Time-Series Forecasting of Mortality Rates using Deep Learning

Ronald Richman FIA FASSA CERA

ronald.richman@qedact.com

Associate Director - QED Actuaries & Consultants

9 September 2020

Paper

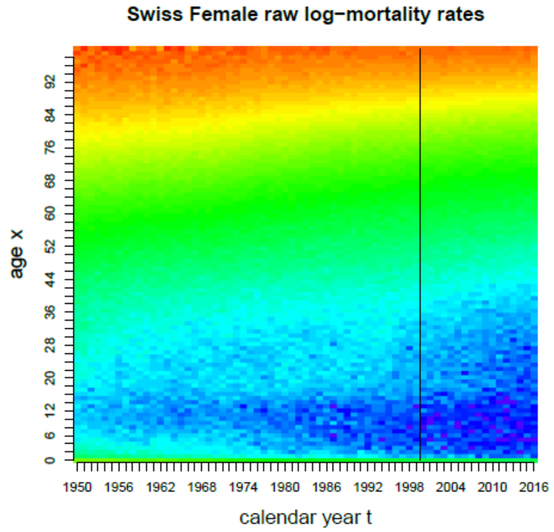
Perla, Francesca and Richman, Ronald and Scognamiglio, Salvatore and Wüthrich, Mario V., Time-Series Forecasting of Mortality Rates using Deep Learning (May 6, 2020). Available at SSRN: https://ssrn.com/abstract=3595426 or http://dx.doi.org/10.2139/ssrn.3595426

Agenda

- **Rationale**
- **Lee-Carter Paradigm for Mortality Modelling**
- **Deep Learning for Mortality Modelling**
- **Time Series Forecasting with Deep Learning**

Rationale

- **High quality mortality data available for multiple countries in the HMD**
- **May require forecasts of mortality for multiple countries simultaneously for modelling**
  - Insurers operating in multiple jurisdictions need to set assumptions for many populations
- **Reasonable to expect that improved forecasts for single countries could be made if all data were utilized to build a model**
  - May be of interest even if mortality is only being forecast for a single country



Agenda

- **Rationale**
- **Lee-Carter Paradigm for Mortality Modelling**
- **Deep Learning for Mortality Modelling**
- **Time Series Forecasting with Deep Learning**

Lee-Carter Model

- **Forecast mortality rates = key inputs into demographic forecasting, life insurance and pensions models**
- **Foundational model for mortality forecasting is the Lee-Carter model (Lee and Carter 1992) (LC model)**
  Many other approaches; within actuarial literature see Cairns, Blake and Dowd (2006) for an approach (CBD model) suited to old-age mortality (model coefficients of logistic model of qx)
- **Mortality over time modeled using:**

$$\log\left(u_{x,t}\right) = a_x + b_x k_t$$

- **i.e. (log) mortality = *average rate + rate of change . time index***
- **Relies on latent variables that must be estimated from data and then multiplied**
  Could use interaction term between the variables Year and Age but this specification would require t.x effects to be fit compared to the t+x effects in the Lee-Carter model.
- **=> use non-linear/PCA regression to estimate the latent terms (Brouhns, Denuit and Vermunt 2002; Currie 2016; Lee and Carter 1992)**

Producing Forecasts

- **Time index k_t estimated for years within sample => need to extrapolate k_t for out-of-sample forecasts**
- **Time series models of varying complexity used to forecast k_t**
- **Two-step process – fit model (a_x , b_x , k_t) and extrapolate - common to other mortality models, such as CBD model**
- **Key judgement in LC model: over what period should the LC model be calibrated so that a_x & b_x appropriate for forecasting period?**
- **<u>Problem 1:</u> If many forecasts are required, e.g. for multiple populations, then a manual process of selecting calibration periods is required if using the LC model**
- **Single population extensions**
  Cohort effect (Renshaw and Haberman 2006)
  Smoothing time series (Currie 2013)

Extending the LC Model

- **What about multiple populations?**

- **Intuition = multi-population mortality forecasting model should produce more robust forecasts**
  Common factors (similar socioeconomic circumstances, shared improvements in public health and medical technology)
  Common trends likely captured with more statistical credibility

- **=> Li and Lee (2005) recommend even if interest is in single series**

Augmented Common Factor (Li and Lee 2005)	$\log(u_{x,t}) = a_x^i + b_x k_t + b_x^i k_t^i$
Common Age Effect (Kleinow 2015)	$\log(u_{x,t}) = a_x^i + b_x k_t^i$

- **<u>Problem 2:</u> Not intended for large scale mortality forecasting - generally applied on smaller sub-set of data => judgment of modeler needed**
  Hard to fit (complex optimization schemes/less known statistical techniques)

- **Which specification is better, when, and why?**

Formalizing the LC Paradigm

- **Mortality modelling:**

$$(t, x, i) \mapsto \log\left(u_{x,t}^{(i)}\right)$$

- **LC paradigm for mortality forecasting for population *i:***

$$\log\left(u_{x,t}^{(i)}\right) = a_x^{(i)} + \left\langle \boldsymbol{b}_x^{(i)}, \boldsymbol{k}_t^{(i)} \right\rangle$$

$$a_x^{(i)} \in \mathbb{R},\ \boldsymbol{b}_x^{(i)}, \boldsymbol{k}_t^{(i)} \in \mathbb{R}^q$$

- **Different models produced by choice of *q***
- ***q = 1 and parameters estimated for each population => LC model***
- ***q > 1 => multi factor LC model***
- ***estimate $b_x^{(i)} = b_x$ i.e. not depending on population I => CAE model***
- ***q=2 and estimate population-specific and overall time effects => ACF model***

LC and Embeddings

- **LC model utilizes step functions:**

$$\log\left(u_{x,t}\right) = g(x) + h(x)i(t),$$

$$g(x) = \begin{cases} a_1 & \text{for } x = 1, \\ a_2 & \text{for } x = 2, \\ \vdots & \\ a_\omega & \text{for } x = \omega, \end{cases}$$

- **For a set of categories P with n_p levels, an embedding of dimension q_P is:**

$$z_{\mathcal{P}} : \mathcal{P} \to \mathbb{R}^{q_{\mathcal{P}}}, \qquad p \mapsto z_{\mathcal{P}}(p).$$

- **LC model can be interpreted using embeddings (see Richman and Wüthrich (2019)):**

$$(x, i) \;\mapsto\; a_x^{(i)} \;\in\; \mathbb{R}.$$

Agenda

- **Rationale**
- **Lee-Carter Paradigm for Mortality Modelling**
- **Deep Learning for Mortality Modelling**
- **Time Series Forecasting with Deep Learning**

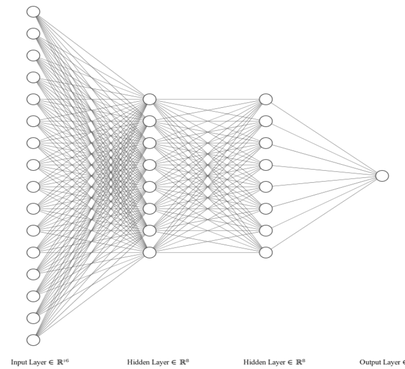
Deep Learning

- **Deep Learning = representation learning technique that automatically constructs hierarchies of complex features to represent abstract concepts**
  Features in lower layers composed of simpler features constructed at higher layers => complex concepts can be represented automatically

- **Typical example of deep learning is feed-forward neural networks, which are multi-layered machine learning models, where each layer learns a new representation of the features.**

