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

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 Generalized Method of Moments (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,

$$C : f(D) \rightarrow Z \tag{1.1}$$

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.

$$NN : R^X \rightarrow R^Y \tag{1.2}$$

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,

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

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 ith jump, and has distribution, $Y_i \sim N\left(\mu_{jump}, \sigma^2_{jump}\right)$. $N_t, t \geq 0$ 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,

$$ANN : R^X \rightarrow R^Y \tag{1.4}$$

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.

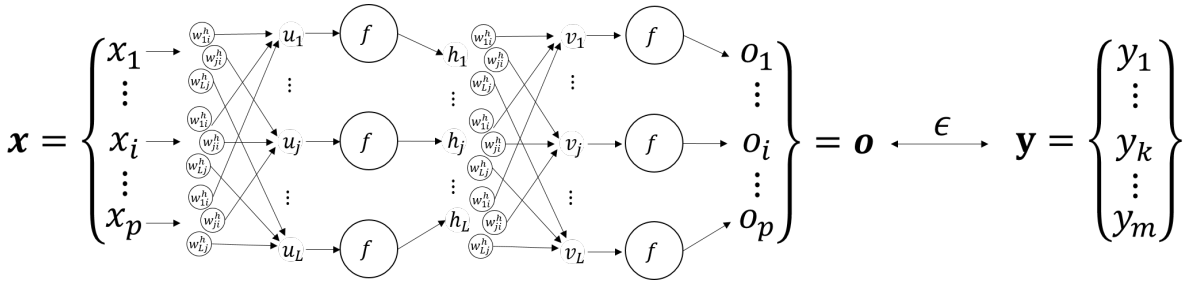


Figure 1.1: Multilayer Perceptron

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, $\mathbf{Z}$

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.

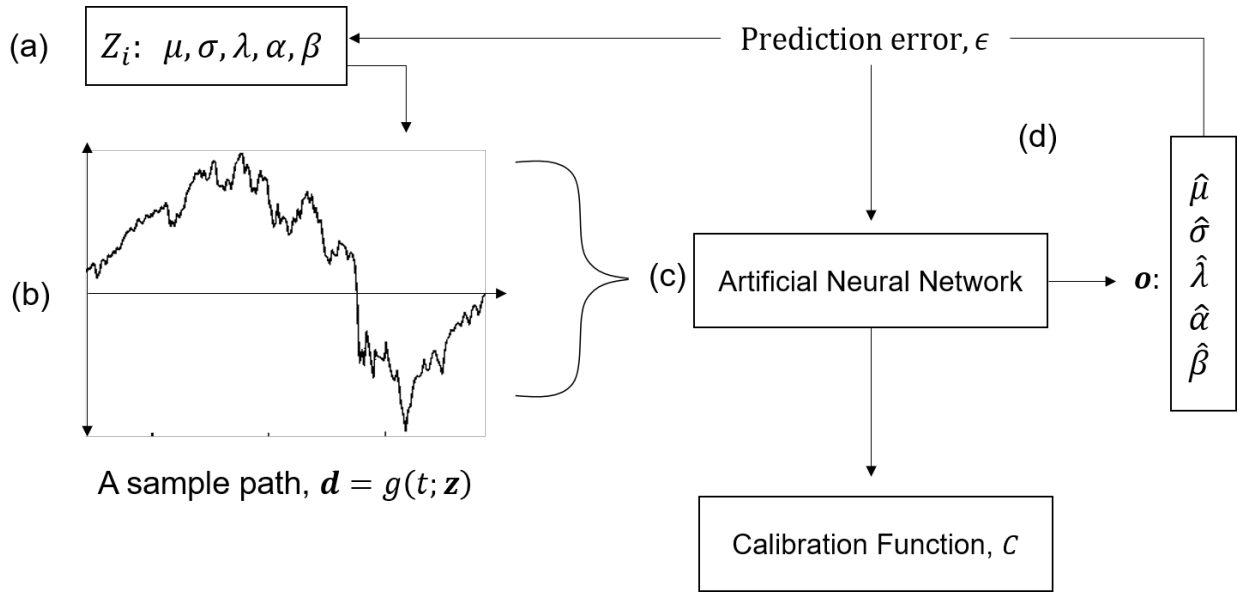


Figure 1.2: The Proposed Calibration Scheme

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-word 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 will be built and trained in Python, using the open source Tensorflow and Keras libraries.

1.5.3 Model Evaluation

The proposed model will be evaluated in terms of the accuracy with which it can predict the true parameters of the stochastic process. This is done by generating sets of parameter values, which are used to simulate sets of corresponding sample paths. These sample paths would then be fed into the calibration function and the results (the parameter estimates) compared to the originally generated parameter set, to obtain a measure of accuracy. The accuracy measure could then be compared against that of other estimation techniques.

Accuracy, however, is not the only evaluation metric. The calibration technique itself (using an ANN to calibrate the stochastic process at hand) will be evaluated by its ability to generalise. This is indicated by the ease with which the technique might deal with different types of stochastic processes (without too many structural changes to the technique itself).

The main focus of this paper remains the definition of a robust methodology that can be followed to approximate the parameters for any stochastic process. A thorough conclusion must report on the potential of the proposed technique to be considered a universal approximation method.

