



UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

A Universal Calibration Scheme for Stochastic Processes using Artificial Neural Networks

by

Rayno Willem Mostert

*Thesis presented in partial fulfilment of the requirements for the
degree of BCommHons(Actuarial Science) in the Faculty of Economic
and Management Sciences at Stellenbosch University*

Study leaders: Mr. Stuart Reid
Mr. Stephen Burgess

2017

Declaration

By submitting this report electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: July 2017

Abstract

Contents

Declaration	i
Abstract	ii
Contents	iii
List of Figures	v
List of Tables	vii
List of abbreviations	viii
1 Introduction	1
1.1 Introduction	1
1.2 Problem Statement	2
1.3 Research Objectives	2
1.4 Importance of the Study	3
1.5 Research Design and Methodology	3
1.5.1 Neural Networks	4
1.5.2 Approximating the Calibration Function, \mathbf{C}	4
2 Literature Review	6
2.1 Calibration of Stochastic Processes	6
2.2 Neural Networks in Modelling and Model Calibration	6
3 Methodology	8
3.1 Artificial Neurons	8
3.1.1 Activation Functions	8
3.2 Convolutional Layer	8
3.2.1 Pooling	10
3.3 Training Artificial Neural Networks	10
3.3.1 Backpropagation	10
3.4 Performance Measurement	11
3.4.1 Coefficient of Determination	11

3.4.2	Average Absolute Percentage Error	11
3.4.3	Mean Squared Error	11
4	Simulation Study	12
4.1	The Merton Jump Diffusion Stochastic Process	12
4.1.1	Simulation	12
4.2	Fully Connected Neural Network	13
4.2.1	Multiple Parameter Prediction Architecture	13
4.2.2	The Dataset	13
4.2.3	Training	13
4.3	Convolutional Neural Network	13
4.3.1	Multiple Parameter Prediction Architecture	13
4.3.2	Dedicated Single Parameter Prediction Architecture	14
4.4	Sensitivity Tests	15
4.5	Conclusion	19
5	Results	21
6	Conclusion and Further Research	24
6.1	Practical Considerations	24
	Bibliography	25
A	Appendix A	28
B	Appendix B	29
B.1	Merton Jump Diffusion Stochastic Process Simulation	29
B.2	Neural Network Models	32

List of Figures

1.1	Multilayer Perceptron	4
1.2	The Proposed Calibration Scheme	5
3.1	Rectifier Activation Function, $f(x)$	9
3.2	Exponential Linear Unit (ELU)	9
3.3	1D Convolutional Layer	9
3.4	2D Convolution Layer	9
3.5	A 2×2 maximum pooling operation.	10
4.1	Multiple Output Prediction ANN	13
4.2	R-Squared values over the training process for the multiple parameter prediction fully connected architecture (ELU activation units).	14
4.3	Multiple Output Prediction CNN	14
4.4	MSE values over the training process for the 8-Layer CNN	14
4.5	R-Squared values over the training process for the multiple parameter prediction convolutional architecture (ELU activation units).	15
4.6	Dedicated Single Parameter Prediction CNN	15
4.7	R-Squared values for the predictions of λ over the training process for the single parameter prediction convolutional architecture (ELU activation units).	16
4.8	The parameter estimates (with 68% confidence interval) of $\hat{\sigma}$, $\hat{\lambda}$, $\hat{\mu}_{jumps}$ and $\hat{\sigma}_{jumps}$, plotted against their actual values ($\sigma = 0.1$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$), while varying the μ parameter in the range $(-1.0, 1.0)$	16
4.9	The deviation (with 68% confidence interval) of the $\hat{\sigma}$ parameter estimate from the actual parameter value, σ , for different values of σ . All the other parameters are kept constant as $\mu = 0.05$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$	17
4.10	The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\lambda}$, $\hat{\mu}_{jumps}$ and $\hat{\sigma}_{jumps}$, plotted against their actual values ($\mu = 0.05$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$), while varying the σ parameter in the range $(0, 0.2)$	18
4.11	The mean deviation (with 68% confidence interval) of the $\hat{\lambda}$ parameter estimate from the actual parameter value, λ , for different values of λ . All the other parameters are kept constant as $\mu = 0.05$, $\sigma = 0.1$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$	18
4.12	The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\mu}_{jumps}$ and $\hat{\sigma}_{jumps}$, plotted against their actual values ($\mu = 0.05$, $\sigma = 0.1$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$), while varying the λ parameter in the range $(0, 0.025)$	19

4.13	The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\mu}_{jumps}$ and λ , plotted against their actual values ($\mu = 0.05$, $\sigma = 0.1$, $\mu_{jumps} = 0.05$ and $\lambda = 0.02$), while varying the σ_{jumps} parameter in the range (0,0.2).	19
4.14	The mean deviation (with 68% confidence interval) of the $\hat{\mu}_{jumps}$ parameter estimate from the actual parameter value, μ_{jumps} , for different values of μ_{jumps} . All the other parameters are kept constant as $\mu = 0.05$, $\sigma = 0.1$, $\lambda = 0.02$ and $\sigma_{jumps} = 0.07$	20
4.15	The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\sigma}_{jumps}$ and λ , plotted against their actual values ($\mu = 0.05$, $\sigma = 0.1$, $\lambda = 0.02$ and $\sigma_{jumps} = 0.07$), while varying the μ_{jumps} parameter in the range (-0.5,0.5).	20
5.1	Various model distributions of the predicted values of μ with true value 0,05.	21
5.2	Various model distributions of the predicted values of σ with true value 0,1.	21
5.3	Various model distributions of the predicted values of λ with true value 0,02.	22
5.4	Various model distributions of the predicted values of μ_{jumps} with true value 0,05.	22
5.5	Various model distributions of the predicted values of σ_{jumps} with true value 0,07.	23

List of Tables

List of abbreviations and/or acronyms

ANN artificial neural network. v, 1–5, 7–9, 11, 14, 16, 17, 22

CNN convolutional neural network. v, 9, 11, 13, 14, 16, 19, 22–24

ELU exponential linear unit. v, 9, 10, 14–16, 22–24

EM expectation maximisation. 7

MLE maximum likelihood estimation. 1, 3, 7

MME generalised method of moments estimation. 1, 3, 7

ReLU rectified linear unit. 9

SDE stochastic differential equation. 1–3, 5, 7, 8

CHAPTER 1

INTRODUCTION

1.1. INTRODUCTION

Stochastic processes are becoming more important to actuaries: they underlie much of modern finance, mortality analysis and general insurance. They are immensely useful because they form the common language of workers in many areas that overlap in actuarial science. It is precisely because most financial and insurance risks involve events unfolding as time passes that models based on processes turn out to be most natural.

— Submission to the Faculty of Actuaries students' society in 1998 (Cairns, Dickson, Macdonald, Waters & Willder; 1998)

Stochastic processes are simply a collection of random variables, usually indexed by time (Barone-Adesi, 2015). They are typically used during the modelling process, in order to describe the evolution of an underlying real-world process. To allow for it to be used within the modelling context, a stochastic process is often expressed by its stochastic differential equation (SDE). Using the SDE, the simulated stochastic process can then be adjusted to best represent the real-world process at hand - whether that be the evolution of the price of a certain stock, the claims on an insurance policy or the mortality rate of a group of policyholders. Oreskes, Shrader-Frechette & Belitz (1994) refer to this procedure of "[manipulating] the independent variables to obtain a match between the observed and simulated distribution or distributions of a dependent variable or variables", as model calibration.

Effectively modelling a real-world process involves two major challenges. Firstly, choosing an appropriate stochastic process with properties that mimic those of the real-world process; and, secondly, finding the most suitable parameters for the relevant SDE. In practice, however, the selection of an appropriate stochastic process is often influenced by the ease with which its parameters can be calibrated. Thus complex stochastic processes, with more complex SDEs, are often substituted for simpler, easily calibrated models. This can lead to the use of models that are subject to simplifying assumptions or possess properties that may not be the best possible representation of reality.

The difficulty associated with model calibration depends on the method of calibration applied. Mongwe (2015) and Honore (1998) describe how common calibration methods, including maximum likelihood estimation (MLE) and the generalised method of moments estimation (MME), can break down under more complex SDEs. MLE, for example, requires the derivation of the likelihood function, which is often difficult in the case of complex SDEs as they can yield unbounded likelihood functions.

Another calibration method that has seen a renaissance in the last decade is that of backpropagation, which has proven to be a powerful gradient descent-based algorithm for calibrating artificial neural networks (ANNs). This paper will explore the ways in which these statistical learning tech-

niques - namely ANNs, calibrated by backpropagation - could, in turn, be applied to the calibration of stochastic processes.

1.2. PROBLEM STATEMENT

Every stochastic process contains a set of parameters, Z , which controls the dynamics of the paths produced by the model. The calibration problem can be framed as a mapping from the observed data, D , or some transformation thereof, to these parameter values, Z . Let C denote the calibration method, then C is necessarily of the form,

$$C : f(D) \rightarrow Z \quad (1.1)$$

An ANN is a collection of interconnected processing units (Teugels & Sundt, 2004), which realises a nonlinear mapping from inputs, R^X to outputs, R^Y .

$$ANN : R^X \rightarrow R^Y \quad (1.2)$$

This mapping is achieved by chaining a sequence of nonlinear multiple regression functions, f , (called activation functions) together in layers (see 1.5.1 below).

From equations (1.1) and (1.2), we can see that an ANN has the ability to estimate the calibration function, C . This assertion is justified by the Universal Approximation Theorem, which states that "standard multilayer feedforward networks are capable of approximating any measurable function to any desired degree of accuracy" (Hornik, Stinchcombe & White, 1989).

The question, however, remains as to what such a network might look like and how it would compare to traditional calibration techniques.

1.3. RESEARCH OBJECTIVES

This paper has the primary objective of researching and testing the viability of ANNs as a universal method of parameter estimation for any stochastic process. As to be seen in the literature review, ANNs have been used extensively in the world of financial modelling, and to some extent in the calibration of simple stochastic processes.

Conceivably, ANNs have the potential to act as a universal calibration method for any stochastic process. This paper aims to investigate whether and how that might function in practice.

This study will consist of two phases. In the first, an ANN will be implemented for a sufficiently complex SDE, for which a likelihood function does exist (for comparative purposes). In the second phase, the accuracy of the network's approximation of the calibration function will be measured and compared against that of other calibration techniques, and an empirical study will attempt to ascertain whether the resulting calibration scheme is robust on real financial data.

1.4. IMPORTANCE OF THE STUDY

An ANN-based approach to parameter estimation could likely provide numerous benefits above the popular MLE and MME approaches. The technique could potentially provide a universal solution to approximate the calibration function, C , for any arbitrarily complex SDE. It does not require the derivation of the likelihood function, which can be difficult. The model drops some of the strong assumptions made by MLE and MME.

Another possible advantage could lie in the scheme's ability to combine the predictive power of ANNs with the descriptive properties of certain stochastic processes. Olden & Jackson (2002) explain that "although in many studies ANNs have been shown to exhibit superior predictive power compared to traditional approaches, they have also been labelled a 'black box' because they provide little explanatory insight into the relative influence of the independent variables in the prediction process". The approach, whereby the ANN is used only to calibrate a more expressive and widely understood model - a stochastic process - might help to remedy this.

