Science World Journal Vol. 16(No 2) 2021
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

DEEP AND PROBABILISTIC LEARNING UNDER UNCERTAINTIES CUM NON-SPHERICAL DISTURBANCES

I. Oloyede

Department of Statistics, University of Ilorin, Nigeria

*Corresponding Author Email Address: oloyede.i@unilorin.edu.ng

ABSTRACT

The study investigates the performances of deep and probabilistic models under uncertainties with non-spherical disturbances inherent in the data. We deemed aleatoric and epistemic uncertainties, the former inherent in the data while the later inherent in the model in probabilistic approach. Loss, mean square error (MSE), mean absolute error (MAE) were adopted to evaluate the performance of the models for training, testing and validating sets. Both multicollinearity and autocorrected error were inherent in the data, there exist negative autocorrected error of magnitude 1.46 and the multicollinearity with magnitude of "inf" that implies imperfect multicollinearity were inherent in the data. Keras Dense layer and Tensor flow probability (tfp) Dense variational layer were adopted. The underlying model were constructed probabilistically to capture aleatoric, epistemic and both. The study observed that the "no uncertainty, classical and aleatoric models behaved well when data were standardised, the magnitude of loss, MAE and MSE reduced by almost 98%, this implies that the accuracy of the parameter were improved, though epistemic and both aleatoric and epistemic uncertainties models depicted poor performances of the model despite their probabilistic nature, this may be due to combination of uncertainty with non-spherical disturbances. The unstandardised data exhibited poor performances in all the models. The study therefore recommended that data should be standardised prior estimation.

Keywords: Aleatoric, Autocorrelation, Epistemic, Error term, Multicollinearity.

INTRODUCTION

Artificial neural network (ANN) is the statistical model built from inspiration from the architecture and cognitive capabilities of biological brain. ANN model have a layer of architecture comprising large neurons in each layer. The first of which is input, the last layer is output. The middle layer is the hidden layer. Each neuron is determined by a non-linear function of the neuron connected with each other, each connection has a weight that is determined from the training data comprising set of input and output pair. Geoffrey *et al.* (2006) developed a fast-learning algorithm for learning multilayer neural network.

Deep learning algorithm (DLA) has the capacity of mapping out perceptions (inputs) with array of output inherent herein hidden layers handling massive large high dimensional inputs, thus the processes are taken place obliviously thereby leading to confused or inaccurate inferences (Alex and Yarin, 2017). In deep learning algorithm, it is common that a person approaches fashion designer to sewing cloth of his or her choice of style, he or she will either describe the style or choose from the existing styles from the designer, the designer will henceforth take the measurement of his/her client, then he will observe the client clearly. This is not sufficient enough; the designer needs to think deeply in order to make good design. Many people have been victim of poor designs due to poor thinking; the client complains bitterly if the designer could not think deeply. It is noteworthy of the inaccurate learning and mapping of deep learning in its attempt to map African-American images which resulted into mapping of it with gorillas (Jessica, 2015). It was reported that DLA with respect to aleatoric uncertainty underperformed in depth regression, but strongly captured in Bayesian deep learning paradigm owing to its probabilistic nature (Alex and Yarin, 2017).

Aleatoric uncertainty is the inherent noise in the dataset, this is so prominent in cross sectional dataset that is characterised with non-spherical disturbances. We infused the simulated dataset with multicollinearity (dependency of the covariates) and autocorrelation of the disturbance term U. The epistemic uncertainty on the other hand refers to the uncertainty inherent in the model. Both uncertainties are captured in Bayesian modelling. We are of the opinion that the presence of both uncertainties will bring about paramistic uncertainty. This type of uncertainty tells us the misbehaviour /misleading of parameters which produce synergistic in lieu of antagonistic and vice-versa. Epistemic uncertainty captures the ignorance in the model which often leads to non-closed form which fitted the dataset well, it is difficult in most cases to ascertain the model that best fit the data well in classical paradigm. More often than none, various distributional and transformation approach had been adopted. Gal (2016) examined both aleatoric and epistemic uncertainties in Bayesian learning paradigm.

