The Canadian Journal of Statistics
Vol. 36, No. 2, 2008, Pages 193–206
La revue canadienne de statistique

Forecasting mortality rates via density ratio modeling

Benjamin KEDEM, Guanhua LU, Rong WEI and Paul D. WILLIAMS

Key words and phrases: Age-specific mortality; autoregression; combined data; forecasting; one-year ahead forecast; predictive distribution; semiparametric method.

MSC 2000: Primary 62M20; secondary 62G07.

Abstract: The authors propose a semiparametric approach to modeling and forecasting age-specific mortality in the United States. Their method is based on an extension of a class of semiparametric models to time series. It combines information from several time series and estimates the predictive distribution conditional on past data. The conditional expectation, which is the most commonly used predictor in practice, is the first moment of this distribution. The authors compare their method to that of Lee and Carter.

Prévision de taux de mortalité par la modélisation d'un rapport de densités

Résumé : Les auteurs proposent une méthode semiparamétrique pour la modélisation et la prévision de la mortalité par tranche d'âges aux États-Unis. Leur approche s'appuie sur la généralisation d'une classe de modèles semiparamétriques au cas de séries chronologiques. Elle exploite l'information provenant de plusieurs séries et estime la loi prédictive à partir du comportement passé. L'espérance conditionnelle, qui sert le plus souvent de prédicteur en pratique, en est le premier moment. Les auteurs comparent leur méthode à celle de Lee et Carter.

1. INTRODUCTION

Since 1900, the United States Government has been collecting mortality data from death registration records assembled by state vital statistics offices. The data are broken down mainly by state, cause, race, gender, and age, and are published in the form of death rates and life expectancies decennially and/or annually for over 100 years. However, the existing electronically documented mortality data are relatively short. In this study we shall use well documented mortality time series from 1970 to 2002 to forecast mortality patterns in the U.S. This gives us relatively short annual age-specific time series, consisting of a little over 30 observations each, stratified by factors such as state, gender, and race. Prediction of future annual death rates based on these time series must take into account their short length.

The objective of the present study is to forecast mortality patterns, using relatively short historical time records, by following a two-stage procedure. First, to each short series we fit a first order autoregressive model, and then, to overcome the problem of short series, the resulting residuals are combined or merged in some fashion to provide estimates of future predictive distributions. Point forecasts, such as the conditional expectation, are obtained as a byproduct.

In this paper we apply a semiparametric forecasting method advanced recently in Kedem, Gagnon & Guo (2005). The method compensates for short individual records by combining them via a density ratio model as described in Section 2. Accordingly, the residuals from several different fitted models are combined in this way in order to estimate the entire future conditional distributions of interest. From this we obtain future conditional probabilities as well as the conditional expectation of future values given past information, the most common predictor. We focus primarily on the prediction of centered annual age-specific log-death rates for the entire U.S. using data from 1970 to 2002.

1.1. U.S. mortality rate data.

The death or mortality rate for age x and year t is the number of people who died at age x during

year t divided by the number of people of age x at the beginning of year t. It is customary to report death rate on a (natural) logarithmic scale. Our data structure has the form of log-death rate denoted as $m(x,t)$ for ages $x = 1, \ldots, 85$, and year $t = 1971, \ldots, 2002$.

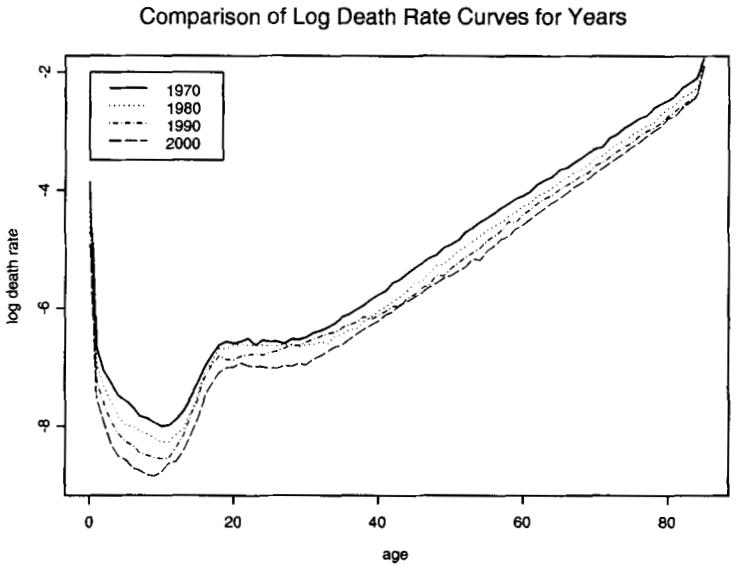

FIGURE 1: Log-death rate $m(x,t)$ as a function of age x for some fixed t.

The plot of $m(x,t)$ as a function of age x resembles a pointed hook with a rather long handle, surprisingly similar to a dentist probe, as seen from Figure 1. Wei, Curtin & Anderson (2003) fitted to these data the eight parameter model of Heligman & Pollard (1980), demonstrating that the model captures well the pointed hook pattern of mortality versus age. Figure 1 also shows that the hook pattern repeats itself year after year persistently, and that in general annual death rates decline, again quite persistently. The decline in death rate for fixed age as a function of time is shown in Figure 2 on a log-scale for several ages. Except for age 0, these time series appear almost as parallel straight lines, but when drawn separately they are much more oscillatory.