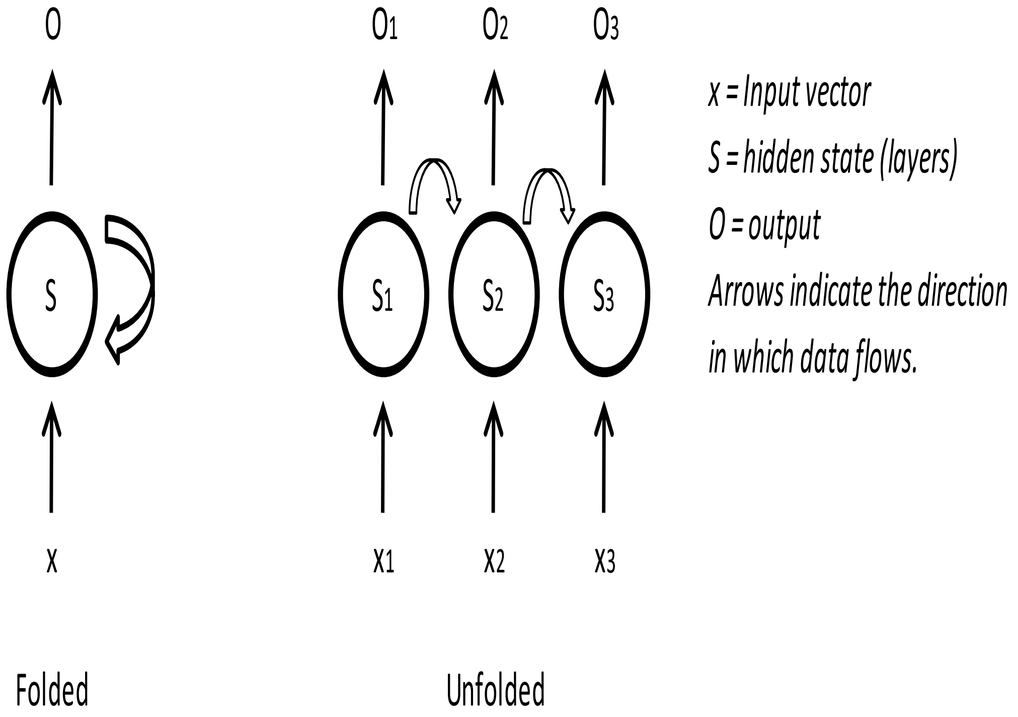
- **The principle: Provide raw data to the network and let it figure out what and how to learn.**

- **Desiderata for AI by Bengio (2009): “Ability to learn with little human input the low-level, intermediate, and high-level abstractions that would be useful to represent the kind of complex functions needed for AI tasks.”**



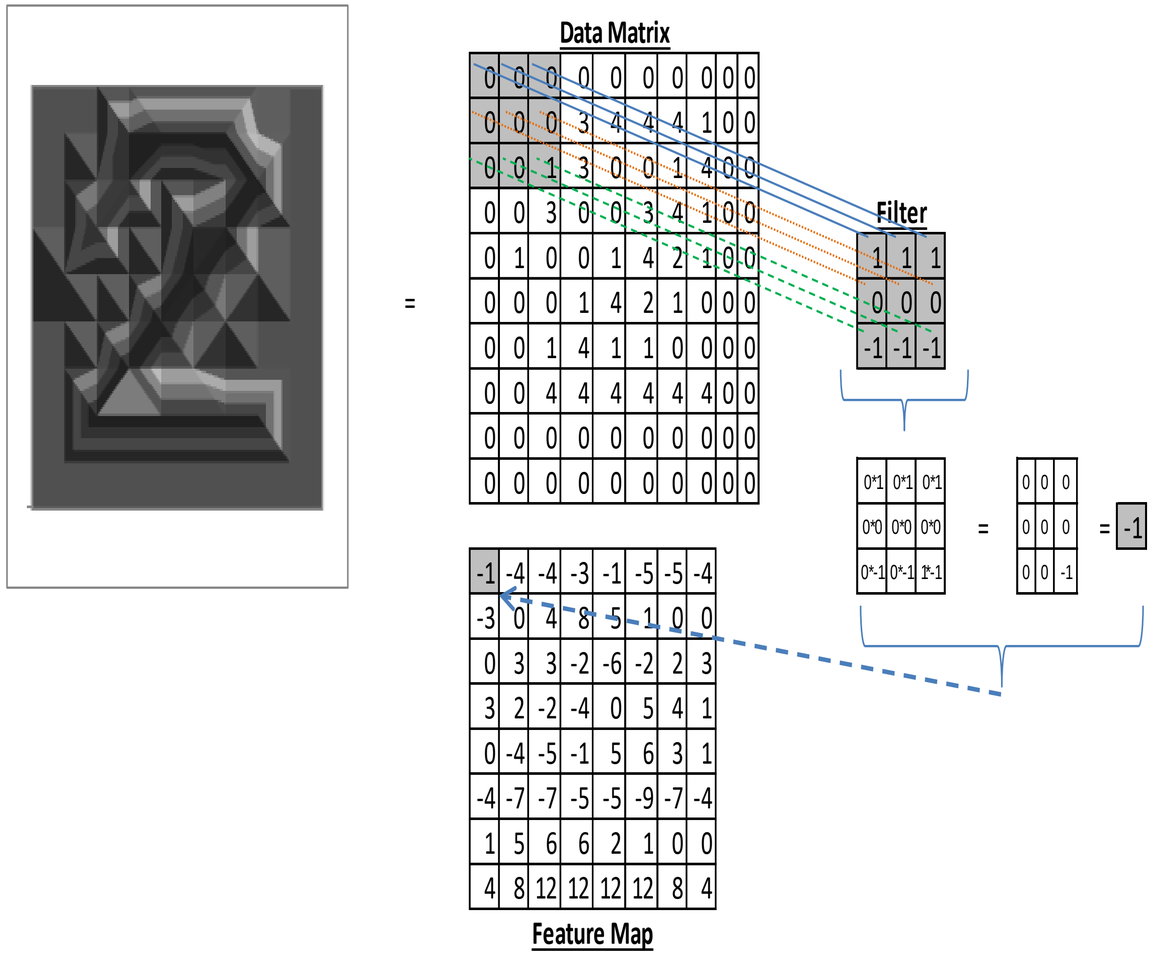
Recurrent NN – Temporal data

- **Data with temporal structure implies that previous observations should influence the current observation**
- **Recurrent network maintains state of hidden neurons over time**
  - Past representation useful for current prediction i.e. network has a ‘memory’
- **Key challenge – difficult to train due to vanishing gradients**
- **Several implementations of the recurrent concept which control how network remembers and forgets state**
  - Long Short Term Memory (LSTM)
  - Gated Recurrent Unit (GRU)
- **RNNs can make predictions at last time step or for each input**



Convolutional NN - Images

- **Prior - features in images are position invariant i.e. can recognize at any position within an image**
  - Also applies to audio/speech and text/time series data
- **Convolutional network is locally connected and shares weights => expresses prior of position invariance**
  - Far fewer parameters than FCN
- **Each neuron (i.e. feature map) in network derived by applying filter to input data**
  - Weights of filter learned when fitting network
  - Multiple filters can be applied
- **Can also be used for time-series applications**



Extending LC – two perspectives

- **Lee Carter model = regression model using features derived from data using PCA**
  CAE + ACF = LC-type regression models with features derived at a regional level

- **<u>Perspective 1:</u> Use a neural network to model the regression problem and let it decide on the feature set**
  LC is a linear model once parameters are known => use a NN to derive non-linear model
  CAE/ACF specify the types of interaction between population-level and regional level parameters => use a NN to derive more predictive specification

- **<u>Perspective 2:</u> use a more general step function formulation to specify the multi-population model**
  LC uses step functions of single dimension for each parameter:

