A Meta-Analytic Framework for Best-Practice Mortality Surface Estimation

Jutta Gampe
Max Planck Institute for Demographic Research
gampe@demogr.mpg.de

Paul H.C. Eilers
Faculty of Social and Behavioural Sciences, Utrecht University
p.h.c.eilers@uu.nl

Magdalena Muszyńska
Terry Sanford Institute of Public Policy, Duke University
magdalena.muszynska@duke.edu

Preliminary draft – please do not cite or quote

Abstract

Similar to the best-practice life expectancy, the best-practice mortality surface over age and time can be estimated. Reliability is a serious issue, because population size across countries varies over several orders of magnitude. Our solution is to estimate a latent mortality distribution, properly accounting for observed magnitudes. This brings a meta-analytic framework to mortality studies, opening many new opportunities to study variability between demographic units in a quantitative way.

1 Introduction

Best-practice life expectancy, that is the highest life expectancy achieved in a particular year, has been rising linearly for almost 150 years, with a slope of about three months per year (Oeppen and Vaupel, 2002). The country in which life expectancy had been highest varied over the years. For women Japan has been leading for the last twenty years, while in the last five years males in Iceland performed best.

High life expectancy is linked to low mortality, but this does not necessarily imply that the leading country automatically performed optimally with respect to mortality at every single age. Mortality may well have been lower in some other country at least for some ages. Corresponding to the notion of best-practice life expectancy we can define the best-practise mortality surface as the mortality surface consisting of the lowest values of mortality at every age over the period of interest. This surface of minimal mortality would give the benchmark in any particular year for any age, and it could be used to compare the performance of different countries relative to what they could have achieved if they had performed optimally. We may be able to assess the contribution of mortality at different ages to gaps between actual performance and the tentative optimum, or we could demonstrate at what ages the countries leading in life-expectancy performed this well, and where there would be room for further improvement. The best-practice surface also could be analyzed like other mortality surfaces: trends could be studied and future mortality levels could be predicted.

A naive approach to estimate the best-practise mortality surface would be to calculate empirical rates from several low-mortality countries and take the observed minimal value of the death rates as the best-practise value. However, this approach is hampered by the fact that low-mortality countries vary considerably in size. For example, the USA with more than 300 million inhabitants, but also Iceland with a total population number of about 300,000, are among the low-mortality countries. Consequently, both the number of people at risk and the number of deaths observed can be quite small, introducing strong variability as well as zero observed mortality for certain ages.

We present a meta-analytic framework that models the observed deaths as outcomes based on a latent mortality distribution, varying over age and time. The observed data are used to estimate this latent distribution. This approach not only allows to handle the problem induced by strongly differing country sizes but it also enables more detailed studies of the mortality distribution than just its minimal value.

2 Data and Model

We assume that information on the number of deaths and the number of individuals exposed to risk is available for J units, which can be countries, like in the application in this paper, or regions within countries or larger geographic entities.

In this application we will use data on deaths and exposures for single years of age and single years from $J = 21$ countries derived from the Human Mortality Database (HMD). We consider ages from 30 to 100 and the period from 1970 to 2000. The countries included in the study were: Austria, Australia, Belgium, Canada, Denmark, England and Wales, Finland, France, Iceland, Italy, Japan, Luxembourg, Netherlands, New Zealand (Non-Maori population), Norway, Portugal, Spain, Sweden, Switzerland, United States, West-Germany.[1]