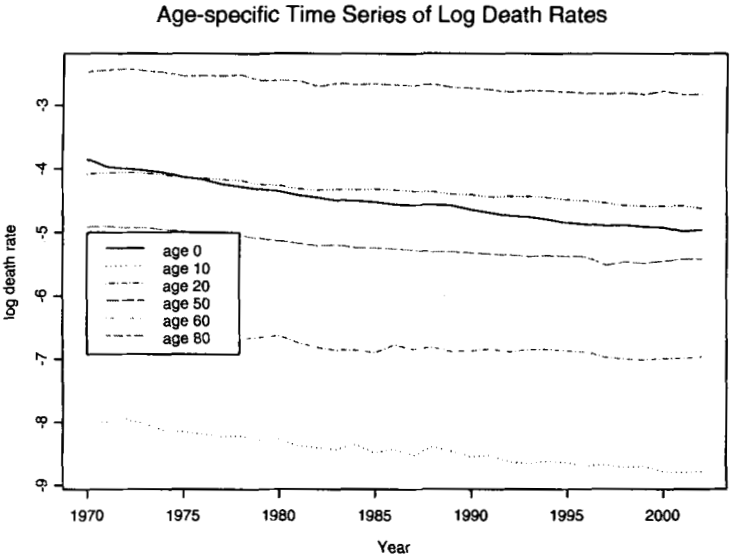

FIGURE 2: Age-specific time series $m(x,t)$ for some fixed x.

Let $d(x,t)$ denote the centered log-death rate matrix, $d(x,t) = m(x,t) - \sum_t m(x,t)/n$. We model $d(x,t)$ instead of $m(x,t)$ in order to compare our method with that of Lee & Carter (1992) who also use centered data. Plots of $d(x,t)$ are shown in Figure 3 as a function of x for some fixed t, and also as a function of t for various fixed ages x. From the plots we see that neighboring time series $d(x,t)$ and $d(x',t)$, where x and x' are close, e.g. ages 60 and 61, behave quite similarly. To compensate for short time records, the semiparametric method combines information from several agewise neighboring time series.

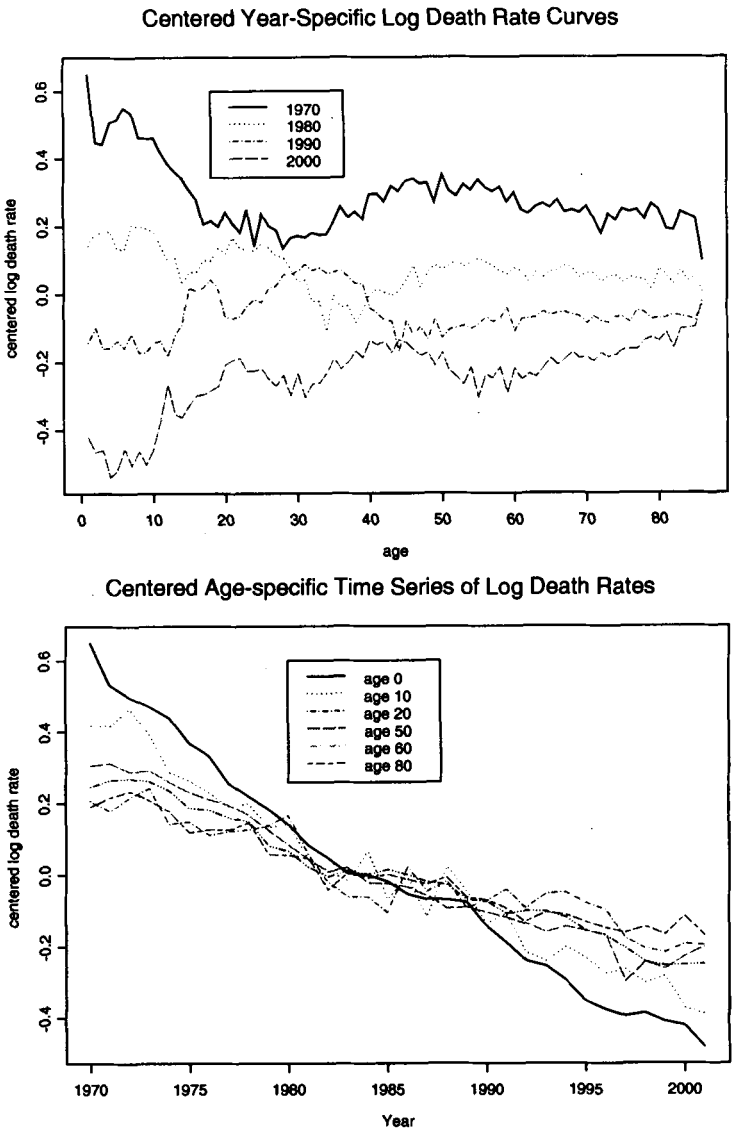

FIGURE 3: Plots of centered log-death rates $d(x,t)$ as a function of x for fixed t (top) and as a function of t for fixed x (bottom).

1.2. The Lee–Carter model.

The model proposed by Lee & Carter (1992) is used by the U.S. Census Bureau as a benchmark for their population forecasts, and its use has been recommended by the two most recent U.S. Social Security Technical Advisory Panels. It also appears to be the dominant method in the academic literature and is used widely by scholars forecasting all-cause and cause-specific

mortality around the world. See Lee (2000) and Lee & Miller (2001). The Lee–Carter model is based on principal components. If m denotes the number of mortality time series, each corresponding to a specific age, the Lee–Carter model searches for the first principal component in m dimensional time series data, and solves for the age and time parameters by using singular value decomposition.

The method presented in this paper is very different and seems appropriate for short range forecasting. Both methods, however, are extrapolative in that future mortality rates are estimated from past rates. Lee & Carter (1992) employed tabulated mortality data available from 1900 to 1987. However we shall compare the two methods using the annual data systematically collected from 1970 to 2002.

