{X|Y}(r_1,r_2),$$

where the functions f_i, h_i, g_i, r_i, $i = 1,2$ *are as in Definition 3.*

Corollary 1. *(i)* $D^{KL}_{X,Y}(p_1,p_2) \geq D^{KL}_X(f_1,f_2)$ *with equality if and only if* $D^{KL}_{Y|X}(h_1,h_2) = 0$*;*

(ii) $D^{KL}_{X,Y}(p_1,p_2) \geq D^{KL}_Y(g_1,g_2)$ *with equality if and only if* $D^{KL}_{X|Y}(r_1,r_2) = 0$*;*

(iii) $D^{KL}_{X,Y}(p_1,p_2) \geq D^{KL}_{Y|X}(h_1,h_2)$ *with equality if and only if* $D^{KL}_X(f_1,f_2) = 0$*;*

(iv) $D^{KL}_{X,Y}(p_1,p_2) \geq D^{KL}_{X|Y}(r_1,r_2)$ *with equality if and only if* $D^{KL}_Y(g_1,g_2) = 0$.

In all above cases equality holds if and only if the normalized values of X, Y *are independent.*

The normalized values of X, Y form two random variables X^*, Y^* with discrete joint mass probability function $p^*_i(x,y) = p_i(x,y)/\sum_x \sum_y p_i(x,y)$ and marginal and conditional probability mass functions as follows $X^* \sim f^*_i$, $Y^*|X^* \sim h^*_i$, $Y^* \sim g^*_i$, $X^*|Y^* \sim r^*_i$. For the random variables X^*, Y^* we have

$$I^{KL}_{X^*,Y^*}(p^*_1,p^*_2) = \sum_x \sum_y p^*_1(x,y) \ln \frac{p^*_1(x,y)}{p^*_2(x,y)}.$$

Proposition 3. *(Weak Additivity) If $h_i(y|x) = g_i(y)$ and consequently $p_i(x,y) = f_i(x)g_i(y)$, $i = 1, 2$, we have that the random variables X^*, Y^*, produced by normalization of X, Y as indicated above, are independent, and it holds that*

$$D^{KL}_{X,Y}(p_1, p_2) = D^{KL}_X(f_1, f_2) + D^{KL}_Y(g_1, g_2) - \xi \ln \eta,$$

where $\xi = \sum_y g_1(y) = \sum_x f_1(x)$ and $\eta = \sum_y g_1(y) / \sum_y g_2(y) = \sum_x f_1(x) / \sum_x f_2(x)$.

It is now easy to see that weak additivity holds if $\sum_x f_1(x) = \sum_x f_2(x)$ or $\sum_y g_1(y) = \sum_y g_2(y)$.

Proposition 4. *(Maximal Information and Sufficiency) Let $Y = T(X)$ be a measurable transformation of X, then*

$$D^{KL}_X(p_1, p_2) \geq D^{KL}_Y(g_1, g_2),$$

with equality if and only if Y is "sufficient", where $p_i = p_i(x)$, $g_i = g_i(y)$, $i = 1, 2$.

Proposition 5. *$D^{KL}(\mathbf{p}, \mathbf{q}) \geq I^{KL}(\mathbf{p}^*, \mathbf{q}^*)$ when one of the following conditions holds: (i) $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i \geq 1$, (ii) $\sum_{i=1}^n p_i > \sum_{i=1}^n q_i$ and $\sum_{i=1}^n p_i \geq 1$, (iii) $\sum_{i=1}^n p_i < \sum_{i=1}^n q_i$ and $\sum_{i=1}^n p_i < 1$.*

As expected the Kullback - Leibler directed divergence $D^{KL}(\mathbf{p}, \mathbf{q})$ involving non - probability vectors $\mathbf{p}$, $\mathbf{q}$, does not share the properties that the traditional Kullback - Leibler directed divergence shares. Under some conditions, some of them are satisfied. More precisely $D^{KL}(\mathbf{p}, \mathbf{q})$, is nonnegative, additive, invariant under sufficient transformations and greater than $I^{KL}(\mathbf{p}^*, \mathbf{q}^*)$. It also satisfies the property of maximal information. So, $D^{KL}(\mathbf{p}, \mathbf{q})$, in general terms, can be regarded as a measure of divergence and therefore can be used for graduating mortality tables as originally proposed by Brockett and Zhang (1986).

