



QED Actuaries & Consultants

LEE AND CARTER GO DEEP LEARNING

25 MAY 2020

Ronald Richman – Associate Director

QED Actuaries & Consultants (Pty) Ltd

T: +27 11 038 3735 | E: Ronald.Richman@qedact.com

GOALS OF THE TALK

- Discuss classical Lee-Carter model and multi-population extensions
- Journey from statistics to machine learning to deep learning
- Understand the problems solved by deep learning and the new options modellers have available
- Focus on mortality modelling with these new tools

AGENDA

- Lee-Carter Paradigm for Mortality Modelling
- From Machine Learning to Deep Learning
- Lee-Carter Neural Network and Application
- Modelling with RNNs
- 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 q_x)

- **Mortality over time modelled using:**

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

- **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?
- Problem 1: 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)

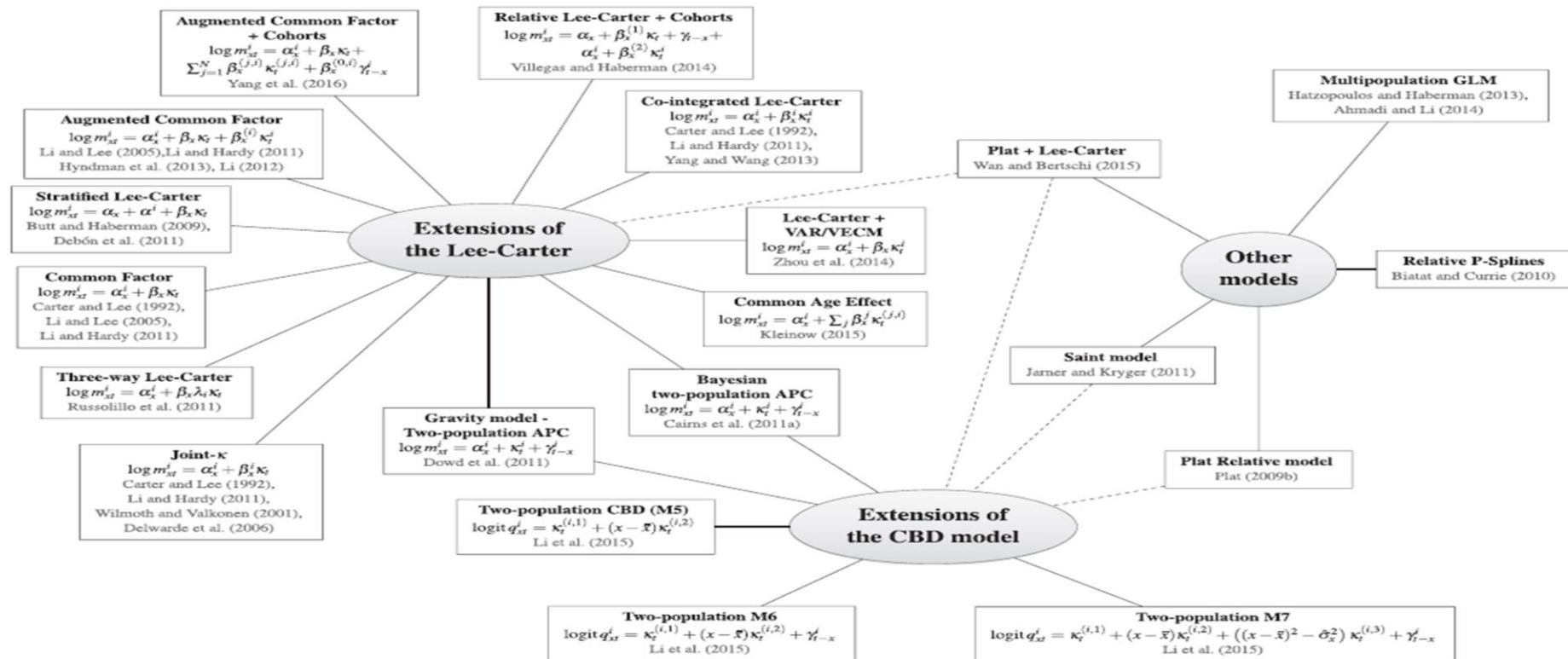
EXTENDING THE LC MODEL

- 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)	$\log(u_{x,t}) = a_x^i + b_x k_t^i$