2. AN APPROACH TO SEMIPARAMETRIC TIME SERIES FORECASTING

Our approach for tackling the problem of short time series is based on a certain "tilt" model studied in several works including Fokianos, Kedem, Qin & Short (2001), Gilbert, Lele & Vardi (1999), Qin (1993), Qin (1998), Qin & Zhang (1997), and Vardi (1982, 1985).

In what follows next, x_{jt} denotes the jth time series depending on covariate vector z_{jt} through model f_j.

2.1. The density ratio model.

Consider the following $m = q + 1$ time series regressions,

$$\begin{aligned} x_{1t} &= f_1(z_{1,t-1}) + \varepsilon_{1t}, \quad t = 1, \ldots, n_1, \\ &\vdots \\ x_{qt} &= f_q(z_{q,t-1}) + \varepsilon_{qt}, \quad t = 1, \ldots, n_q, \\ x_{mt} &= f_m(z_{m,t-1}) + \varepsilon_{mt}, \quad t = 1, \ldots, n_m, \end{aligned} \tag{1}$$

where the n_i are small, the vectors $z_{i,t-1}$ contain past values of covariate time series, and ε_{kt} are independent noise components. Since ε_{kt} cannot be used directly for estimation, our proposed strategy is to first fit model (1) and then use the residuals e_{kt} wherever ε_{kt} are required.

Suppose the ε_{kt} have probability densities,

$$\begin{aligned} \varepsilon_{1t} &\sim g_1(x), \quad t = 1, \ldots, n_1, \\ &\vdots \\ \varepsilon_{qt} &\sim g_q(x), \quad t = 1, \ldots, n_q, \\ \varepsilon_{mt} &\sim g_m(x), \quad t = 1, \ldots, n_m. \end{aligned} \tag{2}$$

Define the reference density $g(x) \equiv g_m(x)$ with $G(x) \equiv G_m(x)$ the corresponding cumulative distribution function. Then we shall assume the density ratio model relative to the reference $g(x)$,

$$\frac{g_j(x)}{g(x)} = \exp\{\alpha_j + \beta_j^\top h(x)\}, \quad j = 1, \ldots, q. \tag{3}$$

This in turn gives the tilt model

$$g_j(x) = e^{\alpha_j + \beta_j^\top h(x)} g(x), \quad j = 1, \ldots, q, \tag{4}$$

with a normalizing constant α_j, vector β_j, and a vector valued distortion or tilt function $h(x)$. Implicitly, α_j is a function of β_j. The distorted densities g_j, the reference g, as well as the α_j and β_j are all unknown, but the distortion function $h(x)$ is assumed to be known and its choice depends on the data.

An important special case of (3) is obtained in the normal case. Assume that $x_1 \sim \mathsf{N}(\mu_1, \sigma_1^2)$ and $x_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$ with densities g_1 and g_2, respectively. Then the density ratio (3) becomes

$$\frac{g_1(x)}{g_2(x)} = \exp\left\{\log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2} + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}, \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)\begin{pmatrix} x \\ x^2 \end{pmatrix}\right\}, \quad (5)$$

with α and $\beta \equiv (\beta_1, \beta_2)^\top$ depending on the normal parameters,

$$\alpha = \log\left(\frac{\sigma_1}{\sigma_2}\right) + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}, \quad \beta = \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}, \frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)^\top,$$

and a two dimensional distortion function $h(x) = (x, x^2)^\top$. Notice that $h(x)$ degenerates to x^2 when $\mu_1 = \mu_2 = 0$, and (5) reduces to

$$g_1(x) = e^{\alpha + \beta x^2} g_2(x) \quad (6)$$

with scalars α, β. The tilt model (6) is useful when the distributions are centered at zero and are symmetric.

2.2. Estimation.

Combine all the residuals from the $q+1$ regressions into a single vector of length $n = n_1 + \cdots + n_q + n_m$,

$$\mathbf{e} = (e_1, \ldots, e_n)^\top \equiv \left((e_{11}, \ldots, e_{1n_1}), \ldots, (e_{q1}, \ldots, e_{qn_q}), (e_{m1}, \ldots, e_{mn_m})\right)^\top. \quad (7)$$

Maximum likelihood estimates for the α_j, β_j, and $G(x)$ can be obtained from the entire vector of residuals (7) by maximizing the likelihood over a class of step cumulative distribution functions with jumps at the values $e_1, \ldots, e_n$. Let $p_i = dG(e_i)$ denote the probability of jump at point e_i. Then the semiparametric likelihood becomes

$$L(\alpha, \beta, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp\{\alpha_1 + \beta_1^\top h(e_{1j})\} \cdots \prod_{j=1}^{n_q} \exp\{\alpha_q + \beta_q^\top h(e_{qj})\}. \quad (8)$$

The likelihood (8) is sometimes referred to as empirical likelihood. To maximize the likelihood (8) we follow a profiling procedure used in Fokianos, Kedem, Qin & Short (2001). First, fix the α and β. Then, subject to the normalization constraint $\sum_{i=1}^{n} p_i = 1$, and the constraints induced by the tilt model (4)

$$\sum_{i=1}^{n} p_i[\exp\{\alpha_j + \beta_j^\top h(e_i)\} - 1] = 0, \quad j = 1, \ldots, q,$$

the p_i which maximize (8) are given by

$$p_i \equiv p_i(\alpha, \beta) = \frac{1}{n_m} \frac{1}{1 + \rho_1 \exp\{\alpha_1 + \beta_1^\top h(e_i)\} + \cdots + \rho_q \exp\{\alpha_q + \beta_q^\top h(e_i)\}}$$

where $\rho_j = n_j/n_m$, $j = 1, \ldots, q$, are the relative series sizes. The final solution for the p_i is

