Deepening Lee-Carter for longevity projections with uncertainty estimation

Mario Marino[1], Susanna Levantesi, Andrea Nigri

[1] Department of Statistics, Sapienza University of Rome
m.marino@uniroma1.it

3^{rd} Insurance Data Science Conference

18 June 2021

Agenda

- Introduction
- Improving the Lee-Carter mortality density forecast
- Numerical application
- Conclusions

Agenda

- Introduction
- Improving the Lee-Carter mortality density forecast
- Numerical application
- Conclusions

Mortality forecasting and Neural Networks

- Predicting mortality continues to be a challenge for demographers and actuaries
- Nowadays, several stochastic mortality models are avalaible
  - Lee-Carter (LC) family
  - Cairns-Blake-Dowd (CBD) family
- Methodological advances in mortality forecasting based on Machine and Deep Learning models
  - Random forest, Gradient Boosting
  - Feed-forward, Convolutional, Recurrent Neural Networks (NN)
- Exploting NN models as predictors, the idea is to create a novel approach: the Mortality Neural Forecasting

Mortality forecasting and Neural Networks

- The present work follows and completes the study in Nigri et al. (2019)
- The approach is the following:
  A. Consider a stochastic mortality model as reference model to fit the observed mortality surface
  B. Forecast future mortality paths by a proper NN model
- The overall mortality model is hybrid, achieving:
  1. Ease of interpretation of age-period-cohort parameters
  2. Accuracy in estimating future mortality outcomes

Model proposal: the LC-LSTM

The LC model as reference model and the LSTM network to improve the LC mortality density forecast

Agenda

- Introduction
- Improving the Lee-Carter mortality density forecast
- Numerical application
- Conclusions

The LC-LSTM model - Formulation

- Reference model: LC Poisson model (Brouhns et al.(2002)).

For $x \in \mathcal{X} = \{0, 1, \dots, \omega\}$ and $t \in \mathcal{T} = \{t_0, t_1, \dots, t_n\}$, we have $D_{x,t} \sim Poi(E^c_{x,t} m_{x,t})$ and

$$\ln m_{x,t} = \alpha_x + \beta_x k_t. \tag{1}$$

- Let $\boldsymbol{\kappa}_{\mathcal{T}} = (k_{t-j})_{t \in \mathcal{T}}$ be the vector of the time lagged k_t, being $j \in \mathbb{N}$ the time lag, we consider:

$$k_t = f_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}; \boldsymbol{\mathcal{W}}\right) + \gamma_t, \tag{2}$$

with $f_{LSTM} : \mathbb{R}^j \to \mathbb{R}$ the LSTM function and $\boldsymbol{\mathcal{W}}$ the weights.

- Over the forecasting horizon $\mathcal{T}' = \{t_n + 1, t_n + 2, \dots, t_n + s\}$, the LC-LSTM model expression is:

$$\ln m_{x,t} = \hat{\alpha}_x + \hat{\beta}_x \left(f_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}'}; \boldsymbol{\mathcal{W}}\right) + \gamma_t\right), \quad \forall t \in \mathcal{T}', \tag{3}$$

with $\hat{\alpha}_x$ and $\hat{\beta}_x$ the estimates of age-dependent parameters.

The LC-LSTM model - Point Predictions

- From a general perspective, the LC time-index values should be interpreted as the realization of the following process:

$$k_t = \varphi\left(\boldsymbol{\kappa}_{\mathcal{T}}\right) + \gamma_t, \quad \forall t \in \mathcal{T}, \tag{4}$$

where the unknown function $\varphi : \mathbb{R}^j \rightarrow \mathbb{R}$ maps the vector $\boldsymbol{\kappa}_{\mathcal{T}}$ to k_t over the time horizon $\mathcal{T}$, unless the noise component.

- The network approximates $\varphi\left(\boldsymbol{\kappa}_{\mathcal{T}}\right)$ according to the time-index history:

$$\hat{k}_t = \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}; \hat{\boldsymbol{\mathcal{W}}}\right) = \mathbb{E}\left(k_t | \boldsymbol{\kappa}_{\mathcal{T}}\right) \tag{5}$$

- Therefore, the LC-LSTM model provides the following point predictions:

$$\ln \hat{m}_{x,t} = \mathbb{E}\left(\ln m_{x,t}\right) = \hat{\alpha}_x + \hat{\beta}_x \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}'}; \hat{\boldsymbol{\mathcal{W}}}\right), \quad \forall t \in \mathcal{T}'. \tag{6}$$

The LC-LSTM model - Uncertainty estimation

- Point predictions could be a poor information and a measure of uncertainty, such as prediction intervals, is necessary

- Exploting the bias-variance trade-off principle, the total variance associated to time-index values is:

$$\sigma^2_{k_t} = \sigma^2_{\hat{k}_t} + \sigma^2_{\gamma} + \mathbb{E}\left[BIAS\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right)^2\right] \qquad (7)$$

where $\sigma^2_{\hat{k}_t}$ is the NN output variance, σ^2_{γ} is the noise variance and $BIAS\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right) = \mathbb{E}\left(\varphi\left(\boldsymbol{\kappa}_{\mathcal{T}'}\right) - \hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right)$.