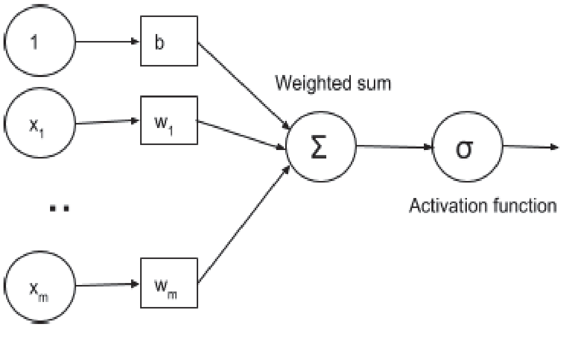
Epistemic uncertainty captures the modelling or augmentation of model weight (parameter) with prior distribution over the likelihood which in most cases arise in ignorance or unknowing approach. Amodei et al., (2016) explored extremely large deep learning model considering more than hundred million of features and 11 hidden layers. Jchmidhuber (2015) examined historical review of deep learning applications. Nicholas and Vadim (2017) claimed that DLA is an algorithm capable of analysis of large set of data of high dimension. They described it as an approach not probabilistic as a result, it suffers the inability of improve estimate when there is noisy data irrespective of the dimension.

Deep learning has revolutionized machine learning which is the nucleus of artificial intelligence (AI). The uncertainty are the major threats to the accuracy of the inferences which in many ways accepted ignorantly as accurate, deep learning failed in this perspective, thus this call for extensive approach in an attempt to have accurate inference.

The related information is search for and redundant information are eliminated. The retained relevant information are connected in layers. Yarin and Zoubin (2015) proposed approach that examined uncertainty theoretically and application wise. Alex and Yarin, (2017) distinguished between aleatoric and epistemic uncertainties in a Bayesian paradigm. Yarin and Zoubin (2015) explored Bayesian deep learning by placing Gaussian prior on the weight of the model with a view of estimating the uncertainty theoretically. The epistemic uncertainty is inherent on the weight cum its prior. In most cases, research cannot ascertain the optimal model that will fit the data well and minimize aleatoric uncertainty that is inherent in the data. They explored practically heteroscedastic loss and accuracy, equally, adopted dropout with the effort of sampling through Monte Carlo algorithm which is used to obtain posterior density. It has been observed that deep learning breakdown in the presence of uncertainty (Keydana, 2018). This then pave way for the adoption of probabilistic modelling which advertently capture uncertainty in the model.

Nicholas and Vadim (2017) claimed that full potential application of Bayesian inference to the field of deep learning algorithm are yet to be fully explored. They added that Bayesian regularization approach adopted in their study provided more advantages in the predictive analytics with non-linear data relationship. Both Bayesian and probabilistic approaches were examined with Kolmogorov's representation of a multiple outcome, the activation function that affine transformation of the dendrite (inputs) which is the weighted sum of elements in the inputs variables.

This study observed loss function with a view to minimize aleatoric uncertainty, the effort gear towards inherent dual disturbance term in the data, thus the data significantly captured aleatoric uncertainty while the model captures epistemic uncertainty. With Bayesian deep learning algorithm, this study obtained both the aleatoric and epistemic uncertainties theoretically and practically in a cross-sectional econometric syndrome. Sequel to section one that cover introductory concept of deep and deep probabilistic learning is the section two which is devoted to deep learning with continuous outcome variable. Section three examined the methodology within the framework of deep probabilistic learning. In section four the study looks into the concluding part of the work



Deep Learning Algorithm of Continuous Outcome

Let $Y = F(X)$ be a mapping of X-input of high dimensional features denote a deep learning input-output mapping, with input space $X = X_1,,,\dots,,,X_P$. The output Y being a continuous random variable.