$$\hat{p}_i = \frac{1}{n_m} \frac{1}{1 + \rho_1 \exp\{\hat{\alpha}_1 + \hat{\beta}_1^\top h(e_i)\} + \cdots + \rho_q \exp\{\hat{\alpha}_q + \hat{\beta}_q^\top h(e_i)\}} \quad (9)$$

obtained by substituting $p_i(\alpha, \beta)$ back into the likelihood (8) and finding maximum likelihood estimators $\hat{\alpha}_j$ and $\hat{\beta}_j$ through profiling. With $I(B)$ the indicator of the event B, the estimated

reference cumulative distribution function is given by

$$\begin{aligned}\widehat{G}(t) &= \sum_{i=1}^{n} \hat{p}_i I(e_i \le t) \\ &= \frac{1}{n_m} \sum_{i=1}^{n} \frac{I(e_i \le t)}{1 + \rho_1 \exp\{\hat{\alpha}_1 + \hat{\beta}_1^\top h(e_i)\} + \cdots + \rho_q \exp\{\hat{\alpha}_q + \hat{\beta}_q^\top h(e_i)\}}. \end{aligned} \quad (10)$$

Smoothing the $\hat{p}_i$ in (9) by a kernel or, alternatively, smoothing increments of $\widehat{G}$ in (10) gives the reference density estimate $\hat{g}$ (Fokianos 2004).

The main point of the semiparametric paradigm discussed here is that the reference cumulative distribution function $G(x)$ is estimated from many samples giving an improved estimate as compared with the empirical cumulative distribution function which is obtained from the reference sample only. This fact has been addressed carefully by several authors. In particular, very general optimality properties of the semiparametric estimates are discussed rigorously in Gilbert (2000). Let $\hat{\boldsymbol{\theta}}_n = (\hat{\alpha}_1, \ldots, \hat{\alpha}_q, \hat{\beta}_1, \ldots, \hat{\beta}_q)^\top$. Then, as Gilbert (2000) has shown, $(\hat{\boldsymbol{\theta}}_n, \widehat{G})$ are asymptotically normal and efficient. Likewise, Zhang (2000b) has shown that quantile estimates obtained by the semiparametric method from both case and control samples are more efficient than estimates that are based on the control sample only, ignoring the case information. More recently, Fokianos (2004) showed that by merging information following the semiparametric paradigm, we obtain improved kernel density estimates with the same bias as the traditional kernel density estimates but with smaller asymptotic variance. Our data analysis below supports this claim. Moreover, merging information in this way can result in powerful tests for distribution equality. See Fokianos, Kedem, Qin & Short (2001), Gagnon (2005), and Kedem & Wen (2007). Regarding the uncertainty in $\widehat{G}$, as shown by Zhang (2000b) and more recently by Lu (2007), $\sqrt{n}\,(\widehat{G} - G)$ converges to a Gaussian process with mean zero and a rather complex covariance structure. In addition, it can be shown that the estimates $\hat{\alpha}$ and $\hat{\beta}$ are asymptotically normal with a covariance structure depending on functionals of G; see Fokianos, Kedem, Qin & Short (2001), Kedem & Wen (2007), Lu (2007), Qin & Zhang (1997), and Zhang (2000a).

We shall apply the semiparametric paradigm in forecasting U.S. mortality rates by combining information, or borrowing strength, from several (agewise) neighboring short U.S. mortality time series.

2.3. Forecasting.

The preceding discussion motivates the following time series forecasting method (Kedem, Gagnon & Guo 2005). Since $x_{m,t+1} = f_m(z_{m,t}) + \varepsilon_{m,t+1}$, and $\varepsilon_{m,t+1} \sim G$, where G is the reference distribution estimated semiparametrically by $\widehat{G}$ as in (10), we estimate the predictive distribution at time $t+1$ conditional on past data $z_{m,t}$ as follows:

$$\begin{aligned} \mathrm{P}\,(x_{m,t+1} \le x \mid z_{m,t}) &= \mathrm{P}\,\{x_{m,t+1} - f_m(z_{m,t}) \le x - f_m(z_{m,t}) \mid z_{m,t}\} \\ &= \mathrm{P}\,\{\varepsilon_{m,t+1} \le x - f_m(z_{m,t}) \mid z_{m,t}\} \\ &= G\{x - f_m(z_{m,t})\} \\ &\approx \widehat{G}\{x - f_m(z_{m,t})\}. \end{aligned} \quad (11)$$

All sorts of point predictors can be obtained from (11). In particular, a one-step ahead predictor for $x_{m,t+1}$ given the past can be approximated by calculating the (conditional) mean of the shifted distribution $\widehat{G}(x - f_m(z_{m,t}))$. Approximate prediction intervals can also be obtained from the estimated distribution (11).

2.4. About independence.

Strictly speaking, the method as outlined above requires independent noise components ε_{kt}, but since the ε_{kt} are replaced by the corresponding residuals in practice, strict independence is not guaranteed. The question is then whether the independence requirement may be relaxed.

In Kedem & Fokianos (2002) it was shown that, subject to some regularity conditions, by using partial likelihood it is possible to bypass independence and extend the generalized linear models methodology to dependent time series. With this in mind, if the likelihood (8) is interpreted in a partial sense, then this suggests the method may still be viable even when residuals are used, as is evident from the present application to mortality rates forecasting, and another very different application to filtered mortality time series reported in Kedem, Gagnon & Guo (2005).

