Validation
of
Volatility Models

Malik Magdon-Ismail
Caltech 136-93
Pasadena, CA 91125
USA

Yaser S. Abu-Mostafa
Caltech 136-93
Pasadena, CA 91125
USA

abstract

In forecasting a financial time series, the mean prediction can be validated by direct comparison with the value of the series. However, the volatility or variance can only be validated by indirect means such as the likelihood function. Systematic errors in volatility prediction have an 'economic value' since volatility is a tradable quantity (e.g., in options and other derivatives) in addition to being a risk measure. We analyze the fidelity of the likelihood function as a means of training (in sample) and validating (out of sample) a volatility model. We report several cases where the likelihood function leads to an erroneous model. We correct for this error by scaling the volatility prediction using a predetermined factor that depends on the number of data points.

Keywords: Validation, Volatility Prediction, Maximum Likelihood

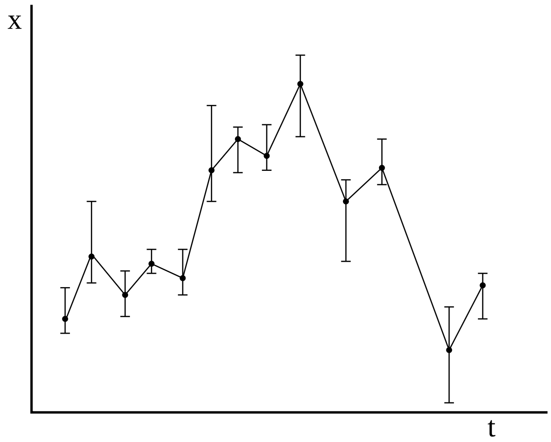

Figure 1.

Introduction

Consider the time series depicted in Fig 1. Each point $x(t)$ is drawn from a distribution whose mean is $\mu(t)$ and variance is $\sigma^2(t)$. While we can only observe $x(t)$, we wish to learn about the mean $\mu(t)$ and the volatility $\tilde{\sigma}(t)$ (a normalized version of $\sigma(t)$). An accurate prediction of the mean tells us about the expected behavior of the time series. An accurate prediction of the volatility is also important, especially in the case of a financial time series. Typically, volatility prediction is used as an explicit measure of risk in static hedging, portfolio selection and margining problems. It serves to place an error bar on the predicted value.

The question arises as to how one can judge various models that are predicting a non-explicit parameter like the variance. The variance falls into the class of non-explicit parameters because, on drawing a random variable from its distribution, no direct information on the variance is conveyed. Depending on the error measure one wishes to choose, the value of the random variable gives direct information on the 'central' value - for example it is the best estimate of the mean using squared error, or the best measure of the median using absolute error. For this reason it is possible for a model to "learn" the mean or median by "training" on the actual value of the random variable using one of these error functions. If we have more than one drawing from the distribution then some direct information does exist on the variance but we would like to consider time series with time-varying distributions, so we only get one drawing from each distribution. As a result, one has to somehow infer information on the variance, and here lies the difficulty with variance prediction. We discuss how this can be done using maximum likelihood, and we take the special case of gaussian noise to illustrate our points.

The most striking result in this paper is that, for any finite number of data points, it is more likely than not that we will pick the worse of two specific models if we use the likelihood function to compare them. It turns out that maximum likelihood will lead to an *underestimation* of the volatility, even when the mean is predicted perfectly. This naturally leads to the question "can we correct for the systematic underestimation?". This allows us to choose from a class of models and then correct for the bias in the method of selection.

Volatility factors into a number of equations in finance. Black and Scholes (1973) derived option pricing models for which the expected future volatility is an important input. Kat (1993) has shown that more accurate volatility prediction will improve the replication efficiency of delta hedging strategies using Black–Scholes hedge ratios, even if the volatility is not constant. Crouchy (1995) shows that for path independent options, the option value depends only on the average volatility while the hedge ratio itself depends on the path of future volatility. The sensitivity of the hedge ratio to short term volatility (we are interested in time-varying volatility) is more of a problem for short term options than long term options. Nonetheless this sensitivity exists and so one would like to have an accurate estimate of the volatility.

$$I_{t-\Delta t} \longrightarrow \boxed{\begin{array}{l}\text{black box}\\ \text{predictor}\end{array}} \longrightarrow \hat{\mu}, \hat{\sigma}$$

Figure 2.

We consider models that predict the mean and variance at time t as in Fig. 2. A variety of techniques exist for predicting variance or volatility. One can use the option prices to compute 'implied' volatilities, such as those derived from the Black-Scholes pricing equations. Another alternative is a multifactor model. Usually combinations of such models work best, but these models all include some constancy in the volatility. Schewert (1989) tests for the constancy of the volatility and strongly rejects this hypothesis. Thus one would like to have models that reasonably account for changing volatility. One can modify the Black-Scholes equations if the volatility is some known function of time by replacing the actual volatility by the average over the remaining life of the option. Autoregressive Conditional Heteroscedasticity (ARCH) type models introduced by Engle(1982) stipulate the conditional variance as a function of past innovations. Generalizations of this are GARCH (Bollerslev, 1986) and EGARCH (Nelson, 1991) where the conditional variance is a function of past innovations and variances. Hull (1987) has tried models that have stochastic parts to them.