The LC-LSTM - estimating $\sigma^2_{\hat{k}_t}$

- To derive $\sigma^2_{\hat{k}_t}$ the conditioned time-index distribution, $\mathbb{P}\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right)$ should be known, but it is not available.

- We could extract it from the data referring to the ensemble technique

- Using bootstrap techniques, multiple training data samples are generated to develop an empirical distribution, $\hat{\mathbb{P}}\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right)$

- The final estimates are then obtained aggregating, by average, the various outputs: bootstrap aggregating or bagging

- In bagging procedures holds that $\mathbb{E}\left[\mathit{BIAS}\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right)^2\right] \to 0$

The LC-LSTM - estimating $\sigma^2_{\hat{k}_t}$

The bagging scheme:

Step 1. Over the series $\boldsymbol{\kappa}_{\mathcal{T}}$, we train the LSTM model obtaining point estimates over $\mathcal{T}'$

Step 2. Generate $B \in \mathbb{N}$ samples of $\boldsymbol{\kappa}_{\mathcal{T}}$ via a proper bootstrap procedure

Step 3. For each sample re-optimize the weights of NN defined in *Step 1*

Step 4. For each NN in *Step 3*, predict the associate point estimate on $\mathcal{T}'$, producing a bootstrap distribution consisting of B point predictions:

$$\hat{\mathbb{P}}\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right) = \left(\hat{k}_t^{(b)} = \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, \hat{\boldsymbol{\mathcal{W}}}^{(b)}\right), b = 1, \ldots, B\right) \tag{8}$$

The LC-LSTM - estimating $\sigma^2_{\hat{k}_t}$

Step 5. From $\hat{\mathbb{P}}\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right)$ calculates estimates of interest by aggregation. Hence, the bagged estimate of the variance $\sigma^2_{\hat{k}_t}$ is:

$$\hat{\sigma}^2_{\hat{k}_t} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, \hat{\boldsymbol{\mathcal{W}}}^{(b)}\right) - \overline{k}_t \right), \qquad (9)$$

where

$$\overline{k}_t = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, \hat{\boldsymbol{\mathcal{W}}}^{(b)}\right)$$

is the bagged estimate for the conditional expectation $\mathbb{E}\left(\hat{k}_t \middle| \boldsymbol{\kappa}_{\mathcal{T}'}\right)$.

The LC-LSTM - estimating σ_γ^2

- Mortality dynamic incorporates an intrinsic randomness not explained by the network: the noise γ
- A NN appropriately trained catches the key input-output data schemes, skimming noisy examples (avoid overfitting)
- Let $\mathcal{T}$ be the training set interval, we deal with the series

$$(\gamma_t)_{t\in\mathcal{T}} = \left(k_t - \hat{k}_t\right)_{t\in\mathcal{T}}$$

  as a proxy of the unwrapped noise by NN
- It helps to evaluate the estimates $\hat{\sigma}_\gamma^2$ as the time-index residual uncertainty over $\mathcal{T}$, spreading the random error over the forecast horizon $\mathcal{T}'$ through a random walk representation

Agenda

- Introduction
- Improving the Lee-Carter mortality density forecast
- Numerical application
- Conclusions

Numerical application

- Countries: Australia, Spain and Japan. Data from HMD, both genders
- Ages: $\mathcal{X} = \{0, \ldots, 99\}$
- Calendar Years: Two periods to check model robustness $\mathcal{T} = \{1950, \ldots, 2018\}$ and $\mathcal{T} = \{1960, \ldots, 2018\}$
- Lag $J = 1$, so that $\boldsymbol{\kappa}_{\mathcal{T}} = (k_{t-1})_{t \in \mathcal{T}}$ and $k_t = f_{LSTM}\left(k_{t-1}; \boldsymbol{\mathcal{W}}\right) + \gamma_t$
- For the bagging scheme: bootstrap from Koissi et al.(2006), with $B = 1000$
- Benchmark: LC Poisson model (Brounhs et al.(2002)), selecting the best ARIMA(p,d,q) model.