Mongwe (2015) presents the example of the Merton jump diffusion process with SDE,

$$d \ln S_t = \left(\mu - \frac{1}{2} \sigma^2 \right) dt + \sigma dB + d \left(\sum_{i=1}^{N_t} Y_i \right) \quad (1.3)$$

where μ is referred to as the drift coefficient and σ as the diffusion coefficient. $B_t, t \geq 0$ is a standard Brownian motion process. Y_i represents the random size of the i th jump, and has distribution, $Y_i \sim N(\mu_{jump}, \sigma_{jump}^2)$. $N_t, t \geq 0$ is a Poisson process with intensity λ .

Note that these parameters provide insight into the characteristics of the stochastic process being modelled. Hence, it could arguably be more enlightening to fit the observed data to an expressive stochastic process which yields explanatory parameters, than simply using a "non-parametric" ANN to model the real-world process entirely. This helps avoid the black-box pitfall commonly associated with ANNs, while still making use of their "superior predictive power". This is of importance, as prudent financial management involves using modelling techniques that are well understood, clearly defined and well documented. ANNs - despite their powerful properties - are often not an acceptable means of modelling in actuarial applications. By using them simply to calibrate complex stochastic processes (which are a suitable and widely acceptable modelling tool), the strengths of ANNs are retained, without the risk associated with a black-box technique.

1.5. RESEARCH DESIGN AND METHODOLOGY

The research design and methodology will give an overview of the model construction process. Firstly, a more thorough description of one of the core components of the scheme - an ANN - is presented. This is followed by an explanation as to how an ANN architecture could be applied to estimating the model calibration function. Lastly, an overview of how the model is to be evaluated follows.

1.5.1 Neural Networks

An Artificial Neural Network is a mathematical model consisting of an interconnected collection of processing units, which realises a nonlinear mapping from inputs R^X to outputs R^Y ,

$$ANN : R^X \rightarrow R^Y \quad (1.4)$$

This mapping is achieved by chaining a sequence of nonlinear multiple regression functions, f , (called activation functions) together in layers. Each input into every activation function is weighted by some value w .

The most common ANN architecture, a multilayer Perceptron, is illustrated in figure 1.1.

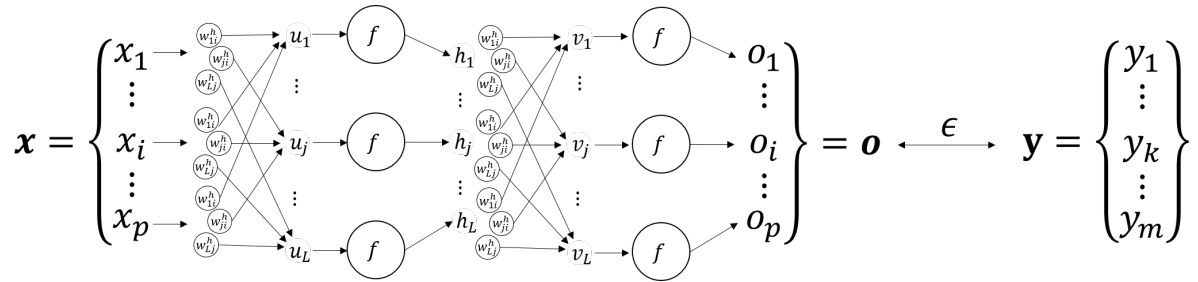


Figure 1.1: Multilayer Perceptron

For a given input, x , and (expected) output, y , the error of the ANN, ϵ , is equal to the distance between the ANN's outputs, o , and the expected outputs. The power of ANNs lies in the fact that they can be trained to minimize this error.

Training the network involves firstly initialising the network with a random set of weights, W . A large set of training data (sets of inputs, X , and desired outputs, Y) is then presented to the network. The optimisation process (often referred to as backpropagation) proceeds by calculating the prediction error, ϵ , for each of these data sets, and then "propagating" this error value backwards through the network so that each weight can be adjusted accordingly. This is achieved by means of automatic differentiation. Automatic differentiation allows us to compute the partial derivative of the error with respect to the weights in the ANN, W (Werbos, 1990).

This process is repeated iteratively over the training data, until the error of the ANN converges or some other stopping criteria is satisfied. At this point, the ANN will have approximated the relation, $\mathbf{X} \rightarrow \mathbf{Y}$.

A number of technical details - such as what inputs are fed into the ANN, what activation function is used, the number of activation functions used, the number of layers used, and the exact function used to estimate the error - have been omitted from this discussion for the sake of brevity.

1.5.2 Approximating the Calibration Function, C

From the previous section, it follows that an ANN should - theoretically - be able to approximate the calibration function defined earlier, $C : f(D) \rightarrow Z$. That is, $NN \approx C$. This can be achieved by:

- i) generating a set of random calibrations, Z

- ii) simulating a set of paths, \mathbf{Z} , using the given SDE for every $Z_i \in \mathbf{Z}$
- iii) extracting a set of "inputs", \mathbf{X} , from the data $\mathbf{X} = f(\mathbf{D})$
- iv) training the ANN to predict the original calibrations, \mathbf{Z} , given the data $\mathbf{X} = f(\mathbf{D})$ as input. This is done by setting the ANN to minimise the error ϵ of its output, \mathbf{O} (which corresponds to an estimate of Z_i) and the true known values of Z_i .

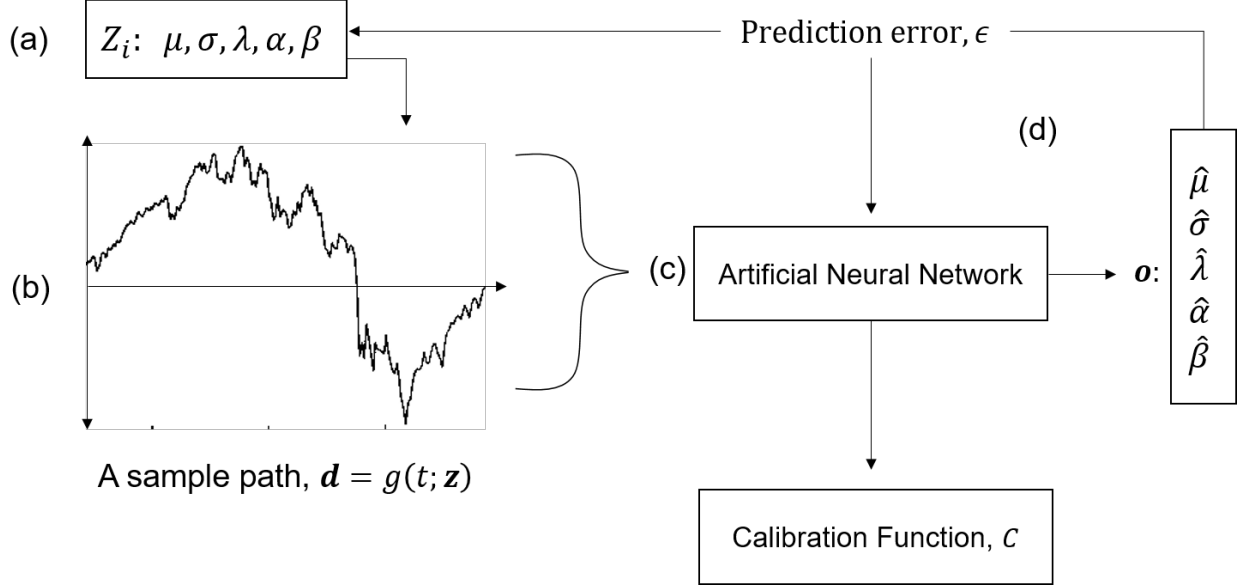


Figure 1.2: The Proposed Calibration Scheme

If a large enough set of calibrations is chosen and enough data is simulated from the SDE using each calibration, the ANN will approximate the desired calibration function, C , for the given SDE.

This trained ANN can then be applied to real-world observations to calibrate the relevant SDE for the observed process. This is done by inputting the data observed from the real-world stochastic process into the trained ANN. The ANN output, \mathbf{o}' , will then be an estimate of the parameters (i.e. a calibration) for the SDE that the ANN was trained on.

In this paper, the ANN was built and trained in Python, using Google's open source Tensorflow™ (Abadi *et al.*, 2015) and Keras (Chollet *et al.*, 2015) libraries. All model training was done using the Adam optimizer (Kingma & Ba, 2014).

A full list of software and packages used in this dissertation is available in appendix A.

CHAPTER 2

LITERATURE REVIEW

The literature discussed throughout this research will fall into mainly two broad categories. Firstly, a review of the common methods proposed for the calibration of stochastic processes, focusing on their limitations, ease of use, universality and accuracy. Secondly, an overview of past applications of ANNs to modelling and model calibration.

2.1. CALIBRATION OF STOCHASTIC PROCESSES

The main argument for a universal neural network approach to model calibration is that traditional methods are often impractical. Numerous academic works substantiate this observation.

Nielsen, Madsen & Young (2000) reviewed the progress made on SDE parameter estimation over the 80s and 90s. They note that the MLE approach does not generalise and, having studied the MME and the efficient method of moments, they explain that both of these methods will result in tests of low power due to the efficiency loss.

More recently, Mongwe (2015) did a study on jump diffusion processes applied to the South African equity and interest rate markets. He reported on the application and accuracy of multiple calibration methods, including the likelihood profiling approach (Honore, 1998), the standard MLE approach, the MLE approach and expectation maximisation (EM). He concluded that both MLE and MME fell short of expectations, and that the likelihood profiling and EM techniques worked best on parameter estimation for jump diffusion processes (each under different restrictions on the parameters) (Mongwe, 2015). These recommendations will be applied when evaluating and comparing the performance of the neural-network calibration approach in this paper to that of existing methods.

2.2. NEURAL NETWORKS IN MODELLING AND MODEL CALIBRATION

Except for the case of Xie, Kulasiri, Samarasinghe & Rajanayaka (2007), the task of calibrating stochastic processes using ANNs has not been thoroughly attempted or documented in the major academic journals examined for this research. Numerous works do however highlight the potential of ANNs in this field. Multiple studies have been done on the use of ANNs in pricing options, and - in particular - to outperform the Black-Scholes model (Yao, Li & Tan; 2000). The results seem to indicate that ANNs outperform the Black-Scholes model in volatile markets, and are particularly useful in these when the "constant σ " assumption underlying the Black-Scholes model is violated.

Tackling the issue of model calibration; Samad & Mathur (1992) investigated the application of ANNs to the calibration of process systems - namely that of first-order process open-loop delay identification. They concluded that ANNs are an attractive solution, as they do not require the subject-specific expertise vital to common engineering approaches, provide high accuracy and prove robust on real-world data.

On the topic of this paper, Xie *et al.* (2007) did an investigation into the feasibility of estimating the parameters both linear and nonlinear SDEs using multilayer perceptron (MLP) networks. Their investigation comprised only of small MLPs with 1-, 2- and 3-hidden-layer, fully connected architectures. They found that, under certain conditions limiting the parameter values of the process, a simple MLP would be able to estimate parameters with high accuracy ($R^2 > 0.93$). They report that this accuracy figure, however, does diminish under noisy conditions and SDEs with high diffusion levels. Another, often overlooked element noted by their research is the importance of the regime used to generate the simulated training data. The paper indicates to a notable increase in accuracy by using a simulation regime that makes use of the same parameters over 5 different Wiener processes, which effectively helps remove the "randomness" and noise from the dataset. They called for more research on the subject, particularly using different ANN architectures. What is notable about the paper presented, is that high accuracy was achieved using a simple network topology. This serves as evidence for the potential of ANNs to act as robust calibration estimators for SDEs.