One would like to have a reliable method for choosing between volatility models. Our goal is the selection of the optimal model given a number of models. Training can be viewed as a generalization of this where one chooses from a *class* of models (e.g., neural networks with a given architecture). We will evaluate maximum likelihood as a selection criterion between models.

1. Defining the Volatility

i. Variability versus Volatility

It is not clear what the causes of volatility are in the financial markets. Suggestions are that it may be caused by the random arrival of information or by the nature of trading. Fama (1965) and French (1980) have tested this issue empirically showing that it is likely to be a combination of these two effects. Financial market indexes seem "too volatile" in that movements seem to not be attributable to any new objective information. It is important to distinguish how "jagged" or "choppy" a time series is from the volatility. One might more accurately refer to the former as the variability. At any given time, if the mean value is known, the tendency of the actual value at that time to wander about this mean value is related to the volatility. In the case where the mean and the variance are constant, the variability will reflect the volatility. We are more concerned with time varying variance so this distinction should be made. A measure of this tendency for the actual outcome to be scattered about the mean is the variance of the value at time t. All that we observe at time t is an outcome from the distribution of the value from which we would like to infer the variance (the standard definition of the volatility is a normalized version of the square root of the variance). Thus we describe the time series by some underlying time varying distribution $f(x \mid I_{t-\Delta t})$ which is the conditional probability density for obtaining a value x at time t given the information available at $t - \Delta t$. We are interested in $E_f[x]$, a prediction for the outcome and $\sigma^2 = E_f[(x - E_f[x])^2]$ a measure of the (squared) volatility. E_f denotes expectation with respect to $f(x, t)$.

ii. Correspondence with the Black-Scholes Volatility Models

In modeling the time variation of a stock price, Black-Scholes assume an Ito Process (Ito, 1951) of the form

$$\mathrm{d}S = \tilde{\mu} S \mathrm{d}\,t + \tilde{\sigma} S \epsilon \sqrt{\mathrm{d}\,t} \tag{1}$$

$\epsilon \sim N(0,1)$, S is the instrument price. $\tilde{\sigma}$ is defined as the volatility and is the standard deviation of the proportional change in the stock price in unit time. It is this $\tilde{\sigma}$ that enters into the calculation of the hedge ratios, option prices, etc.

$$\frac{\Delta S}{S} \sim N(\tilde{\mu}\Delta t, \tilde{\sigma}\sqrt{\Delta t}) \tag{2}$$

or $\Delta S \sim N(S\tilde{\mu}\Delta t, S\tilde{\sigma}\sqrt{\Delta t})$. The model we have is that at time t the probability density of S is given by $f(S \mid I_{t-\Delta t})$. In particular, $I_{t-\Delta t}$ includes the information $S_{t-\Delta t}$. Hence, the Black-Scholes model falls into this class of models.

$$S_t \sim N(S_{t-\Delta t}(1+\tilde{\mu}\Delta t), \tilde{\sigma} S_{t-\Delta t}\sqrt{\Delta t}) \tag{3}$$

The variance σ^2 is related to the volatility $\tilde{\sigma}$ by

$$\tilde{\sigma} = \frac{\sigma}{S_{t-\Delta t}\sqrt{\Delta t}} \tag{4}$$

So to calculate option prices, hedge ratios, etc., according to the Black-Scholes prescription, it suffices to know σ^2, the variance. With this consideration in mind, we now restrict our analysis to the prediction of variance.

2. Setting Up the Problem of Variance Estimation

We start by introducing the notation that will be used throughout the paper.

$$\begin{aligned} \sigma &\equiv \text{an actual variance} \\ \hat{\sigma} &\equiv \text{a predicted variance} \\ \mu &\equiv \text{an actual mean} \\ \hat{\mu} &\equiv \text{a predicted mean} \\ d &\equiv \text{data drawn from the actual distribution} \\ x \sim N(\mu,\sigma) &\equiv x \text{ has a gaussian(normal) distribution with mean } \mu \text{ and variance } \sigma^2 \end{aligned} \tag{5}$$

i. Basic Setup

We will consider the case of noisy time series, financial series being a special case. The problem is set up in the following way. Given the history of information (including the full history of values of the time series), there exists some conditional probability distribution for the next value. We label the time series variable x, then

$$f(x_t \mid I_{t-\Delta t}) = \begin{array}{l}\text{probability density function for } x_t, \\ \text{the value of the series at time t}\end{array} \tag{6}$$

where $I_{t-\Delta t}$ is the information available at time $t - \Delta t$. Usually, $I_{t-\Delta t}$ is taken as the past few values of the variable x. Ideally, we would like to know what f is, but we will focus on the first two moments of f, as the first moment is the best prediction of the value and the second moment is related to the volatility. So we are interested in

$$\mu(I_{t-\Delta t}) \qquad \text{and} \qquad \sigma(I_{t-\Delta t}) \tag{7}$$

A model consists of a "black box" that takes as input $I_{t-\Delta t}$ and outputs $\{\hat{\mu}^t, \hat{\sigma}^t\}$, predictions of the mean and variance for time t (Fig 2). A collection of models consists of a set of such pairs of functions $\{\hat{\mu}_i^t, \hat{\sigma}_i^t\}_{i=1}^M$. We will drop the t superscript when the context is clear. The index i refers to which model we are talking about. In this paper, we are not concerned with exactly what goes into the black box of Fig 2. All we know is that we are given a set of models (e.g., GARCH, Neural Networks,...) that take the input $I_{t-\Delta t}$ and provide estimates of the mean and variance as output.