3 Graduation via the Cressie - Read Power Divergence

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

Cressie and Read (1984) defined a power divergence between two probability vectors $\mathbf{p}^*, \mathbf{q}^*$ by

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

where λ is a real valued parameter. The values at $\lambda = 0, -1$ are defined by continuity. For $\lambda \to 0$, we have $I^0(\mathbf{p}^*, \mathbf{q}^*) = \sum_{i=1}^n p_i^* \ln \frac{p_i^*}{q_i^*}$, which is the Kullback - Leibler directed

divergence. The power divergence has the properties of other measures of divergence such as nonnegativity, symmetry, continuity, nonadditivity and strong nonadditivity. We note that divergence (2) is a *directed divergence* (Cressie and Read, 1984).

Cressie and Read (1984) also used the family of power divergence statistics, for goodness of fit purposes. If $\mathbf{x} = (x_1, x_2, ..., x_k)^T$ is a random vector of counts following the multinomial distribution with parameters $m, \mathbf{p}^*$, where $\mathbf{p}^* = (p_1^*, p_2^*, ..., p_k^*)^T$ is the vector of cell probabilities and $\sum_{i=1}^{k} x_i = m$ and $\sum_{i=1}^{k} p_i^* = 1$ and $\hat{\mathbf{p}}^* = (\hat{p}_1^*, \hat{p}_2^*, ..., \hat{p}_k^*)^T$ is the maximum likelihood estimator of $\mathbf{p}^*$ under the $H_0 : \mathbf{p} \in P$, then the family of *power divergence statistics* is defined as

$$2nI(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} x_i \left[\left(\frac{x_i}{n\hat{p}_i^*} \right)^{\lambda} - 1 \right], \tag{3}$$

where λ 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 $2nI(\lambda)$ as $\lambda \to 0$ and $\lambda \to -1$, respectively.

It can be easily seen (Read and Cressie, 1988) that $2nI(\lambda)$ given by (3) is equal to (i) the X^2 statistic for $\lambda = 1$, (ii) the G^2 statistic for $\lambda \to 0$, (iii) the modified likelihood ratio statistic for $\lambda \to -1$, (iv) the Freeman - Tukey statistic F^2 for $\lambda = -(1/2)$ and (v) the Neyman - modified X^2 for $\lambda = -2$. As an alternative to the X^2 and G^2 statistics, Cressie and Read (1984) proposed the power divergence statistic with $\lambda = 2/3$ which lies between of them.

3.1 Power directed divergence without probability vectors

We have already mentioned that in the problem of graduation we have not probability vectors. So we have to define the directed divergence of order λ for non - probability vectors.

Definition 4. *We define as*

$$D^{CR}(\mathbf{p}, \mathbf{q}) = \frac{1}{\lambda(\lambda+1)} \sum_{i} p_i \left[\left(\frac{p_i}{q_i} \right)^{\lambda} - 1 \right], \lambda \in R$$

the Cressie - Read directed divergence of order λ between two non - probability vectors $\mathbf{p}$ and $\mathbf{q}$, where $\sum_i p_i \neq 1$ and $\sum_i q_i \neq 1$.

Now we have to see if this measure has information theoretic and divergence properties. In the sequel we will assume that $\lambda \neq 0$ and $\lambda \neq -1$.

Lemma 2. *For the Cressie - Read directed divergence involving non - probability vectors $\mathbf{p}, \mathbf{q}$, it holds that*

$$D^{CR}(\mathbf{p}, \mathbf{q}) = \left(\sum_i p_i \right) k^{\lambda} \left[I^{CR}(\mathbf{p}^*, \mathbf{q}^*) - \frac{1 - k^{\lambda}}{k^{\lambda}} \frac{1}{\lambda(\lambda+1)} \right],$$

where $I^{CR}(\mathbf{p}^*, \mathbf{q}^*)$ is the Cressie - Read directed divergence between two probability vectors $\mathbf{p}^*, \mathbf{q}^*$ and $k = \sum_i p_i / \sum_i q_i$.