Mortality at age a in year t is denoted by $\mu(a,t)$ and we assume that $\mu(a,t)$ varies across units according a distribution with density $f_{a,t}(m)$. That is, mortality $\mu(a,t)$ is itself considered to be a random variable, having a latent distribution, which can only be inferred indirectly. Mortality $\mu_j(a,t)$ in any of the J units is a realization from this density $f_{(a,t)}(m)$. To make inference on the distribution of $\mu(a,t)$ an immediate solution would be the following: If $y_j(a,t)$ is the number of deaths in unit j at age a in year t and $n_j(a,t)$ denotes the corresponding exposures, then we could estimate the $\mu_j(a,t)$ by the empirical rates

$$\hat{\mu}_j(a,t) = \frac{y_j(a,t)}{n_j(a,t)}, \tag{1}$$

and therefrom derive mean, variance or other sample statistics of interest. One drawback of this strategy is that the accuracy of the empirical rates in the different units is not taken into account. This is particularly relevant for sample extremes, where units with small exposures and/or number of events may lead to high or low mortality estimates due to higher variability

[1]The HMD currently covers a total number of 32 countries, some of which (from Eastern and South Eastern Europe and Taiwan) were not included here. These countries offer shorter series of data, and if we we are interested in the lower tail of the mortality distribution this omission will be negligible. If, however, were interested, e.g., in the variability of mortality across Europe we should include these countries as well.

of the estimates (1). As an example Figure 1 shows female mortality estimates at age 85 in 1980. Two small countries, Iceland and Luxembourg, have death rates close to the extremes but show large standard errors, and we want to take this information into account when making inference on $\mu(a, t)$.

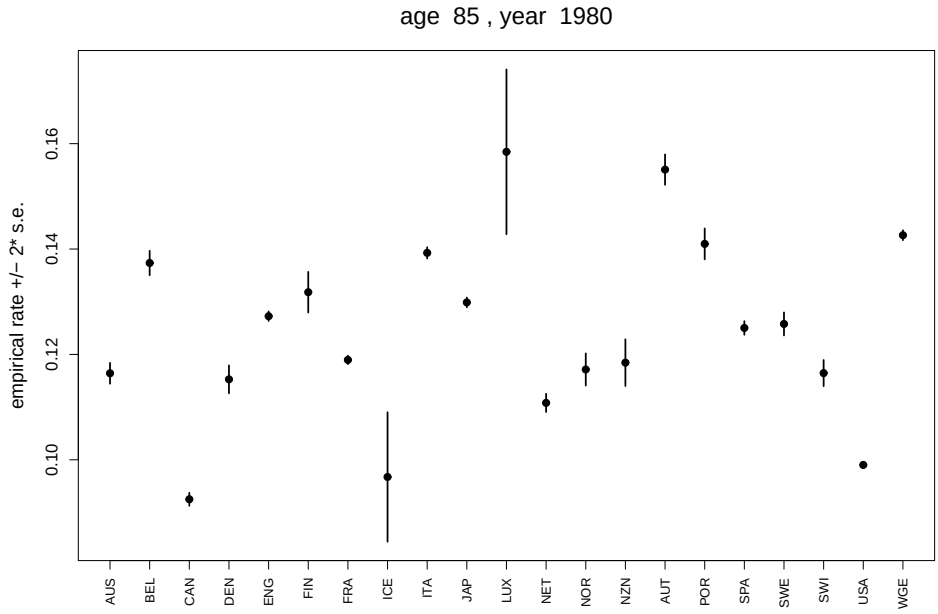

Figure 1: Death rates for females at age 85 in year 1980 for the $J = 21$ countries included in this study. Shown are estimates and ± 2 standard errors.

In this paper we therefore suggest a different approach. For ease of presentation we focus on one age and one year and drop the dependence on a and t in the notation in the following. We consider a discrete distribution for μ with a dense grid of mass-points $m_k, k = 1, \ldots, K$, and probability masses $p_k = P(\mu = m_k)$. Naturally, the p_k sum to one. The grid can be equally spaced on the scale of the m_k, but usually equidistant values on the log-scale $\eta_k = \ln m_k$ will be more appropriate, especially for small values of μ.