More directly, the independence question was investigated in Kedem, Gagnon & Guo (2005) by means of an extensive simulation applied to the bivariate linear system

$$\begin{aligned} x_t &= a_1 x_{t-1} + a_2 y_{t-1} + \varepsilon_t, \\ y_t &= b_1 x_{t-1} + b_2 y_{t-1} + \eta_t, \end{aligned} \tag{12}$$

$t = 1, \ldots, N$, with independent Gaussian noise components $\varepsilon_t \sim \mathsf{N}(0, \sigma_1^2)$ and $\eta_t \sim \mathsf{N}(0, \sigma_2^2)$ satisfying the density ratio model with

$$g_\varepsilon(x) = e^{\alpha + \beta x^2} g_\eta(x). \tag{13}$$

Then, estimating G when all the parameters in (12) and (13) were known and using known independent noise components ε_t, η_t, and also when all the parameters were estimated (once with $N = 50$ and once with $N = 500$) and using residuals $\hat{\varepsilon}_t, \hat{\eta}_t$, gave nearly the same $\widehat{G}$, and hence nearly identical forecasts. This suggests that there are situations where the quality of prediction of the semiparametric method is not necessarily affected much by the use of the observed residuals.

3. ONE YEAR AHEAD PREDICTION OF U.S. MORTALITY

3.1. A two-stage procedure.

Define $a_k = \sum_t m(k,t)/n$. As mentioned above, we analyze the centered log-death rate matrix $d(k,t)$, $d(k,t) = m(k,t) - a_k$. For each fixed age k, consider the annual time series of centered log death-rates from 1970 to 2001. Thus $t = 1, \ldots, 32$.

Write $x_{kt} = d(k,t)$. First, to each such time series we fit the first order autoregressive model with drift c_k,

$$x_{kt} = b_k x_{k,t-1} + c_k + \varepsilon_{kt}, \quad k = 1, \ldots, q, m. \tag{14}$$

The drift parameter is added in order to capture a downward trend observed in the age-specific centered log death-rate time series as exemplified in Figure 3. Accordingly, the functions f_k in the system (1) are given by $f_k(x_{k,t-1}) = b_k x_{k,t-1} + c_k$. The coefficient b_k and the drift c_k are estimated by least squares, and in our application the ε_{kt} are replaced by the residuals derived from the model (14).

Next, we choose a density ratio model for the residuals. Data analysis shows that the residuals corresponding to model (14) are centered around zero and that their histograms resemble those obtained from small normal samples. This motivates the distortion model (6) with $h(x) = x^2$.

We consider the mth residual sample $(e_{m1}, \ldots, e_{mn_m})$ as the reference where each component has distribution function G and density g. Similarly, assume that each residual component of the vector $(e_{k1}, \ldots, e_{kn_k})$ has distribution G_k and density g_k, $k = 1, \ldots, q$. Following the above semiparametric paradigm, and combining it with insight gained from histograms of residuals from (14), we assume the density ratio relationship,

$$g_k(x) = e^{\alpha_k + \beta_k x^2} g(x), \quad k = 1, \ldots, q. \tag{15}$$

An application of the semiparametric procedure to the combined data **e** defined by (7) gives the semiparametric estimate $\widehat{G}$ for the reference distribution. Similarly, from (10) and (15) we obtain the estimated distribution function of the kth sample $\widehat{G}_k$ from which the predictive distribution is computed by

$$\mathrm{P}\,(x_{k,t+1} \le x \mid z_{kt}) \approx \widehat{G}_k(x - b_k x_{kt} - c_k). \tag{16}$$

3.2. Data analysis.

We consider 85 age-specific time series of log-death rates (all-cause) for ages $1, \ldots, 85$, where the age category 85 includes ages 85 and older. To simplify the analysis, this grouping or lumping of ages 85 and older had to be done at some point and we chose, somewhat arbitrarily, age 85 as a threshold. However, the data file does have the specificity to subpartition this category to obtain a more detailed mortality prediction. Mortality at age 0 is not considered in the present analysis due to its behavior which is very different from that at other ages. See Figure 2.

From the previous discussion, the assumption that the density ratio model (15) holds for time series groups corresponding to neighboring ages seems reasonable. Indeed, in retrospect our data analysis lends credence to this assumption. In our analysis, therefore, we apply the semiparametric method by combining information from each of the age groups, consisting of five ages each and dubbed "5-age," 1–5, 6–10, ..., 81–85, a total of 17 groups, where the time series "in the middle" of each group is taken as the reference. For example, in the group 1–5, the time series of age 3 is taken as the reference, meaning that the relevant distribution from this time series serves as the reference distribution for the group. We applied the semiparametric model separately to each group to estimate the reference distribution and the corresponding distorted distributions to obtain predicted mortality curves.

As an illustration, consider the age group 31–35 from 1970 to 2001. As mentioned before, we chose a quadratic distortion function $h(x) = x^2$ due to the rough symmetry of the residuals around zero, resembling the behavior of normal residuals. There are altogether five residual samples, and the sample of residuals from age 33 is considered as the reference. The actual conditional point predictions of log death-rate in 2002 for the age group 31–35 are obtained from (16) by computing the first moments of the shifted predictive distributions $\widehat{G}_k$, $k = 31, \ldots, 35$, respectively, with $\widehat{G}_{33}$ as the reference. This analysis is repeated for all 17 groups. The 2002 prediction results for all ages are compared with the true 2002 centered log-rates in the tables and figures below.

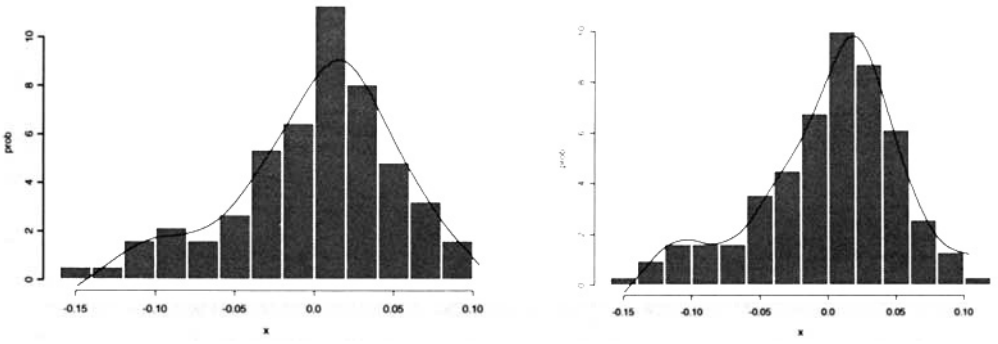

FIGURE 4: Estimated reference probability density function of age 33 from the combined data **e** for the 3-age group 32–34 (left), and the 5-age group 31–35 (right), respectively.