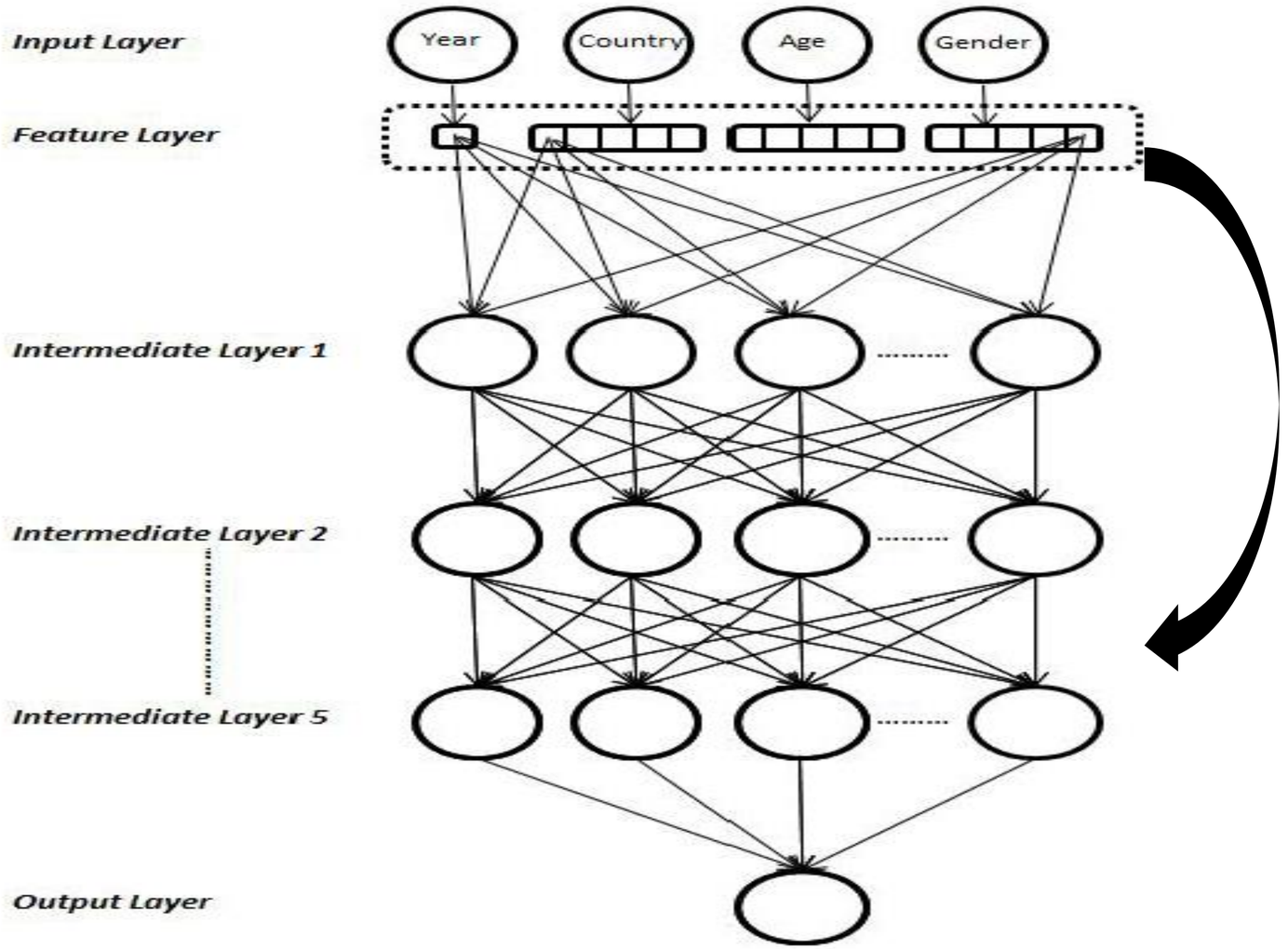
$$\log\left(u_{x,t}\right) = g(x) + h(x)i(t),$$

$$g(x) = \begin{cases} a_1 & \text{for } x = 1, \\ a_2 & \text{for } x = 2, \\ \vdots & \\ a_\omega & \text{for } x = \omega, \end{cases}$$

  Use NN to derive multi-dimensional vectors for each parameter using an embedding layer

Lee-Carter Neural Network

- **Multi-population mortality forecasting model (Richman and Wüthrich 2018)**
- **Supervised regression on HMD data (inputs = Year, Country, Age; outputs = m_x)**
- **5 layer deep FCN**
- **Generalizes the LC model in both ways mentioned before**
- **Note that no time-series forecasting is done**
  - Year enters the model as a numerical variable.
  - Forecasts made by predicting using values for Year beyond the range of input data



Multi-population results

- **Results of comparing the models (LC/ACF/CAE/LCNN)**
- **Best performing model is deep neural network…**
- **…produces the best out-of-time forecasts 51 out of 76 times**
- **For purposes of large scale mortality forecasting, deep neural networks dramatically outperform traditional single and multi-population forecasting models**

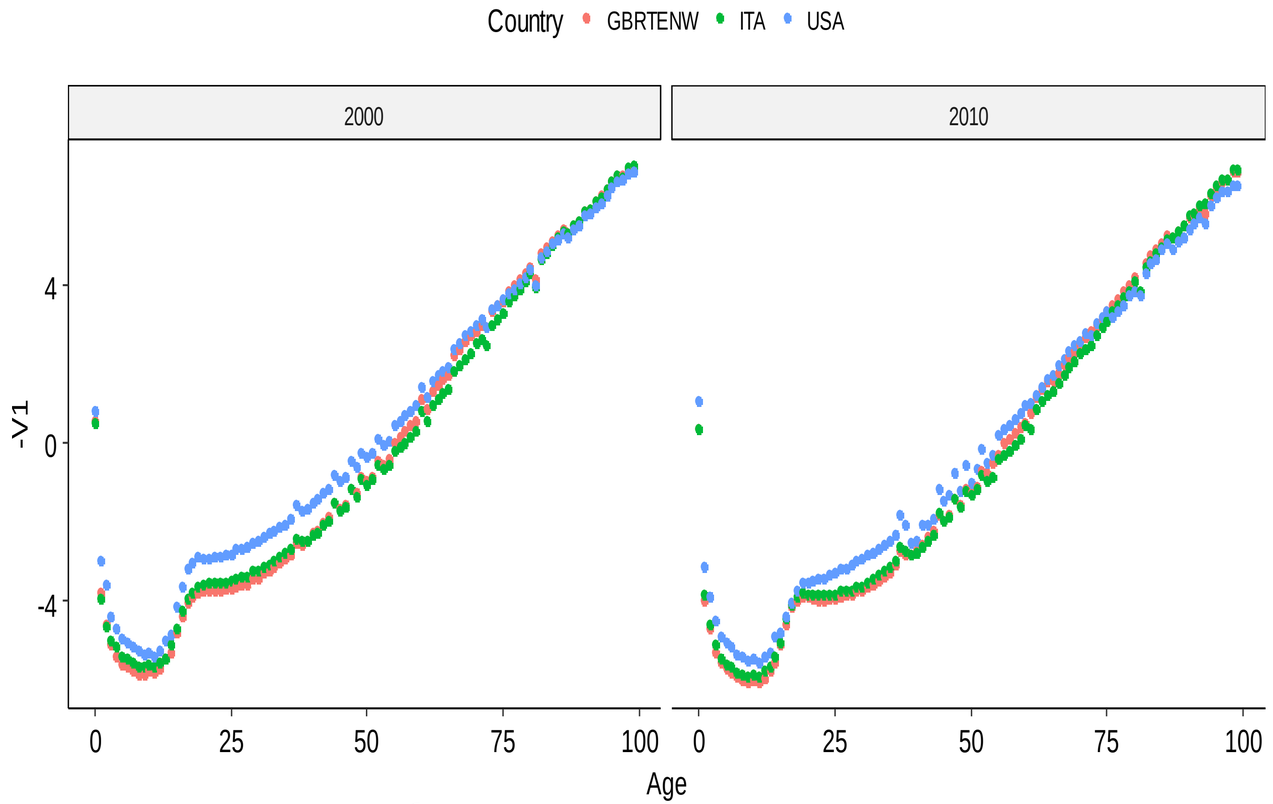
	Model	Average MSE	Median MSE	Best Performance
1	LC_SVD	5.50	2.48	19
2	ACF_SVD_region	3.46	2.50	36
3	ACF_SVD_country	7.30	4.77	9
4	ACF_BP	6.12	3.00	12