ii. Choosing Between the Models

In order to choose between the models, one requires some 'validation' data. Since our goal is to predict the mean μ and the variance σ^2. We would ideally be given n data points which consist of a series of inputs and the actual values of the mean and variance corresponding to those inputs,$\{I_{t-\Delta t}^\alpha, \mu^\alpha, \sigma^\alpha\}_{\alpha=1}^n$. For each model i, one then compares its predictions $\{\hat{\mu}_i^\alpha, \hat{\sigma}_i^\alpha\}$ with the actual values and then chooses that model that performed 'best' on the validation data. More formally, one constructs an error measure for model i

$$E_i = \sum_{\alpha=1}^n \mathcal{E}(I_{t-\Delta t}^\alpha, \hat{\mu}_i^\alpha, \hat{\sigma}_i^\alpha, \mu^\alpha, \sigma^\alpha)$$

and chooses the model yielding the lowest value for the error. $\mathcal{E}$ is a measure of how bad a prediction was on a given data point. The most familiar error measure is probably the squared difference error

$$E_i = \sum_{\alpha=1}^n [(\hat{\mu}_i^\alpha - \mu^\alpha)^2 + (\hat{\sigma}_i^\alpha - \sigma^\alpha)^2] \tag{8}$$

Unfortunately one does not usually have access to the actual values of the mean and variance. All one usually has are data points ($\{d_\alpha\}_{\alpha=1}^n$) drawn from the distributions $f(x_t \mid I_{t-\Delta t}^\alpha)$. Based on the data, one has to construct an error measure

$$E_i = \sum_{\alpha=1}^{n} \mathcal{E}(I_{t-\Delta t}^\alpha, \hat{\mu}_i^\alpha, \hat{\sigma}_i^\alpha, d_\alpha)$$

that evaluates the estimates $\hat{\mu}$ and $\hat{\sigma}$ without knowing the actual μ and σ. The goal is to evaluate this error measure for the models and pick the model that minimizes the error. The question now arises as to what kind of error measure to take.

iii. Using Maximum Likelihood to Choose Between Models

If we know the functional dependence of $f(x_t \mid I_{t-\Delta t})$ on the parameters that we are trying to predict, then after observing the data we can evaluate the likelihood that the data occurred under the assumption that model i is correct. Given the likelihoods of the different models, we will choose that model that has the highest likelihood

$$l(\vec{d} \mid \text{model } i) = \prod_{\alpha=1}^{n} f(d_\alpha \mid I_{t-\Delta t}^\alpha) \tag{9}$$

where we have assumed that the data have been drawn independently from their respective distributions. The likelihood is the physically important quantity because it relates to a probability. A more mathematically useful quantity is the log ($likelihood$) because it yields easily to the law of large numbers for many data points, as it converts the product into a sum. In order to evaluate the right hand side of equation (9) for a given model, we need to make some assumption about f. For the analysis that follows we will assume that f is gaussian. A similar analysis could be done for any other assumption on f. Thus

$$f(d \mid I_{t-\Delta t}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(d-\mu)^2}{2\sigma^2}} \tag{10}$$

where μ and σ depend on $I_{t-\Delta t}$. An equivalent way to formulate this model (say) for the price of a stock is

$$S_t = \mu + \sigma N(0,1) \tag{11}$$

To reproduce the Black-Scholes model, one would choose $\mu = S_{t-\Delta t}(1+\tilde{\mu}\Delta t)$ and $\sigma = \tilde{\sigma} S_{t-\Delta t}\sqrt{\Delta t}$. In particular, we see that Black-Scholes models satisfy our gaussian assumption on f. The likelihood that the data occurred under a particular model i is given by

$$l(\vec{d} \mid \text{model } i) = \prod_{\alpha=1}^{n} \frac{1}{\sqrt{2\pi \hat{\sigma}_\alpha^{i\,2}}} e^{-\frac{(d_\alpha - \hat{\mu}_\alpha^i)^2}{2\hat{\sigma}_\alpha^{i\,2}}} \tag{12}$$

Maximizing the likelihood, or equivalently maximizing the Log(*likelihood*) is the criterion that we will use to differentiate between the models. As an example, consider the training of Neural Networks to predict the variance using maximum likelihood as the objective function to optimize (Weigend, 1995). In this case, the models i correspond to all the functions that the Neural Network can implement and we are choosing one of them using maximum likelihood as the criterion.

3. Analysis of the Maximum Likelihood Scheme

i. Expected Value of the Likelihood and Log(*likelihood*)

We begin by considering the advantages of the maximum likelihood scheme. Maximum likelihood is widely used for parameter estimation (Valavanis,1959). Here we will analyse maximum likelihood through its expectation. It is well known that for a sample of data drawn from the same distribution, the maximum likelihood predictors are $\hat{\mu} = \bar{d}$, the sample mean and $\hat{\sigma}^2 = \overline{d^2} - \overline{d}^2$, the sample variance, and according to the law of large numbers, these estimates approach the true values with probability 1. Suppose we do not have many data points. We can look at the expectation of the likelihood and the Log(*likelihood*) instead.