For a given value of mortality m_k the number of deaths y_j in unit j is a Poisson variable with mean $\nu_j = n_j m_k$. That is

$$w_{jk} = P(y_j \,|\, m_k) = \frac{\exp\{-n_j m_k\}\,(n_j m_k)^{y_j}}{y_j!} = \alpha_j \,\exp\{-n_j m_k\}\, m_k^{y_j} \tag{2}$$

with

$$\alpha_j = n_j^{y_j}/y_j! \tag{3}$$

independent of k. The marginal distribution of the y_j is therefore

$$P(y_j) = \sum_{k=1}^{K} w_{jk} p_k. \tag{4}$$

To estimate the mixing distribution $p_k, k = 1, \ldots, K$, the EM-algorithm (Dempster et al., 1977) is a natural choice.

The E-step results from the

$$P(m_k \,|\, y_j) = \frac{P(y_j \,|\, m_k)\, p_k}{P(y_j)} = \frac{w_{jk}\, p_k}{\sum_l w_{jl} p_l}. \tag{5}$$

(The constants α_j appear as factors in both numerator and denominator of (5) and cancel out.)

In the M-step we obtain $p_k^{(s+1)}$ from the current values $p_k^{(s)}$ as

$$p_k^{(s+1)} = \sum_{j=1}^{J} \frac{1}{J} \frac{w_{jk}\, p_k^{(s)}}{\sum_{l=1}^{K} w_{jl}\, p_l^{(s)}} \tag{6}$$

(see e.g. Aitkin (1996)).

This procedure is not limited to observations from a Poisson distribution but can be used more generally in mixtures of generalized linear models (Aitkin, 1999). Consequently, it is also possible to study the mixture of Binomial variables if probabilities of death $q(a,t)$ instead of death rates $\mu(a,t)$ are to be modeled.

Without any further restrictions on the p_k the EM-algorithm will converge to the nonparametric maximum likelihood estimate (NPMLE) of the mixing distribution of $\mu(a,t)$ (Laird, 1978). Usually only a few mass-points carry positive probabilities, leading to a rather spiky and far from smooth mixing distribution. Furthermore, convergence of the EM-algorithm typically is very slow.

To get round both drawbacks Eilers (2007) introduced the following strategy. In each iteration a smoothing step is introduced. That is, starting from the current values of the $p_k^{(s)}$ steps (5) and (6) are performed as before. However, before the resulting

$$\tilde{p}_k^{(s+1)} = \sum_{j=1}^{J} \frac{1}{J} \frac{w_{jk}\, p_k^{(s)}}{\sum_{l=1}^{K} w_{jl}\, p_l^{(s)}} \tag{7}$$

are introduced into the next step of the EM-iteration they are smoothed by an additional smoothing step:

$$p^{(s+1)} = S_\lambda(\tilde{p}^{(s+1)}) \tag{8}$$

with $p^{(s+1)} = (p_1^{(s+1)}, \ldots, p_K^{(s+1)})'$ and $\tilde{p}^{(s+1)}$ accordingly. The smoothing function $S_\lambda(.)$ depends on an additional parameter λ that controls the amount of smoothness introduced in this step. Naturally, the smoothing step should preserve the property $\sum_k p_k^{(s+1)} = 1$ of the mixing distribution.

2.1 The smoothing sub-step

The smoothing is performed by applying a discrete Whittaker smoother (Eilers, 2003). The function $S_\lambda(.)$ solves, for a given value of the smoothing parameter λ, the following penalized least-squares problem (for simplicity we drop the iteration index $s+1$ here):

$$S_\lambda(p) = \arg\min_p \left\{ (p - \tilde{p})'(p - \tilde{p}) + \lambda^2\, p' D_2' D_2 p + 2\lambda\, p' D_1' D_1 p \right\}, \tag{9}$$

where the matrices D_1 and D_2 calculate first and second order differences of the elements of p, respectively, i.e.,

$$D_1 = \begin{pmatrix} -1 & 1 & 0 & \\ & \ddots & \ddots & \\ & 0 & -1 & 1 \end{pmatrix} \qquad D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & 1 & -2 & 1 \end{pmatrix}.$$

Equation (9) implies that we determine the vector of the mixing distribution p to be inserted into the next E-step as the one, which is close to the outcome of the most recent M-step (6) — this is implemented by the sums of squares $(p-\tilde{p})'(p-\tilde{p})$ — however, variation between neighboring elements of p is restricted by the penalty terms $\lambda^2\, p' D_2' D_2 p + 2\lambda\, p' D_1' D_1 p$. The larger the value of λ, the stronger the effect of the penalization and the smoother the resulting vector p will be. A value of $\lambda = 0$ would imply no smoothing and the unmodified EM-algorithm would be performed.