Proposition 6. *(The nonnegativity property)*

$$D^{CR}(\mathbf{p}, \mathbf{q}) \geq 0,$$

if one of the following conditions holds:

$$(i)\ \sum_i p_i = \sum_i q_i;\ (ii)\ \sum_i p_i > \sum_i q_i \ and\ \lambda \notin (-1, 0);$$

$$(iii)\ \sum_i p_i > \sum_i q_i \ and\ m < I^{CR}(\mathbf{p}^*, \mathbf{q}^*);\ (iv) \sum_i p_i < \sum_i q_i \ and\ \lambda \in (-1, 0);$$

$$(v) \sum_i p_i < \sum_i q_i \ and\ m < I^{CR}(\mathbf{p}^*, \mathbf{q}^*),\ where\ m = \frac{1 - k^\lambda}{k^\lambda} \frac{1}{\lambda(\lambda + 1)}.$$

Equality holds if one of the following conditions holds:

$$(a) \sum_i p_i = \sum_i q_i \ and\ \mathbf{p} = \mathbf{q};\ (b) \sum_i p_i > \sum_i q_i \ or\ \sum_i p_i < \sum_i q_i \ and\ m = I^{CR}(\mathbf{p}^*, \mathbf{q}^*).$$

Proposition 7. *$D^{CR}(\mathbf{p}, \mathbf{q}) \geq I^{CR}(\mathbf{p}^*, \mathbf{q}^*)$ when one of the following conditions holds: (i) $\sum_i p_i = \sum_i q_i$, (ii) $\sum_i p_i > \sum_i q_i$ and $\lambda \notin (-1, 0)$, (iii) $\sum_i p_i < \sum_i q_i$ and $\lambda \in (-1, 0)$. Equality holds if $m = I^{CR}(\mathbf{p}^*, \mathbf{q}^*)$ independently of the value of λ, where m as in Proposition 6.*

Definition 5. *(Bivariate Divergence) In the framework of Definition 2 we define the Cressie - Read directed divergence between two bivariate non - probability functions p_1, p_2 as*

$$D^{CR}_{X,Y}(p_1, p_2) = \frac{1}{\lambda(\lambda + 1)} \sum_x \sum_y p_1(x, y) \left[\left(\frac{p_1(x, y)}{p_2(x, y)} \right)^\lambda - 1 \right].$$

Definition 6. *(Conditional Divergence) In the framework of Definition 3 we set*

$$D^{CR}_{Y|X=x}(h_1, h_2) = \frac{1}{\lambda(\lambda + 1)} \sum_y h_1(y|x) \left[\left(\frac{h_1(y|x)}{h_2(y|x)} \right)^\lambda - 1 \right]$$

and

$$\begin{aligned} D^{CR}_{Y|X}(h_1, h_2) &= E_X \left[D^{CR}_{Y|X=x}(h_1, h_2) \right] \\ &= \frac{1}{\lambda(\lambda + 1)} \sum_x f_1(x) \sum_y h_1(y|x) \left[\left(\frac{h_1(y|x)}{h_2(y|x)} \right)^\lambda - 1 \right], \end{aligned}$$

for the variable X, and $D^{CR}_{X|Y}(r_1, r_2)$ is defined in an analogous way.

Proposition 8. *(Weak additivity) If* $h_i(y|x) = g_i(y)$ *and consequently* $p_i(x,y) = f_i(x)g_i(y)$, $i = 1, 2$, *we have that the random variables* X^*, Y^*, *which are the "standardized" values of* X, Y, *are independent, then*

(a) $D^{CR}_{X,Y}(p_1, p_2) = D^{CR}_X(f_1, f_2) + D^{CR}_Y(g_1, g_2) + p_{1\bullet\bullet}\eta^\lambda\lambda(\lambda+1)I^{CR}_{X^*}(f_1^*, f_2^*)I^{CR}_{Y^*}(g_1^*, g_2^*) + p_{1\bullet\bullet}\left(1-\eta^\lambda\right)\frac{1}{\lambda(\lambda+1)}$, *where* $p_{i\bullet\bullet} = \sum_x \sum_y p_i(x,y),\ i = 1, 2,$

$(b) D^{CR}_{X,Y}(p_1, p_2) = D^{CR}_X(f_1, f_2) + D^{CR}_Y(g_1, g_2)$ *if* $\eta = 1$ *and if one of the marginal pairs* $(f_1^*, f_2^*), (g_1^*, g_2^*)$ *are identical where* $\eta = p_{1\bullet\bullet}/p_{2\bullet\bullet}$.