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 unpractical. 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 generalised method of moments 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 MME 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 kth neuron is given by:

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

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. Glorot *et al.* (2011) however showed that ReLUs yield better performance. ReLUs employ the activation function $f(x) = max(0, x)$.

Clevert, Unterthiner *&* Hochreiter (2015) introduced the "exponential linear unit", which provided 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.

$$f(x) = \begin{cases} x & if \quad x > 0 \\ \alpha\left(\exp(x) - 1\right) & if \quad x \leq 0 \end{cases} \tag{3.2}$$

3.2. CONVOLUTIONAL LAYER

Convolutional Neural Networks (CNN's) are a variation of the traditional multi-layer perceptron architecture. Like ordinary feed-forward networks, they consist of neurons with trainable weights and biases. Unlike feed-forward 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.

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

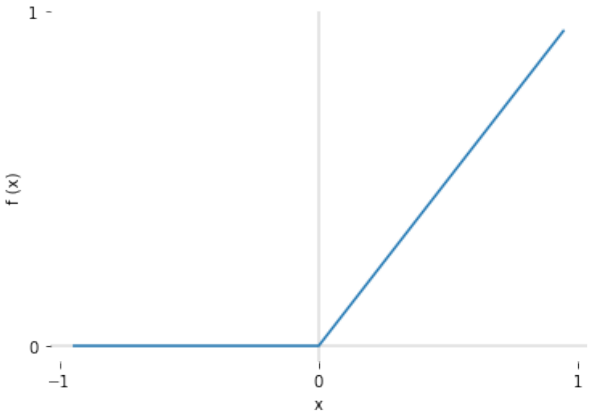


Figure 3.1: Rectifier Activation Function, $f(x)$.

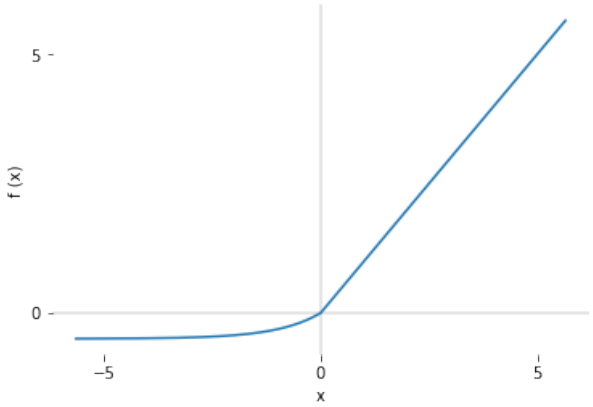


Figure 3.2: Exponential Linear Unit (ELU) Activation Function, with $\alpha = 0.5$.

certain way to ascertain what they might detect.

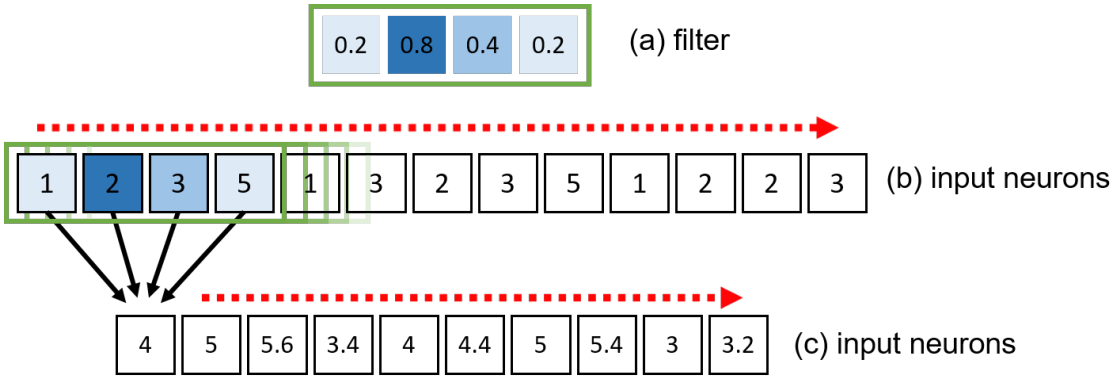


Figure 3.3: A One-Dimensional Convolution Layer

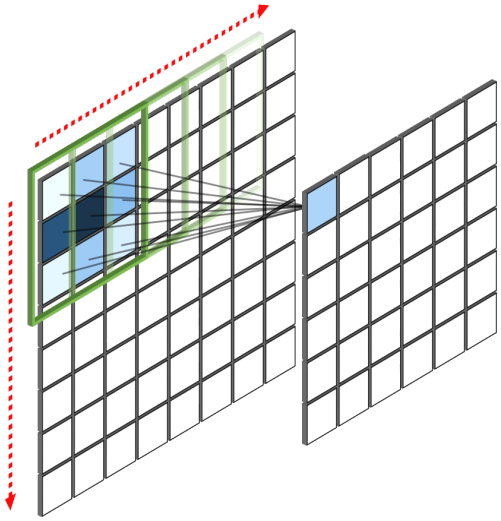


Figure 3.4: A Two-Dimensional Convolution Layer

3.2.1 Pooling

Convolutional Network 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.