The combination of first and second order penalty terms is due to the fact that we want to estimate a mixing distribution whose individual elements p_k have to be non-negative. This is properly taken care of by the second penalty term $2\lambda\, p' D_1' D_1 p$, as had been demonstrated by Eilers and Goeman (2004). Also, the definition of the penalty guarantees that $\sum_k p_k$ will equal one whenever based on a vector $\tilde{p}$ whose elements sum to one.

Solving this penalized least-squares problem leads to a system of linear equations for p

$$(I + \lambda^2 D_2' D_2 + 2\lambda D_1' D_1)\, p = \tilde{p}, \tag{10}$$

which can be easily solved for given values of λ.

The practical implementation of this algorithm consists of the following steps:

- A dense uniform grid of mass-points is chosen on the log-scale $\eta_k = \ln m_k$. The starting values for the mixing distribution are commonly chosen as a uniform distribution $p_k^{(0)} = P(m_K) = 1/K, k = 1, \ldots, K$.

- The values of the w_{jk}, see equation (2), are determined, usually neglecting the values of α_j. These values only have to be calculated once during the whole procedure.

- The marginal probabilities are calculated from (4) and inserted into (5). The M-step (6) is performed.

- For the smoothing sub-step a value of λ has to be chosen and currently has to be

selected by the user. Practically this is done by inspecting results for a grid of λ-values. Generally the optimal values of the smoothing parameter depends on the mass-point grid, the number of observations included, and on the variance of the latent distribution to be estimated. In our applications commonly values of λ between 10 and 70 showed good results.

- For the chosen value of λ the system (10) is solved and the resulting p inserted into the next EM-iteration.

- The iterations are continued until the maximum difference between elements of $p^{(s)}$ and $p^{(s+1)}$ is below a threshold ε. In our application we commonly chose

$$\max_k |p_k^{(s+1)} - p_k^{(s)}| < 1e^{-6}.$$

  In all applications we performed the number of necessary iterations never exceeded seven, demonstrating the efficiency introduced by the additional smoothing step.

The resulting final estimate $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_K)'$ is a discrete but smooth approximation to the latent density $f_{a,t}(m)$ of $\mu(a,t)$, from which further parameters of interest, such as the mean, variance or low-percentage quantiles can be derived.

3 Application

Figure 2 illustrates our procedure for the example given in Figure 1. The upper panel shows the scaled likelihoods for the $J = 21$ observation. While large countries show very narrow peaks, the likelihood functions for Iceland and Luxembourg are rather broad. The bottom panel gives the estimated (non-normalized) densityfor $a = 85$ and $t = 1980$ for the latent mortality distribution (on log-scale).

In the following Figure 3 the estimated latent mortality distributions are given for the age range of 50 to 100 in years 2000. As the curves are given on log-scale, the linear upward