Giebel & Rainer (2013) proposed a novel calibration method for time series that dynamically adapts the parameters of a stochastic model by using small MLP networks (2-layer). These ANNs use data from the past n observations to inform an updated parameter value. They then used the updated stochastic process to forecast the time series one day ahead, while updating the parameters at each time step as they proceed through the time series. They argue that updating the parameters of the stochastic process at each time step is more realistic, due to investors often weighting recent observations as more relevant than aged data. Using this method, different weights can be assigned to data from different dates in the past. The calibration scheme proposed in this paper could potentially add great value to the technique described by Giebel and Rainer, as it would allow the periodic recalibration of more complex processes, and the incorporation of many additional parameters.

CHAPTER 3

METHODOLOGY

3.1. ARTIFICIAL NEURONS

In Chapter 1, ANNs were introduced. Figure 1.1 introduces 3 types of trainable neurons: weights (denoted by w), biases (u), and activation functions (f). Hence, every individual neuron multiplies the set of inputs, x_i by their corresponding weights w_i and adds some bias, u - the sum of which is then fed through an activation function f . The result is propagated forward in the network to the next set of neurons. Mathematically, the output of the k th neuron is given by:

$$\mathbf{o}_k = f \left(\sum_i [\mathbf{w}_{ki} \mathbf{x}_i] + \mathbf{u}_k \right) \quad (3.1)$$

3.1.1 Activation Functions

The activation function f is used to add non-linearity to the architecture. Without it, the model would be little more than a multivariate linear model. There are a number of choices for this activation function. The most popular being rectified linear units (ReLUs), exponential linear units (ELUs), logistic sigmoid- and hyperbolic tangent functions.

Prior to the work of Glorot, Bordes & Bengio (2011), logistic sigmoid and hyperbolic tangent functions were the commonest activation functions in neural network architectures. However, Glorot *et al.* (2011) showed that ReLUs yield better performance. ReLUs employ the activation function:

$$f(x) = \max(0, x) \quad (3.2)$$

Clevert, Unterthiner & Hochreiter (2015) introduced the "exponential linear unit", which provide even better performance than ReLUs, both in terms of learning speed as well as generalisation potential. An ELU with parameter $\alpha > 0$ has activation function:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha (\exp(x) - 1) & \text{if } x \leq 0 \end{cases} \quad (3.3)$$

3.2. CONVOLUTIONAL LAYER

Convolutional Neural Networks (CNN's) are a variation of the traditional multi-layer perceptron architecture. Like ordinary fully connected networks, they consist of neurons with trainable weights and biases. Unlike fully connected networks, these neurons are grouped into sets of small filters. With every forward pass, the filters are sequentially convolved across the extend of the input volume. This produces an activation map containing the result of the "filtration" at every point of the input volume.

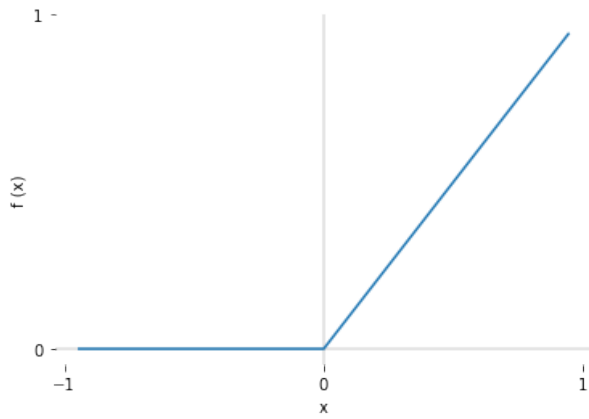


Figure 3.1: Rectifier Activation Function (ReLU), $f(x)$.

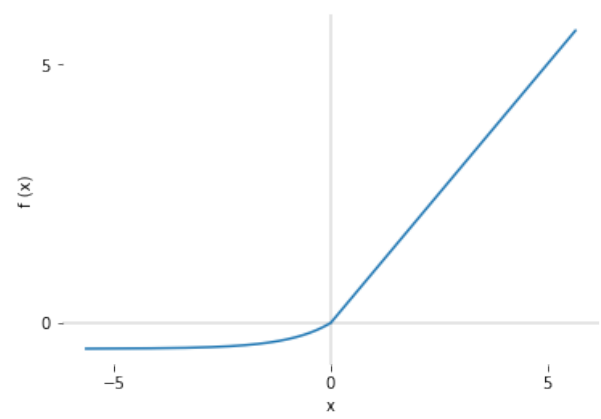


Figure 3.2: Exponential Linear Unit (ELU) Activation Function, with $\alpha = 0.5$.

In an image-recognition setting, the network will typically train filters that activate on the detection of visual features such as edges or a blotch of colour (cs2). However, in this study, there is no definite way to ascertain what they might detect.

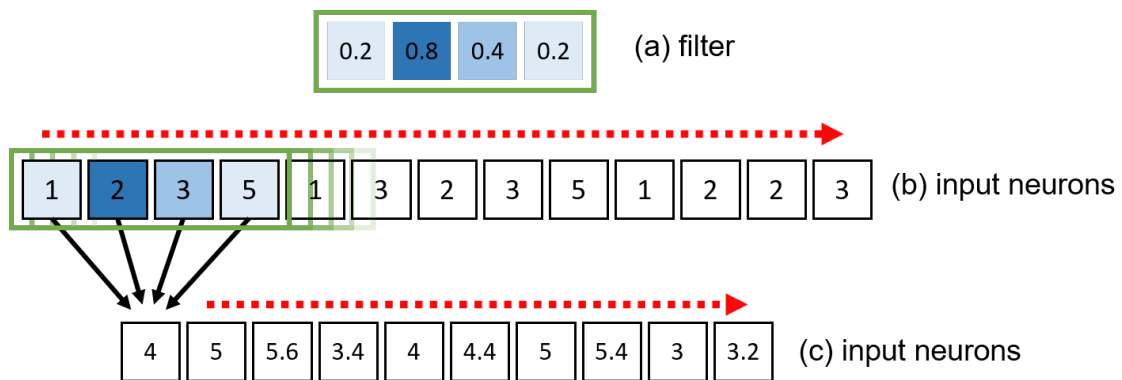


Figure 3.3: A One-Dimensional Convolution Layer

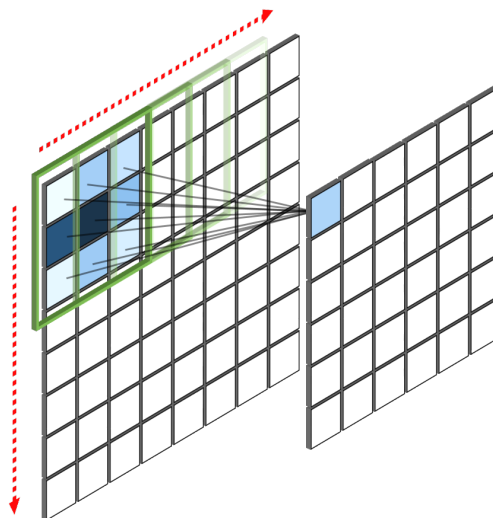


Figure 3.4: A Two-Dimensional Convolution Layer

3.2.1 Pooling

CNN architectures often feature pooling layers, which aggregate the outputs of multiple preceding neurons into a single feature, which is then propagated forward through the network. Two prevalent pooling operations are subsampling (where the mean of the preceding neuron outputs are transmitted to the subsequent layer), as well as maximum pooling - which propagates the maximum output from a set of preceding neuron outputs, to the next layer. The empirical results of Scherer, Müller & Behnke (2010) show that "a maximum pooling operation significantly outperforms subsampling operations". In the convolutional network implementations from this dissertation, extensive use was made of maximum pooling operations. Note that - unlike the filters of a convolutional layer - the weights of an average- or maximum pooling operation cannot be adjusted during training, and hence the pooling layers do not form part of the set of trainable network layers.

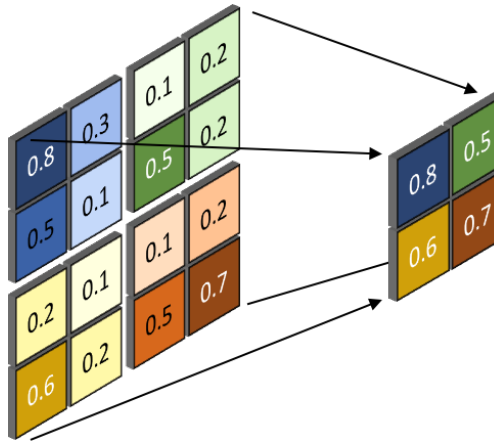


Figure 3.5: A 2×2 maximum pooling operation.

3.3. TRAINING ARTIFICIAL NEURAL NETWORKS

Before any supervised learning approach can be utilised for inference, it needs to undergo an optimisation process known as "training". Training involves adjusting the appropriate parameters of the network (often referred to as the trainable parameters), in order to minimise the error between the output of the model, and the desired output. The error is defined by some selected loss function, L . The artificial neurons introduced in section 3.1 are "trained" using a technique known as backpropagation.

3.3.1 Backpropagation

Backpropagation consists of two main steps: a propagation step and a parameter adjustment step. In turn, the propagation step consists of both forward- and backward passes.

In the forward propagation step, the inputs are fed through the model to produce a set of output activations, \mathbf{o}' . These outputs are then compared to the desired outputs \mathbf{y}' , and an error value ϵ' is obtained using some loss function L . That is, $\epsilon' = L(\mathbf{o}', \mathbf{y}')$. These errors are then propagated backwards through the network to generate a set of gradients for each artificial neuron. What

follows is the process of gradient descent: the gradient defines a direction (increase or decrease) as well as a magnitude in which to adjust the individual parameter (usually a weight or bias) in order to minimise the error, ϵ' . In the parameter adjustment step, the trainable parameters of the model are each adjusted in the opposite direction of the gradient defined for that parameter, so as to minimise this error.

These steps are repeated until some stopping criteria is reached.

3.4. PERFORMANCE MEASUREMENT

3.4.1 Coefficient of Determination

In measuring the performance of the ANN, we will use the the coefficient of multiple determinations, R^2 , between the actual parameter values, y , and the ANN-predicted parameter values \hat{y} . R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (y_i - \bar{y})^2} \quad (3.4)$$

where y is the actual parameter value, \hat{y} is the predicted parameter value, \bar{y} is the mean parameter value, and m is the size of the sample. Any estimate that is more accurate than the sample mean would result in an R^2 value of greater than zero. An R^2 value of 1 would indicate a perfect fit.

3.4.2 Average Absolute Percentage Error

Another model evaluation metric is the average absolute percentage error (AAPE).

$$AAPE = 100 \cdot \frac{1}{m} \sum_{i=1}^m \frac{|y_i - \hat{y}_i|}{y_i} \quad (3.5)$$

where y is the actual parameter value, \hat{y} is the predicted parameter value, and m is the size of the sample.

3.4.3 Mean Squared Error

A natural loss function to consider in the optimisation procedure concerned in this dissertation is that of *mean squared error* (MSE). MSE measures the squared mean deviation of the predicted values (yielded by the model under consideration) from the actual observed values.

$$MSE = \sum_{i=1}^m (y_i - \hat{y}_i)^2 \quad (3.6)$$

where y is the actual observed parameter (output) value, \hat{y} is the predicted parameter (output) value, and m is the size of the sample.

CHAPTER 4

SIMULATION STUDY

It is important to note that when building a neural network for a specific modelling exercise, one has little prior knowledge of what the model should look like. Rough guidelines do exist, for example that Recurrent Neural Network architectures are often used for time series problems (SOURCE?) or that convolutional neural networks (CNNs) are well suited to image recognition tasks (SOURCE?). Beyond these vague guidelines however, little evidence exists to inform the potential properties that a network might need.