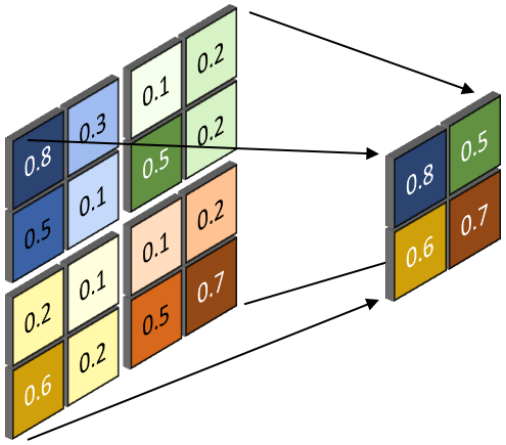


Figure 3.5: A $2x2$ maximum pooling operation.

3.3. PERFORMANCE MEASUREMENT

3.3.1 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, $\mathbf{y}$, and the ANN-predicted parameter values $\hat{\mathbf{y}}$. R^2 is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{m} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{m} \left(y_i - \overline{y}\right)^2} \tag{3.3}$$

.

where y is the actual parameter value, $\hat{y}$ is the predicted parameter value, $\overline{y}$ 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.

3.3.2 Average Absolute Percentage Error

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

$$AAPE = 100 \cdot \frac{1}{m} \sum_{i=1}^{m} \frac{\left| y_i - \hat{y}_i \right|}{y_i} \tag{3.4}$$

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

3.3.3 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.

$$MSE = \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2 \tag{3.5}$$

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 guidlines do exist, for example that Recurrent Neural Network architectures are often used for time series problems (SOURCE?) or that Convolutional Neural Networks 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,

$$dS_t = \mu S_t dt + \sigma S_t dW_t + S_t dJ_t \tag{4.1}$$

where

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

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:

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

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 FEED-FORWARD 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.

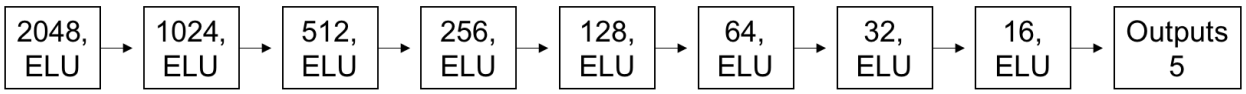


Figure 4.1: 9-Layer Fully Connected Feed-forward ANN

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).

4.3. 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.

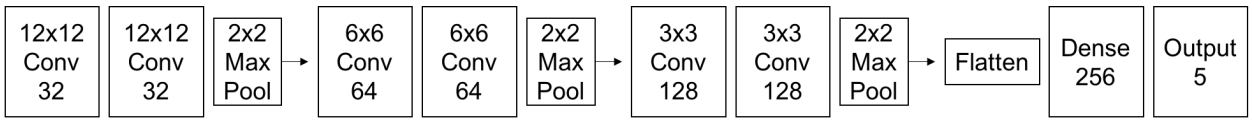


Figure 4.2: 8-Layer Convolutional Neural Network

Experiments were performed using ELU activation functions.

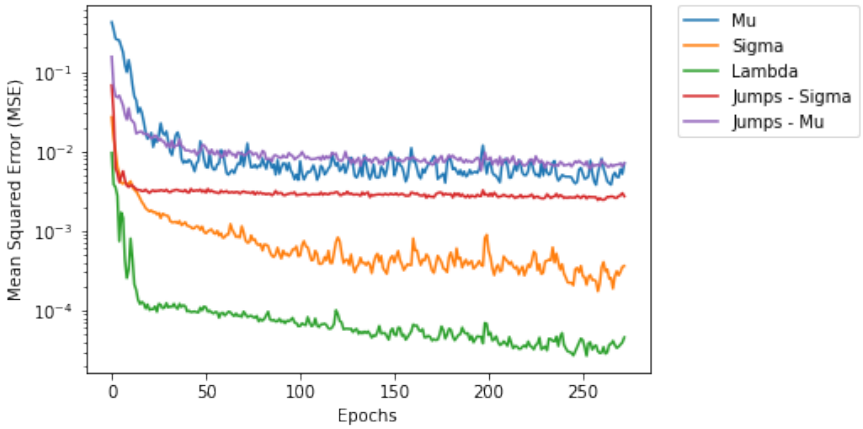
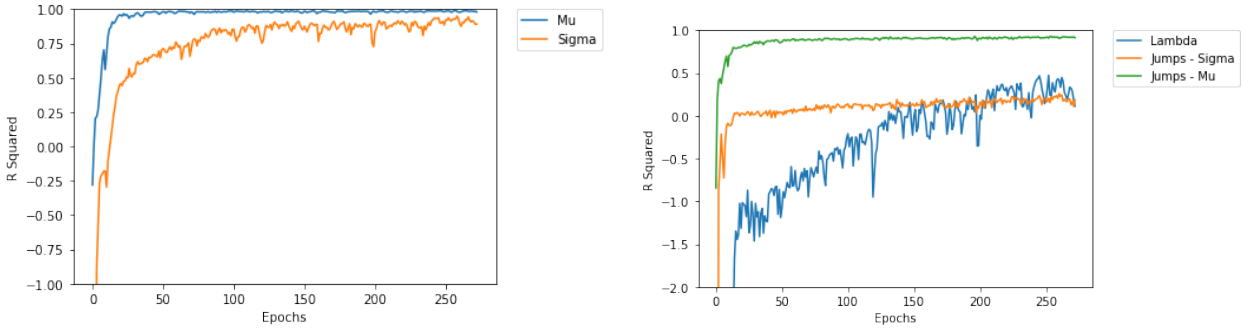


Figure 4.3: MSE values over the training process for the multiple parameter prediction convolutional architecture. Each epoch involves 10 iterations of a simulated batch of 150 randomly selected process parameters.