Proposition 9. *(Maximal Information and Sufficiency) Let* $Y = T(X)$ *be a measurable transformation of* X, *then*

$$D^{CR}_X(p_1, p_2) \geq D^{CR}_Y(g_1, g_2),$$

when $b > 1$, *where* $b = \left(\sum_x p_1(x) / \sum_x p_2(x)\right)^\lambda$, *with equality if and only if* Y *is "sufficient", where* $p_i = p_i(x)$, $g_i = g_i(y)$, $i = 1, 2$.

We have already seen that the power directed divergence $D^{CR}(\mathbf{p}, \mathbf{q})$, under some conditions is nonnegative, additive, greater than $I^{CR}(\mathbf{p}^*, \mathbf{q}^*)$ and invariant under sufficient transformations. It also shares the property of maximal information. So, we can regard $D^{CR}(\mathbf{p}, \mathbf{q})$ as a measure of divergence and therefore use it for graduation purposes.

3.2 Graduation via power divergence

Based on the above we can now apply the power divergence to the problem of actuarial graduation. In order to obtain the graduated values v_x, we minimize the Cressie - Read divergence

$$D^{CR}(\mathbf{v}, \mathbf{u}) = \frac{1}{\lambda(\lambda+1)} \sum_x v_x \left[\left(\frac{v_x}{u_x}\right)^\lambda - 1\right]$$

for given λ subject to $\mathbf{v} \geq \mathbf{0}$ and $g_i(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T\mathbf{D}_i\mathbf{v} + \mathbf{b}_i^T\mathbf{v} + c_i \leq 0$, $i = 1, 2, ..., r$, where $\mathbf{D}_i$ is a positive semidefinite matrix for each i and $\mathbf{b}_i$, c_i are constants. We have already mentioned that the constraints (i) - (v) given in Section 2.1 may be written in the form of $g_i(\mathbf{v})$ with the consequence to have $r = 32$ constraints. These constraints represent smoothness, monotonicity, convexity, equality in the number of deaths and equality in the total age of death. The minimization is done for various values of the parameter λ and we choose as the best graduation the one that gives satisfactory results for smoothness and fit. In this way we can interpret the resulting series of the graduated values, as the series which satisfy the constraints and is least distinguishable in the sense of the Cressie - Read directed divergence from the series of the crude values $\{u_x\}$.

It is obvious that if we choose $\lambda = 0$, we perform graduation through the Kullback - Leibler directed divergence that Zhang and Brockett (1987) described. In the following section we provide numerical illustration.

4 Numerical Investigation

The problem of graduation 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, we will use four different data sets of death probabilities. The first data set is the one that Brockett and Zhang (1986) use. We denote this data set by BZ86. The second one comes from London (1985, p. 162), originally from Miller (1949), and will be denote it by L85. The third one comes from the Actuarial Society of Hong Kong, is available on the Internet (www.actuaries.org.hk) refers to males and will be denoted by HK01M. The last one also comes from the same Society, refers to females and will be denoted by HK01F. The above-mentioned data sets are of different size. Especially, the BZ86 data set consists of 15 death probabilities for ages 70 to 84 while the L85 data set consists of 20 death probabilities belonging to ages 75 to 94. From HK01M we have used 16 death probabilities for ages 70 to 85 while from HK01F we have taken 20 death probabilities for ages 70 to 89.

We have performed several graduations for each data set, using different values of the parameter λ and the constraints of smoothness, monotonicity, convexity and the two actuarial constraints. Among them are the values 1, 0, -1, $-(1/2)$, and -2, which give the X^2 statistic, Kullback - Leibler divergence, modified likelihood ratio statistic, Freeman - Tukey statistic F^2 and Neyman - modified X^2, respectively. We also use the value $2/3$ that Cressie and Read (1984) proposed. We note that the value of M in the first constraint, is different in each set, and it is computed through graduation by the Whittaker - Henderson method, except for the BZ86 data set where we have used the value that Brockett and Zhang (1986) use.