	Model	Average MSE	Median MSE	Best Performance
1	LC_SVD	5.50	2.48	33
2	CAE_SVD	4.76	2.35	13
3	CAE2_SVD	12.01	1.79	14
4	CAE2_BP	5.59	3.46	16

	Model	Average MSE	Median MSE	Best Performance
1	LC_SVD	5.50	2.48	7
2	LC_ACF_region	3.46	2.50	10
3	ACF_BP	6.12	3.00	4
4	CAE_BP	5.59	3.46	4
5	DEEP	2.68	1.38	51

Features in last layer of network

- **Representation = output of last layer (128 dimensions) with dimension reduced using PCA**
- **Can be interpreted as relativities of mortality rates estimated for each period**
- **Output shifted and scaled to produce final results**
- **Generalization of Brass Logit Transform where base table specified using NN (Brass 1964)**

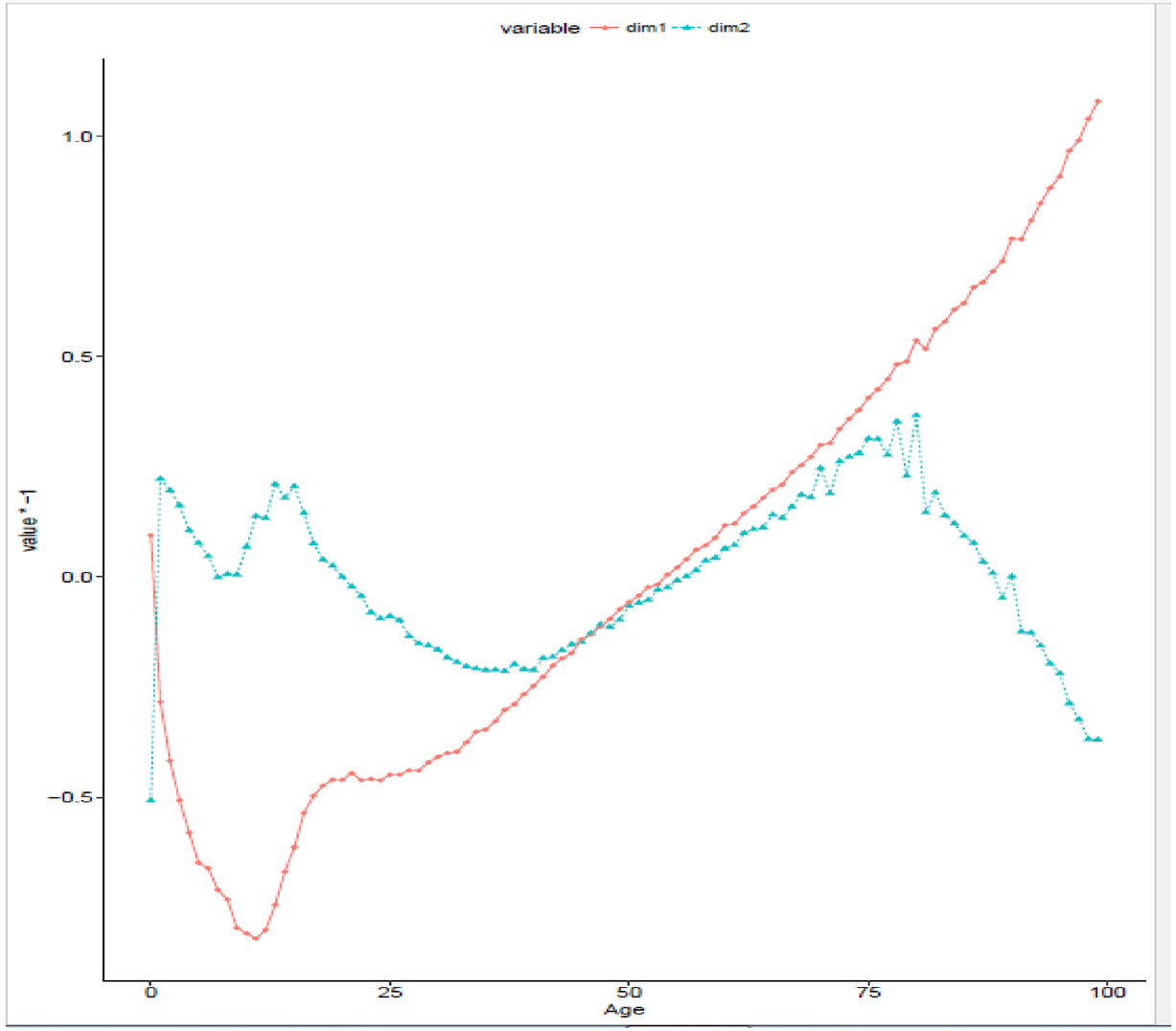


$y_x = a + b * z_x^{Ref}$, where:

y_x = *logit of mortality at age x*
a,b = regression coefficients
z_x^{Ref} = *logit of reference mortality*

Learned embeddings

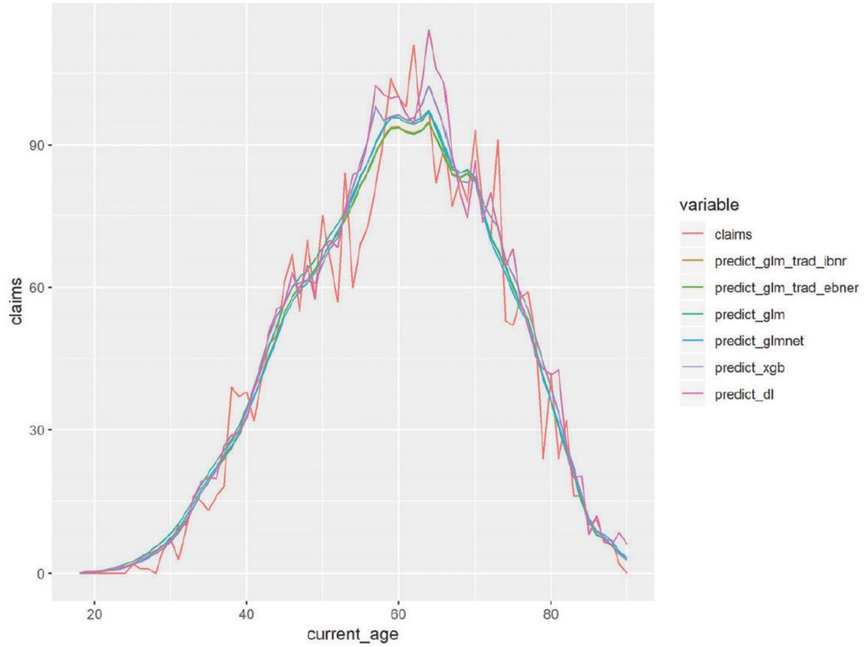
- **Age embeddings extracted from LCNN model**
- **Five dimensions reduced using PCA**
- **Age relativities of mortality rates**
- **In deeper layers of network, combined with other inputs to produce representations specific to:**
  - Country
  - Gender
  - Time
- **First dimension of PCA is shape of lifetable**
- **Second dimension is shape of child, young and older adult mortality relative to middle age and oldest age mortality**



Application to Insurance Data

- **Applied the same network structure to data from a reinsurer in Rossouw and Richman (2019)**
  - Consists of mortality and morbidity rates from 4 contributing companies over ~15 years
  - Trained on 5 years of data and forecast 4 years
- **Compared results of NN to several other models (LASSO/GBM) using Poisson deviance as criterion**
- **NN beat other models but had bias at portfolio level (see Wüthrich (2019))**
- **Debiased results:**
  - Poisson deviance: 22 836
  - AvE: 99.7%