Consider the likelihood function as a function of the data d_α. Thus it is itself a random variable, for which the distribution is known given the distribution of the data point d_α. One can calculate 'the expectation of $l(d_\alpha)$'=$\langle l_i \rangle$ and 'the expectation of $log\,(l(d_\alpha))$'=$\langle log\, l_i \rangle$ for model i. For simplicity in notation, we will drop the α index.

$$\begin{aligned} log\langle l_i \rangle &= -\frac{1}{2}\left[\frac{(\mu - \hat{\mu}_i)^2}{\sigma^2 + \hat{\sigma}_i^2} + log(\sigma^2 + \hat{\sigma}_i^2) + log\, 2\pi\right] \\ \langle log\, l_i \rangle &= -\frac{1}{2}\left[\frac{(\mu - \hat{\mu}_i)^2 + \sigma^2}{\hat{\sigma}_i^2} + log(\hat{\sigma}_i^2) + log\, 2\pi\right] \end{aligned} \tag{13}$$

The first expression tends toward the second in the limit $\sigma^2 \to 0$, as should be expected. Both the likelihood and log(likelihood) are convex independently in the variables $\hat{\mu}_i$ and $\hat{\sigma}_i^2$. However, as functions of two variables, they are not convex.

$$\begin{aligned} \max_{\hat{\mu}_i, \hat{\sigma}_i^2}[log\langle l_i \rangle] \quad &\Rightarrow \quad \hat{\mu}_i \to \mu,\ \hat{\sigma}_i^2 \to 0 \\ \max_{\hat{\mu}_i, \hat{\sigma}_i^2}[\langle log(l_i) \rangle] \quad &\Rightarrow \quad \hat{\mu}_i \to \mu,\ \hat{\sigma}_i^2 \to \sigma^2 \end{aligned} \tag{14}$$

We see that the expectation of the Log(*likelihood*) is maximized when the predicted mean and variance are equal to the true mean and variance. So if we expect the observed value of the Log(*likelihood*) to be close to its expected value, which will be true with enough data points, then it seems reasonable to maximize the Log(*likelihood*) in order to predict $\{\hat{\mu}, \hat{\sigma}\}$. For this reason we use maximizing the Log(*likelihood*) as our criterion for choosing between models.

The expectation of the likelihood itself is not maximized at the correct values. Its maximum is when $\hat{\sigma} \to 0$. This rules out likelihood itself as a comparator because its expected behavior is not desirable. What about the Log(*likelihood*) ? To investigate this issue we first define what it means for a model to be 'better' than another. Suppose we have two models, model 1 and model 2 with

$$\begin{aligned} & \mid \hat{\mu}_1 - \mu \mid > \mid \hat{\mu}_2 - \mu \mid \\ \text{and} \qquad & \mid \hat{\sigma}_1 - \sigma \mid > \mid \hat{\sigma}_2 - \sigma \mid \end{aligned} \tag{15}$$

We will then say that model 1 is worse than model 2. * The hope now is that if model 1 is worse than model 2 then its expected Log(*likelihood*) should also be worse. Unfortunately one can find common situations where this is not the case, i.e., model 1 has a higher expected Log(*likelihood*) though it is worse. This becomes evident from Fig. 3.

As $\hat{\sigma}_i^2 \to 0$, $\langle log\, l_i \rangle \to -\infty$ so we can make model 2's variance large while sending model 1's variance to zero and fulfill the condition. Thus we see that even log(*likelihood*) will lead to an expected worse choice in such situations. However we note that training (say neural networks) by maximizing log(*likelihood*)(Weigend,1995) using small perturbations will select 'better' models on the average because the maximum is unique. Note that by initially choosing a model with lower log(*likelihood*), training could be faster owing to the asymmetry of the curve about the actual variance. If the neural network cannot implement the maximum of the curve then training may stop at a worse value than is necessary, owing to this asymmetry.

From this analysis, we see that a model predicting a smaller variance can have an expected Log(*likelihood*) that is lower than a 'worse' model which is predicting a higher variance. What about the possibility of a 'worse' model that is predicting a lower variance? Is there some sense in which we might go wrong and pick the wrong model? This will be the subject of the next section. In particular we will compare the ground truth model to a 'worse' model predicting a lower variance.

ii. Probability of Choosing the Wrong Model

* One might consider the alternate measure $e = \mid \log[\frac{\hat{\sigma}^2}{\sigma^2}] \mid$. With this measure it is also possible to find two models (1,2) with $e_1 < e_2$ and $\langle \log l_1 \rangle < \langle \log l_2 \rangle$.