It is also interesting to compare the results from the 3-age group 32–34 with the 5-age group 31–35. Figure 4 shows the histograms and overlaid estimated reference density of age 33 ob-

tained from the combined data e for the 3-age and 5-age groups. Since we combined more information in the 5-age group there is a noticeable improvement in the density fit.

For age group 31–35, the estimated tilted cumulative distribution functions $\widehat{G}_k(x)$ from (15), each estimated from $5 \times 32 = 160$ residuals, and the corresponding empirical cumulative distribution functions, each from 32 residuals, are shown in Figure 5 for ages 31, 32, 33, 34 (the cumulative distribution function for age 35 is not plotted). Since more information is used (or combined) in deriving the $\widehat{G}_k(x)$ than used in obtaining the empirical distributions, the $\widehat{G}_k(x)$ are smoother as is evident from the figure. So, in some sense, the semiparametric cumulative distribution functions are smooth versions of the corresponding empirical distributions.

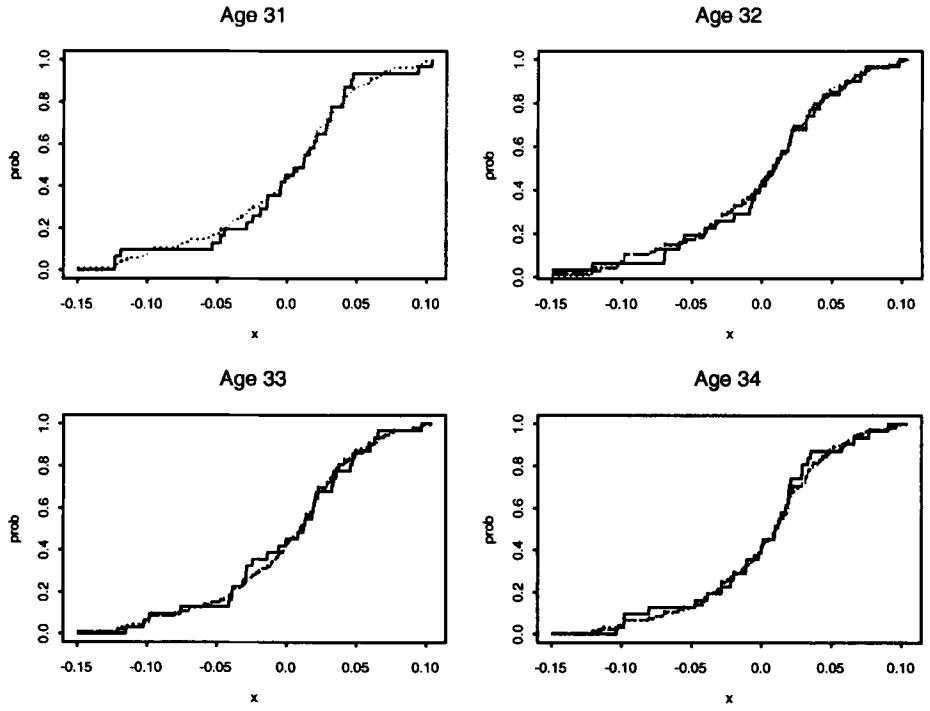

FIGURE 5: Comparison of the empirical (solid line) and estimated (dotted line) cumulative distribution functions for the indicated ages. The estimated cumulative distribution function for age 35 is not shown.

The corresponding 5-age smoothed probability density functions $\hat{g}_k(x)$ (solid lines) and their related histograms are shown in Figure 6 for ages 31, 32, 33, 34. For the sake of comparison, for ages 32, 33, 34 the figure also depicts the 3-age smoothed $\hat{g}_k(x)$ (dotted lines). The estimated probability density function for age 35 is not shown. The plots point to the consistency of the method in that the 3-age and 5-age estimates are not far apart.

Once the estimated reference $\widehat{G}(x)$ and the estimated distributions $\widehat{G}_k(x)$ are obtained, we apply (16) to approximate the probability distribution of the one-year-ahead centered log-death rate in 2002 for the age group 31–35. As a point predictor we use the mean of the predictive distribution, that is, the conditional expectation. The corresponding 95% confidence interval is also derived from the estimated predictive distribution.

Table 1 gives the prediction results only (to save space) for ages $5, 10, 20, \ldots, 80$. For each indicated age it gives the semiparametric prediction for 2002 and the corresponding prediction interval (PI), as well as the Lee–Carter prediction. Comparison by mean square error (MSE) between the two methods is given in Table 2. Generally speaking, compared with the Lee–Carter method, the semiparametric method improves the prediction as measured by MSE. The improvement of the semiparametric method is more noticeable for age groups which display

more steady and gradual change of death rate as in age groups 31–50 and 71–85. From Table 2, the overall prediction MSE from the semiparametric method in the three cases reported there are 0.104, 0.187, 0.249, compared, respectively, to 0.297, 0.619, 0.645 from the Lee–Carter method. The most significant improvement is for the age groups 31–50 and 71–85, whereas both methods perform quite similarly for all other age groups as we see from the table.