Model	Poisson deviance	Actual vs. predicted
glm_trad_ibnr	22 944	99.3%
glm_trad_ebner	22 947	99.5%
glm	22 883	97.0%
glmnet	22 826	98.3%
xgb	22 822	95.9%
dl	22 799	93.8%



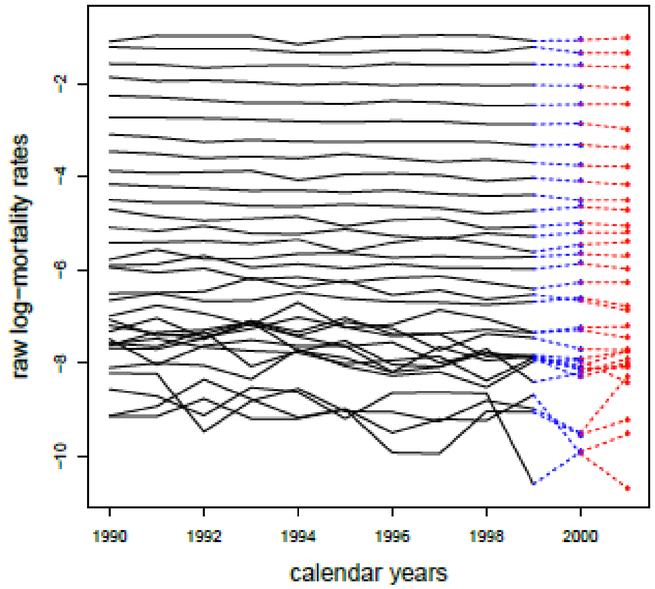
LC go Machine Learning: RNNs

- **Mortality rates have a time-series structure, e.g. Swiss Female mortality in 1990-2001**

- **Can NNs exploit time-series structure to forecast mortality directly?**
  RNNs seem to be a natural choice due to sequential processing
  LCNN does not rely on time-series structure directly

- **Swiss mortality rates forecast using RNNs in Richman and Wüthrich (2019)**
  Models trained on data 1950-1999
  Forecasts made for 2000-2016
  Models fit for each gender separately

- **To reduce volatility of input data, matrices of rates at ages *x*-*x+4* fed into networks to forecast rates at age *x+2***



	in-sample		out-of-sample	
	female	male	female	male
LSTM3 ($T = 10$, $(\tau_0, \tau_1, \tau_2, \tau_3) = (5, 20, 15, 10)$)	2.5222	6.9458	0.3566	1.3507
GRU3 ($T = 10$, $(\tau_0, \tau_1, \tau_2, \tau_3) = (5, 20, 15, 10)$)	2.8370	7.0907	0.4788	1.2435
LC model with SVD	3.7573	8.8110	0.6045	1.8152

Extending the RNN model

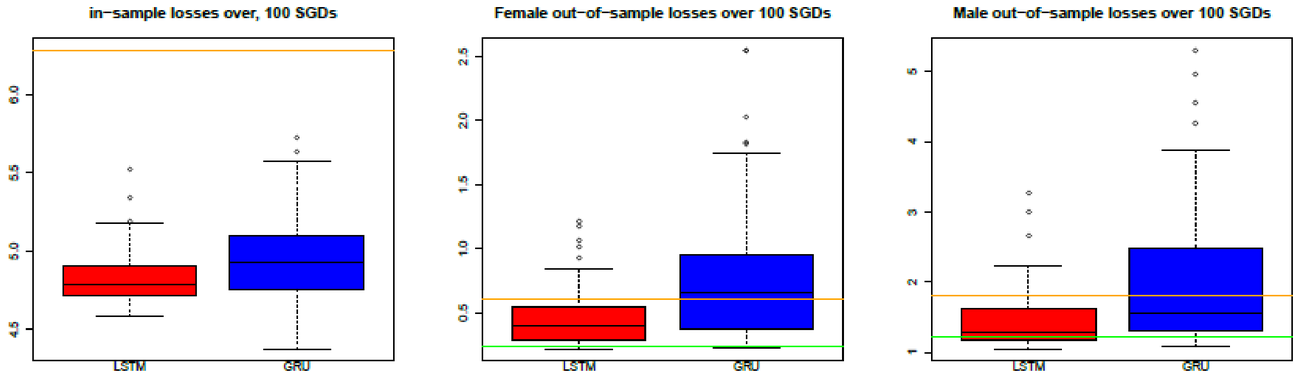
- **NN models flexible enough to incorporate extensions easily, e.g., joint modelling of both genders**
  Gender incorporated explicitly using dummy-coding and implicitly through rates input to the network

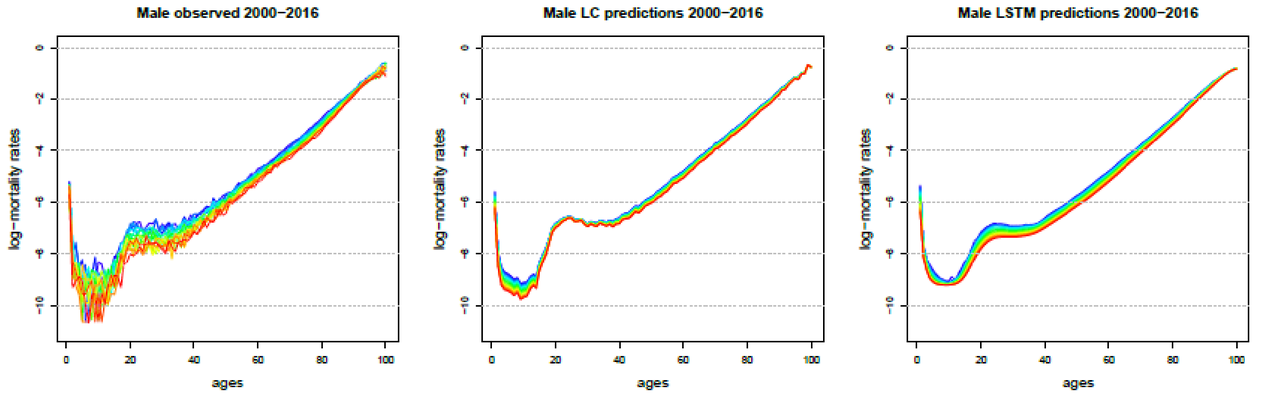
- **Results improved versus single gender models**
  LSTM model beats the GRU model.

- **Direct forecasting using RNNs leads to unstable results, which are much improved via model averaging (ensembling)**

- **Ensemble model captures improvements, particularly in young adult mortality significantly better than LC.**

	in-sample	out-of-sample	
	both genders	female	male
LSTM3 ($T = 10$, $(\tau_0, \tau_1, \tau_2, \tau_3) = (5, 20, 15, 10)$)	4.7643	0.3402	1.1346
GRU3 ($T = 10$, $(\tau_0, \tau_1, \tau_2, \tau_3) = (5, 20, 15, 10)$)	4.6311	0.4646	1.2571
LC model with SVD	6.2841	0.6045	1.8152