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



(a) σ, μ **(b)** $\lambda, \sigma_{jumps}, \mu_{jumps}$

Figure 4.4: R-Squared values over the training process for the multiple parameter prediction convolutional architecture (ELU activation units).

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 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}.

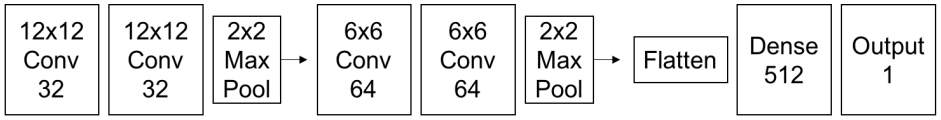


Figure 4.5: 6-Layer Convolutional Neural Network

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 4.4b - exhibits a reluctance to converge to a desired level of accuracy. The dedicated single architecture defined in figure 4.5 was implemented and trained to predict only the single parameter value λ per sample path.

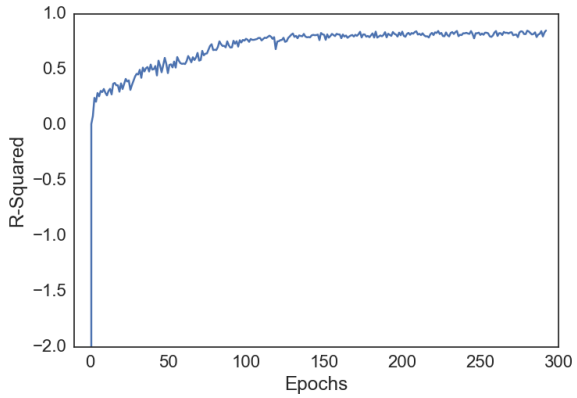


Figure 4.6: R-Squared values for the estimates of λ over the training process for the single parameter prediction convolutional architecture (ELU activation units).

The result is a much quicker and smoother convergence to an acceptable level of accuracy, as visible in figure 4.6. Compare this to the very rough convergence of λ in figure 4.4b. Quicker convergence leads to less training time and ultimately easier use of the model.

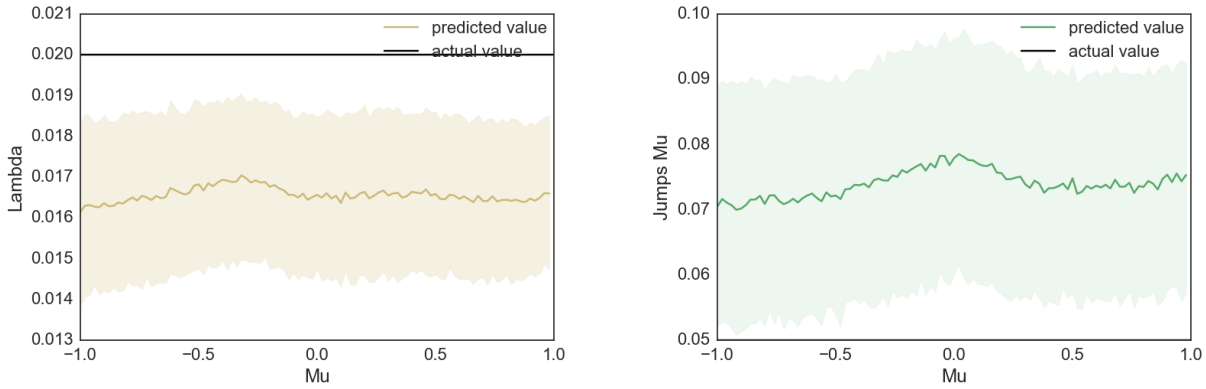
4.4. PARAMETER INTERACTIONS

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 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 4.7: 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 4.7a). This is clear throughout the investigation - λ is almost always underestimated. This isn't something to be expected, since ANN's 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 4.7b). As with the estimate of λ, one would expect and ANN to be able to easily correct for this clear bias.

Variations in σ

As to be expected, larger volatility tends to widen the confidence interval around a particular parameter estimate. This is particularly clear in figure 4.8, 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 4.9a, 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 4.10)