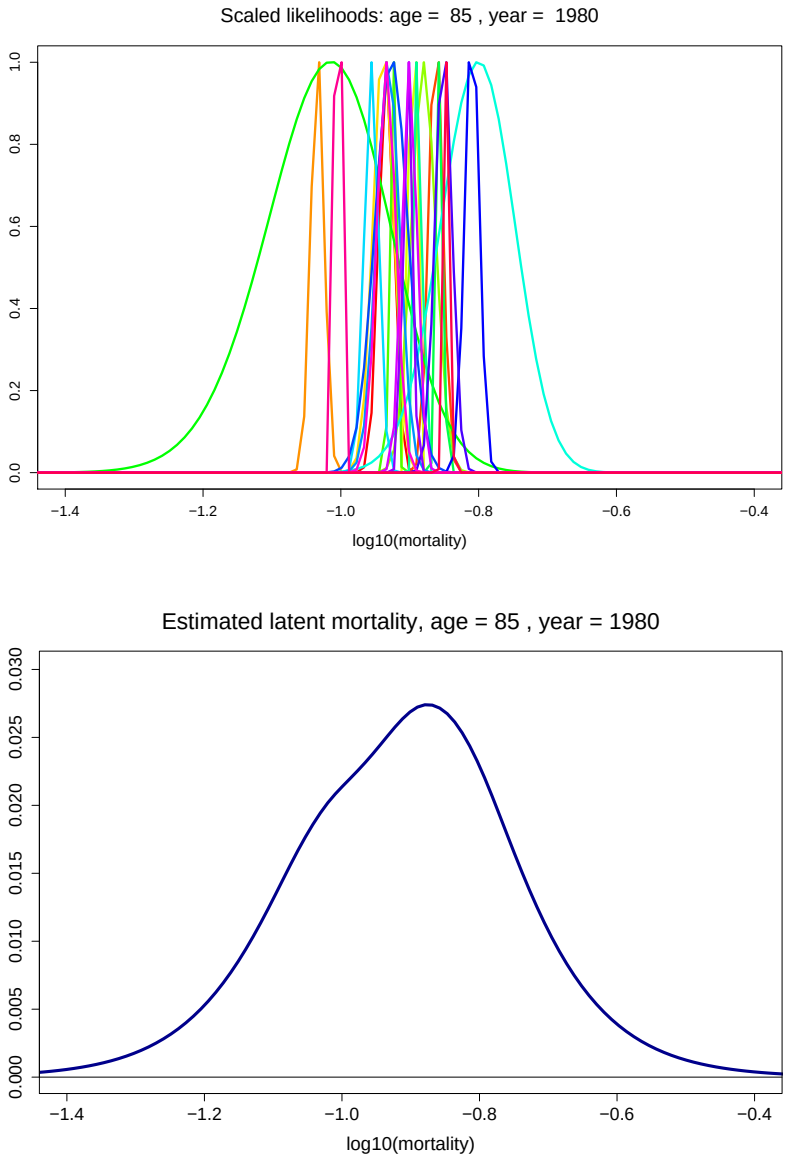

Figure 2: Scaled likelihoods (top panel) for 21 countries, and estimated latent mortality distribution (bottom panel).

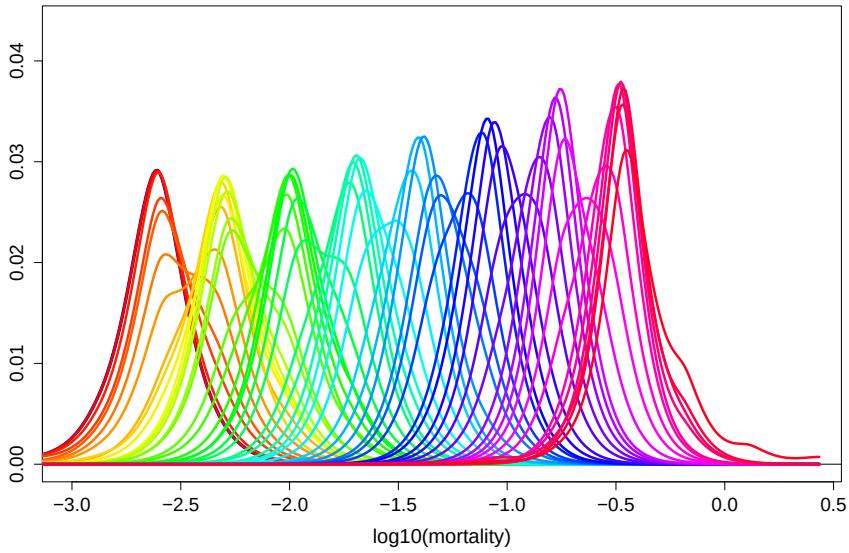

Figure 3: Estimated latent mortality distributions for ages 50 to 100, in year 2000.