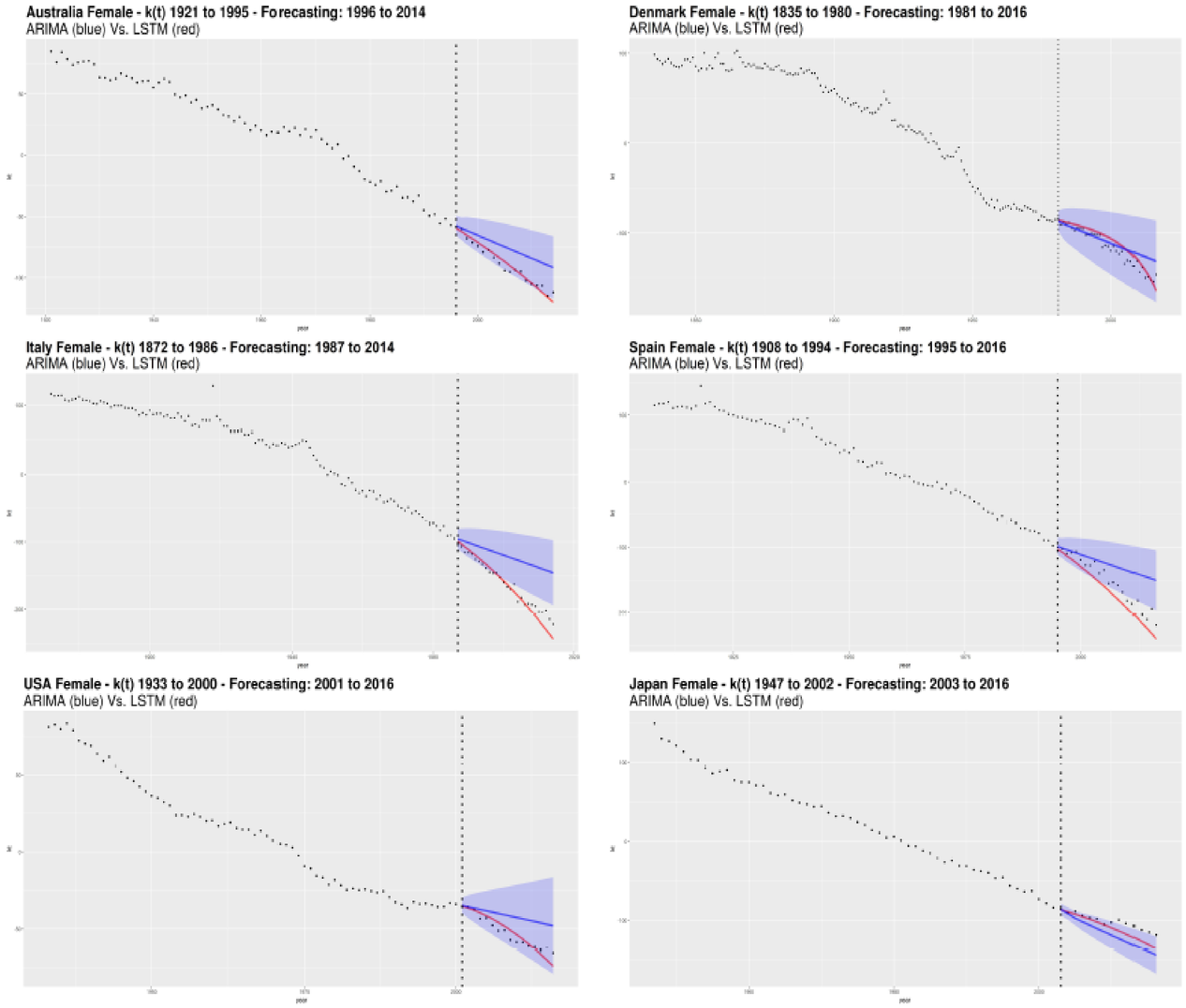




Combining LC + RNNs

- **Deep Learning Integrated Lee–Carter Model of Nigri, Levantesi, Marino, Scognamiglio & Perla, F. (2019)**
- **Apply SVD to derive a_x, b_x, k_t of LC model**
- **Instead of ARIMA, forecast k_t using RNN**

Country	Male		Female	
Australia	*MAE*	*RMSE*	*MAE*	*RMSE*
κ_t ARIMA	24.75	28.04	13.95	15.55
κ_t LSTM	2.57	3.24	3.12	3.83
Denmark	*MAE*	*RMSE*	*MAE*	*RMSE*
κ_t ARIMA	11.10	16.10	7.99	10.70
κ_t LSTM	2.97	3.84	6.62	8.18
Italy	*MAE*	*RMSE*	*MAE*	*RMSE*
κ_t ARIMA	55.41	63.46	41.12	45.74
κ_t LSTM	4.63	5.48	8.04	10.69
Spain	*MAE*	*RMSE*	*MAE*	*RMSE*
κ_t ARIMA	20.24	26.33	26.10	33.95
κ_t LSTM	7.69	8.96	15.61	17.67
the USA	*MAE*	*RMSE*	*MAE*	*RMSE*
κ_t ARIMA	8.39	9.48	10.81	12.54
κ_t LSTM	2.31	2.86	3.32	4.18
Japan	*MAE*	*RMSE*	*MAE*	*RMSE*
κ_t ARIMA	9.61	10.50	15.12	17.52
κ_t LSTM	4.71	5.24	7.89	10.23



Agenda

- **Rationale**
- **Lee-Carter Paradigm for Mortality Modelling**
- **Deep Learning for Mortality Modelling**
- **Time Series Forecasting with Deep Learning**

Processing Time Series with DL

- **RNN models still specialized for single population - can this be expanded to the multi-population case?**
- **Furthermore, can the excessive volatility of the RNN calibrations be reduced?**
- **In the LCNN model, we have the following regression function learned by the network:**

$$(t, x, i) \;\mapsto\; \log\left(u_{x,t}^{(i)}\right)$$

- **Hypothesis: can neural network models designed for directly processing sequential data outperform more general network architectures applied to sequential data?**
- **i.e. we wish to map directly from observed mortality rates of many populations to a time feature:**

$$U^{(i)} \;\mapsto\; \boldsymbol{k}_t^{(i)} \;\in\; \mathbb{R}^q$$

- **Addressed in Perla, Richman, Scognamiglio and Wüthrich (2020)**

Defining the Model

- **For modelling, work with scaled version of mortality rates:**

$$\boldsymbol{y}_{t_0+T+1}^{(i)} = \frac{\log(\boldsymbol{u}_{t_0+T+1}^{(i)}) - y_0}{y_1 - y_0} \in [0,1]^d.$$

where y_1 is the maximum of the observed log rates, and y_0 is the minimum.