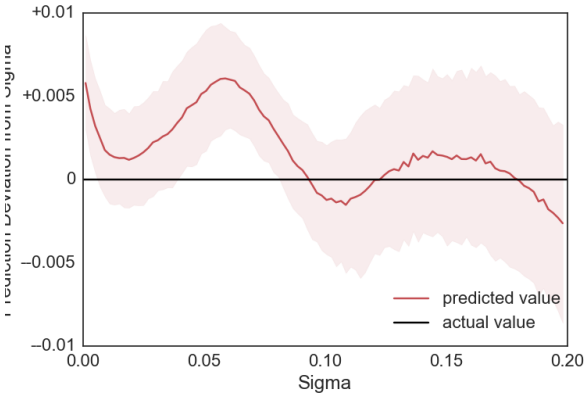
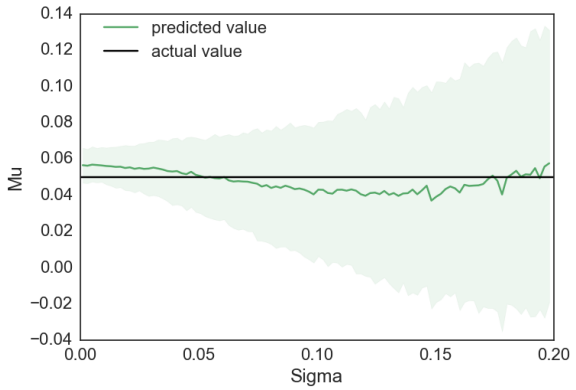
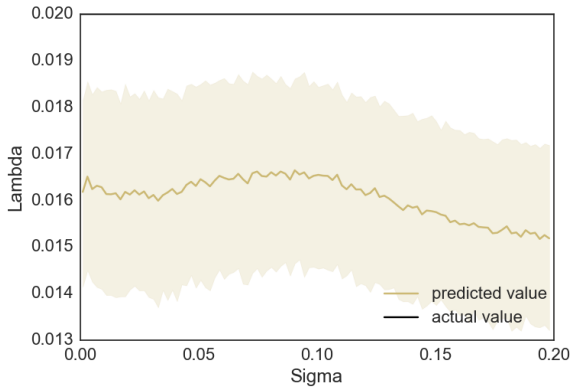


Figure 4.8: The deviation (with 68% confidence interval) of the $\hat{\sigma}$ parameter estimate from the actual parameter value, σ, for different values of σ. All the other parameters are kept constant as $\mu = 0.05, \lambda = 0.02, \mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$.



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



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

Figure 4.9: 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 4.11a 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 it's estimate of σ_{jumps}.

In figure 4.11b, 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 $\hat{\lambda}$.

As to be expected, larger σ_{jumps} tend to lower the confidence associated with the estimate, $\hat{\mu}_{jumps}$ (Figure 4.12a).

Figire 4.12b illustrates how σ_{jumps} puts upward pressure on the estimate of λ. This might indicate

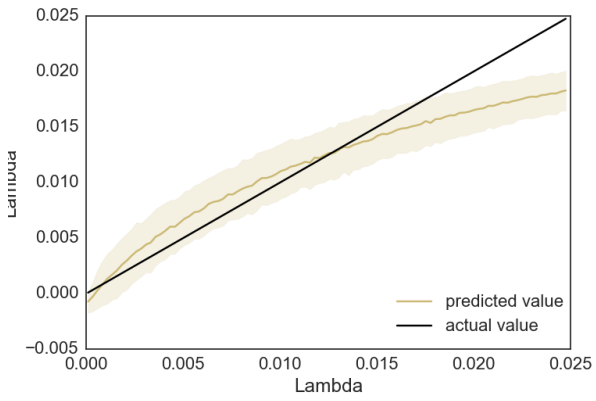
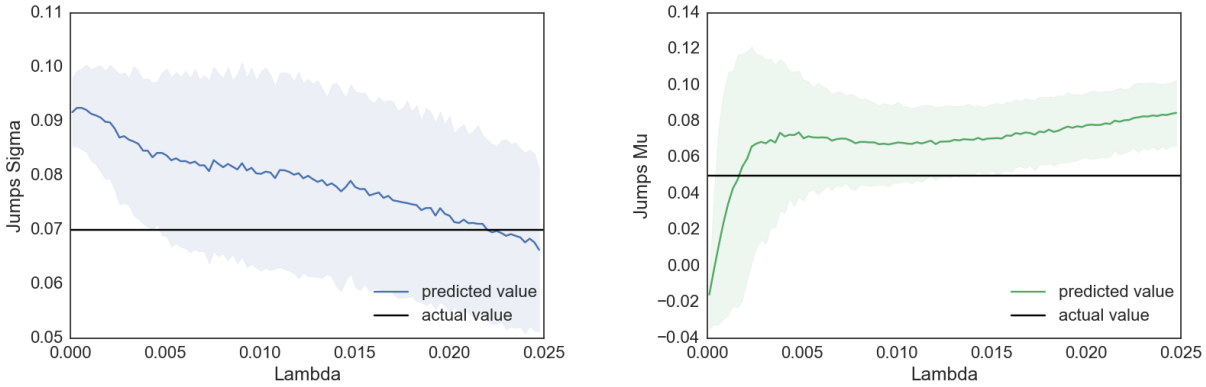


Figure 4.10: The mean deviation (with 68% confidence interval) of the $\hat{\lambda}$ parameter estimate from the actual parameter value, λ, for different values of λ. All the other parameters are kept constant as $\mu = 0.05, \sigma = 0.1, \mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$.