4.1. THE MERTON JUMP DIFFUSION STOCHASTIC PROCESS

The Merton Jump Diffusion Stochastic process, presented in the seminal work of Merton (1976), aimed to address the limitations of the Geometric Brownian Motion process. It has the stochastic differential equation,

$$dS_t = \mu S_t dt + \sigma S_t dW_t + S_t dJ_t \quad (4.1)$$

where

$$J_t = \sum_{j=1}^{N_t} (V_j - 1) \quad (4.2)$$

is a compound Poisson process. V_j are independent, identically distributed positive random variables representing the jump sizes. $N_t, t \geq 0$ is a Poisson process with intensity λ , which is independent of J_t and W_t .

To obtain the log returns, we can derive the function $f(t, S_t) = \ln S_t$ using Itô's formula:

$$d \ln S_t = \frac{1}{S_t} dS_t - \frac{1}{2S_t^2} (dS_t)^2 = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma dW_t + dJ_t \quad (4.3)$$

In this dissertation we will have V_j follow a log-normal distribution with parameters μ_{jumps} and σ_{jumps} .

4.1.1 Simulation

The neural network models will be trained on the log-returns of the simulated processes. These processes will be simulated using a random set of parameters (within certain bounds).

The simulated parameters, μ , σ , λ , μ_{jumps} and σ_{jumps} will be constrained to the following bounds: $\mu \in [-1, 1]$, $\sigma \in [0.001, 0.2]$, $\lambda \in [0.0001, 0.025]$, $\sigma_{jumps} \in [0.001, 0.2]$, and $\mu_{jumps} \in [-0.5, 0.5]$.

4.2. FULLY CONNECTED NEURAL NETWORK

4.2.1 Multiple Parameter Prediction Architecture

The first study was done on a standard 9-layer fully connected multi-perceptron architecture, trained to predict all five parameters of the Merton Jump Diffusion process at once.

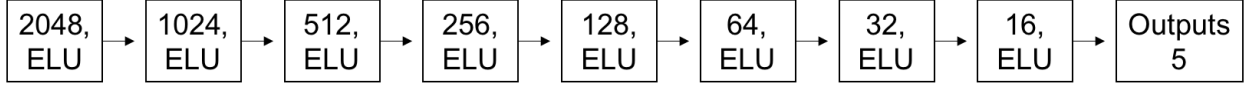


Figure 4.1: 9-Layer Fully Connected Feed-forward ANN

The experiments were performed using ELU activation functions in the network.

4.2.2 The Dataset

The dataset on which the Fully Connected ANN was trained, was created using the following steps:

- 1) Random sets of parameters, \mathbf{z}_i , with $(\mu \in [-1, 1], \sigma \in [0.001, 0.2], \lambda \in (0, 0.025], \mu_{jumps} \in [-0.5, 0.5]$ and $\sigma_{jumps} \in [0.001, 0.2])$ were uniformly generated.
- 2) The daily log returns from a Merton Jump Diffusion stochastic process (equation 4.3) were simulated using these randomly generated parameter sets, \mathbf{z}_i , as parameters.
- 3) For each set of returns, the first 20 sample moments, as well as the autocorrelations up to the first 40 lags were calculated.
- 4) For every parameter set, its corresponding set of moments and autocorrelations were fed into the input layer of the ANN described in subsection 4.2.1 above. The ANN was then trained, using the process of backpropagation, to produce an estimate, \mathbf{o}_i , of the original set of parameters, \mathbf{z}_i . This was done by minimizing the mean squared error between the vectors \mathbf{z}_i and \mathbf{o}_i (see the diagram in figure 1.2).

4.2.3 Training

Figures 4.2a and 4.2b illustrate the R-Squared values throughout the training process of the fully connected architecture.

4.3. CONVOLUTIONAL NEURAL NETWORK

4.3.1 Multiple Parameter Prediction Architecture

The second study was done on a fairly standard 8-layer convolutional architecture, which produced estimates for all five parameters.

Experiments were performed using ELU activation functions.

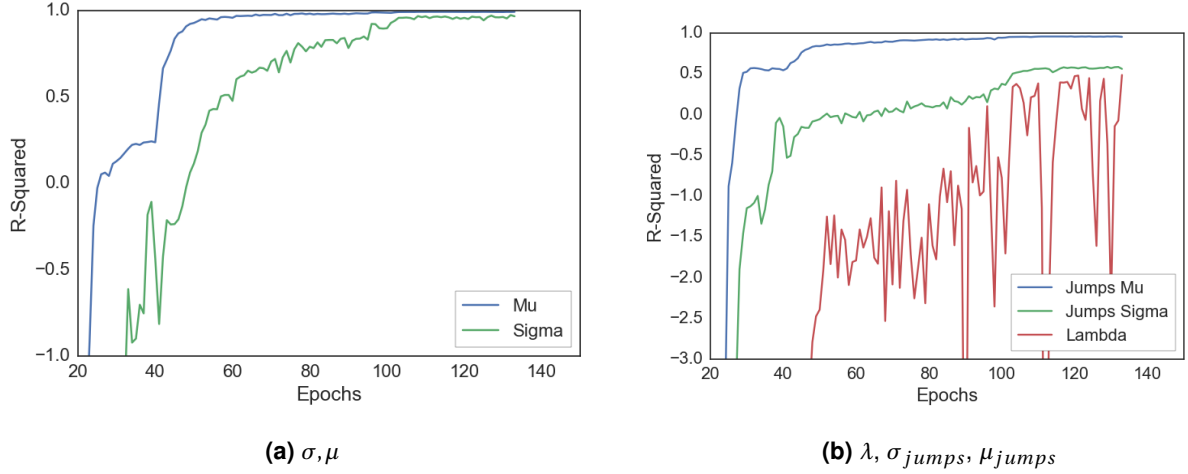


Figure 4.2: R-Squared values over the training process for the multiple parameter prediction fully connected architecture (ELU activation units).

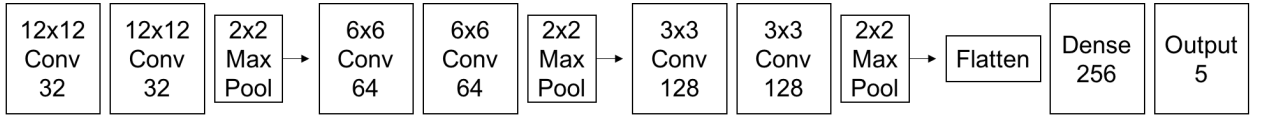


Figure 4.3: 8-Layer CNN

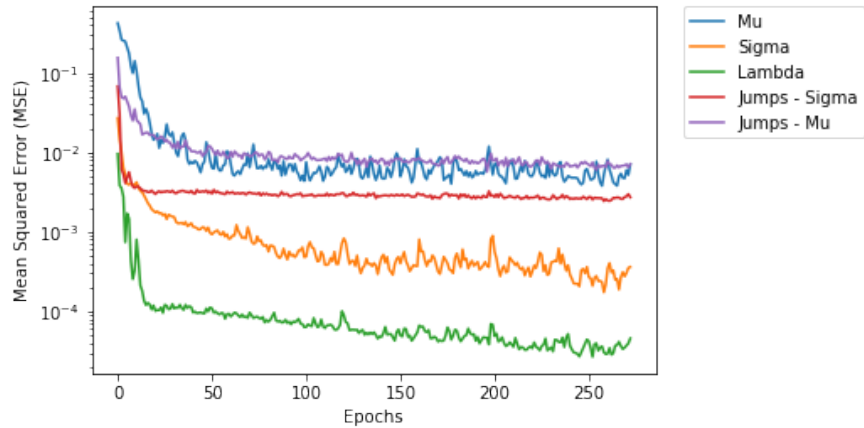


Figure 4.4: MSE values over the training process for the multiple parameter prediction convolutional architecture. Each epoch involves 10 iterations of a simulated batch of 150 randomly selected process parameters.

The reasons for the large difference in MSE values are related to the nature of the parameters. For example μ (μ) - the drift of the Geometric Brownian Motion component of the process - can take on values between -0.5 and 0.5 , while λ_{jumps} (λ_{jumps}) - the probability of a jump occurring at any given point, can only take on values between 0 and 0.003 .

4.3.2 Dedicated Single Parameter Prediction Architecture

In training the multiple output model, one might notice a slight oscillation in the accuracy of the parameters. At higher levels of accuracy, as the network becomes more accurate at predicting one output, it might become less accurate for another. There seems to exist a payoff, whereby the

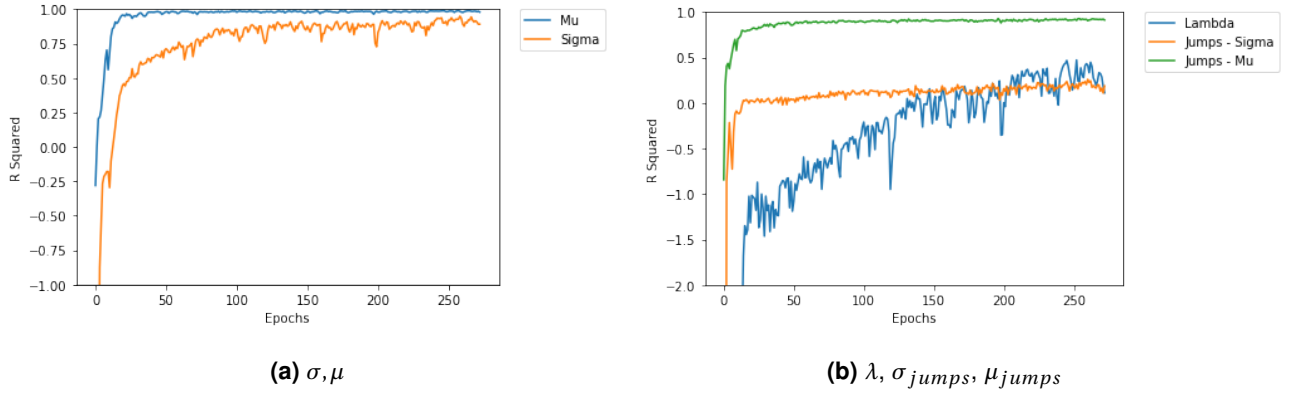


Figure 4.5: R-Squared values over the training process for the multiple parameter prediction convolutional architecture (ELU activation units).

accuracy reduces as the prediction accuracy for another increases.

This raises the question of a dedicated network architecture for every parameter. A slight variation on the architecture used in 4.3.1 above was implemented. The structure was scaled down to six trainable layers, with a larger penultimate layer. The architecture only outputs an estimate for a single parameter. Hence, using this architecture, separate networks will be used to individually estimate the values of each of the parameters, μ , σ , λ , μ_{jumps} and σ_{jumps} .

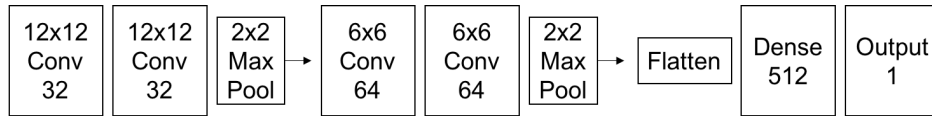


Figure 4.6: 6-Layer CNN

The following sections will discuss the convergence of the dedicated models in comparison to the multiple output prediction model. The final prediction accuracy of the model will be discussed in chapter 5.

Lambda

The most obvious parameter estimation issue in the multiple output model of 4.3.1, exists with λ , which - as clearly visible in figure 4.5b - exhibits a reluctance to converge to a desired level of accuracy. The dedicated single architecture defined in figure 4.6 was implemented and trained to predict only the single parameter value λ per sample path.