- **For region r and gender g, can extend LC paradigm to:**

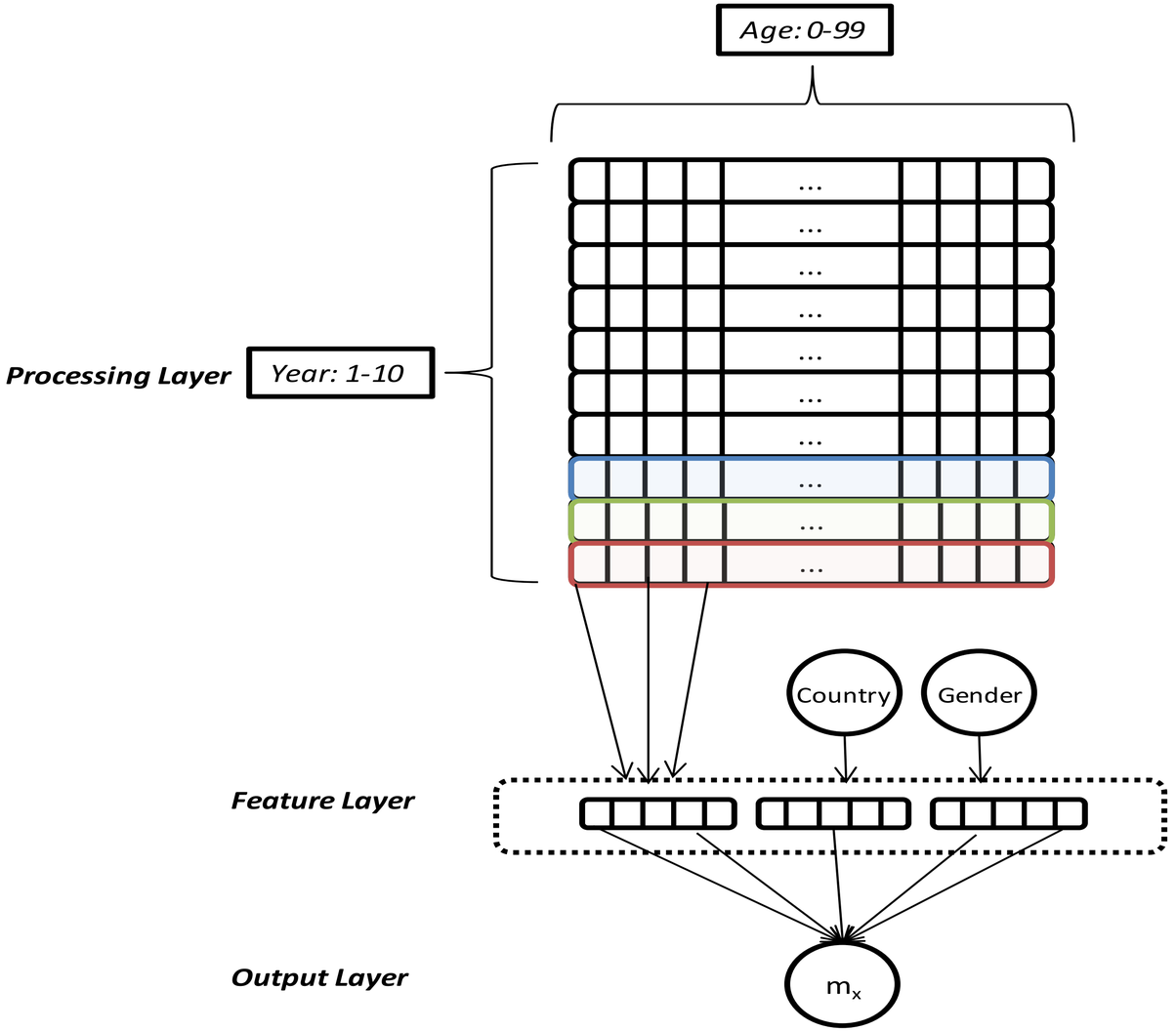
$$\sigma^{-1}\left(\widehat{y}_{x,t_0+T+1}^{(r,g)}\right) = \underbrace{w_{x,0} + \left\langle W_x^{\mathcal{R}}, z_{\mathcal{R}}(r)\right\rangle + \left\langle W_x^{\mathcal{G}}, z_{\mathcal{G}}(g)\right\rangle}_{a_x} + \underbrace{\left\langle W_x^{f}, \boldsymbol{z}_f(U_{t_0}^{(i)})\right\rangle}_{b_x.k_t}$$

where:

- $\boldsymbol{z_R(r)}$ **is a region embedding and** $\boldsymbol{W_x^R}$ **is a region coefficient**
- $\boldsymbol{z_G(g)}$ **is a gender embedding and** $\boldsymbol{W_x^G}$ **is a gender coefficient**
- $\boldsymbol{z_f(U_{t0}^i)}$ **is a representation learned using a neural network with input =** $\boldsymbol{U_{t0}^i}$ **(matrix of past mortality rates) and** $\boldsymbol{W_x^f}$ **is a trend age adjustment coefficient**

Model Structure

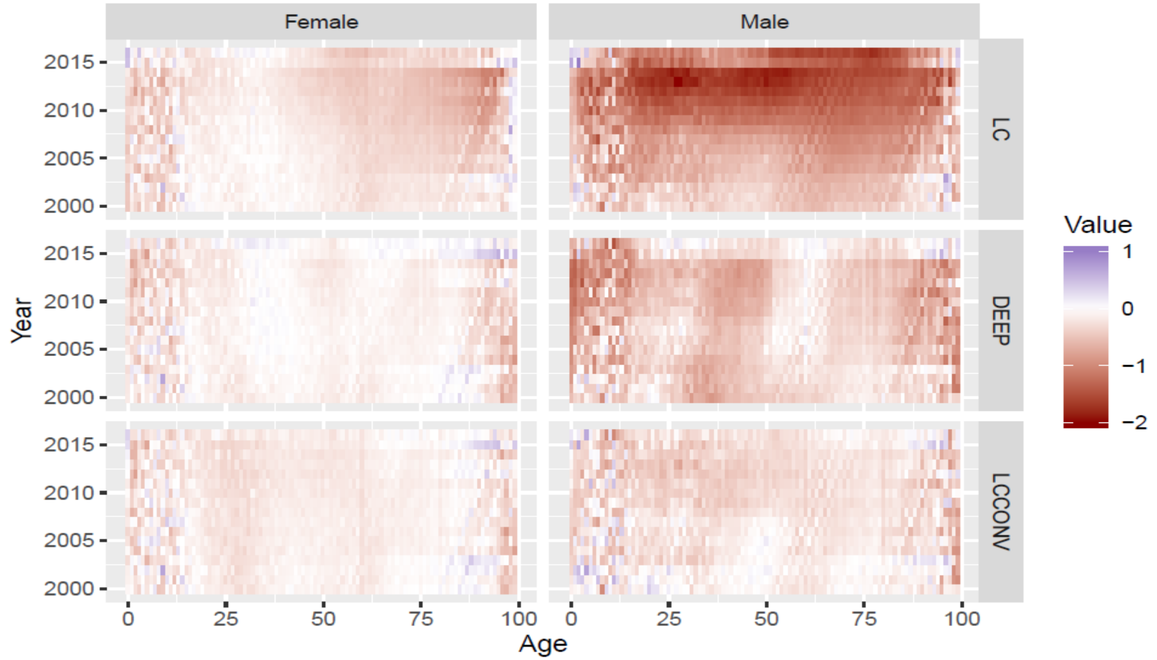
- **Similar to LCNN model…**
- **… however, time variable replaced with outputs of a NN processing layer**
- **Diagram shows a CNN being used to derive the mortality trend**
- **Can also use an RNN**
- **CNNs appear to perform better than RNNs**
- **LCCONV model forecasts mortality rates in a single step…**
- **… no time series forecasting required.**
- **Forecasts derived by adding single year forecasts to matrix of rates, and reapplying the model.**



Forecast Results – HMD (1)

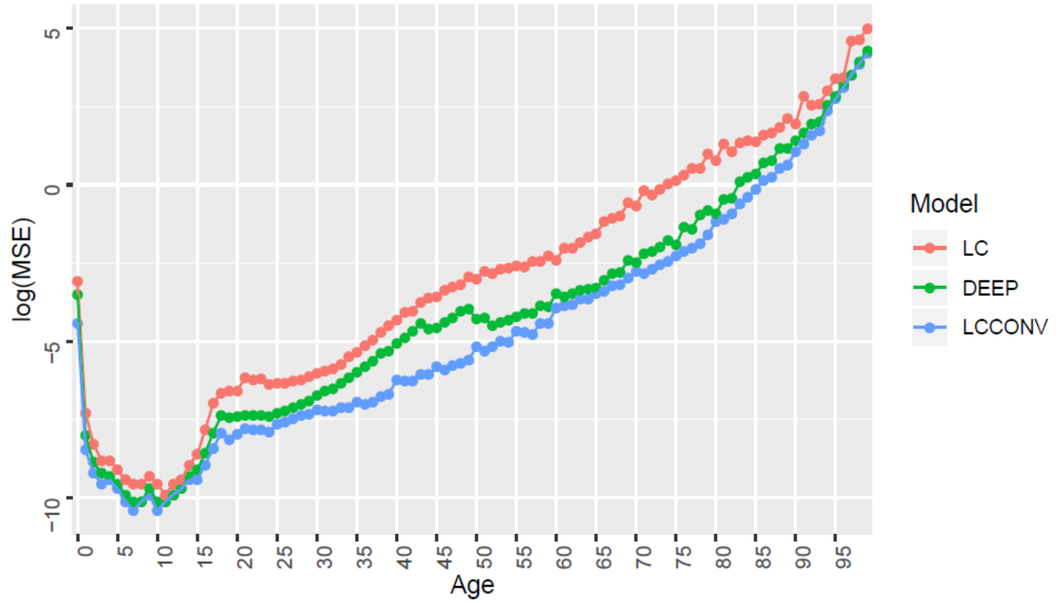
- **The CNN model (LCCONV) achieves better performance versus the LC model on 75/76 populations in the HMD**
- **LCCONV beats the LCNN model in an extra 8 populations and achieves a substantially lower out-of-sample MSE**
- **Residual plot shows that model is substantially better for males, whereas the performance is similar for females**
- **Using RNNs to process the data, instead of predicting also leads to good performance**