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

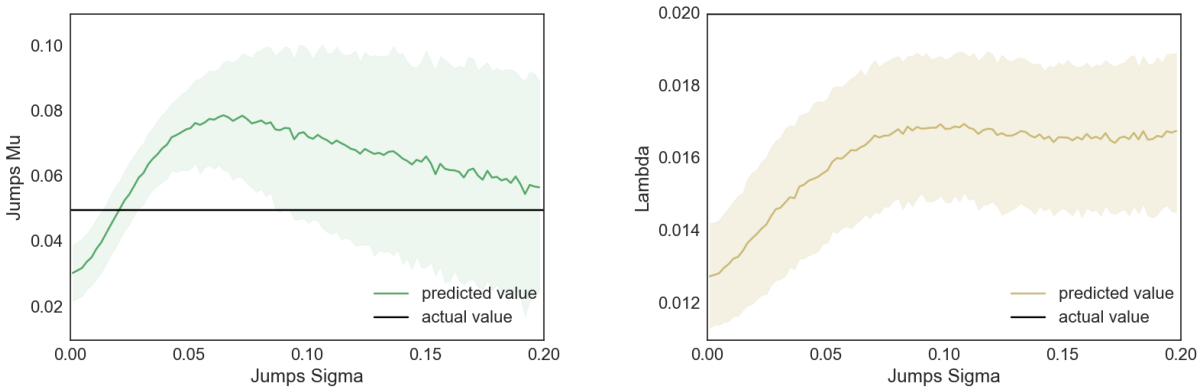
(b) The mean $\hat{\mu}_{jumps}$ parameter estimate plotted against the actual parameter value, $\mu_{jumps} = 0.05$, for different values of λ.

Figure 4.11: 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 $(0, 0.025)$.

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

(b) The mean λ parameter estimate plotted against the actual parameter value, $\lambda = 0.2$, for different values of σ_{jumps}.

Figure 4.12: 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 $(0, 0.2)$.

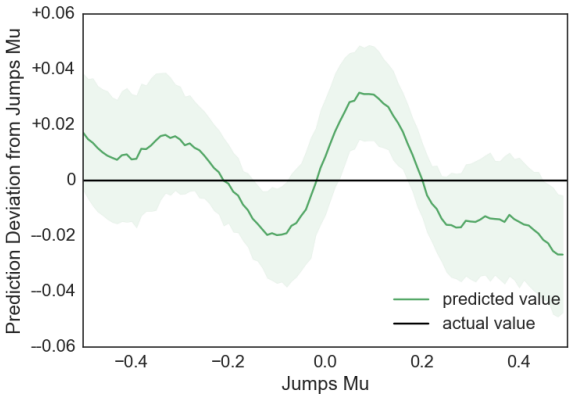
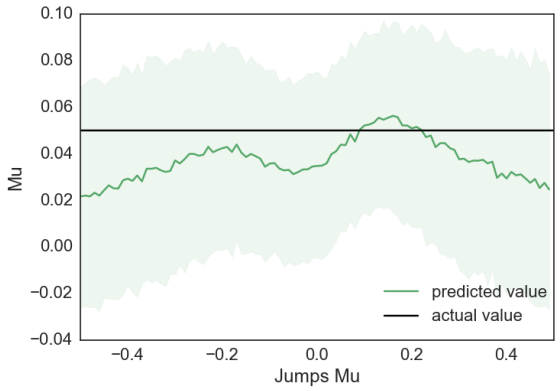
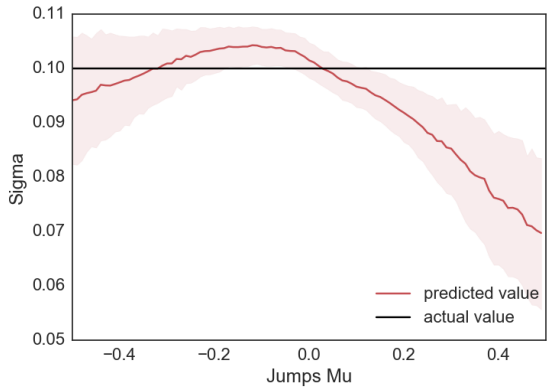


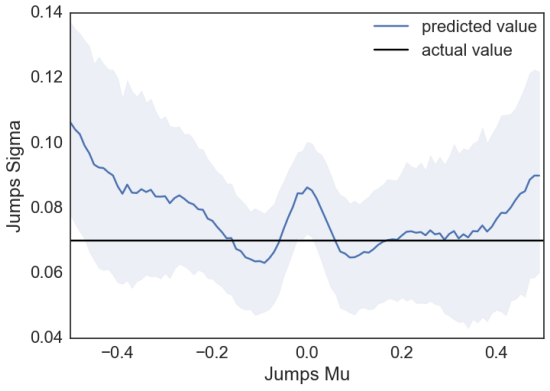
Figure 4.13: The mean deviation (with 68% confidence interval) of the $\hat{\mu}_{jumps}$ parameter estimate from the actual parameter value, μ_{jumps}, for different values of μ_{jumps}. All the other parameters are kept constant as $\mu = 0.05, \sigma = 0.1, \lambda = 0.02$ and $\sigma_{jumps} = 0.07$.