The result is a much quicker and smoother convergence to an acceptable level of accuracy, as visible in figure 4.7. Compare this to the very rough convergence of λ in figure 4.5b. Quicker convergence leads to less training time and ultimately easier use of the model.

4.4. SENSITIVITY TESTS

Due to the nature of the parameters involved in the Merton Jump Diffusion process, one could expect interactions between parameter estimates. For example, increasing the μ_{jumps} parameter might cause "confusion", since a model could "interpret" larger jumps as a higher rate of volatility

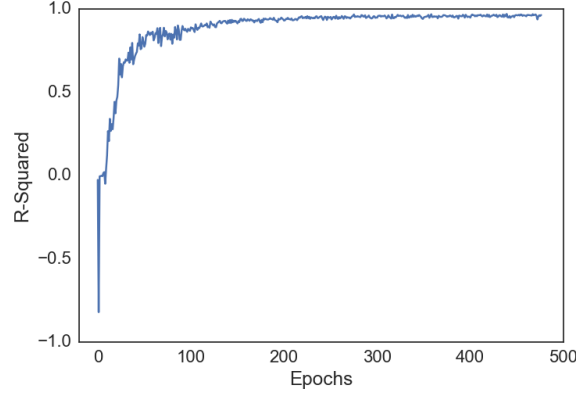
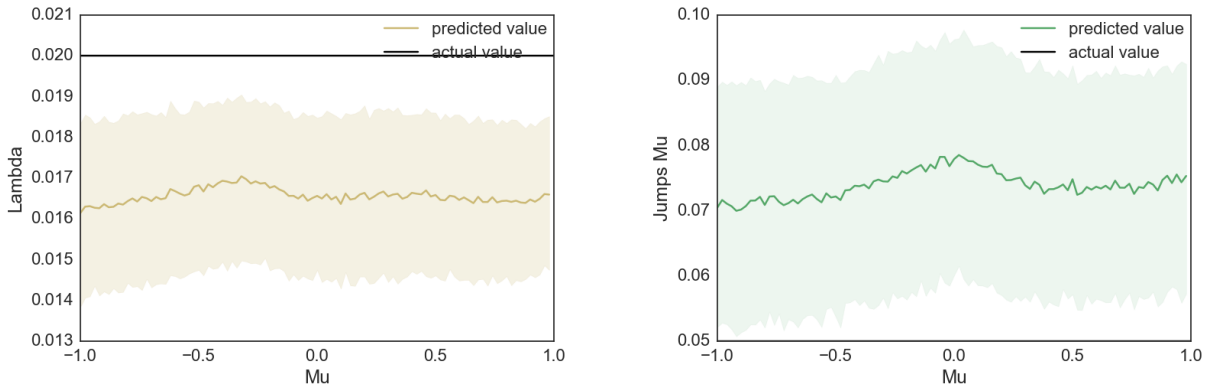


Figure 4.7: R-Squared values for the estimates of λ over the training process for the single parameter prediction convolutional architecture (ELU activation units).

from the Geometric Brownian Motion σ component, and hence produce higher values of σ . What follows is an investigation into how sensitive the individual parameter estimates are with respect to changes in the magnitudes of the other parameters.

This investigation was performed using 1000 simulated sample paths from a Merton Jump Diffusion process with parameters $\mu, \sigma, \lambda, \mu_{jumps}$ and σ_{jumps} . With each test, a single parameter was selected to be varied during the training process. The effect of this parameter change on the ANN performance was then monitored by reading the estimates of each of the other parameters and comparing them to their actual values.

Variations in μ



(a) The λ parameter estimate plotted against the actual parameter value, $\lambda = 0.02$, for different values of μ .

(b) The $\hat{\mu}_{jumps}$ parameter estimate plotted against the actual parameter value, $\mu_{jumps} = 0.05$, for different values of μ .

Figure 4.8: The parameter estimates (with 68% confidence interval) of $\hat{\sigma}$, $\hat{\lambda}$, $\hat{\mu}_{jumps}$ and $\hat{\sigma}_{jumps}$, plotted against their actual values ($\sigma = 0.1$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$), while varying the μ parameter in the range $(-1.0, 1.0)$.

Of all the parameters involved in the Merton Jump Diffusion process, μ arguably has the least interaction with the estimates of the other parameters.

It is rather strange that the model exhibits a tendency to consistently underestimate the value of λ (Figure 4.8a). This is clear throughout the investigation - λ is almost always underestimated. This isn't something to be expected, since ANNs have the property of easily being able to correct for bias. Similarly, all other parameters (besides μ) kept constant, the model seems to consistently overestimate the value of μ_{jump} (Figure 4.8b). As with the estimate of λ , one would expect and ANN to be able to easily correct for this clear bias.

Variations in σ

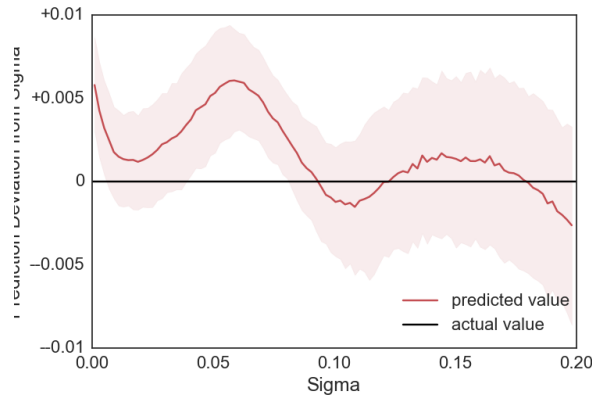


Figure 4.9: The deviation (with 68% confidence interval) of the $\hat{\sigma}$ parameter estimate from the actual parameter value, σ , for different values of σ . All the other parameters are kept constant as $\mu = 0.05$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$.

As to be expected, larger volatility tends to widen the confidence interval around a particular parameter estimate. This is particularly clear in figure 4.9, where the confidence regarding the estimate of $\hat{\sigma}$ falls, as σ increases. The most clear-cut illustration of this can be seen in the estimates of $\hat{\mu}$ in figure 4.10a, where there is a clearly visible widening of the confidence interval as sigma increases.

Variations in λ

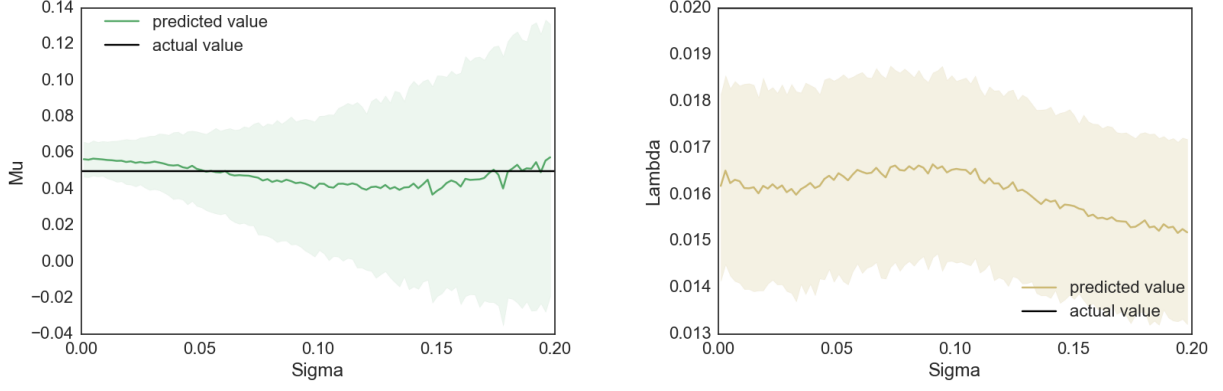
The multiple output prediction model shows a slight overestimation of $\hat{\lambda}$ for small values of λ , and a slight underestimation for larger values of λ (Figure 4.11)

Figure 4.12a shows an interesting relationship between the value of $\hat{\sigma}_{jumps}$ and λ . The model estimate of $\hat{\sigma}_{jumps}$ shows an inversely proportional relationship to the value of λ . Larger λ also seems to make the model "less confident" in its estimate of σ_{jumps} .

In figure 4.12b, we see that the opposite is true for the estimate of $\hat{\mu}_{jumps}$. As might be expected, small λ values tend to suffocate the estimate of $\hat{\mu}_{jumps}$. It is conceivable that lower values of λ might result in less jumps, which could be "misinterpreted" as a lower mean jump size.

Variations in σ_{jumps}

The value of σ_{jumps} has notable effects on the errors of the estimates of $\hat{\mu}_{jumps}$ and $\hat{\lambda}$.



(a) The mean $\hat{\mu}$ parameter estimate plotted against the actual parameter value, $\mu = 0.05$, for different values of σ . (b) The mean $\hat{\lambda}$ parameter estimate plotted against the actual parameter value, $\lambda = 0.02$, for different values of σ .

Figure 4.10: The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\lambda}$, $\hat{\mu}_{jumps}$ and $\hat{\sigma}_{jumps}$, plotted against their actual values ($\mu = 0.05$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$), while varying the σ parameter in the range (0,0.2).

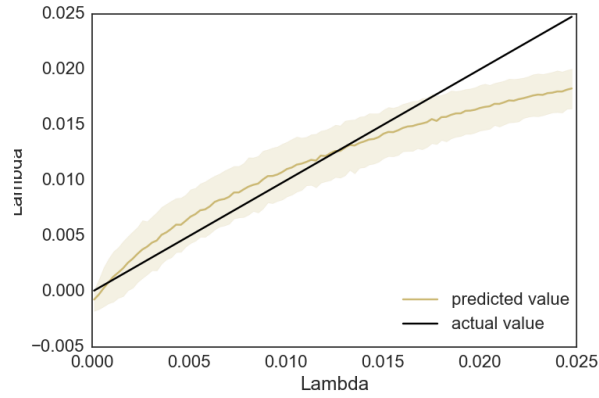


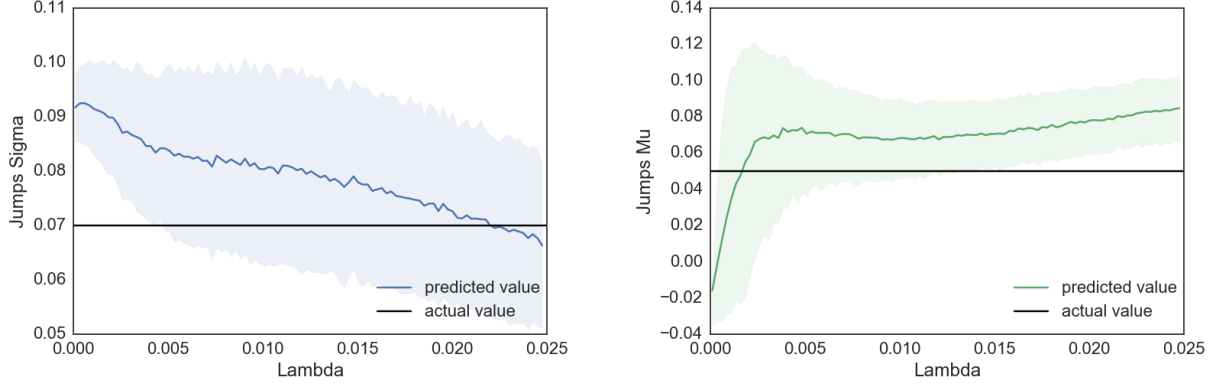
Figure 4.11: The mean deviation (with 68% confidence interval) of the $\hat{\lambda}$ parameter estimate from the actual parameter value, λ , for different values of λ . All the other parameters are kept constant as $\mu = 0.05$, $\sigma = 0.1$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$.