- **Problem 2: 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?**

COMPLEXITY: MULTI-POPULATION MORTALITY MODELLING

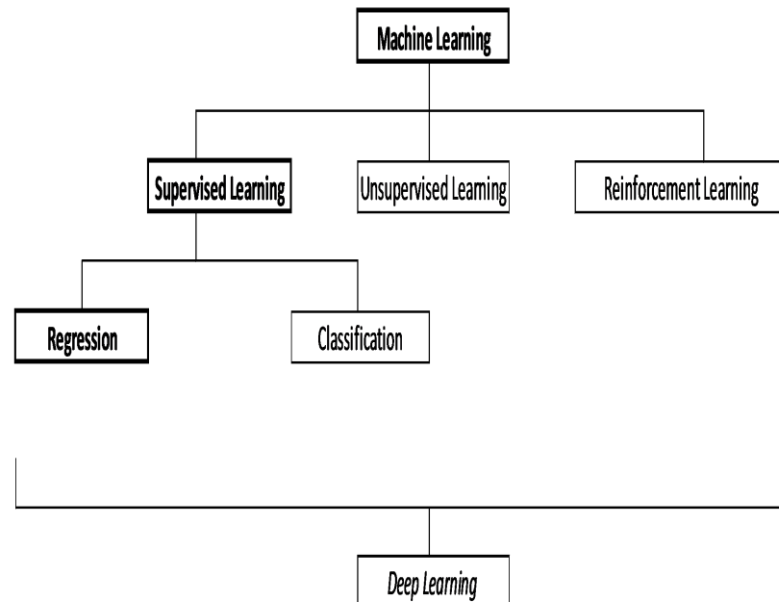


AGENDA

- Lee-Carter Paradigm for Mortality Modelling
- From Machine Learning to Deep Learning
- Lee-Carter Neural Network and Application
- Modelling with RNNs
- Time Series Forecasting with Deep Learning

MACHINE LEARNING

- **Machine Learning** is “the study of algorithms that allow computer programs to automatically improve through experience” (Mitchell 1997)
- **Machine learning is an approach to the field of Artificial Intelligence**
Systems trained to recognize patterns within data to acquire knowledge (Goodfellow, Bengio and Courville 2016).
- **Earlier attempts to build AI systems = hard code knowledge into knowledge bases ... but doesn't work for highly complex tasks e.g. image recognition, scene understanding and inferring semantic concepts (Bengio 2009)**
- **ML Paradigm – feed data to the machine and let it figure out what is important from the data!**
Deep Learning represents a specific approach to ML.



SUPERVISED LEARNING

- Supervised learning = application of machine learning to datasets that contain features and outputs with the goal of predicting the outputs from the features (Friedman, Hastie and Tibshirani 2009).
- Feature engineering - Suppose we realize that Claims depends on Age² => enlarge feature space by adding Age² to data. Other options – add interactions/basis functions e.g. splines

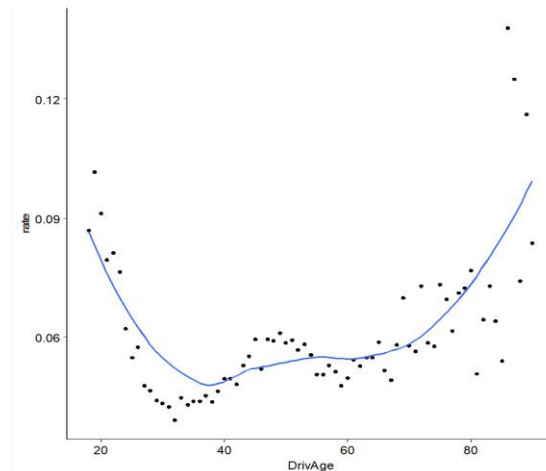
y (outputs)

X (features)

```
> freMTPL2freq
```

	IDpol	ClaimNb	Exposure	Area	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Density	Region	DrivAge_2
1:	1	1	0.100000000	D	5	0	55	50	B12	Regular	1217	R82	3025
2:	3	1	0.770000000	D	5	0	55	50	B12	Regular	1217	R82	3025
3:	5	1	0.750000000	B	6	2	52	50	B12	Diesel	54	R22	2704
4:	10	1	0.090000000	B	7	0	46	50	B12	Diesel	76	R72	2116
5:	11	1	0.840000000	B	7	0	46	50	B12	Diesel	76	R72	2116

678009:	6114326	0	0.002739726	E	4	0	54	50	B12	Regular	3317	R93	2916
678010:	6114327	0	0.002739726	E	4	0	41	95	B12	Regular	9850	R11	1681
678011:	6114328	0	0.002739726	D	6	2	45	50	B12	Diesel	1323	R82	2025
678012:	6114329	0	0.002739726	B	4	0	60	50	B12	Regular	95	R26	3600
678013:	6114330	0	0.002739726	B	7	6	29	54	B12	Diesel	65	R72	841



ISSUES WITH ML APPROACH

- **In many domains, traditional approach to designing machine learning systems relies on human input for feature engineering or model specification.**
- **Three arguments against traditional approach:**

