Reply to reviewers

We would like to thank the editor and the reviewers for their time, feedback and suggestions. We tried hard to address all their comments and we consider that the manuscript has greatly improved thanks to your suggestions.

We have uploaded a marked text highlighting the parts of the text where changes were made.

Summary

The manuscript introduces the maximum entropy mortality model where forecasting of future mortalities can be based on extrapolation of a finite number of statistical moments for the age-at-death distribution using a multivariate time series model. As a case study, the 'England and Wales' male mortality experience for years 1960-2016 and ages 0-95 is considered. The model is shown to outperform (in the sense of out-of-sample predictive power w.r.t. ME, MAE, etc.) classic approaches such as the Lee and Carter (1992) model.

Comments and questions

The manuscript addresses a relevant issue for actuarial practice and contains an interesting empirical study.

I have the following comments and questions:

1. I suggest you shorten Section 2.1 and Section 2.2 of the manuscript considerably, hereby heightening the mathematical level of exposition. Also, Appendix 5.1 should be omitted as it can be assumed well known to your audience. Thank you. We have eliminated several paragraphs, particularly section 2.2 of details of the entropy and the section in the appendix where the time series model is described in detail.

2. There appears to be a number of minor typos (e.g. p. 4 line 49, p. 7 line 34, and p. 11 line 60 - in all three cases the verb is missing a past tense `d'). Additional proof reading would be beneficial. The text has undergone English editing.

3. In Section 1 you provide an interesting discussion on the use of age-at-death distributions compared to mortality rates for longevity forecasting. If I understand your arguments correctly, you conclude that extrapolation methods based on death frequencies are advantageous to methods based on mortality rates (p. 5 top). Yes.

   How do you reach this conclusion? This is the main hypothesis that we test in our manuscript, namely that dx-models outperform mx-models in forecasting. We test this by studying the performance of the 6 mortality forecasting models assessed over 169 scenarios using population data from 10 developed countries and using 6 different accuracy measures. Our main conclusion is that no model can outperform all the other models over all the scenarios, however we noticed a tendency of the methods based on death frequencies to exhibit a superior predictive power as shown in Table 4.3. We have made clear now in the introduction of Section 1 that this is the hypothesis that we want to test in our manuscript.

4. What do α and ω of equation (1) refer to? Is this the range of the distribution? Yes. clarifying sentence was added in the manuscript right after equation 1.

5. In Section 2.1 and in the beginning of Section 2.3, μ refers to the true moment, while later in Section 2.3 (p. 11 top), μ now refers to the 'observed numerical values'. What do you mean with the observed values? We have reformulated the expression in the manuscript to avoid confusing language.

   Are these the later forecasts of empirical moments? The *MaxEnt* algorithm uses a given number of moments to approximate a density function at a given point in time. In the context presented in this article, for obtaining the fitted values of the MEM one would use the observed empirical moments, and for getting the forecast values (future distributions) the moments generated from a stochastic process must be used. Note that the density-estimation-part of the model, applied for each year of data separately, works in the same manner regardless of the input data (observed empirical values or predicted moments from a random-walk with drift process).

   Therefore, these can be one or the other. At this point in the article we are methodologically discussing only a density estimation algorithm without referring to forecasting.
   What about $\tilde{\mu}$ in equation (16)? Is this an empirical central moment? Yes, this is an empirical central moment used in fitting the time series model. We have clarifyied this in the text.

6. There appears to be an error in equation (15) - compare with equation (2.7) of Mead and Papanicolaou (1984).
   Equation (15) and others, are shown in a different form compared with Mead and Papanicolaou (1984) in order to avoid the angle bracket notation ⟨ ⟩. The reason for this is the decrease in popularity and usage of a mathematical notation that might cause confusions. Equation (15) indicates that μ_n is found if equation (14) is differentiated with respect to λ_n and set equal to zero.

7. Based on which observations and predictions are you computing the out-of-sample performance? We are using the observed and model-specific-forecast-life-expectancy values over all ages to obtain a matrix of errors (ME, MAE etc.) in each scenario. The accuracy of the forecast in the scenario for a given model is the average of all the values in the matrix.

   You are mentioning and the use of six different accuracy measures and the aggregation (by averaging) over all scenarios (appearing by moving the evaluation windows 1 year forward). But how do you arrive at a value for a specific accuracy measure and scenario? For a specific country and model multiple scenarios of equal dimension (fitted/forecasted years) are explored in order to account for the robustness of the models. Usually 18 scenarios per country. The aggregation is done by averaging the results over all scenarios for each model/accuracy measure/country.

   In the tables included in the article we are showing aggregated measures, they are easier to convey the message of which measure performs the best.

What are the observation(s) and prediction(s) you compare? Observed and forecast life expectancy levels at all ages, from birth until age 95. This is shown in the figures as well and is denoted in the legend with “Validation Set”.

I suggest you add details providing a precise explanations. The following clarifying paragraph is now added in section 3.3:
For each scenario and model we are computing a matrix of accuracy errors. The overall accuracy of the forecast for a specific model–scenario is the average of all the values in the corresponding matrix. For a given population multiple scenarios of equal dimension (see section 3.4) are explored in order to account for the robustness of the models. The aggregation is done by averaging the results over all scenarios for each model/accuracy measure/country.

8. In footnote 2, you write that the indicators evaluated in this manuscript are the life expectancies. Does this in any way relate to my comment/question 7)? Yes, we have replied to your question (7) and with this also clarified (8).

9. You do not discuss cohort effects, see e.g. Renshaw and Haberman (2006). How does the MEM model compare to extrapolation based methods for mortality rates allowing for cohort effects? Can the MEM model capture cohort effects? I strongly suggest you to elaborate on these aspects. Furthermore, including a cohort model, e.g. the extension of the Lee-Carter model from Renshaw and Haberman (2006), in the model comparison would significantly improve the practical relevance of the model comparison.

   This is a very good point for which we are grateful to the reviewer. We extended our study by evaluating the Renshaw-Haberman (2006) model too. Please find the results and discussion in the revised manuscript. Thanks.

10. In Section 4 you conclude that the main advantage of the MEM model is that age specific trends are no longer based on the assumption of constant changes in mortality (as e.g. in the Lee-Carter model). How do you reach this conclusion? Under the Lee-Carter setting the future mortality levels are determined by extrapolation, most often in a linear fashion, of a univariate time index. This results in projections that show constant changes in mortality that do not hold over long periods. The MEM relies a multivariate approach to extrapolating the location and shape measures (variance, skewness, …) of the age-at-death distribution. When the shape measures are translated into mortality dynamics the results will be anything but constant (accelerating, decelerating or negative changes, etc.).
    Could another (main) advantage of the MEM model not be that it is able to capture cohort effects?
    When we designed the method, we did not try to model or capture cohort effects. Our approach is purely statistical. However, by checking the results of our study we learned that MEM can capture cohort effects more accurately than most of the models. And this is due to the approach of modelling mortality in a compositional framework (i.e. if a life is saved at age x it must be placed back in the death-distribution at later age, say x+t). We have now included this among the advantages of the model and thank the reviewer for directing us to this advantage of the proposed model.

References (besides those cited in the manuscript)

Renshaw and Haberman (2006) A cohort-based extension to the Lee-Carter model for mortality reduction factors. Insurance: Mathematics and Economics 38, pp. 556 - 570.
Included.

Additional note:
Compared to the previous version of the manuscript the results in the presented tables and figures suffered minor changes. The changes are not affecting their initial interpretability, nor the performance of the models compared to each other. These are due to the adoption of a uniform approach to dealing with the jump-off adjustment in the first years of forecast for all models.