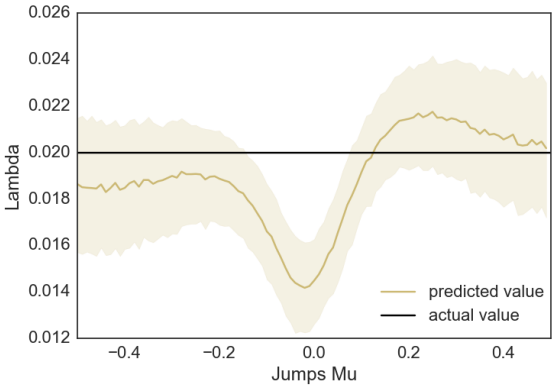
(a) The mean $\hat{\mu}$ parameter estimate plotted against the actual parameter value, $\mu = 0.05$, for different values of μ_{jumps}.



(b) The mean σ 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}.



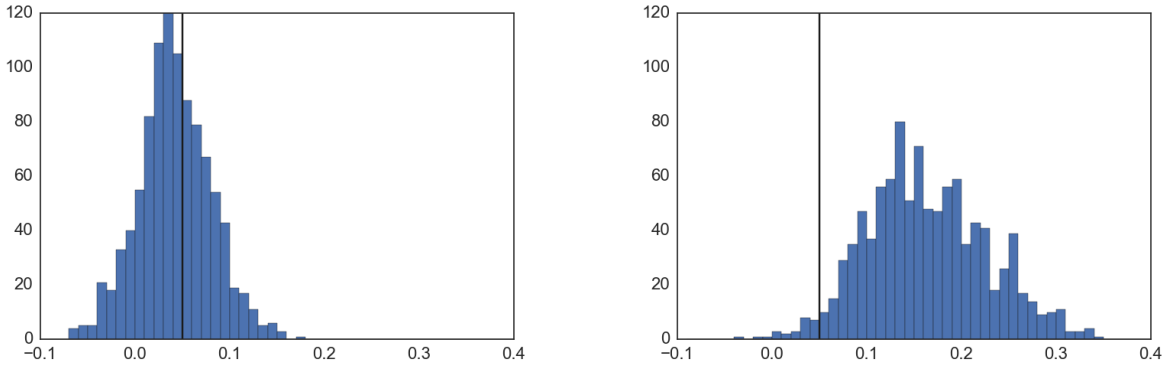
(d) The mean λ parameter estimate plotted against the actual parameter value, $\lambda = 0.2$, for different values of μ_{jumps}.

Figure 4.14: 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 $(-0.5, 0.5)$.

CHAPTER 5
RESULTS

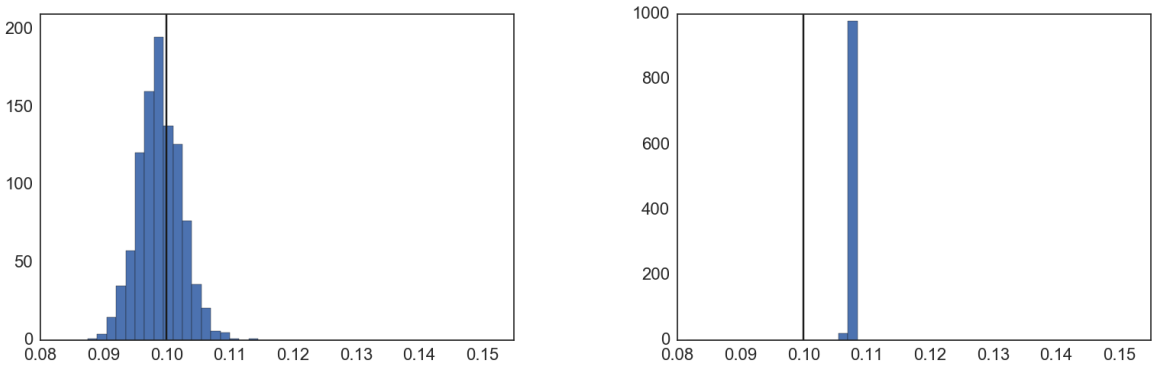
First, the individual accuracy of the selected architectures were investigated. 1000 simulated sample paths from a Merton Jump Diffusion process with parameters, $\mu = 0.05$, $\sigma = 0.1$, $\lambda = 0.02$, $\mu_{jumps} = 0.05$ and $\sigma_{jumps} = 0.07$ were used. The (already trained) NN models 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) a multiple output convolutional ANN model as defined in section 4.3.1 above,
b) a fully connected NN as DEFINED IN SECTION INSERT HERE, and
c) a dedicated single output convolutional ANN model as defined in section 4.3.2 above.



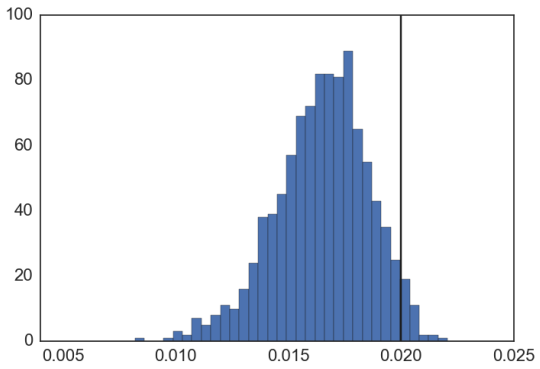
(a) Convolutional Architecture - Multiple Output - ELU **(b)** Fully Connected Architecture - Multiple Output - ELU