Complexity – which are the relevant features to extract/what is the correct model specification? Difficult with very high dimensional, unstructured data such as images or text. (Bengio 2009; Goodfellow, Bengio and Courville 2016)

Expert knowledge – requires suitable prior knowledge, which can take decades to build (and might not be transferable to a new domain) (LeCun, Bengio and Hinton 2015)

Effort – designing features is time consuming/tedious => limits scope and applicability (Bengio, Courville and Vincent 2013; Goodfellow, Bengio and Courville 2016)

- **Complexity is not only due to unstructured data. Many difficult problems of model specification arise when performing actuarial/demographic tasks at a large scale:**

Mortality forecasting for multiple populations

REPRESENTATION LEARNING

- Representation Learning = ML technique where algorithms automatically design features that are optimal (in some sense) for a particular task
- Traditional examples are PCA (unsupervised) and PLS (supervised):

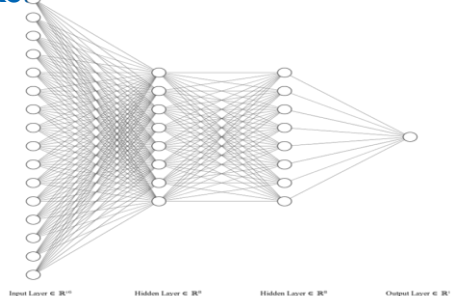
PCA produces features that summarize directions of greatest variance in feature matrix

PLS produces features that maximize covariance with response variable (Stone and Brooks 1990)

- Feature space then comprised of learned features which can be fed into ML/DL model
- BUT: Simple/naive RL approaches often fail when applied to high dimensional data

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.”**



AGENDA

- Lee-Carter Paradigm for Mortality Modelling
- From Machine Learning to Deep Learning
- Lee-Carter Neural Network and Application
- Modelling with RNNs
- Time Series Forecasting with Deep Learning

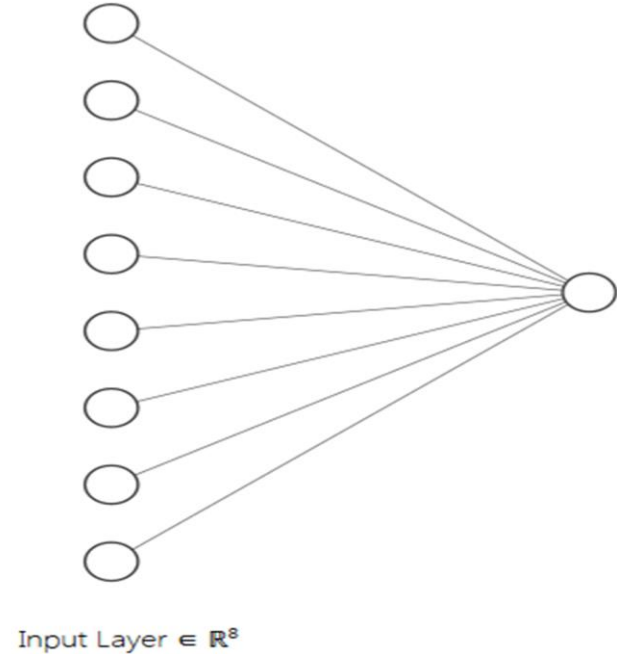
SINGLE LAYER NN = LINEAR REGRESSION

- **Single layer neural network**

Circles = variables

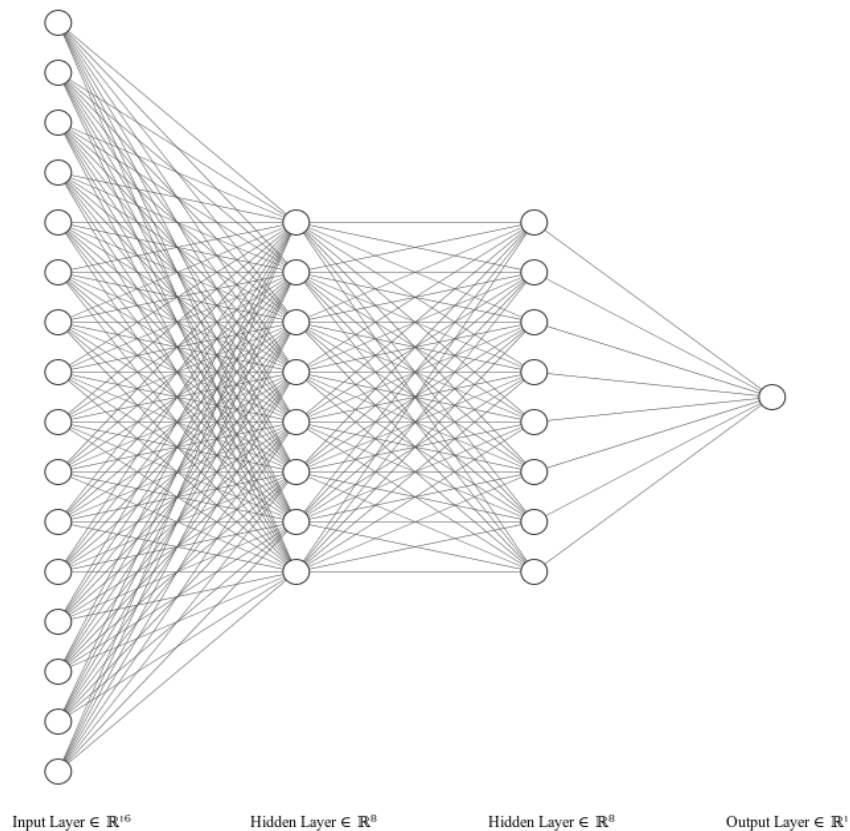
Lines = connections between inputs and outputs