A set of hidden layers are inherent in the mapping function, then we have affine activation function f_i given by: $f_i^{W,b} = f_i\left(\sum_{j=1}^{N_j} W_{ij}\, z_j + b_i\right)$ where W and z stand as weight matrix and inputs of the ith layers, we add the error structures of both heteroscedastic and auto-correlated error. Gaussian prior distribution $w \sim N(0,I)$, this is referred to as epistemic uncertainty, this study therefore modified the prior with auto-correlated error which are embedded in the data generation syndrome. This is known as Bayesian neural network where possible weight parameters are average out which is different from classical point of view where the weight parameter is obtained iteratively.

$$p\left(y|f^{w}(X)\right) \qquad (1)$$

The posterior density is obtained by conjugating the prior with the likelihood

$$p(w|X,y) \cong p(y|X,w)p(w)/p(y|X) \qquad (2)$$

$p(y|X)$ is known as normalization which is assumed to be unity. In deep leaning regression we specify the log likelihood as in [9]

$$-logp\left(y\middle|f^{\widehat{w}},(X_i)\right) \propto \frac{1}{2\sigma^2} \parallel y_i - f^{\widehat{w}}(X_i) \parallel^2 + \frac{1}{2} log\sigma^2 \qquad (3)$$

To capture the distribution of error structure in our study, we model σ^2 which capture noise in the data and model, both heteroscedastic and autocorrected error are inherent in σ^2 embedded in Gaussian likelihood. Dropout is infused as a variation Bayesian approximation which is the aggregate of two Gaussian with small variance in one and zero mean in the other [9] VBA can be expressed as

$$l(\theta,p) = -\frac{1}{N}\sum_{i=1}^{N} logp\left(y_i|f^{\widehat{w}},(X_i)\right) + \frac{1-p}{2N} \parallel \theta \parallel^2. \qquad (4)$$

P is the dropout probability. Since aleatoric uncertainty is captured in the dataset, thus epistemic uncertainty is captured by the predictive variance as

$$var(y) \approx \sigma^2 + \frac{1}{T}\sum_{i=1}^{T} f^{\widehat{w}}(X)' f^{\widehat{w}}(X_i) - E(y)' E(Y), \qquad (5)$$

The predictive mean of Epistemic uncertainty $E(Y) = \frac{1}{T}\sum_{i=1}^{T} f^{\widehat{w}}(X)$, in the above equation, σ^2 part representing the noise (uncertainty) inherent in the data.

METHODOLOGICAL DESIGN

Let Dataset D be denoted by: $[X_i, y_i]_{i=1}^{N}$

$$y(x) = X'w + \varepsilon \qquad (6)$$

where $w \in R^d$ represents Parameters and $\varepsilon = \rho\varepsilon_{t-1} + u_t$, assume $u_t \overset{iid}{\sim} N(0,\sigma^{-2})$ of which the error process is stationary at $/\rho/$<1, thus the covariance matrix of ε is expressed as: $\sigma^{-2}\Omega$ where

$$\Omega = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & . & . & . & \rho^{T-1} \\ \rho & 1 & \rho & . & . & . & \rho^{T-2} \\ \rho^2 & \rho & 1 & . & . & . & \rho^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \vdots & \vdots & \vdots & 1 \end{pmatrix} \qquad (7)$$

ε is autocorrected error infused into the data cum the X-variables that are collinear of which are not of full rank, where the regressors perfectly multicollinearity with variance inflation factor on magnitude “inf”. The study injected both the $\sigma^{-2}\Omega$ and $\lambda(\beta'\beta+m)$ into the model (likelihood) to capture both autocorrected error and multicollinearity. Both non-spherical disturbances were inherent in the data.

Let $y = f(x){+}\in$ such that $f(x) = x'w$ where x is the input vector, w denotes vector of weight (parameters) of a linear deep learning model. f is the transfer function while y represents the output.