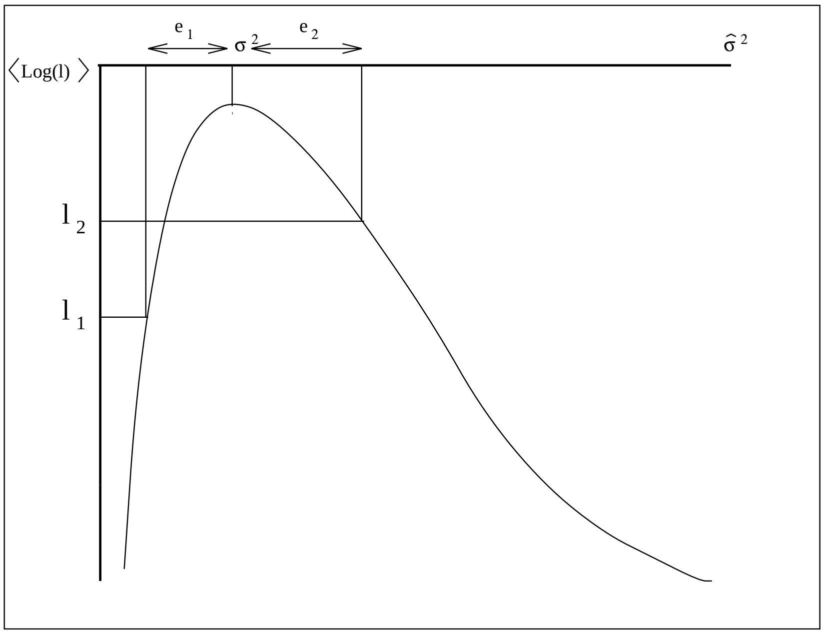

Figure 3. Plot of $\langle log\, l_i \rangle$ as a function of $\hat{\sigma}_i^2$. The curve is not symmetric about the true variance. In the figure is shown how model 2 has a higher expected likelihood despite having a higher prediction error $(e_2 > e_1)$

Suppose that you have two models, model 1 worse than model 2 as described above. One can ask whether it is possible that model 1 will be picked more often than model 2. This is distinct from the analysis of the expected value because it is possible for the expected value to be greater for model 1 despite the fact that the probability that model 1's Log($likelihood$) is higher is very small. In this case one might still be content because most of the time one will be choosing the better model even though its expected Log($likelihood$) is lower. Unfortunately, we show that this too is not the case.

In making a choice between two models, one compares their likelihoods and chooses the model for which the likelihood is higher. One can therefore ask the question: which model has a higher probability of being chosen? If model 1 has higher likelihood less than half the time then we say that it is the 'inferior' model in the sense that we will choose model 2 as our estimate of the ground truth model more often than we would choose model 1. Once again we can ask the question: Is it possible that one might pick a 'worse' model (in the sense described in the previous section) more often than one might pick a 'better' model? To answer this question we will compare a 'worse' model with the ground truth model. We shall see that this will lead to some striking results.

Set up the problem in the following way. We have n data points $\{I_{t-\Delta t}, d_\alpha\}$ where each of the d_α have been drawn from a (possibly *different*) distribution $\sim N(\mu_\alpha, \sigma_\alpha)$. We will consider the case where $\hat{\mu}^i_\alpha = \mu_\alpha$ (perfect prediction of the mean), while the variances predicted by the models are given by

$$\hat{\sigma}^i_\alpha = \lambda^i_\alpha \sigma_\alpha, \quad \lambda^i_\alpha \geq 0 \tag{16}$$

thus the models are parametrized by $\lambda^i_\alpha, \alpha = 1, \ldots, n$. The likelihood of model i with parameters $\vec{\lambda}_i$ is

$$l(\vec{d} \mid \vec{\lambda}_i) = \prod_\alpha N_{d_\alpha}(\mu_\alpha, \lambda^i_\alpha \sigma_\alpha) \tag{17}$$

We are going to compare the two models $\{\lambda_\alpha = 1\}$, the perfect model, and $\{\lambda_\alpha\}$. We choose the model for which the likelihood of the data is highest. We seek the probability $(P_{\vec{\lambda}})$ that $l_{\vec{\lambda}} \geq l_{\vec{1}}$ where $l_{\vec{1}}$ is the likelihood of the ground truth model and $l_{\vec{\lambda}}$ is the likelihood of model $\vec{\lambda}$.

Now treat the data outcome $\vec{d}$ as a random variable $\vec{x}$. Then this likelihood itself is a random variable, depending on the parameter $\vec{\lambda}$, so we can ask the question: what is the probability that the random variable with parameter $\vec{\lambda}$ is greater than the random variable with parameter $\vec{\lambda} = \vec{1}$, i.e., we are asking for the frequency with which the 'worse' model is chosen over the *actual* model. This probability is given by

$$P_{\vec{\lambda}} = \text{Prob}[l_{\vec{\lambda}} \geq l_{\vec{1}}] = \int\limits_{\{l_{\vec{\lambda}} \geq l_{\vec{1}}\}} \mathrm{d}^n\mathrm{x}\, l(\vec{x} \mid \vec{\lambda} = \vec{1}) \tag{18}$$

For the case $\vec{\lambda} = \lambda\vec{1}$, the result is calculated in Appendix A along with various asymptotic properties. Fig. 4 summarizes the result in a plot of P_λ versus λ. For λ slightly less than 1, P_λ is greater than $\frac{1}{2}$. This means that a worse model will be chosen over the actual model more than $\frac{1}{2}$ the time. *This holds for any number of data points.* The smallest value that λ can be with $P_\lambda \geq \frac{1}{2}$ $(\lambda_{min}(n))$ is tabulated for various n in table 1.

Referring to table 1, we note that for n=100, a model with a 0.7% error in the prediction of σ will be chosen more often than the actual model. This is a fairly significant error when dealing with a tradable quantity.