- **Input layer holds the variables that are input to the network...**
- **... multiplied by weights (coefficients) to get to result**
- **Single layer neural network is a GLM**



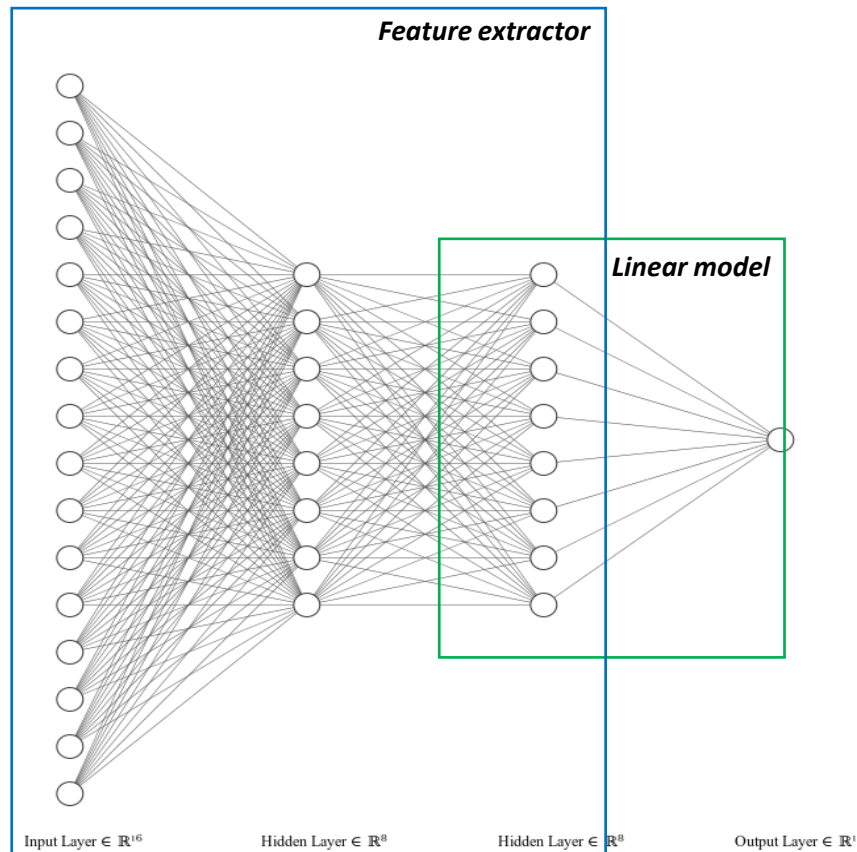
DEEP FEEDFORWARD NET

- Deep = multiple layers
- Feedforward = data travels from left to right
- Fully connected network (FCN) = all neurons in layer connected to all neurons in previous layer
- More complicated representations of input data learned in hidden layers - subsequent layers represent regressions on the variables in hidden layers



NEURAL NETS GENERALIZES GLM

- Intermediate layers = representation learning, guided by supervised objective.
- Last layer = (generalized) linear model, where input variables = new representation of data
- No need to use GLM – strip off last layer and use learned features in, for example, XGBoost
- Or mix with traditional method of fitting GLM



SPECIALIZED ARCHITECTURES

- Most modern examples of DL achieving state of the art results on tasks rely on using specialized architectures i.e. not simple fully connected networks
- Key principle - Use architecture that expresses useful priors (inductive bias) about the data
=> Achievement of major performance gains

Embedding layers – categorical data (or real values structured as categorical data)

Autoencoder – unsupervised learning

Convolutional neural network – data with spatial/temporal dimension e.g. images and time series