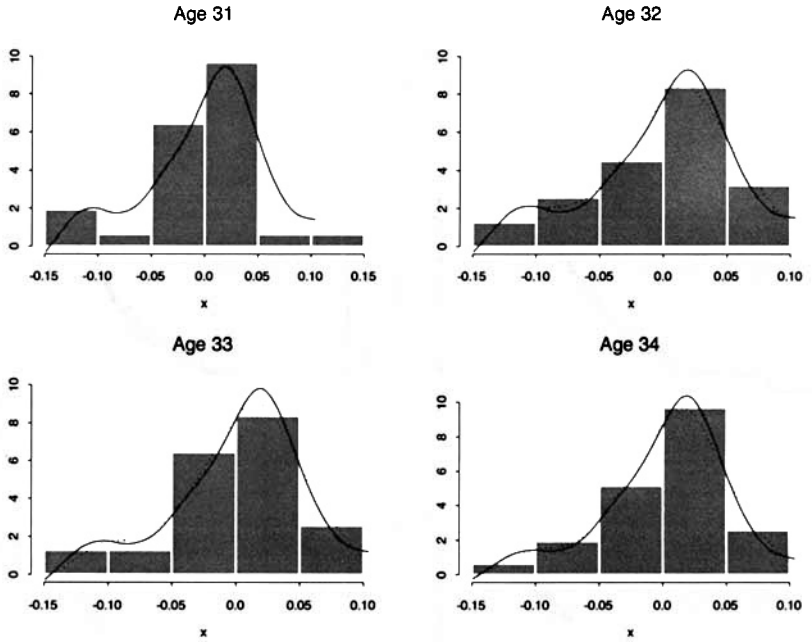

FIGURE 6: Histograms and overlaid estimated probability density functions for 3-age group 32–34 (dotted line) and 5-age group 31–35 (solid line). The estimated probability density function for age 35 is not shown.

In the above data analysis we combined information from non-overlapping 5-age groups. The analysis was repeated for the general poulation case by using a sliding window of overlapping 5-age groups, each time moving up by a single year. Interestingly, the MSE results were very close to those reported in Table 2, replacing the SP row 0.104, 0.050, 0.015, 0.030, 0.009 by 0.105, 0.051, 0.014, 0.031, 0.008. This suggests that the choice of the reference time series within an age group may be arbitrary.

TABLE 1: Prediction comparison between the semiparametric and Lee–Carter methods for 2002 for some ages. The first two rows give the 95% PI bounds for the semiparametric forecasts, and the rest are the predictions from the semiparametric method (SP), true values in 2002, and the prediction from the Lee–Carter (LC) method.

Age	5	10	20	30	40	50	60	70	80
Lower	−8.781	−8.933	−7.072	−6.977	−6.276	−5.473	−4.628	−3.776	−2.905
Upper	−8.599	−8.671	−6.870	−6.755	−6.094	−5.362	−4.557	−3.695	−2.774
SP	−8.699	−8.819	−6.997	−6.858	−6.178	−5.431	−4.601	−3.749	−2.842
True	−8.639	−8.785	−6.970	−6.868	−6.172	−5.416	−4.622	−3.733	−2.824
LC	−8.661	−8.835	−7.023	−6.810	−6.252	−5.534	−4.615	−3.752	−2.874

From Table 2, we see that the MSE from the semiparametric method is lower for the general population case than in both the female and white female cases. This is not surprising since

more data are available from the total population, whereas in the other two cases we deal with subpopulations. The fact that the age group 1–30 has a larger MSE than that from the other groups is due to the large variation of the data associated with that age group.

Since our mortality data are truncated at age 85, we cannot calculate traditional life tables from the death rate forecasts. Instead we provide in Table 3 a comparison between the true and predicted (by our method) number of survivors by age and sex out of 100,000 live births. The true values and their forecasts are close.

TABLE 2: Mean square error of (all-cause) prediction from the semiparametric (SP) and Lee–Carter (LC) methods for the general population, female, and white female.

	Age group	1–85	1–30	31–50	51–70	71–85
General	SP model	0.104	0.050	0.015	0.030	0.009
	LC model	0.297	0.078	0.180	0.026	0.013
Female	SP model	0.187	0.121	0.026	0.032	0.008
	LC model	0.619	0.226	0.341	0.027	0.025
W. Female	SP model	0.249	0.176	0.031	0.033	0.007
	LC model	0.645	0.257	0.329	0.041	0.019

4. TWO-YEAR AHEAD FORECASTING

So far we have discussed one-year ahead prediction. However, our one-step procedure can be extended to multi-year ahead forecasting. One way to proceed is to use the predicted values from previous steps when making long term predictions. Thus in two-year ahead forecasting we use the previous one-year ahead forecasts, and proceed as above. The prediction error may get amplified through each additional step even if minor deviations of prediction from true values occur. The results from this procedure are reported in Table 4(a). Again, the overall MSE is lower for the semiparametric method as compared with the Lee–Carter method.

A second procedure for forecasting j-years ahead is to extend the above one-step ahead forecasting method to residuals resulting from time series regression models where the present at time t is regressed on observed values up to and including $t - j$. Thus in the present case, to get two-year ahead mortality forecasts we use (14) with the modification that x_{kt} is regressed on $x_{k,t-2}$. The MSE from this method is reported in Table 4(b). Once more, the overall MSE is lower for the semiparametric method as compared with the Lee–Carter method. The disadvantage of this procedure is that some data are lost due to the larger time lags.

5. CONCLUDING REMARKS

We have used a two-stage forecasting semiparametric procedure suitable for short time series to obtain forecasts of U.S. age-specific mortality rates. To estimate conditional predictive distributions, the method combines short time series by appealing to a density ratio model. Point predictors as well as future probabilities can be obtained from the estimated conditional distributions. A comparison with the well known Lee–Carter singular value decomposition method points to the potential of the semiparametric method. In general the semiparametric method provides more precise short term prediction as compared with the Lee–Carter procedure.