shift of the modes of the curves basically gives the exponential increase of mortality with age. Furthermore the densities get more peaked with higher ages (on log-scale). Together with the increase in means the variance for $\mu(a, t)$ increases over age though. The wave-like pattern of the distributions over age appears in all other years, too, and presumably is introduced by some interpolation schedule performed in the HMD. This peculiar pattern makes it obvious that additional smoothing over the age-axis would be advantageous to make mortality distributions for neighboring ages more alike.

Changes in the latent distribution can be compared more easily by looking at image plots. For the years 1970 and 2000 these are given in Figure 4, for all ages between 30 and 100. As it can clearly be seen the wavy pattern of the densities is well reflected in the image plot as well. In the plot for the year 2000 the lines of the 5%- and the 10%-quantile of the latent mortality distribution are given. It is evident that the regular pattern observed also has an impact on the trajectory of these quantiles over age, thus making an approach that jointly smoothes the mixing distributions also over age even more desirable.

Finally, the latent distributions all are almost symmetric and closely resemble a Normal density. As an alternative to the smoothing step we can fit a normal distribution, which has the advantage that it can be completely characterized by just two parameters, mean and

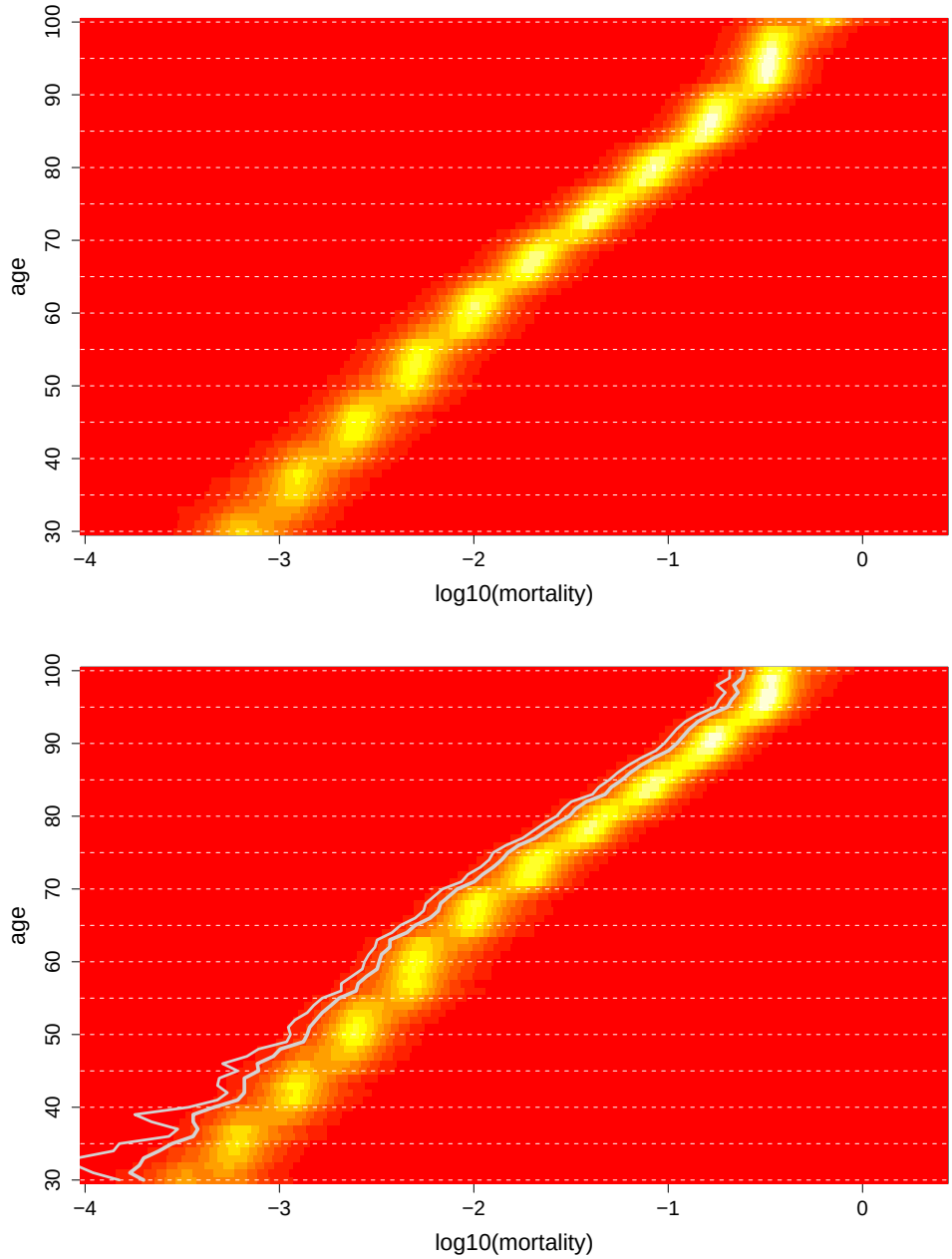

Figure 4: Image plot of the latent mortality distributions for ages 30 to 100, in year 1970 (top) and in year 2000 (bottom). Grey lines in the image for 2000 indicate the 5% and 10% quantiles of the latent mortality distributions estimated independently over ages.

standard deviation. Low-percentage quantiles for $\mu(a,t)$ can then be directly derived from the corresponding log-Normal distributions.

4 Discussion and Outlook

We presented an algorithm for estimation of latent mortality distributions for a group of countries. While straightforward “best-practice" would only look at the lower (observed) extreme values, our approach offers quantification of many aspects of variability of mortality. Once latent distributions are estimated for a range of years and ages, one can compute summary statistics, like quantiles and inter-quantile ranges and study them for meaningful and interesting demographic patterns.