Recurrent neural network – data with sequential/temporal structure

Skip connections – makes training neural networks easier

EMBEDDING LAYER – CATEGORICAL DATA

- One hot encoding expresses the prior that categories are orthogonal => similar observations not categorized into groups

	Actuary	Accountant	Quant	Statistician	Economist	Underwriter
Actuary	1	0	0	0	0	0
Accountant	0	1	0	0	0	0
Quant	0	0	1	0	0	0
Statistician	0	0	0	1	0	0
Economist	0	0	0	0	1	0
Underwriter	0	0	0	0	0	1

- Traditional actuarial solution – credibility (i.e. mixed models)

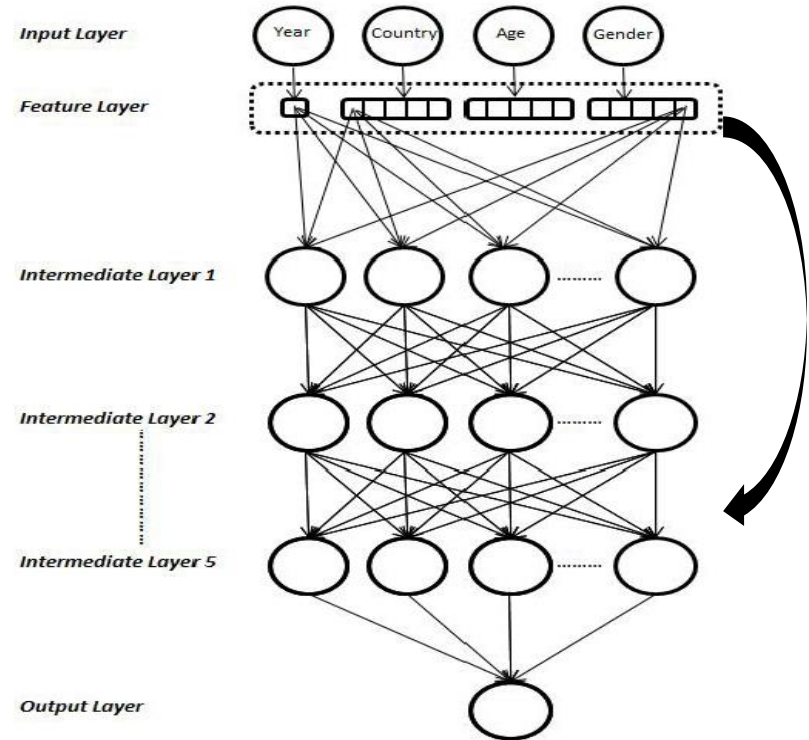
- Embedding layer prior – related categories should cluster together:

Learns dense vector transformation of sparse input vectors and clusters similar categories together

	Finance	Math	Statistics	Liabilities
Actuary	0.5	0.25	0.5	0.5
Accountant	0.5	0	0	0
Quant	0.75	0.25	0.25	0
Statistician	0	0.5	0.85	0
Economist	0.5	0.25	0.5	0
Underwriter	0	0.1	0.05	0.75

EXAMPLE – LEE-CARTER NEURAL NET

- **Multi-population mortality forecasting model (Richman and Wüthrich 2018)**
- **Supervised regression on HMD data (inputs = Year, Country, Age; outputs = mx)**
- **5 layer deep FCN**
- **Generalizes the LC model by letting the NN specify the model**
- **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 forecast 51 out of 76 times**
- **For purposes of large scale mortality forecasting, deep neural networks dramatically outperform traditional single and multi-population forecasting models**

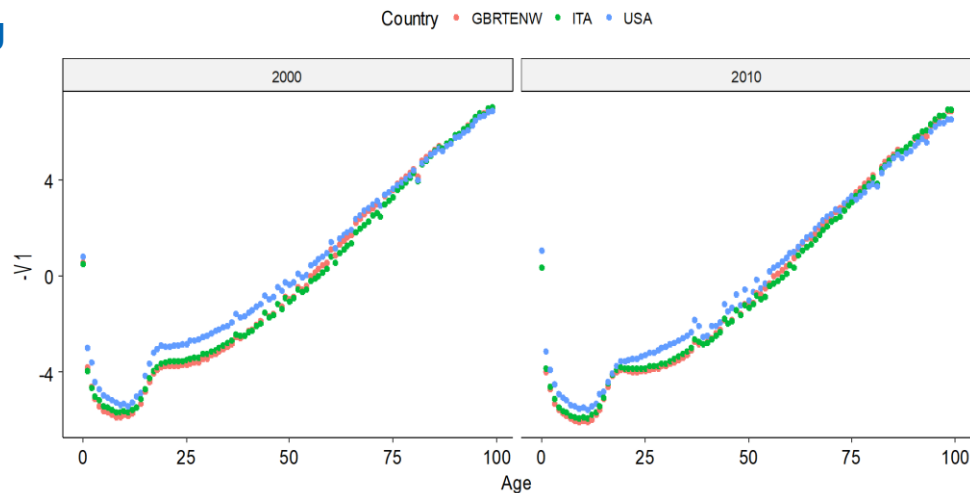
	Model	Average MSE	Median MSE	Best Performance
1	LC_SVD	5.50	2.48	19
2	ACF_SVD_region	3.46	2.50	36
3	ACF_SVD_country	7.30	4.77	9
4	ACF_BP	6.12	3.00	12