As to be expected, larger σ_{jumps} tend to lower the confidence associated with the estimate, $\hat{\mu}_{jumps}$ (Figure 4.13a).

Figure 4.13b illustrates how σ_{jumps} puts upward pressure on the estimate of λ . This might indicate that the model is not sure as to whether "more jumps" just means "greater volatility" in the jump sizes.

Variations in μ_{jumps}

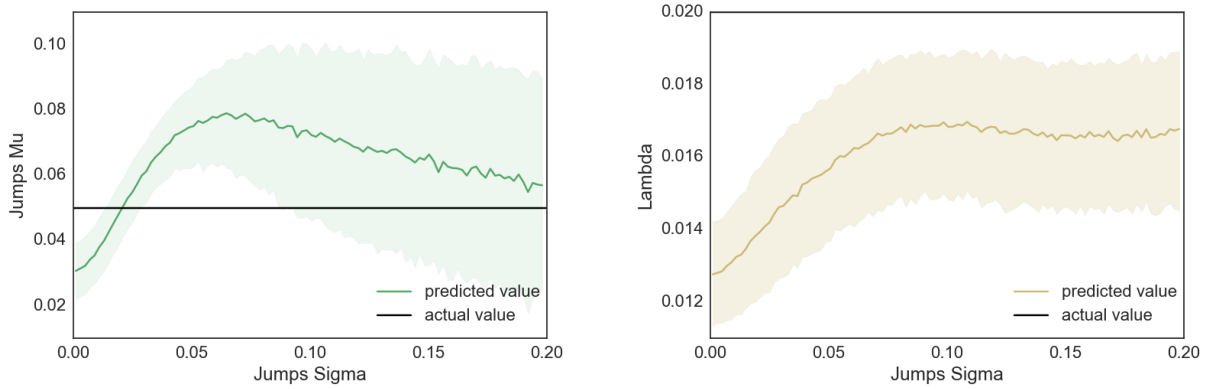
Variations in μ_{jumps} notably affect almost all of the parameter estimates produced by the CNN Multiple Output model.



(a) The mean $\hat{\sigma}_{jumps}$ parameter estimate plotted against the actual parameter value, $\sigma_{jumps} = 0.07$, for different values of λ .

(b) The mean $\hat{\mu}_{jumps}$ parameter estimate plotted against the actual parameter value, $\mu_{jumps} = 0.05$, for different values of λ .

Figure 4.12: The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\mu}_{jumps}$ and $\hat{\sigma}_{jumps}$, plotted against their actual values ($\mu = 0.05$, $\sigma = 0.1$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$), while varying the λ parameter in the range (0,0.025).



(a) The mean $\hat{\mu}_{jumps}$ parameter estimate plotted against the actual parameter value, $\mu_{jumps} = 0.05$, for different values of σ_{jumps} .

(b) The mean $\hat{\lambda}$ parameter estimate plotted against the actual parameter value, $\lambda = 0.02$, for different values of σ_{jumps} .

Figure 4.13: The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\mu}_{jumps}$ and $\hat{\lambda}$, plotted against their actual values ($\mu = 0.05$, $\sigma = 0.1$, $\mu_{jumps} = 0.05$ and $\lambda = 0.02$), while varying the σ_{jumps} parameter in the range (0,0.2).

4.5. CONCLUSION

While a multiple output model might be more convenient to build and use, it was illustrated that single output models are quicker to train and more robust. The author's study also found that, while a fully connected architecture based on the moments and autocorrelations of the process performed well for the simpler parameters such as μ and σ , its estimates of λ were less consistent. Figure 4.2b demonstrates the difficulty of training the fully connected model to estimate the λ parameter.

In particular for the λ parameter, a single-output convolutional model was quicker to train than either of the multiple output models (figure 4.7).

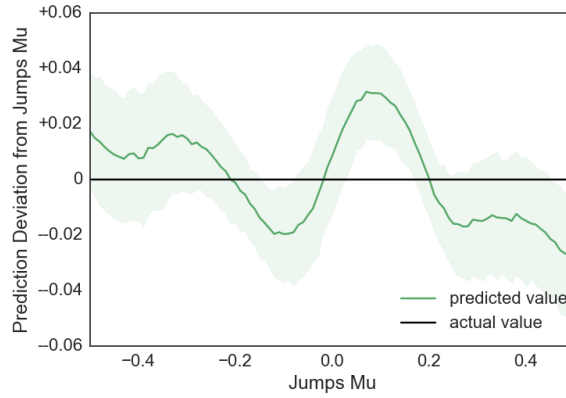
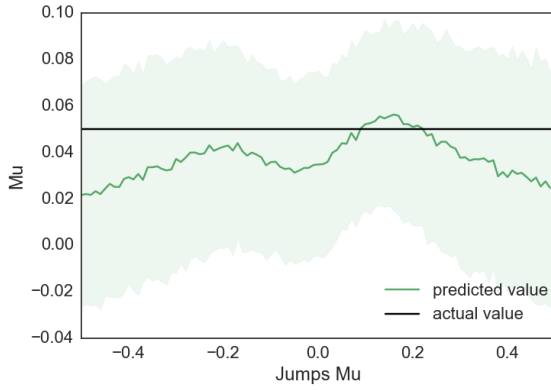
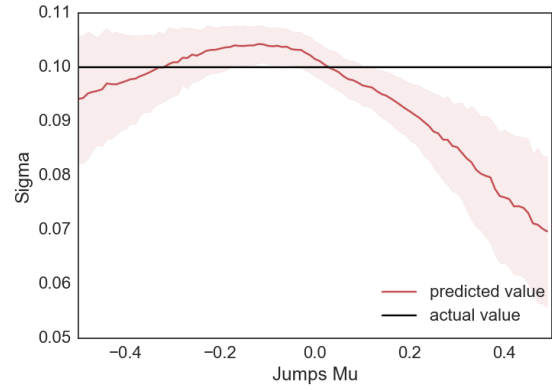


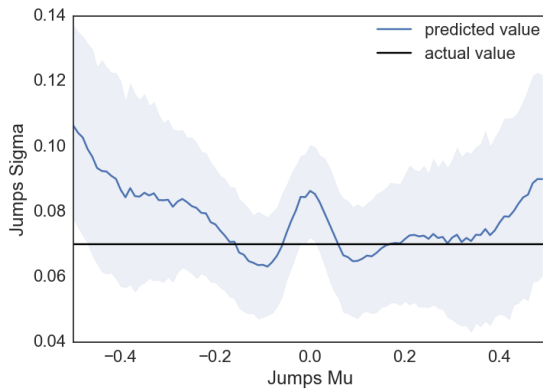
Figure 4.14: The mean deviation (with 68% confidence interval) of the $\hat{\mu}_{jumps}$ parameter estimate from the actual parameter value, μ_{jumps} , for different values of μ_{jumps} . All the other parameters are kept constant as $\mu = 0.05, \sigma = 0.1, \lambda = 0.02$ and $\sigma_{jumps} = 0.07$.



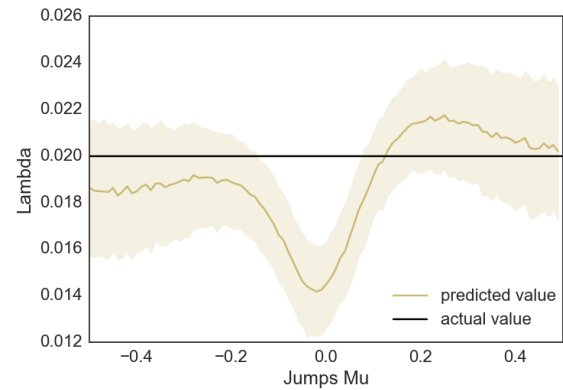
(a) The mean $\hat{\mu}$ parameter estimate plotted against the actual parameter value, $\mu = 0.05$, for different values of μ_{jumps} .



(b) The mean $\hat{\sigma}$ parameter estimate plotted against the actual parameter value, $\sigma = 0.1$, for different values of μ_{jumps} .



(c) The mean $\hat{\sigma}_{jumps}$ parameter estimate plotted against the actual parameter value, $\sigma_{jumps} = 0.07$, for different values of μ_{jumps} .



(d) The mean $\hat{\lambda}$ parameter estimate plotted against the actual parameter value, $\lambda = 0.02$, for different values of μ_{jumps} .

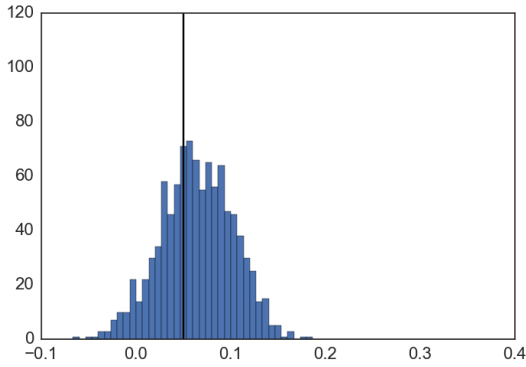
Figure 4.15: The mean parameter estimates (with 68% confidence interval) of $\hat{\mu}$, $\hat{\sigma}$, $\hat{\sigma}_{jumps}$ and $\hat{\lambda}$, plotted against their actual values ($\mu = 0.05, \sigma = 0.1, \lambda = 0.02$ and $\sigma_{jumps} = 0.07$), while varying the μ_{jumps} parameter in the range $(-0.5, 0.5)$.

CHAPTER 5

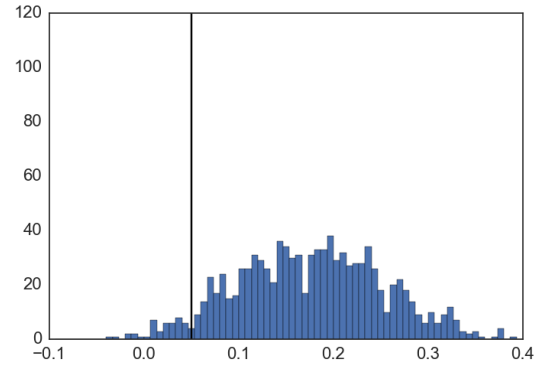
RESULTS

First, the individual accuracy of the selected architectures were investigated. 1000 simulated sample paths from a Merton Jump Diffusion process with parameters, $\mu = 0.05$, $\sigma = 0.1$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$ were used. The (already trained) NN models were then used to predict these original parameters, given the set of 1000 sample paths. Figures 5.1, 5.2, 5.3, 5.4, and 5.5 show the parameter estimation results for $\mu, \sigma, \lambda, \mu_{jumps}$ and σ_{jumps} using:

- a multiple output convolutional ANN model as defined in section 4.3.1 above,
- a fully connected NN as defined in section 4.2.1, and
- a dedicated single output convolutional ANN model as defined in section 4.3.2 above.