The method we used is non-Bayesian. Bayesian methods for forecasting in short time series are available, a useful special case of which is discussed by De Oliveira, Kedem & Short (1997), and Kedem & Fokianos (2002). Interestingly, there too the prediction is based on a predictive distribution, but the method is very different.

Death rates drop rapidly from infants to children, thus, combining data from age zero with other ages to form an age group is less appealing. It seems preferable to employ methods suitable for univariate time series to forecast the annual mortality for age zero. When monthly infant death rates are available, the semiparametric method can be applied to this age group separately.

TABLE 3: Number of survivors by age and sex, out of 100,000 born alive, from both semiparametric forecasts and true values in 2002.

	Forecast			True		
Age	Total	Male	Female	Total	Male	Female
0	100000	100000	100000	100000	100000	100000
1	99311	99231	99371	99298	99217	99360
5	99182	99085	99256	99174	99076	99252
10	99107	99004	99186	99098	98992	99184
15	99013	98890	99108	99000	98875	99104
20	98685	98425	98914	98662	98400	98902
25	98219	97717	98679	98192	97691	98662
30	97746	97031	98394	97722	97006	98384
35	97189	96275	98012	97171	96249	98009
40	96384	95221	97425	96386	95228	97422
45	95231	93739	96550	95216	93733	96520
50	93558	91581	95293	93515	91553	95220
55	91205	88625	93446	91128	88521	93381
60	87762	84429	90632	87629	84211	90570
65	82616	78258	86313	82484	77986	86328
70	75218	69571	79978	75148	69339	80074
75	65081	57967	71052	65014	57710	71164
80	51665	43306	58630	51680	43142	58758
85	35348	26897	42244	35442	26938	42330

For convenience, we chose to fit to the mortality rate time series the AR(1) model (14). This of course is only one possibility and there are other choices. For example, we could set the coefficient b_k in (14) to be 1, or use an AR(2) model. A model which provides a better fit could reduce the prediction error.

TABLE 4: Prediction MSE from the semiparametric (SP) and Lee–Carter (LC) methods for two-year ahead forecasting. (a) Predicted one-year ahead forecasts are used. (b) Autoregression lagged by 2.

	Age group	1–85	1–30	31–50	51–70	71–85
(a)	SP model	0.180	0.128	0.019	0.026	0.007
	LC model	0.389	0.088	0.246	0.033	0.021
(b)	SP model	0.211	0.132	0.048	0.025	0.005
	LC model	0.389	0.088	0.246	0.033	0.021

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers and several colleagues at the National Center for Health Statistics for useful comments and suggestions which have been incorporated in the paper. This work was supported by CDC/NCHS.

REFERENCES

V. De Oliveira, B. Kedem & D. A. Short (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, 92, 1422–1433.

K. Fokianos (2004). Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society Series B*, 66, 941–958.

K. Fokianos, B. Kedem, J. Qin, & D. A. Short (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43, 56–65.

R. E. Gagnon (2005). *Certain Computational Aspects of Power Efficiency and State Space Models.* Doctoral dissertation, Dept. of Mathematics, University of Maryland, College Park.

P. B. Gilbert (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *The Annals of Statistics*, 28, 151–194.

P. B. Gilbert, S. R. Lele & Y. Vardi (1999). Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. *Biometrika*, 86, 27–43.

L. Heligman & J. H. Pollard (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107, 659–671.

B. Kedem & K. Fokianos (2002). *Regression Models for Time Series Analysis.* Wiley, New York.

B. Kedem, R. E. Gagnon & H. Guo (2005). Time Series Prediction Via Density Ratio Modeling. Unpublished manuscript, Dept. of Mathematics, University of Maryland, College Park.

B. Kedem & S. Wen (2007). Semi-parametric cluster detection. *Journal of Statistical Theory and Practice*, 1, 49–72.

R. Lee (2000). The Lee–Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, 4, 80–93.

R. D. Lee & L. R. Carter (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87, 659–671.

R. Lee & T. Miller (2001). Evaluating the performance of the Lee–Carter approach to modeling and forecasting mortality. *Demography*, 38, 537–549.

G. Lu (2007). *Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and its Application to Mortality Forecasting.* Doctoral dissertation, Dept. of Mathematics, University of Maryland, College Park.

J. Qin (1993). Empirical likelihood in biased sample problems. *The Annals of Statistics*, 21, 1182–1196.

J. Qin (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85, 619–630.

J. Qin & B. Zhang (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609–618.

Y. Vardi (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10, 616–620.

Y. Vardi (1985). Empirical distribution in selection bias models. *The Annals of Statistics*, 13, 178–203.

R. Wei, R. L. Curtin & R. Anderson (2003). Modeling U.S. mortality data for building life tables and further studies. *2003 Joint Statistical Meeting Proceedings, Biometrics Section*, 4458–4464.

B. Zhang (2000a). M-estimation under a two sample semiparametric model. *Scandinavian Journal of Statistics*, 27, 263–280.

B. Zhang (2000b). Quantile estimation under a two-sample semi-parametric model. *Bernoulli*, 6, 491–511.

Received 21 May 2007
Accepted 5 November 2007

Benjamin KEDEM: bnk@math.umd.edu
Guanhua LU: ghlu@math.umd.edu
Department of Mathematics
University of Maryland
College Park, MD 20742, USA

Rong WEI: rrw5@cdc.gov
Paul D. WILLIAMS: pdougwilliams@verizon.net
Office of Research and Methodology
National Center for Health Statistics
Hyattsville, MD 20782, USA