The likelihood of the deep learning model
$p(y|X,w) = \prod_{i=1}^{n} p\,(y_i|X_i,w)$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_n^2\Omega}} exp\left(\frac{-(y_i - X_i{\prime}w)^2 + \lambda(\beta'\beta+m)}{2\sigma_n^2\Omega}\right) \quad (8)$$

$$= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_n^2\Omega)^{n/2}} exp\left(\frac{-(|y_i - X_i{\prime}w|)^2 + \lambda(\beta'\beta+m)}{2\sigma_n^2\Omega}\right) \quad (9)$$

$$= N(X'w + \lambda I\ \ ,\sigma_n^2 I\Omega) \quad (10)$$

In Bayesian paradigm, the prior is specified over the parameters space. The prior for the w is assumed to have zero mean and covariance matrix $w = N\left(0,\Sigma_p\right)$. The posterior density is expressed as: $p(w|y,X) = \frac{p(y|X,w)p(w)}{p(y|X)}$ where p(y|X) is normalizing constant or marginal likelihood which is assumed to be unity $\int p(y|X) = 1$

$$p(w|y,X) = \int p(y|X,w)p(w) \quad (11)$$

Completing the squares have

$$p(w|y,X) \propto exp\left(-\frac{1}{2\sigma_{n\Omega}^2}(y - X'w)^2 + \lambda(\beta'\beta + m)\right) exp\left(-\frac{1}{2}w'\Sigma_p^{-1}w\right) \quad (12)$$

$$\propto exp\left(-\frac{1}{2}(w-\overline{w})'\left(\frac{1}{\sigma_n^2\Omega}XX' + \lambda I + \Sigma_P^{-1}\right)(w-\overline{w})\right) \quad (13)$$

where $\overline{w} = \sigma_n^{-2}\Omega(\sigma_n^2XX' + \lambda I + \Sigma_P^{-1})^{-1}Xy$. The prior is Gaussian with mean $\overline{w}$ and covariance

$$A^{-1} = \sigma_n^2X\Omega X + \lambda I + \Sigma_p^{-1}, \quad (14)$$

thus we have

$$p(w|y,X) \propto N\left(\overline{w} = \frac{1}{\sigma_n^2}A^{-1}Xy, A^{-1}\right) \quad (15)$$

RESULTS AND INTERPRETATION

The study used normal distribution with scale of 1, with tfp.layers.DistributionLambda which return an instance of with tfd.Distribution. The model in anyway cannot capture data structure due to its complexity. The data has dual non-spherical disturbances (multicollinearity and autocorrelation of error term, the multicollinearity is severe with variance inflation factor of *inf* where autocorrelation of error term is negative with value of 1.46 that is less than 1.5), mere probabilistic model without uncertainty may not be able to capture the inherent uncertainty. The second model focused on modelling aleatoric uncertainty; this type of uncertainty inherent in the data coupled with the associated disturbances. Aleatoric the variability is inherent in the data which make impossible to predict target perfectly. The inherent variability in the data inhibit the certainty of the parameters.

Epistemic: the noise inherent in the data implies that certainty of parameters of underlying process of linear relationship between regressors and regresand is doubtful. Epistemic uncertainty can be reduced if data I increased, this is impossible in case of aleatoric. In epistemic, standard keras is replaced with Tensorflow probability DenseVariational layer. The tfp layer uses a variational posterior Q(w) over the weight to denote the uncertainty in the weight(parameters). The dense layer regularises the posterior Q(w) in order to close to prior p(w) which models the uncertainty in the underlying process (Pavel, et al., 2019). Both uncertainty: an extra output is added to TFP DenseVariational to capture aleatoric uncertainty in order to model the scale of the target distribution.

Table 1: Depicting RMSE, MAE, LOSS AND MSE Of Training, Testing And Validating Unstandardized Datasets For The Different Models In One Epoch.

	No Uncertainty	Aleatoric	Epistemic	Both	Frequentist
RMSE Val	14.23	14.37	14.31	14.67	13.71
loss	102.11	207.84	100.53	182.32	196.40
MAE	11.83	11.70	11.89	11.13	11.43
MSE	199.28	201.19	206.32	176.04	196.40
RMSE Test	10.11	14.42	10.03	13.50	14.01

From the above table, the study observed in validation data, that both aleatoric and epistemic uncertainties models have the highest root mean square error of magnitude 14.67 while the classical approach which only considered non-spherical disturbances without modelling uncertainty has lowest root mean squares error with magnitude 14.23. The classical adopted keras *Dense layer* while the probabilistic approach made use of *DenseVariational* model uncertainty cum non-spherical disturbances that inherent in the data. Epistemic-uncertainty in probabilistic model has minimum root mean square error and loss for the test data and training data respectively while aleatoric uncertainty has the highest mean square error and loss for test and training data respectively. The study recorded that epistemic uncertainty recorded the highest magnitude of mean square error and mean absolute error with both uncertainties having the lowest MSE and MAE for the trading data. It can be inferred from the study the both modelling uncertainty and data infusing with non-spherical disturbances inhibit the perfect and accuracy of the parameters and test of hypothesis and standard error that make use of the parameters.