4. Correcting for the Maximum Likelihood Prediction Error

Up to now we have diagnosed some problems with the maximum likelihood scheme. We now try to tackle the problem of compensating for this error using our understanding of

n	$\lambda_{min}(n)$	$P_{max}(n)$	$\langle\frac{1}{\lambda}\rangle$=$\alpha_{correc}$
1	0.4937	0.6831	∞
2	0.7072	0.6321	1.7725
5	0.8729	0.5842	1.1894
10	0.9349	0.5594	1.0837
20	0.9670	0.5422	1.0397
50	0.9866	0.5268	1.0153
100	0.9934	0.5186	1.0076
500	0.9986	0.5088	1.0015
1000	0.9993	0.5059	1.0008

Table 1: Showing the results for the analysis of the probability that model $\vec{\lambda}$ is chosen more often than the actual model.For $\lambda_{min}(n) \leq \lambda \leq 1$, $P_\lambda \geq \frac{1}{2}$. $\langle\frac{1}{\lambda}\rangle$ is the expected correction factor α_{correc} for the new prediction given the maximum likelihood model prediction ($\hat{\sigma} \rightarrow \alpha_{correc}\hat{\sigma}$). $P_{max}(n)$ is the maximum value that P_λ can take and occurs for $\lambda \rightarrow 1^-$

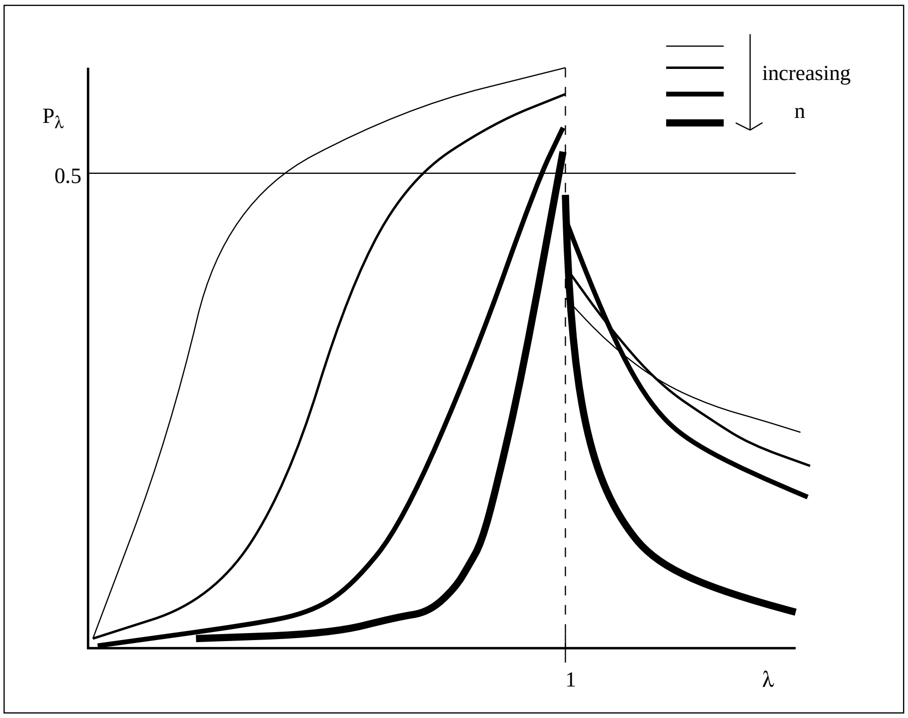

Figure 4. Plot of P_λ as a function of λ. For n even we get the result $P_\lambda = 1 - e^{-\beta n}\sum_{k=0}^{\frac{n}{2}-1}\frac{(\beta n)^k}{k!}$, $\beta = \frac{\lambda^2 \log \lambda}{\lambda^2-1}$. The result for n odd can be obtained numerically. Regions where the curve exceeds $\frac{1}{2}$ are regions where the worse model is picked more often than the actual one.

how it fails. The models that are more probable than the actual model are models which underestimate the variance (assuming the mean is predicted perfectly). One expects that the model chosen using maximum likelihood will have a systematic bias for predicting a variance lower than the actual variance. Suppose that we have a model that is predicting $\hat{\sigma} = \lambda\sigma$ on the average. We would like to correct our prediction arrived at using maximum likelihood methods by multiplying our prediction by some correction factor to get a better prediction.

$$\hat{\sigma} \rightarrow \alpha_{correc}\hat{\sigma} \tag{19}$$

In order to calculate α_{correc} we need the probability distribution for obtaining a particular model with parameter λ. The method of compensating will depend on the the exact method that was used to arrive at the model. Here we will consider the case that seems appropriate to neural networks that are trained on the data using maximum likelihood. The general philosophy will become apparent.

The problem is set up in the following way. The method by which a model is picked is by comparison of likelihood on a data set that contains some number of data points n. The model with the highest likelihood is then chosen. This model will then serve as the predictor and so given a prediction, we would like to correct it by some factor.

i. Probability Distribution Over λ

We consider the following model for learning to predict the variance. We have the class of models that are parametrised by λ. Training proceeds in the following way. Start with a random λ and perturb λ a little toward the higher range and toward the lower range. So we are dealing with the three λ's $\{\lambda - \frac{d\lambda}{2}, \lambda, \lambda + \frac{d\lambda}{2}\}$. Now using the data we compare the likelihood of these three models and choose the model that produces the highest likelihood. We continue the process until no change in λ results. One can now decrease the step size if desired, to attain the necessary accuracy. In this way, training of the model proceeds in such a way that the model ends up picking that λ that maximized the likelihood function on the data set. In actuality the learning proceeds not by varying λ but by varying the parameters of the model (in the case of neural networks, these are the weights). One might question why the model's λ should be the same for all the possible inputs. This will not be the case in general, but we are asking what the correction factor is *on the average*. To calculate this we look at the probability distribution that we end up with a model with a certain λ (for which the correction factor is $\frac{1}{\lambda}$).