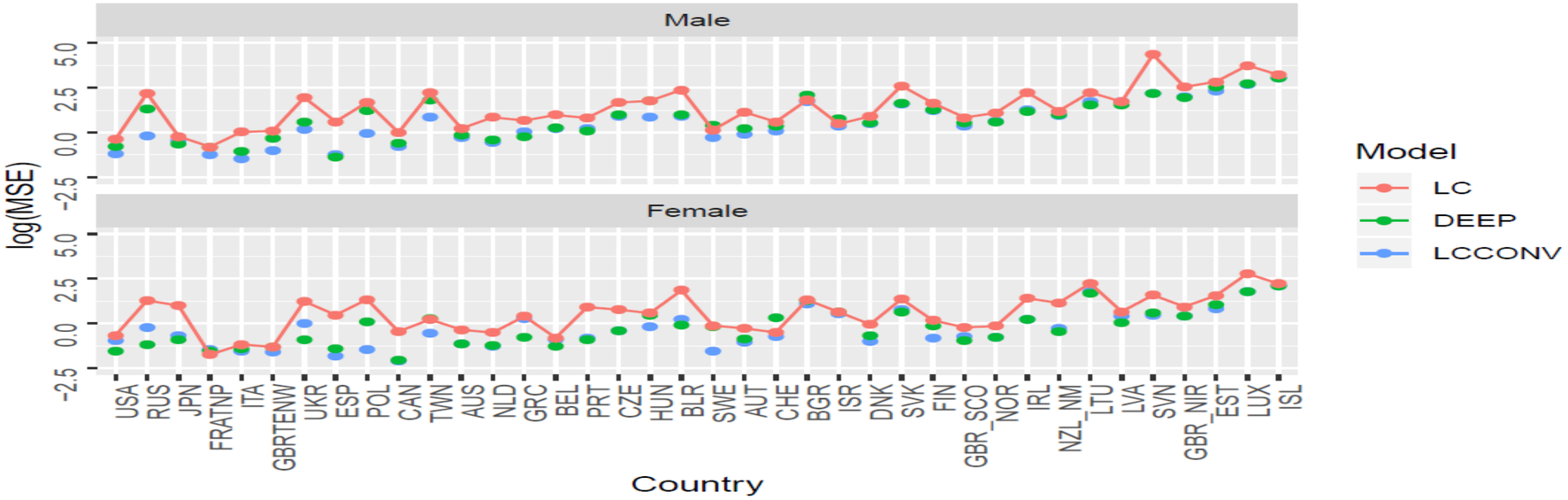
model	test_loss	ensemble MSE	# populations
LCCONV	2.27	2.24	75/76
LCCONV_tanh	2.62	2.58	61/76
LCCONV_relu	3.26	3.10	57/76
LCLSTM1	2.86	2.54	69/76
LCLSTM1_tanh	3.32	3.03	58/76
LCLSTM1_relu	3.33	3.25	52/76
LCLSTM2	2.43	2.32	74/76
LCLSTM2_tanh	2.36	2.27	75/76
LCLSTM2_relu	3.44	3.11	56/76
DEEP	2.83	2.53	67/76



Forecast Results – HMD (2)

- **Populations sorted by size – LC model error increases with size of population**
- **Similar results for NN models, however, lower than LC model in most instances**
- **LCCONV lower than LCNN mainly on large male populations**



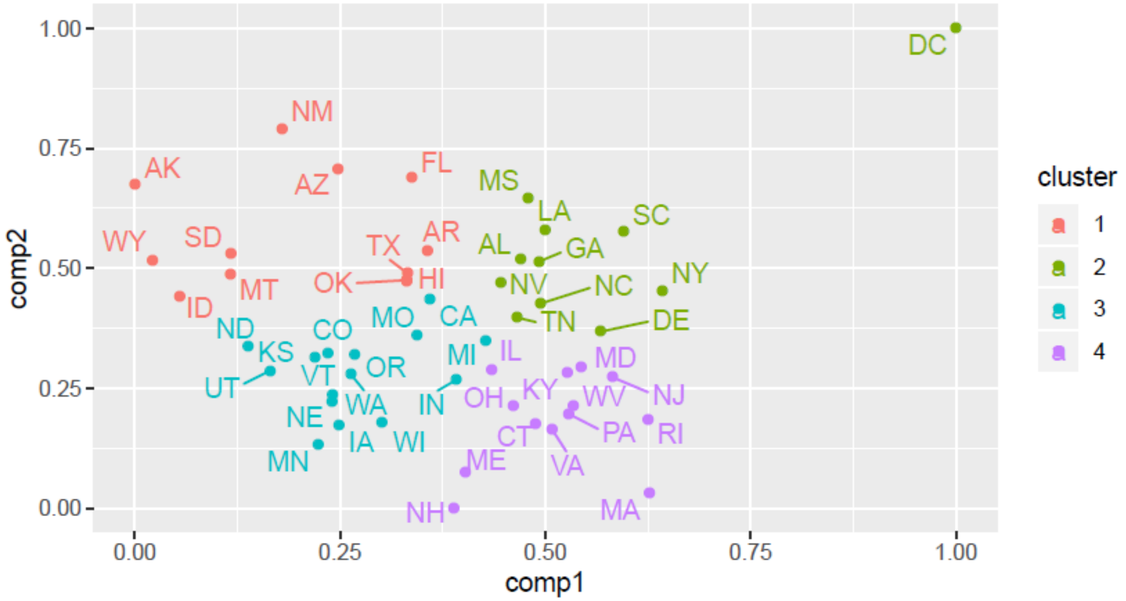


Forecast Results – USMD

- **Does the LCNN model generalize beyond the HMD?**
- **Fit model to the USMD (50 states + DC)**
- **Beats LC model 101/102 times**
- **RNN models more competitive on USMD than on HMD**

model	test loss	ensemble MSE	# populations
LCCONV	0.50	0.49	101/102
LCCONV_tanh	0.55	0.54	100/102
LCCONV_relu	0.74	0.73	86/102
LCLSTM1	0.63	0.58	98/102
LCLSTM1_tanh	0.66	0.53	99/102
LCLSTM1_relu	1.20	1.10	52/102
LCLSTM2	0.50	0.46	102/102
LCLSTM2_tanh	0.53	0.50	101/102
LCLSTM2_relu	1.09	0.96	62/102

- **Cluster 1 - includes states located in the south of US (Arizona, Florida, New Mexico, Texas) / US states geographically distant from the US zone such as Alaska and Hawaii.**
- **Cluster 2 - groups countries in the southeast zone (Louisiana, Georgia, South Carolina, Mississippi and Alabama)**
- **Cluster 3 – central (Colorado, Kansas, Utah and Missouri) and the northwest (Washington, Oregon, Montana and**
- **North Dakota)**
- **Cluster 4- The last cluster groups countries in the northeast of US (Pennsylvania, Connecticut, Massachusetts, Virginia and West Virginia).**



Conclusion

- **Deep learning models provide new opportunities for modelling mortality**
- **Potential gains of moving from traditional specification of models for mortality**
- **Some models can directly process mortality rates to produce forecasts with increased accuracy over more general models…**
- **…. however finding an optimal model architecture can be challenging**
- **Future research should address:**

  reasons for high variability of RNN models applied to forecast mortality rates

  uncertainty bounds on predictions

References

- **See** https://gist.github.com/RonRichman/655cca0dd79afcd20b33d3131c537414