Table 2: Depicting MAE, LOSS AND MSE OF Training and Validating Unstandardized Datasets for the Different Models in 100 Epoch

	No uncertainty	Aleatoric	Epistemic	Both	Frequentist
LOSS	95.61	194.79	99.76	205.93	177.95
MAE	11.41	11.42	11.59	11.55	10.85
MSE	189.76	189.71	199.44	198.87	177.95
Val_loss	105.48	218.27	110.80	227.50	188.73
Val_mae	13.24	13.29	13.38	13.36	12.28
Val_mse	209.40	209.52	220.51	218.69	188.73

The above table described loss, mse and mae of both training and validating data the study observed that both uncertainties have the highest loss, mae, mse in both training and validating data whereas no uncertainty has minimum loss in both training and validating data, the classical has minimum mse and mae both in training and validating data.

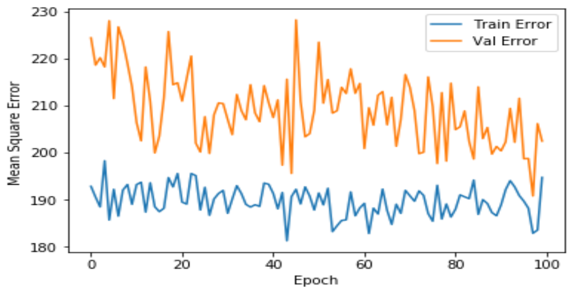


Fig. 1 showing training and validation error for no uncertainty model of unstandardised datasets

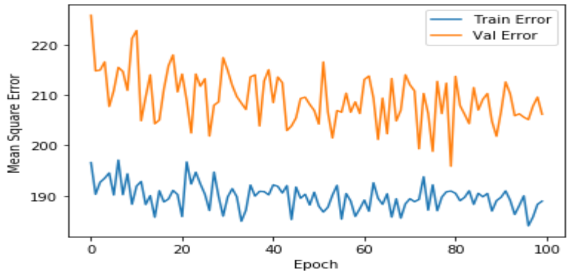


Fig. 2: showing training and validation error for aleatoric uncertainty model of unstandardized datasets

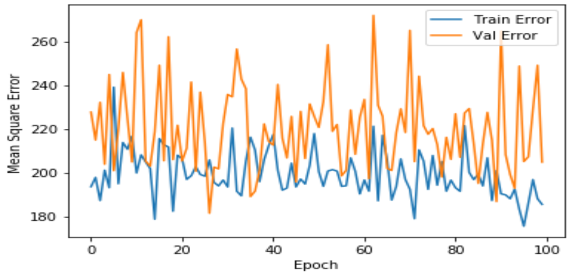


Fig. 3: showing training and validation error for epistemic uncertainty model of unstandardized datasets

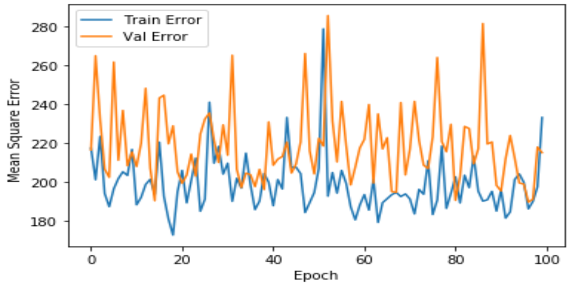


Fig. 4 showing training and validation error for both aleatoric and epistemic uncertainties model of unstandardized datasets