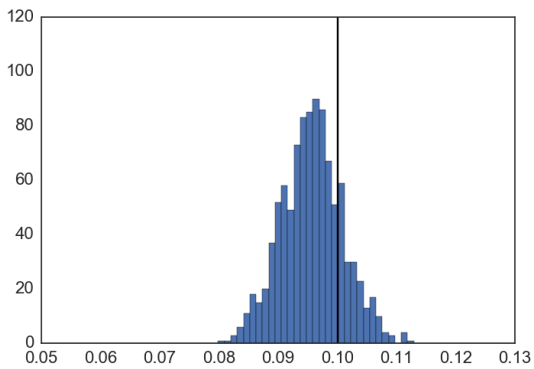


(a) CNN Architecture - Multiple Output - ELU

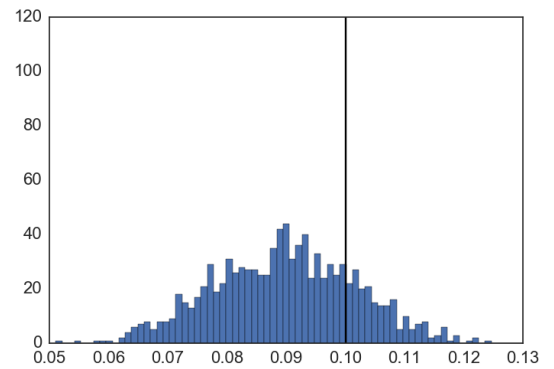


(b) Fully Connected Architecture - Multiple Output - ELU

Figure 5.1: Various model distributions of the predicted values of μ with true value 0,05.

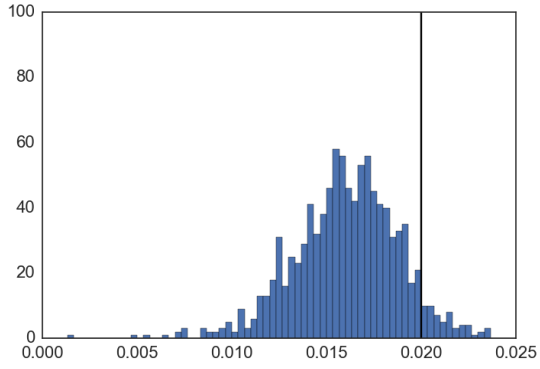


(a) CNN Architecture - Multiple Output - ELU

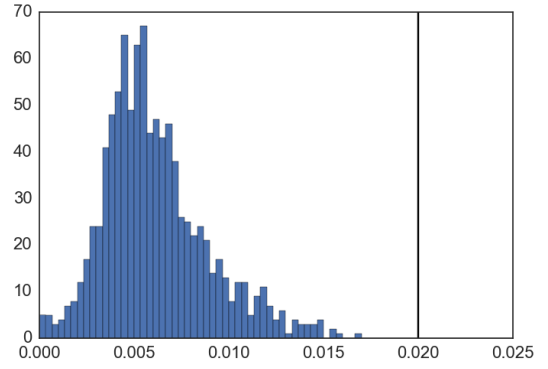


(b) Fully Connected Architecture - Multiple Output - ELU

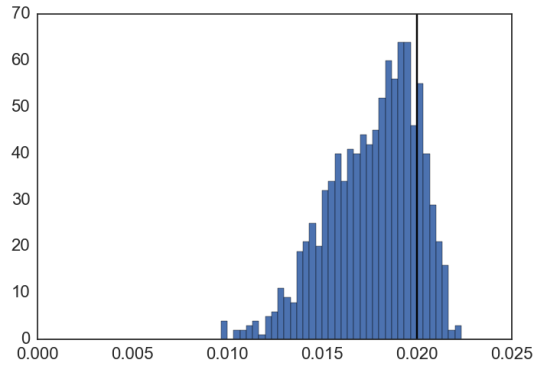
Figure 5.2: Various model distributions of the predicted values of σ with true value 0,1.



(a) CNN Architecture - Multiple Output - ELU

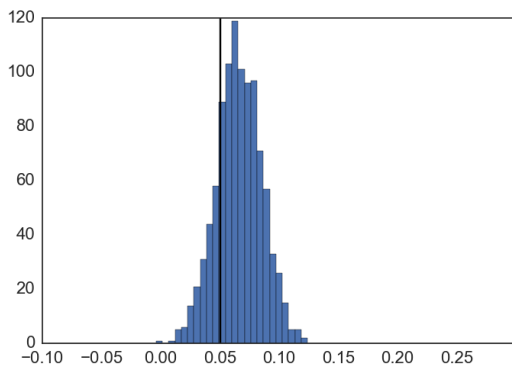


(b) Fully Connected Architecture - Multiple Output - ELU

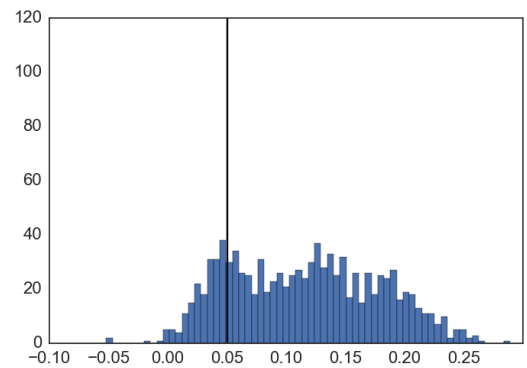


(c) CNN Architecture - Single Output - ELU

Figure 5.3: Various model distributions of the predicted values of λ with true value 0,02.

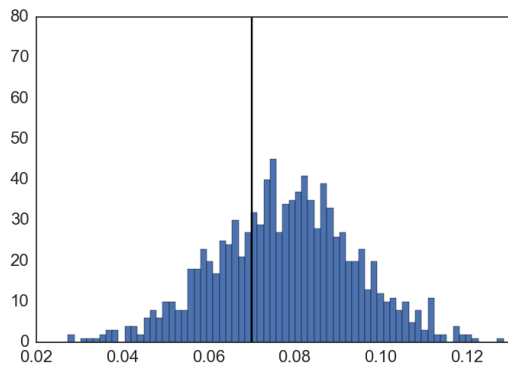


(a) CNN Architecture - Multiple Output - ELU

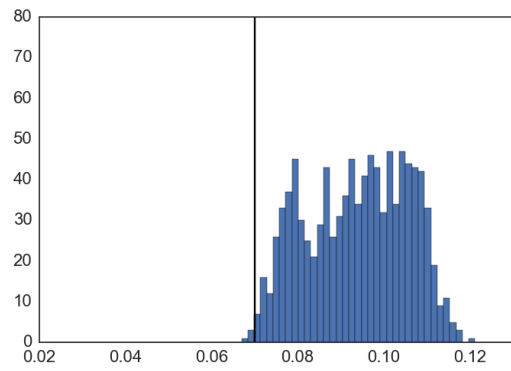


(b) Fully Connected Architecture - Multiple Output - ELU

Figure 5.4: Various model distributions of the predicted values of μ_{jumps} with true value 0,05.



(a) CNN Architecture - Multiple Output - ELU



(b) Fully Connected Architecture - Multiple Output - ELU

Figure 5.5: Various model distributions of the predicted values of σ_{jumps} with true value 0,07.

CHAPTER 6

CONCLUSION AND FURTHER RESEARCH

This project investigated the ability of ANNs to estimate the parameters of a sufficiently complex SDE. Chapter 1 introduced the problem of calibrating SDEs, and discussed the limitations of existing approaches. Chapter 2 presented the literature on similar studies. The concepts used throughout the dissertation were explained in chapter 3, as was the methodology used in the simulation study. The simulation study itself, together with the proposed ANN architectures were presented in chapter 4. Chapter 5 reported on the results of the study and compared them to those obtained using traditional parameter estimation techniques such as MLE, MME, EM and likelihood profiling.

6.1. PRACTICAL CONSIDERATIONS

It must be noted that - at the time of writing - building and training ANNs is a time consuming and computationally intensive process. Overall, the author spent more than 300 GPU-hours on NVIDIA® Tesla® K80 devices implementing, training and testing the architectures presented in this dissertation. However, the progress made in the last year alone indicate that building and training ANNs are becoming more widely adopted, and - with that - come improvements in hardware and software tools.

Bibliography

Cs231n convolutional neural networks for visual recognition.

Available at: <http://cs231n.github.io/convolutional-networks/>

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Available at: <http://tensorflow.org/>

Barone-Adesi, G. 2015. *Stochastic Processes*. Wiley Encyclopedia of Management. John Wiley and Sons, Ltd. ISBN 9781118785317.

Available at: <http://dx.doi.org/10.1002/9781118785317.weom040071>

Cairns, A., Dickson, D., Macdonald, A., Waters, H. & Willder, M. 1998. Stochastic processes: learning the language. *Faculty of Actuaries students' society*.

Chollet, F. *et al.* 2015. Keras. <https://github.com/fchollet/keras>.

Clevert, D., Unterthiner, T. & Hochreiter, S. 2015. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289.

Available at: <http://arxiv.org/abs/1511.07289>

Giebel, S. & Rainer, M. 2013. Neural network calibrated stochastic processes: forecasting financial assets. *Central European Journal of Operations Research*, 21(2):277–293.

Glorot, X., Bordes, A. & Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323.

Honore, P. 1998. Pitfalls in estimating jump-diffusion models.

Hornik, K., Stinchcombe, M. & White, H. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Available at: <http://www.sciencedirect.com/science/article/pii/0893608089900208>

Hunter, J.D. 2007. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Jones, E., Oliphant, T., Peterson, P. *et al.* 2001–. SciPy: Open source scientific tools for Python. [Online; accessed <today>].

Available at: <http://www.scipy.org/>

Kingma, D.P. & Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Available at: <http://arxiv.org/abs/1412.6980>

Merton, R.C. 1976. Option pricing when underlying stock returns are discontinuous. ID: 271671271671.

Available at: <http://www.sciencedirect.com.ez.sun.ac.za/science/article/pii/S0304405X76900222>

Mongwe, W.T. 2015. No title. *Analysis of equity and interest rate returns in South Africa under the context of jump diffusion processes*.

Nielsen, J.N., Madsen, H. & Young, P.C. 2000. Parameter estimation in stochastic differential equations: An overview. ID: 271897271897.

Available at: <http://www.sciencedirect.com/science/article/pii/S1367578800900178>

Olden, J.D. & Jackson, D.A. 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1-2):135–150.

Available at: <http://www.sciencedirect.com/science/article/pii/S0304380002000649>

Oliphant, T. *et al.*. 2001–. Numpy: Open source array tools for Python. [Online; accessed <today>].

Available at: <http://www.numpy.org/>

Oreskes, N., Shrader-Frechette, K. & Belitz, K. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646.

Available at: <http://www.jstor.org/stable/2883078>

Pérez, F. & Granger, B.E. 2007. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29. ISSN 1521-9615.

Available at: <http://ipython.org>

Reid, S.G. 2015. Random walks down wall street, stochastic processes in python.

Available at: <http://www.turingfinance.com/random-walks-down-wall-street-stochastic-processes->

Rossum, G. 1995. Python reference manual. Technical Report, Amsterdam, The Netherlands, The Netherlands.

Samad, T. & Mathur, A. 1992. Parameter estimation for process control with neural networks. ID: 271876271876.

Available at: <http://www.sciencedirect.com/science/article/pii/S0888613X9290008N>

Scherer, D., Müller, A. & Behnke, S. 2010. *Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition*, pages 92–101. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-15825-4.

Available at: http://dx.doi.org/10.1007/978-3-642-15825-4_10

Seabold, S. & Perktold, J. 2010. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Teugels, J.L. & Sundt, B. 2004. *Encyclopedia of actuarial science*. Hoboken, NJ: Hoboken, NJ : John Wiley and Sons. Includes bibliographical references and index.

Waskom, M., Botvinnik, O., drewokane, Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J.B., Warmerhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans, C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A. & Lee, A. 2016. seaborn: v0.7.1 (june 2016).