Numerical application - Learning process

- Setting $T = 2000$ as forecasting year, the series $\boldsymbol{\kappa}_{\mathcal{T}}$ is splitted in:

$$\begin{aligned} \text{TRAINING SET:} \quad \mathcal{TR} &= (k_t | k_{t-1})_{t=t_0,\dots,T} \\ \text{TESTING SET:} \quad \mathcal{TS} &= (k_t | k_{t-1})_{t=T+1,\dots,t_n}, \end{aligned} \tag{10}$$

where $t_0 = \{1950, 1960\}$.

- To validate the model we divide the $\mathcal{TR}$ set into a sub-training set and in a validation set, considering the splitting rule $80\% - 20\%$:

$$\begin{aligned} \text{SUB-TRAINING SET:} \quad \mathcal{TR}^{\text{sub}} &= (k_t | k_{t-1})_{t=t_0,\dots,T^{\text{sub}}} \\ \text{VALIDATION SET:} \quad \mathcal{VS} &= (k_t | k_{t-1})_{t=T^{\text{sub}}+1,\dots,T} \end{aligned} \tag{11}$$

- Tuning by grid search

Numerical application - Performance metrics

- Accuracy metrics for point predictions, with $\hat{y} = \{\hat{k}; \ln \hat{m}\}$

$$RMSE_{(y)} = \sqrt{\frac{\sum_{t=t_n+1}^{t_n+s} (y_t - \hat{y}_t)^2}{s-1}}$$

- Quality metrics for prediction interval, with $\hat{y} = \{\hat{k}; \ln \hat{m}\}$

Prediction Interval Coverage Probability

$$PICP_{(y)} = \frac{1}{s-1} \sum_{t=t_n+1}^{t_n+s} \mathbf{1}_{\left\{\hat{y}_t \,\in\, [\hat{y}_t^L, \hat{y}_t^U]\right\}},$$

Mean Prediction Interval Width

$$MPIW_{(y)} = \frac{1}{s-1} \sum_{t=t_n+1}^{t_n+s} \hat{y}_t^U - \hat{y}_t^L.$$

with $\mathbf{1}_{\{\cdot\}} = 1$ if $\hat{y} \in [y^L, y^U]$, and $\mathbf{1}_{\{\cdot\}} = 0$ otherwise.

- k_t performance metrics for each training period. Forecasting years: 2001-2018.

Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		RMSE	$PICP_{(k)}$	$MPIW_{(k)}$	*RMSE*	$PICP_{(k)}$	$MPIW_{(k)}$	*RMSE*	$PICP_{(k)}$	$MPIW_{(k)}$	*RMSE*	$PICP_{(k)}$	$MPIW_{(k)}$
Australia	ARIMA	9.514	1	**53.503**	3.861	1	25.195	5.138	1	**47.485**	3.637	1	25.089
	LSTM	**4.280**	1	32.865	**3.790**	1	**39.478**	**1.970**	1	28.143	**2.659**	1	**37.433**
Japan	ARIMA	3.743	1	21.503	**10.084**	0.556	20.767	4.647	1	17.392	9.790	0.500	12.409
	LSTM	**2.228**	1	**43.784**	18.014	**1**	**53.431**	**2.069**	1	**28.209**	**5.818**	**1**	**30.701**
Spain	ARIMA	14.038	0.333	19.354	**6.215**	1	21.394	13.071	0.333	17.343	5.805	1	20.747
	LSTM	**8.625**	**1**	**35.424**	7.471	1	**60.373**	**9.983**	**0.778**	**23.340**	**4.357**	1	**28.141**

- In $m_{x,t}$ performance metrics for each training period. Forecasting years: 2001-2018.

$x = 45$

Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$
Australia	LC	0.227	**1**	**0.534**	**0.091**	0.944	0.267	0.175	**1**	**0.478**	**0.084**	0.944	0.265
	LC-LSTM	**0.110**	0.944	0.295	0.142	0.944	**0.407**	**0.116**	0.944	0.280	0.097	**1**	**0.394**
Japan	LC	0.071	0.667	**0.180**	0.255	0	0.173	**0.063**	0.722	0.150	0.155	0.056	0.105
	LC-LSTM	**0.062**	**0.722**	0.143	**0.077**	**0.444**	**0.254**	0.073	**0.944**	**0.243**	**0.061**	**0.667**	**0.115**
Spain	LC	0.200	0.333	0.153	**0.104**	0.611	0.179	0.228	**0.333**	0.136	**0.067**	0.722	0.174
	LC-LSTM	**0.161**	**0.556**	**0.276**	0.502	**0.944**	**0.489**	**0.205**	0.278	**0.215**	0.073	**0.944**	**0.259**