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

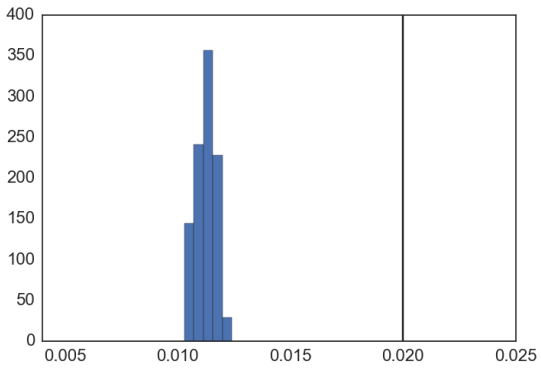


(a) Convolutional Architecture - Multiple Output - ELU **(b)** Fully Connected Architecture - Multiple Output - ELU

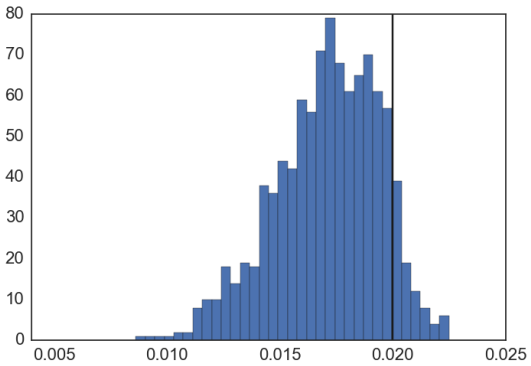
Figure 5.2: Various model distributions of the predicted values of σ with true value $0,1$.



(a) Convolutional Architecture - Multiple Output - ELU



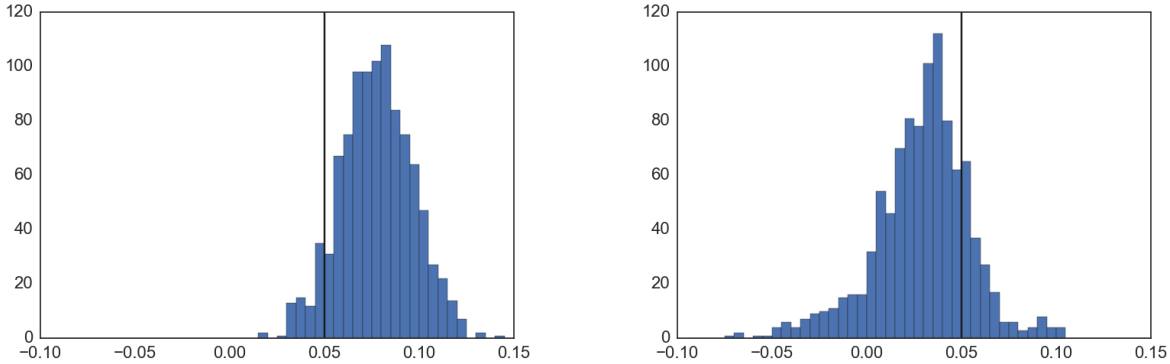
(b) Fully Connected Architecture - Multiple Output - ELU



(c) Convolutional Architecture - Single Output - ELU

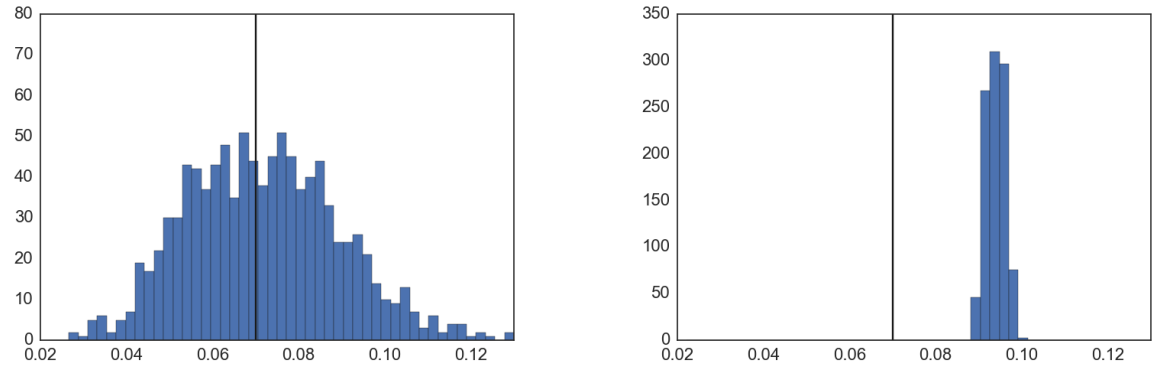
Figure 5.3: Various model distributions of the predicted values of λ with true value $0,02$.

Architecture	No. Parameters	Activation Function	Accuracy



(a) Convolutional 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) Convolutional 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 $0,07$.

Bibliography

Cs231n convolutional neural networks for visual recognition.
Available at: http://cs231n.github.io/convolutional-networks/

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*.

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: //www.sciencedirect.com/science/article/pii/0893608089900208

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/0304405X76900222

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: //www.sciencedirect.com/science/article/pii/S0304380002000649

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

Samad, T. *&* Mathur, A. 1992. Parameter estimation for process control with neural networks. ID: 271876271876.
Available at: http://www.sciencedirect.com/science/article/pii/0888613X9290008N

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.

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

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: //www.sciencedirect.com/science/article/pii/S0305048399000663