Available at: <https://doi.org/10.5281/zenodo.54844>

Werbos, P.J. 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Xie, Z., Kulasiri, D., Samarasinghe, S. & Rajanayaka, C. 2007. The estimation of parameters for stochastic differential equations using neural networks. *Inverse Problems in Science and Engineering*, 15(6):629–641. Doi: 10.1080/17415970600907429.

Available at: <http://dx.doi.org.ez.sun.ac.za/10.1080/17415970600907429>

Yao, J., Li, Y. & Tan, C.L. 2000. Option price forecasting using neural networks. *Omega*, 28(4):455–466.

Available at: <http://www.sciencedirect.com/science/article/pii/S0305048399000663>

APPENDIX A

LIST OF SOFTWARE USED

IPython	A command shell for interactive computing.	Pérez & Granger (2007)
Keras	A high-level neural networks API, written in Python running on top of TensorFlow.	Chollet <i>et al.</i> (2015)
Matplotlib	Python 2D plotting library.	Hunter (2007)
Numpy	Package for scientific computing with Python.	Oliphant <i>et al.</i> (2001–)
Python	Programming language.	Rossum (1995)
Scipy	Python-based open-source software for mathematics, science, and engineering.	Jones <i>et al.</i> (2001–)
Seaborn	Python visualization library based on matplotlib.	Waskom <i>et al.</i> (2016)
Statsmodels	Python module that provides classes and functions for the estimation of many different statistical models.	Seabold & Perktold (2010)
Tensorflow	An open-source software library for Machine Intelligence.	Abadi <i>et al.</i> (2015)
Turing Finance	Python implementations of various stochastic processes.	Reid (2015)

APPENDIX B

SOURCE CODE

B.1. NEURAL NETWORK MODELS

```

1 import tensorflow as tf
2 import keras
3 from keras.models import Model, Sequential
4 from keras.layers import Input, Dense, Flatten, LSTM, MaxPooling2D, Input
5 from keras.layers.convolutional import Conv2D
6 from keras.optimizers import Adam
7 from keras import backend as K
8
9 def r2(y_true, y_pred):
10     SSE = K.sum(K.square(y_true - y_pred))
11     SST = K.sum(K.square(y_true - K.mean(y_true)))
12     return 1 - SSE / SST
13
14 def fullyconnected_multiple_ELUs():
15     input_1 = Input(batch_shape = (None, 60))
16
17     layer1 = Dense(2048, activation='elu')(input_1)
18     layer2 = Dense(1024, activation='elu')(layer1)
19
20     layer3 = Dense(512, activation='elu')(layer2)
21     layer4 = Dense(256, activation='elu')(layer3)
22
23     layer5 = Dense(128, activation='elu')(layer4)
24     layer6 = Dense(64, activation='elu')(layer5)
25
26     layer7 = Dense(32, activation='elu')(layer5)
27     last_layer = Dense(16, activation='elu')(layer7)
28
29     output1 = Dense(1, name="sigma")(last_layer)
30     output2 = Dense(1, name="mu")(last_layer)
31     output3 = Dense(1, name="jump_sigma")(last_layer)
32     output4 = Dense(1, name="jump_mu")(last_layer)
33     output5 = Dense(1, name="lambda")(last_layer)
34
35     feedforward = Model(input = input_1, output=[output1, output2, output3, output4,
36     output5])
37
38     feedforward.compile(loss='mean_squared_error', optimizer='adam', metrics=[r2, '
39     mean_absolute_percentage_error'])
40
41     return feedforward
42
43 def fullyconnected_multiple_ReLUs():
44     input_1 = Input(batch_shape = (None, 60))
45
46     layer1 = Dense(2048, activation='relu')(input_1)

```

```

45     layer2 = Dense(1024, activation='relu')(layer1)
46
47     layer3 = Dense(512, activation='relu')(layer2)
48     layer4 = Dense(256, activation='relu')(layer3)
49
50     layer5 = Dense(128, activation='relu')(layer4)
51     layer6 = Dense(64, activation='relu')(layer5)
52
53     layer7 = Dense(32, activation='relu')(layer5)
54     last_layer = Dense(16, activation='relu')(layer7)
55
56     output1 = Dense(1, name="sigma")(last_layer)
57     output2 = Dense(1, name="mu")(last_layer)
58     output3 = Dense(1, name="jump_sigma")(last_layer)
59     output4 = Dense(1, name="jump_mu")(last_layer)
60     output5 = Dense(1, name="lambda")(last_layer)
61
62     feedforward = Model(input = input_1, output=[output1, output2, output3, output4,
63     output5])
64
65     feedforward.compile(loss='mean_squared_error', optimizer='adam', metrics=[r2, '
66     mean_absolute_percentage_error'])
67
68     return feedforward
69
70 def covnet_multiple_ELUs_8_layers():
71
72     input_1 = Input(shape = (40, 50, 1))
73
74     layer1 = Conv2D(32, (12, 12), padding='same', activation='elu')(input_1)
75     layer2 = Conv2D(32, (12, 12), padding='same', activation='elu')(layer1)
76     layer3 = MaxPooling2D(pool_size=(2,2))(layer2)
77
78     layer4 = Conv2D(64, (6, 6), padding='same', activation='elu')(layer3)
79     layer5 = Conv2D(64, (6, 6), padding='same', activation='elu')(layer4)
80     layer6 = MaxPooling2D(pool_size=(2,2))(layer5)
81
82     layer7 = Conv2D(128, (3, 3), padding='same', activation='elu')(layer6)
83     layer8 = Conv2D(128, (3, 3), padding='same', activation='elu')(layer7)
84     layer9 = MaxPooling2D(pool_size=(2,2))(layer8)
85
86     flatten = Flatten()(layer9)
87     last_layer = Dense(256, activation='elu')(flatten)
88     output1 = Dense(1, name="sigma")(last_layer)
89     output2 = Dense(1, name="mu")(last_layer)
90     output3 = Dense(1, name="jump_sigma")(last_layer)
91     output4 = Dense(1, name="jump_mu")(last_layer)
92     output5 = Dense(1, name="lambda")(last_layer)
93
94     convnet_mo_elu = Model(input = input_1, output=[output1, output2, output3, output4
95     , output5])
96
97     convnet_mo_elu.compile('adam', 'mean_squared_error', metrics=['
98     mean_absolute_percentage_error', r2])

```

```

95
96     return convnet_mo_elu
97
98 def covnet_single_ReLUs_6_layers():
99     convnet_lambda = Sequential()
100     convnet_lambda.add(Conv2D(32, (12, 12), input_shape=(40, 50, 1), padding='same',
101         activation='relu'))
102     convnet_lambda.add(Conv2D(32, (12, 12), padding='same', activation='relu'))
103     convnet_lambda.add(MaxPooling2D(pool_size=(2,2)))
104
105     convnet_lambda.add(Conv2D(64, (6, 6), padding='same', activation='relu'))
106     convnet_lambda.add(Conv2D(64, (6, 6), padding='same', activation='relu'))
107     convnet_lambda.add(MaxPooling2D(pool_size=(2,2)))
108
109     convnet_lambda.add(Flatten())
110     convnet_lambda.add(Dense(512, activation='relu'))
111     convnet_lambda.add(Dense(1))
112
113     convnet_lambda.compile('adam', 'mean_squared_error', metrics=[
114         'mean_absolute_percentage_error', r2])
115
116     return convnet_lambda

```


B.2. MERTON JUMP DIFFUSION STOCHASTIC PROCESS SIMULATION

The source code in this section (B.1) was provided by Reid (2015).

```

1 import numpy as np
2 from numpy import random as nrand
3 import math
4 import random
5
6 class ModelParameters:
7     """
8     Encapsulates model parameters
9     """
10
11     def __init__(self,
12                 all_time, all_delta, all_sigma, gbm_mu,
13                 jumps_lamda=0.0, jumps_sigma=0.0, jumps_mu=0.0):
14
15         # This is the amount of time to simulate for
16         self.all_time = all_time
17
18         # This is the delta, the rate of time e.g. 1/252 = daily, 1/12 = monthly
19         self.all_delta = all_delta
20
21         # This is the volatility of the stochastic processes
22         self.all_sigma = all_sigma
23
24         # This is the annual drift factor for geometric brownian motion
25         self.gbm_mu = gbm_mu
26
27         # This is the probability of a jump happening at each point in time
28         self.lamda = jumps_lamda
29
30         # This is the volatility of the jump size
31         self.jumps_sigma = jumps_sigma
32
33         # This is the average jump size
34         self.jumps_mu = jumps_mu
35
36
37 def random_model_params():
38     return ModelParameters(
39         # Fixed Parameters
40         all_time=2000,
41         all_delta=0.00396825396,
42
43         # Random Parameters
44         all_sigma = nrand.uniform(0.001,0.2),
45         gbm_mu = nrand.uniform(-1,1),
46         jumps_lamda=nrand.uniform(0.0001,0.025),
47         jumps_sigma=nrand.uniform(0.001, 0.2),
48         jumps_mu=nrand.uniform(-0.5,0.5),
49     )
50
51 def brownian_motion_log_returns(param):

```

```

52     """
53     This method returns a Wiener process. The Wiener process is also called Brownian
54     motion. For more information
55     about the Wiener process check out the Wikipedia page: http://en.wikipedia.org/
56     :param param: the model parameters object
57     :return: brownian motion log returns
58     """
59     sqrt_delta_sigma = math.sqrt(param.all_delta) * param.all_sigma
60     return nrand.normal(loc=0, scale=sqrt_delta_sigma, size=param.all_time)
61
62 def geometric_brownian_motion_log_returns(param):
63     """
64     This method constructs a sequence of log returns which, when exponentiated,
65     produce a random Geometric Brownian
66     Motion (GBM). GBM is the stochastic process underlying the Black Scholes options
67     pricing formula.
68     :param param: model parameters object
69     :return: returns the log returns of a geometric brownian motion process
70     """
71     assert isinstance(param, ModelParameters)
72     wiener_process = np.array(brownian_motion_log_returns(param))
73     sigma_pow_mu_delta = (param.gbm_mu - 0.5 * math.pow(param.all_sigma, 2.0)) * param
74     .all_delta
75     return wiener_process + sigma_pow_mu_delta
76
77 def jump_diffusion_process(param):
78     """
79     This method produces a sequence of Jump Sizes which represent a jump diffusion
80     process. These jumps are combined
81     with a geometric brownian motion (log returns) to produce the Merton model.
82     :param param: the model parameters object
83     :return: jump sizes for each point in time (mostly zeroes if jumps are infrequent)
84     """
85     assert isinstance(param, ModelParameters)
86     s_n = time = 0
87     small_lamda = -(1.0 / param.lamda)
88     jump_sizes = np.zeros(param.all_time)
89     while s_n < param.all_time:
90         s_n += small_lamda * math.log(random.uniform(0, 1))
91         for j in range(0, param.all_time):
92             if time * param.all_delta <= s_n * param.all_delta <= (j + 1) * param.
93             all_delta:
94                 jump_sizes[j] += random.normalvariate(param.jumps_mu, param.
95                 jumps_sigma)
96                 break
97         time += 1
98     return jump_sizes
99
100 def geometric_brownian_motion_jump_diffusion_log_returns(param):
101     """
102     This method constructs combines a geometric brownian motion process (log returns)
103     with a jump diffusion process
104     (log returns) to produce a sequence of gbm jump returns.

```

```
97 :param param: model parameters object
98 :return: returns a GBM process with jumps in it
99 """
100 assert isinstance(param, ModelParameters)
101 jump_diffusion = jump_diffusion_process(param)
102 geometric_brownian_motion = geometric_brownian_motion_log_returns(param)
103 return np.add(jump_diffusion, geometric_brownian_motion)
```