	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)



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

y_x = logit of mortality at age x

a, b = regression coefficients

z_x^{Ref} = logit of reference mortality

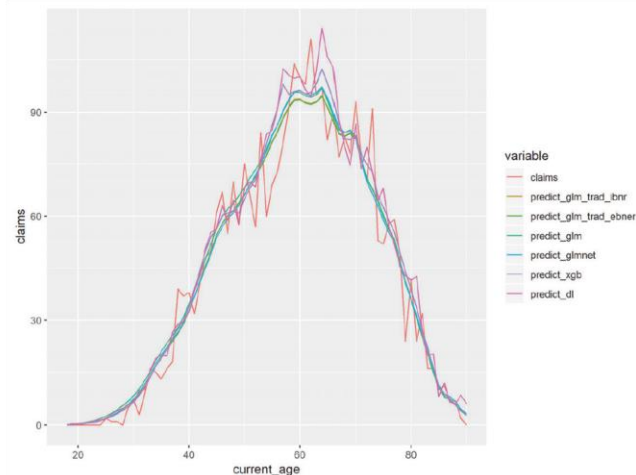
APPLICATION TO INSURANCE DATA

- 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%

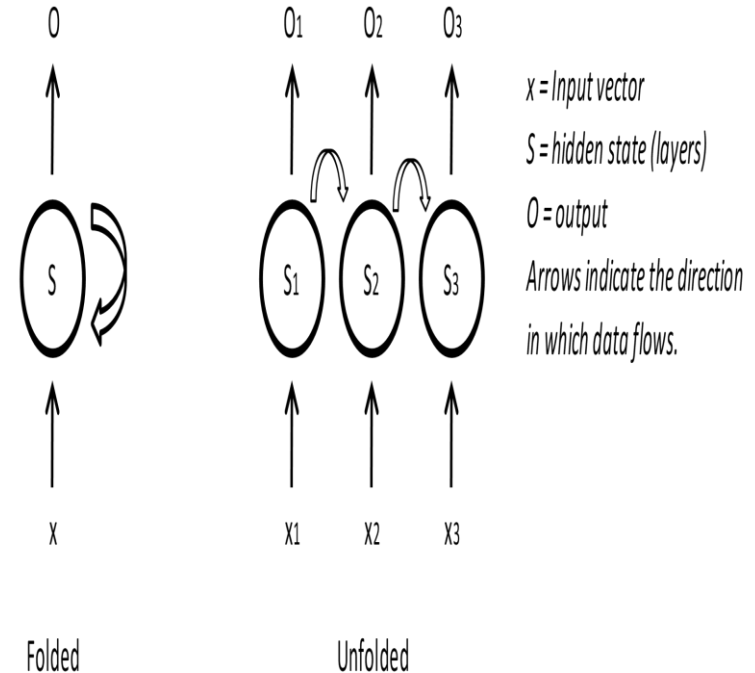


AGENDA

- Lee-Carter Paradigm for Mortality Modelling
- From Machine Learning to Deep Learning
- Lee-Carter Neural Network and Application
- Modelling with RNNs
- Time Series Forecasting with Deep Learning

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**



LEE & CARTER GO DEEP 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

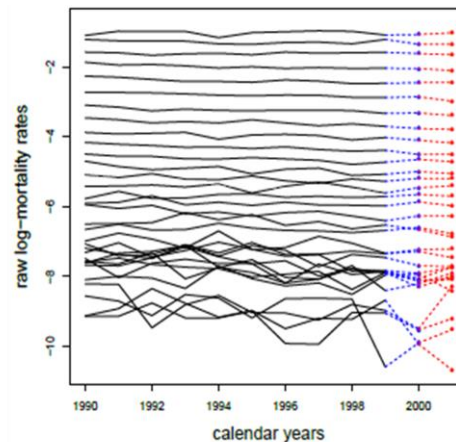
- Swiss mortality rates forecast using RNNs in Richman and Wüthrich (2019)