It is expected and logical that different choice of the parameter λ leads to different graduated values. We have already mentioned that the two basic elements of graduation are smoothness and goodness of fit. So, in order to compare the several graduations for each data set, we computed, after the graduation, the measure $F = \sum_{x=1}^{n} w_x(u_x - v_x)^2$, used by Whittaker - Henderson. We note that as weights we used $w_x = \frac{l_x}{v_x(1-v_x)}$, where l_x is the number of people at risk in the age x. The measure $\sum_{x=1}^{n-3}(\Delta^3 v_x)^2$ was used for measuring the smoothness of the graduated values.

The value of the smoothness measure S computed after the Whittaker - Henderson graduation as well as the average value of smoothness $\bar{S}$ taken from several graduations using power divergences are given in Table 1. It is obvious that both methods give almost the same value for the smoothness measure S.

In Figure 1, we have plotted the value of the smoothness measure S versus the value of the parameter λ, with S depicted in the y-axis and λ in the x-axis. The blue line in each plot denotes the value of M in the smoothness constraint. We can see that apart from the BZ86 data set, the other three follow the same pattern. When $-\infty < \lambda < -1$, S takes

Data set	S via Whittaker-Henderson	S via Power Divergences
BZ86	0.0009	0.0006
L85	0.0000355697	0.0000313201
HK01M	0.0000248294	0.0000215776
HK01F	0.0000148469	0.0000123451

Table 1: Value of smoothness measure

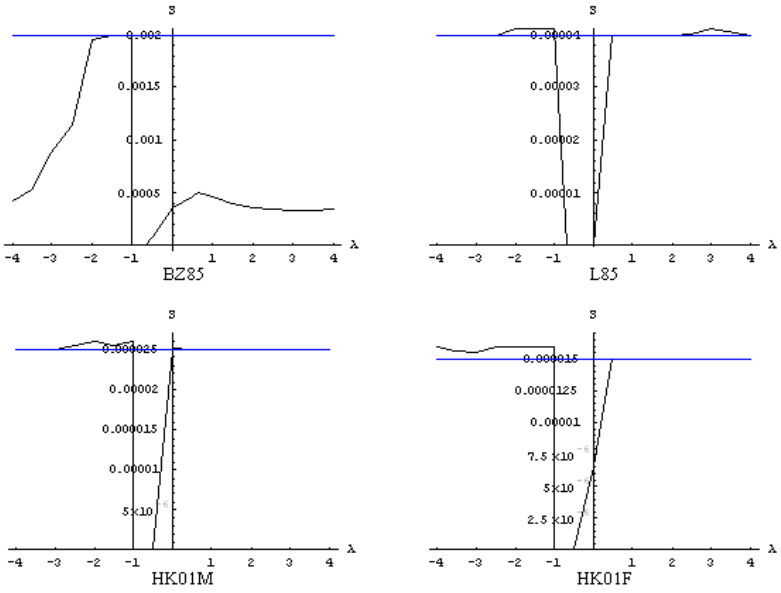

Figure 1: Smoothness S versus λ

a value near the value of M. Then, when $-1 < \lambda < -0.5$ S takes a very small value almost equal to zero and then for the remaining values of λ, it also takes a value near the value of M. So, for values of λ, between -1 and -0.5, the method oversmooths the data.

In Figure 2, we present the analogous plots concerning the measure of fit F,with F in the y-axis and λ in the x-axis. We can also see a same pattern for the L85, HK01M and HK01F data sets. For values of λ smaller than -1, the measure of fit increases, till its maximum value. This means that graduation is not acceptable as the graduated values depart too far from the crude values. When λ takes a value almost equal to -1, F decreases and it is stabilized for the remaining values of λ.

5 Conclusions

Although there are a lot methods of graduation, 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 to use. It also depends on the environment (setting) of the problem, its constraints and the purpose of graduation.