What is the probability that we stop at the value λ? We consider the region $[\lambda, \lambda+d\lambda]$ and derive the probability density over λ, which we can use to calculate various desired properties like expectations. The detailed derivation of this probability density is in Appendix B. The final result we get for this probability density is

$$P(\lambda)d\lambda = \frac{n^{\frac{n}{2}} e^{-n\frac{\lambda^2}{2}}}{\Gamma(\frac{n}{2})} \left(\frac{\lambda^2}{2}\right)^{\frac{n}{2}-1} \lambda d\lambda \tag{20}$$

We show a plot of $P(\lambda)$ for various n in Fig. 5

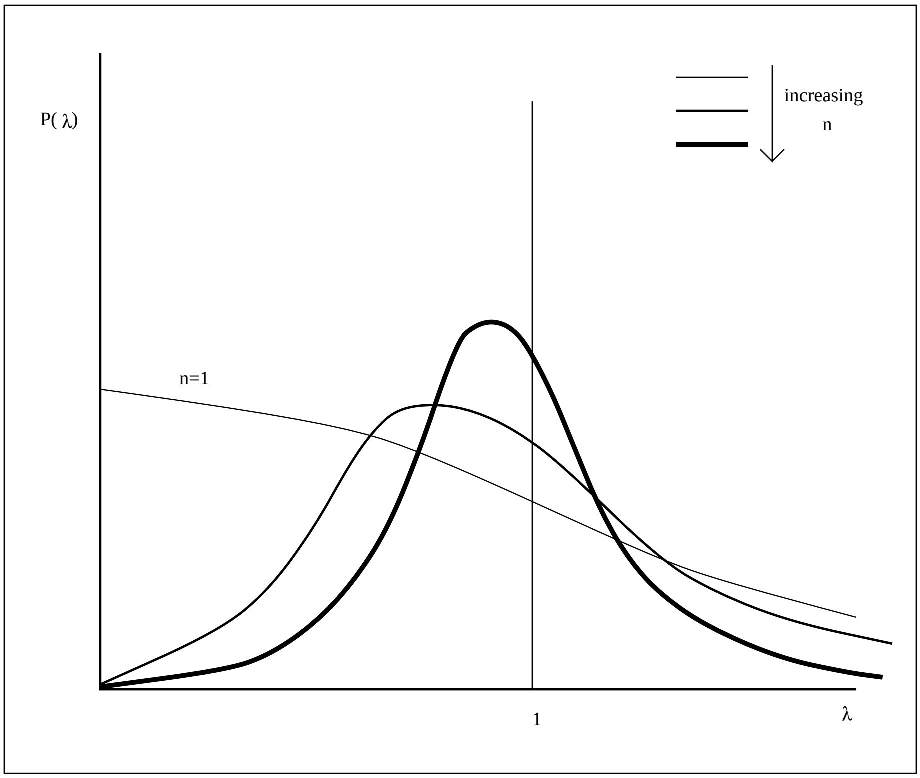

Figure 5. Plot of $P(\lambda)$ for various values of n. As $n \to \infty$ $P(\lambda)$ tends to a gaussian centered at 1. When n=1, $P(\lambda)$ is finite at the origin.

Thus given the data set, we settle on a model λ that maximizes the likelihood. The probability distribution for this is $P(\lambda)$. Knowing the properties of $P(\lambda)$, can we change the prediction in some deterministic way in order to improve our expected performance. More quantitatively we write $\sigma = \alpha_{correc}\hat{\sigma}{=}\alpha_{correc}\lambda\sigma$. So

$$\alpha_{correc} = \frac{1}{\lambda} \tag{21}$$

We are interested in $\langle \frac{1}{\lambda} \rangle$, the expected correction factor. One might also be interested in knowing, on the average, by what factor we are off in the prediction of σ, i.e., $\langle \lambda \rangle$. Note that $\langle \frac{1}{\lambda} \rangle \geq \frac{1}{\langle \lambda \rangle}$. These quantities are easily calculated from $\mathrm{P}(\lambda)$ using the identity (B-8)

$$\begin{aligned} \alpha_{correc} = \left\langle \frac{1}{\lambda} \right\rangle &= \sqrt{\left(\frac{n}{2}\right)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \\ \frac{1}{\langle \lambda \rangle} &= \sqrt{\left(\frac{n}{2}\right)} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} \end{aligned} \tag{22}$$

Values of the correction factor for various values of n are given in table 1. Note that for $n = 100$ the correction is about 0.8% which is not trivial.