It is fruitful to draw the parallel with meta-analysis in medical statistics, the area for which the algorithm was originally developed. The latent distribution can be interpreted as a prior distribution in an empirical Bayes sense. One can compute empirical posterior distributions for individual countries, which can be used to quantify uncertainties and to improve individual mortality estimates by “shrinking". Our units observation are countries, but in larger countries the model might be useful to study states or provinces.

Currently the choice of the smoothing parameter is done subjectively. Automatic ways to select the value of λ clearly have to be developed. Furthermore, the estimation of the latent mixing distribution is performed independently over age and time. It is however reasonable to assume that the distribution of mortality $\mu(a,t)$ varies smoothly over age and also, possibly with exception of years with specific events such as severe epidemics, over time. Hence it would be natural to ‘join’ neighboring mixing distributions by an additional penalty to allow for smooth variation of the densities $f_{a,t}(m)$ over age or time.

References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing 6*, 251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*, 117–128.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood estimatin from incomplete data via the EM algorithm (with discussion). *Journal of the ROyal Statistical Sciety B 39*, 1–38.

Eilers, P. H. (2003). A perfect smoother. *Analytical Chemistry 75*(14), 3631–3636.

Eilers, P. H. (2007). Data exploration in meta-analysis with smooth latent distributions. *Statistics in Medicine 26*, 3358–3368.

Eilers, P. H. and J. Goeman (2004). Enhancing scatterplots with smooth densities. *Bioinformatics 20*(5), 623–628.

HMD. Human Mortality Database. University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Jornal of the American Statistical Association 73*, 805– 811.

Oeppen, J. and J. W. Vaupel (2002). Broken limits to life expectancy. *Science 296*, 1029–1031.