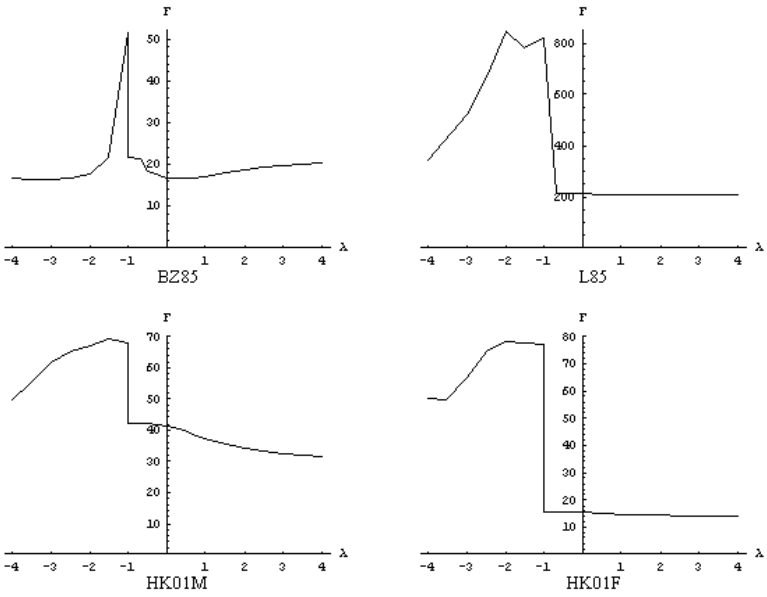

Figure 2: Goodness of fit F versus λ

After a theoretical evaluation of the Kullback - Leibler $D^{KL}(\mathbf{p}, \mathbf{q})$ divergence involving non - probability vectors, which Brockett and Zhang (1986) use, for purposes of graduation, we have concluded that this measure shares some of the properties of the Kullback - Leibler directed divergence $I^{KL}(\mathbf{p}^*, \mathbf{q}^*)$. Under some conditions, $D^{KL}(\mathbf{p}, \mathbf{q})$ is nonnegative, additive and invariant under sufficient transformations. Thus, we can regard $D^{KL}(\mathbf{p}, \mathbf{q})$ as a measure of information, and consequently use it in the problem of graduation.

Furthermore, other divergence measures, such as the power divergence indices, can be used for graduation. We proved that in the case that it does not involve probability vectors, it shares some of the properties that the power divergence with probability vectors shares. More specifically, under some conditions it is nonnegative, additive, greater than $I^{CR}(\mathbf{p}^*, \mathbf{q}^*)$ and invariant under sufficient transformations. So, we can regard $D^{CR}(\mathbf{p}, \mathbf{q})$ as a measure of divergence and therefore it can be used in the problem of graduation.

In the numerical illustration, minimizing the power divergence for various values of λ gave equivalent results, in terms of smoothness, to those of other methods of graduation such as the Whittaker - Henderson method. We can obtain different values for the smoothness index S by changing the value of M in the first constraint. However, we cannot say which value of the parameter λ is the best for graduation. Values of λ smaller than -1 give unacceptable results as far as goodness of fit is concerned and as such they should be avoided.

6 References

Brockett, P.L. (1991). Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications, Transactions of the Society of Actuaries 43, 73 - 114.

Benjamin, P. and Pollard, J.H. (1992). The Analysis of Mortality and Other Actuarial Statistics, Butterworth - Heinemann, London, 6th edition.

Brockett, P.L. and Zhang, J. (1986). Information Theoretical Mortality Graduation, Scandinavian Actuarial Journal, 131 - 140.

Cressie, N.A.C and Read, T.R.C. (1984). Multinomial Goodness-of-Fit Tests, Journal of Royal Statistical Society, B, Vol. 46, No. 3, 440 - 464.

Haberman, S. (1998). Actuarial Methods, Encyclopedia of Biostatistics, 1, (Eds., P. Armitage and Th. Colton), 37 - 49, John Wiley & Sons, New York.

Kullback, S. (1959). Information Theory and Statistics, John Wiley & Sons, New York.

London, D. (1985). Graduation: The Revision of Estimates, ACTEX Publications, Winsted, Connecticut.

Miller, M.D. (1949). Elements of Graduation, Actuarial Society of America, New York.

Read, T.R.C and Cressie, N.A.C. (1988). Goodness - of - Fit Statistics for Discrete Multivariate Data, Springer - Verlag, New York.

Zhang, J. and Brockett, P.L. (1987). Quadratically Constrained Information Theoretic Analysis, SIAM Journal of Applied Mathematics 47, no 4, 871 - 885.