$x = 65$

Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$
Australia	LC	0.157	1	**0.672**	0.061	0.944	0.283	0.106	1	**0.623**	0.058	1	0.293
	LC-LSTM	**0.056**	1	0.371	0.061	**1**	**0.431**	**0.043**	1	0.365	**0.052**	1	**0.436**
Japan	LC	0.054	**1**	**0.177**	0.160	0.444	0.178	0.063	0.833	0.161	0.151	0.333	0.128
	LC-LSTM	**0.035**	0.944	0.141	**0.077**	**1**	**0.262**	**0.029**	**1**	**0.261**	**0.028**	**1**	**0.141**
Spain	LC	0.097	0.278	0.157	0.079	0.778	0.206	0.106	0.222	0.158	0.073	0.889	0.229
	LC-LSTM	**0.060**	**1**	**0.285**	**0.66**	**1**	**0.568**	**0.080**	**0.889**	**0.249**	**0.068**	**0.944**	**0.340**

$x = 85$

Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$
Australia	LC	**0.053**	0.944	**0.344**	**0.032**	1	0.191	**0.039**	0.944	**0.319**	0.033	1	0.194
	LC-LSTM	0.056	0.944	0.190	0.033	1	**0.292**	0.049	0.944	0.187	**0.026**	1	**0.289**
Japan	LC	**0.030**	**0.889**	**0.134**	**0.050**	**0.778**	0.142	0.040	0.944	0.133	**0.071**	0.444	0.115
	LC-LSTM	0.034	0.778	0.107	0.171	0.500	**0.209**	**0.029**	0.944	**0.215**	0.080	0.444	**0.126**
Spain	LC	0.082	0.333	0.113	**0.059**	0.611	0.122	0.086	0.278	0.116	0.057	0.833	0.150
	LC-LSTM	**0.052**	**1**	**0.204**	0.447	**1**	**0.335**	**0.066**	**0.944**	**0.183**	**0.048**	**1**	**0.223**

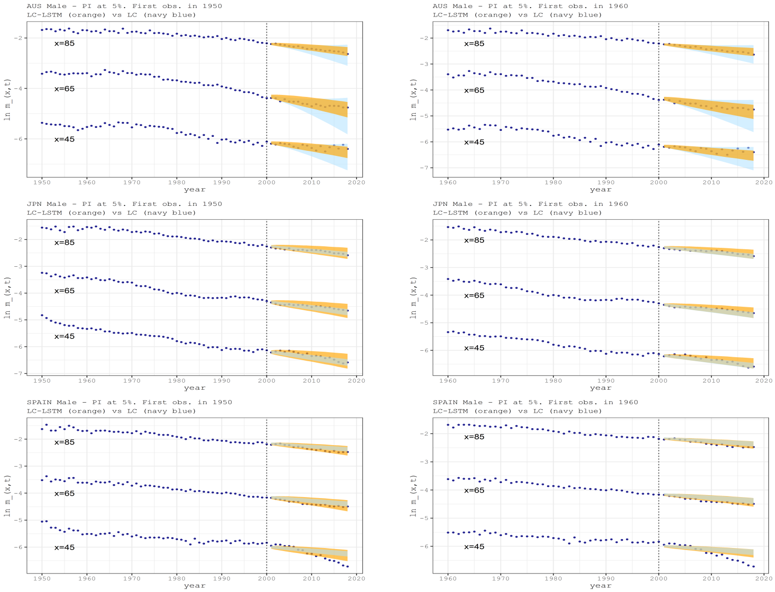

Figure 1: MALE PI ($\alpha = 5\%$). Forecasting period: 2001-2018. Training period: 1950-2000 (left), 1960-2000 (right).

Numerical application - Long Term Forecasts

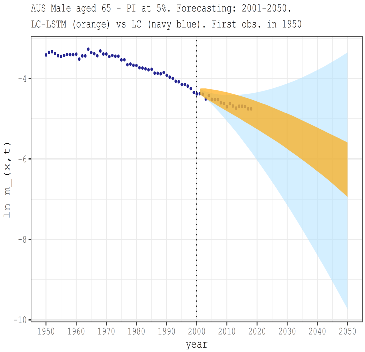

Figure 2: AUSTRALIA MALE ($\alpha = 5\%$). Age 65. Forecasting period: 2001-2050. Training period: 1950-2000 (left), 1960-2000 (right).

Agenda

- Introduction
- Improving the Lee-Carter mortality density forecast
- Numerical application
- Conclusions

Conclusions

- The LC-LSTM seems to be an effective improvement of the canonical LC model predictive ability, in terms of both point and interval predictions

- The proposed model reflects important features, also in the long-run, as:
  - Robustness w.r.t. to the training period
  - Biologically consistency
  - Plausibility in uncertainty levels

- *The LC-LSTM model poses a compromise between the interpretation of the mortality phenomenon and high precision in anticipating its future realizations*

THANKS FOR YOUR ATTENTION!