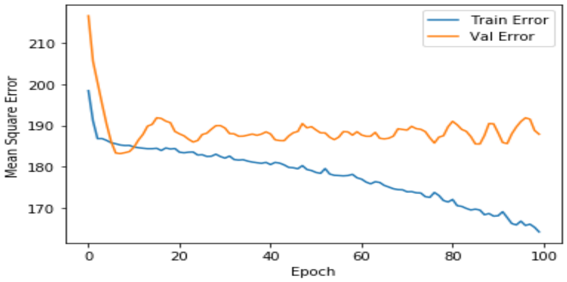


Fig. 5: showing training and validation error for frequentist model of unstandardized datasets

From figures 1 to 5, it is observed that training error decrease slowly as the epoch increases, validating error stay above the training error, the study observed that effect of non-spherical disturbances and uncertainty in the model have great impact in the outcome of the study.

Table 3: Depicting RMSE, MAE, LOSS AND MSE of Training, Testing and Validating Standardized Datasets for the Different Models in One Epoch

	No Uncertainty	Aleatoric	Epistemic	Both	Frequentist
RMSE Val	1.25	1.20	13.04	14.11	1.09
LOSS	1.46	1.61	71.01	39.22	1.01
MAE	1.15	0.92	9.31	6.67	0.82
MSE	1.74	1.27	137.70	74.82	1.01
RMSE Test	1.21	1.27	8.43	6.26	1.01

Table 4: Depicting MAE, LOSS AND MSE of Training and Validating unstandardized Datasets for the Different Models in 100 Epoch

	No Uncertainty	Aleatoric	Epistemic	Both	Frequentist
LOSS	1.41	1.54	507.48	1483.06	1.02
MAE	1.11	0.95	20.14	20.23	3.32
MSE	1.95	1.45	1014.84	1031.58	2.61
Val_loss	1.43	1.69	514.15	1212.08	1.12
Val_mae	1.14	1.04	21.34	20.85	2.61
Val_mse	1.98	1.58	1026.02	936.25	3.32

The above table described loss, mse and mae of both training and validating data, the study observed that both uncertainties have the highest loss, mae, mse in both training and validating data except epistemic model that has highest mse in validating data whereas classical has minimum loss both at training and validating data. Aleatoric uncertainty has minimum mae and mse for both the training and validating data.

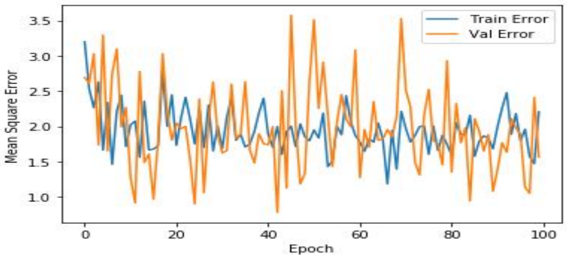


Fig. 6: showing training and validation error for both no uncertainties model of standardized datasets

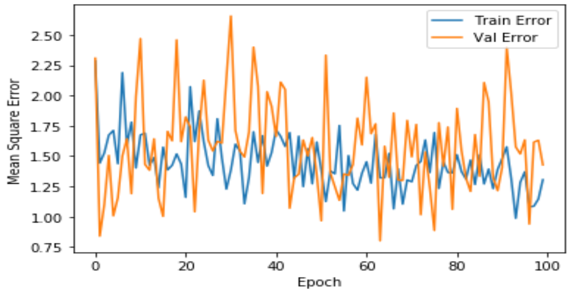


Fig. 7 showing training and validation error for aleatoric uncertainty model of standardized datasets

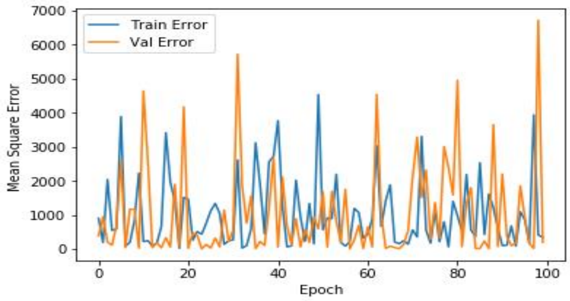


Fig. 8 showing training and validation error for epistemic uncertainty model of standardized datasets