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

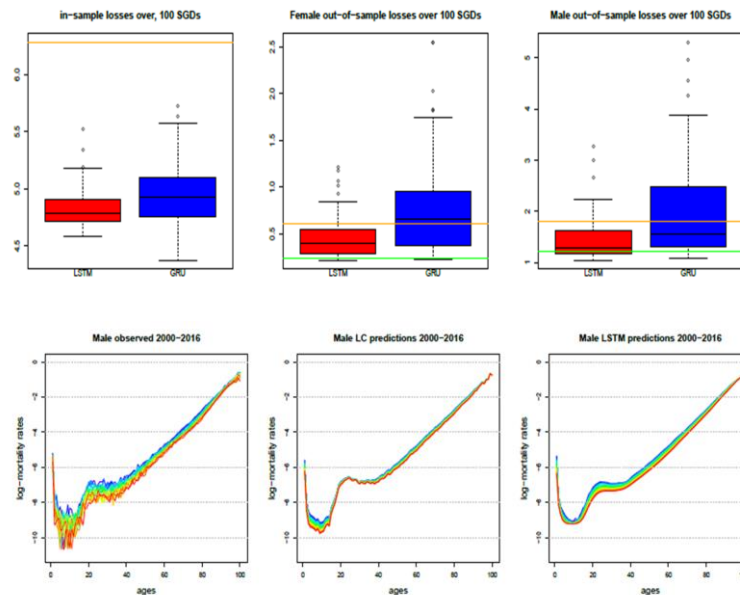
- **Results improved versus single gender models**

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 both genders	out-of-sample female	out-of-sample 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



AGENDA

- Lee-Carter Paradigm for Mortality Modelling
- From Machine Learning to Deep Learning
- Lee-Carter Neural Network and Application
- Modelling with RNNs
- Time Series Forecasting with Deep Learning

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

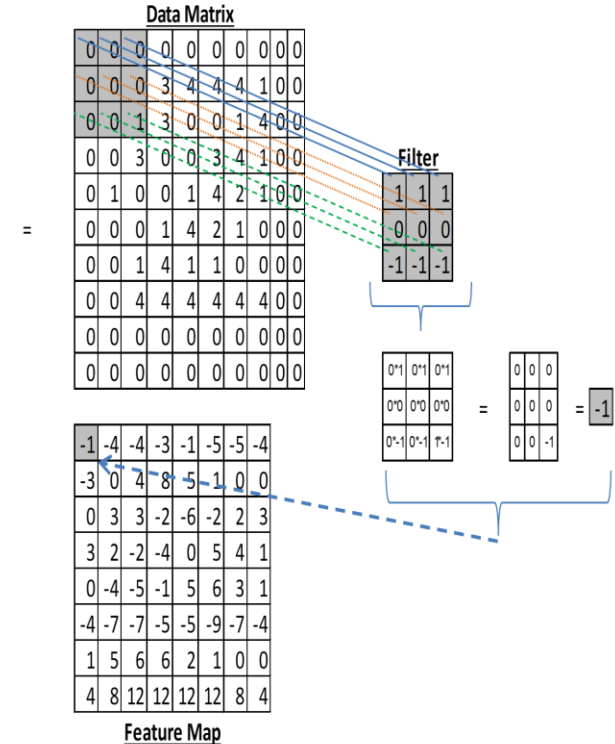
- **Convolutional network is locally connected and shares weights => expresses prior of position invariance**

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**



PROCESSING TIME SERIES WITH DL

- RNN model is 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:

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

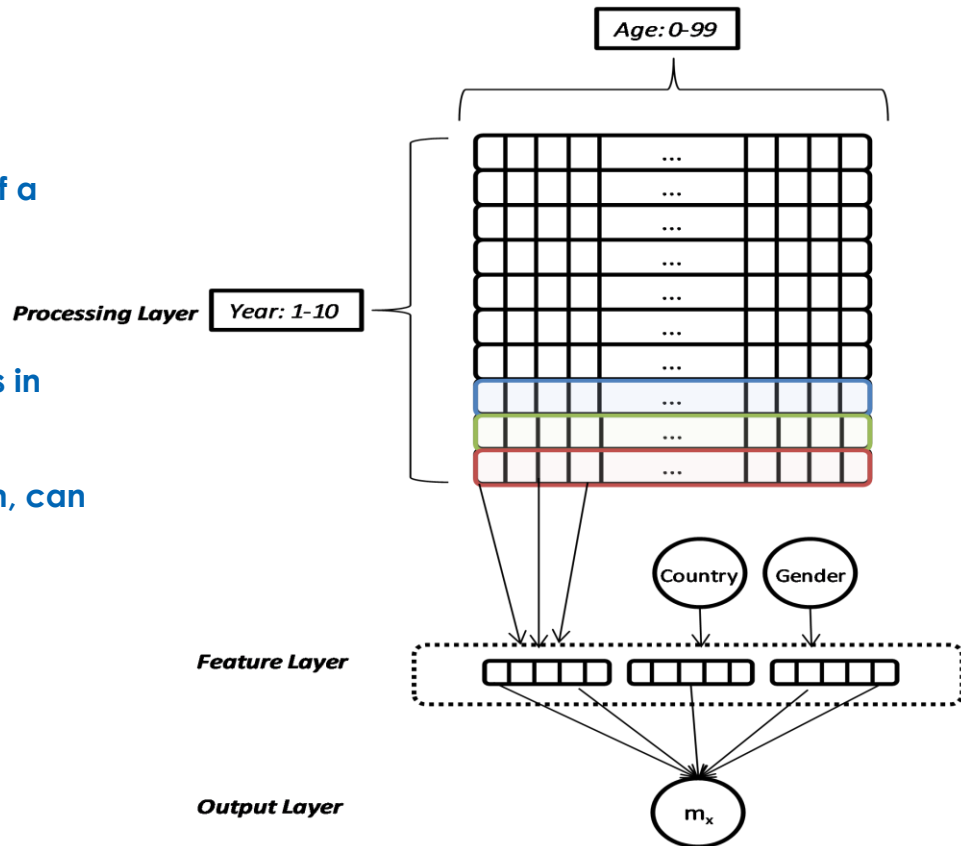
- Addressed in Perla, Richman, Scognamiglio and Wüthrich (2020)
- **Finding: deep networks do not perform well in this context**
Since mortality rates have changed in relatively simple manner, deep networks might overcomplicate the modelling

MODEL STRUCTURE

- Similar to LCNN model...
- ... however, time variable replaced with outputs of a NN processing layer
- CNNs appear to perform better than RNNs
- Best performance achieved with no non-linearities in the model
- Since features are used immediately for prediction, can be interpreted as an extended LC model:

$$\log(\mu_{x,t}^{p,g}) = \widehat{a}_x^p + \widehat{a}_x^g + b_x \cdot \widehat{k}_t$$

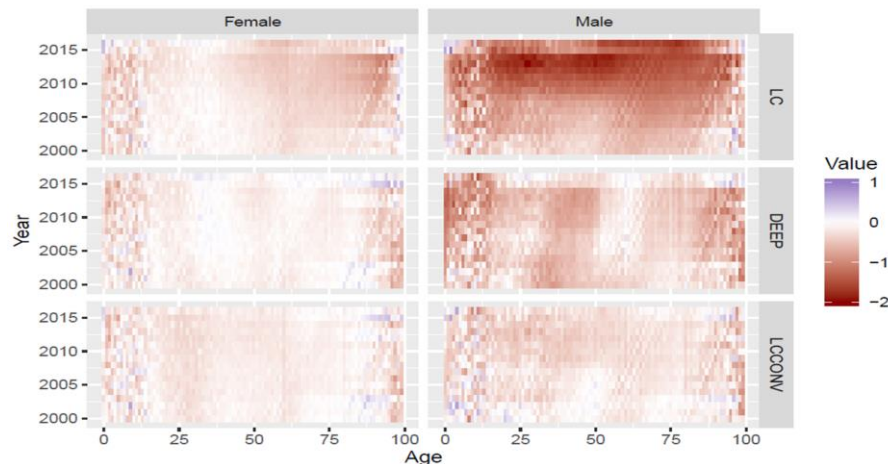
- First terms equivalent to $\mathbf{a}_x$
- Last term equivalent to $\mathbf{b}_x \cdot \mathbf{k}_t$
- Judging by number of parameters, this model is simpler than the LCNN model



FORECAST RESULTS

- The CNN model (LCCONV) achieves better performance versus the LC model on 75/76 populations in the HMD
- Unadjusted model also generalized well – beat LC on 101/102 populations in the USMD
- 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



CONCLUSION

- Deep learning models provide new opportunities for modelling mortality
- Learned representations from deep neural networks often have readily interpretable meaning
- Emphasis on predictive performance => 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>

THANK YOU – ANY QUESTIONS?



QED Actuaries & Consultants Pty Ltd

1st Floor, Hunts End Office Park,
38 Wierda Road West, Sandton,
South Africa

QED Actuaries & Consultants Mauritius Ltd

The POD, Vivea Business Park,
Moka,
Mauritius

QED Actuaries & Consultants Kenya Ltd

Nairobi Garage, Cove Court,
Watermark Business Park, Karen, Nairobi,
Kenya

QED Actuaries Nigeria Ltd

Workstation Offices, No. 7 Ibiyinka Olorunbe Close,
Victoria Island, Lagos,
Nigeria