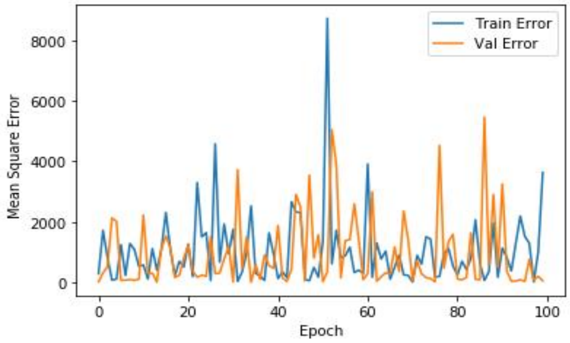


Fig. 9 showing training and validation error for both aleatoric and epistemic uncertainties model of standardised datasets

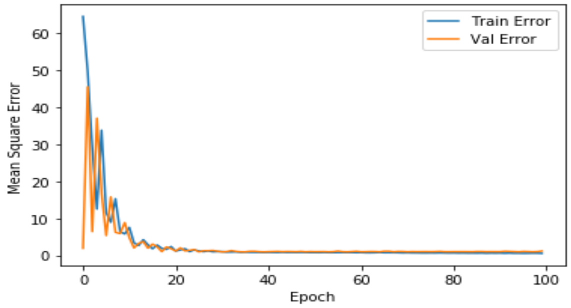


Fig. 10 showing training and validation error for frequentist model of standardised datasets

Conclusion

The study compared the frequentist and probability approaches to capture uncertainty and non-spherical disturbances inherent in the data and model. Both keras Dense layer and tfp DenseVariational layer were adopted. The underlying model were constructed probabilistically to capture aleatoric, epistemic and both. Measurement criteria were used to evaluate the performance of the underlying model. The study observed that the no uncertainty, classical and aleatoric behave well when data are standardised, the magnitude of loss, mae and mse reduced by almost 98%, this implies that the accuracy of the parameter will be certain, though epistemic and both uncertainties depict poor performance of the model despite their probabilistic nature, this may be due to combination of uncertainty with non-spherical disturbances. The outcome of the study is analogous with Keydana (2018) and Gal (2016).

REFERENCES

Alex K. and Yarin G., (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in neural information processing systems, Pp 5574–5584.

Amodei D., Sundaram A. , Rishita A., Jingliang B., Eric B., Carl C., Jared C., Bryan C., Qiang C., Guoliang C. et al. , (2016). Deep speech 2: End-to-end speech recognition in English and mandarin. In International Conference on Machine Learning, Pp 173–182.

Gal Y.(2016). “uncertainty in deep learning” *Ph.D thesis*, university of Cambridge.

Geoffrey E.H, Simon O,and Yee-Whye T,(2006). “Journal Neural Computation archive, Vol. 18(7), Pp 1527 – 1554, MIT Press Cambridge, MA, USA, doi>10.1162/neco.2006.18.7.1527

Jessica G.,(2015). “Google Photos Labelled Black People” *‘Gorillas’*, U.S.A .

Keydana (2018). TensorFlow for R: You sure? A Bayesian approach to obtaining uncertainty estimates from neural networks. Retrieved from https://blogs.rstudio.com/tensorflow/posts/2018-11-12-uncertainty_estimates_dropout.

Nicholas G.P. and Vadim O.S., (2017). Deep Learning a Bayesian Perspective, arvix:1706.00473 v4.(stat.Mc)

Pavel S., Chris S., Jacob B, Joshua V.D., and the TensorFlow Probability team (2019). Regression with Probabilistic Layers in TensorFlow Probability, https://blog.tensorflow.org/2019/03/regression-with-probabilistic-layers-in.html

Schmidhuber J.(2015). Deep Learning in neural networks: An overview. Neural networks, 61: 85-117.

Yarin G. and Zoubin G.,(2015). Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158,.