We summarize the correction method, assuming that the models are predicting the mean well: Train the models using maximum likelihood and arrive at a model to be used for future predictions. Now given a new test input, obtain this model's prediction and correct by the correction factor described above.

ii. Example: Training Neural Networks Using maximum Likelihood

In this example the model is a neural network with 131 weights that are involved in the training of the model to predict σ (Weigend, 1995). The input variable which we have called $I_{t-\Delta t}$ is $x \in [0, \pi]$. The mean $\mu(x)$ and variance $\sigma^2(x)$ are functions of x and the model is a mapping $x \to \begin{bmatrix} \hat{\mu}(x) \\ \hat{\sigma}^2(x) \end{bmatrix}$. The network was trained on n examples by altering the weight in the direction that increased the likelihood that the data occurred under the model. The final network arrived at is used as the final model. It is to this network that we wish to apply the correction factor. A question arises as to what the effective number of data points (n_{eff}) is. Each parameter can be regarded as being trained on some of the data points. So n_{eff} should be approximately proportional to the number of examples, $n_{eff} \sim \gamma n$. Using this relationship we can get a theoretical prediction to compare with the experimental value. The results are summarized in table 2*

The agreement between the theoretical values and the experimental values seems convincing especially as the number of data points increases. Also note that the variance in the experimental values is relatively small implying that the network has indeed settled on a model with almost a constant parameter of λ. How one would calculate n_{eff} for a general

* We thank Zehra Cataltepe of the Learning Systems Group at Caltech for use of the results from her experiments in verifying the method of Weigend(1995). $\langle \alpha_{correc} \rangle$ was computed from the results only where σ was larger than a threshold because where σ is small, the behaviour is erratic. The mean as expected was learned well, so we can apply the analysis above where we assumed the mean was being predicted exactly.

n	$\langle\lambda\rangle$		$\langle\frac{1}{\lambda}\rangle$=$\alpha_{correc}$	
	expt.	theory	expt	theory
50	0.68±0.18	0.56±0.07	1.25±0.18	1.98±0.58
100	0.66±0.04	0.70±0.07	1.24±0.03	1.27±0.12
150	0.67±0.05	0.77±0.06	1.22±0.04	1.16±0.06
300	0.88±0.14	0.86±0.04	1.07±0.09	1.07±0.02
500	0.92±0.05	0.91±0.03	1.04±0.03	1.04±0.01

Table 2: Comparison of the theoretical results that are predicted from the induced probability distribution over λ with the experimentally obtained values for learning the variance with a neural network. The theoretical predictions were obtained by assuming a proportional relationship between n_{eff} and n, fitting γ to the data using the expected theoretical form. For $\langle\lambda\rangle$, γ was 0.006±0.002 and for$\langle\frac{1}{\lambda}\rangle$, γ was 0.044±0.014. It is interesting to compare $\frac{1}{\gamma}$ to the number of weights=N_w. $\frac{1}{N_w}$ = 0.008.

class of models with a given learning algorithm is not yet obvious. In our discussion we have assumed that the class of models is good enough to implement the various models with parameter λ. Exactly how we search through this space and how the models are parametrized are expected to affect n_{eff}.

5. Conclusions

It seems appropriate to summarize the path that we have followed in this paper. We started out by setting up a framework for comparing between models. In this framework, we used maximum likelihood to compare between models and we found that this leads to choosing wrong models. The results of the maximum likelihood analysis can be summarized by the picture in Fig. 6 .

We find a systematic underestimation of the variance in time series analogous to that of a sample variance. However when the mean $\mu(t)$ is given, this underestimation persists. When the mean is predicted well, we attempt to correct for the systematic underprediction of the variance by multiplying by a correction factor, α_{correc}. We find that this correction factor is economically significant even for a large number of data points. In this way we are able to choose a model from a class of models that were trained on different data, and then improve that model using the correction factor. Unfortunately this will not work when the mean is not being predicted well. In this case, the variance will tend to be over predicted to compensate for the bias. Thus one would like to have a way to predict the variance without

	IN SAMPLE	OUT OF SAMPLE
FINITE DATA	Choosing the model that maximizes the likelihood will yield a model that systematically predicts lower variance, even if the mean is predicted well.	Probability of choosing the wrong model is $> \frac{1}{2}$ for some 'worse' models.
EXPECTED VALUE	NOT APPLICABLE	Possible to find models 1,2 with model 1 worse than 2 but $\langle log\, l_1 \rangle > \langle log\, l_2 \rangle$.

Figure 6: Diagram depicting what could go wrong with the Maximum likelihood scheme in the 3 possible cases.

having to predict the mean. This is a direction open for future work as is the question of combining models that have been trained on different data. It is necessary, however, to have a "good" way to compare models before one even starts to think about combining models.

6. Acknowledgments

We would like to thank Dr. Amir Atiya, Joseph Sill and Zehra Cataltepe for helpful discussion.

APPENDIX A

In this appendix we derive the technical results used in section 3. The a priori model that will be assumed is gaussian. It will be shown that there are instances where using maximum likelihood to compare two models will lead to choosing the 'wrong' model most of the time. Suppose we have a data point(d_α) drawn from the actual distribution. Given a model $\{\hat{\mu}_\alpha,\hat{\sigma}_\alpha\}$ the likelihood of the data is

$$l(d_\alpha \mid \hat{\mu}_\alpha, \hat{\sigma}_\alpha) \equiv N_{d_\alpha}(\hat{\mu}_\alpha, \hat{\sigma}_\alpha) = \frac{1}{\sqrt{2\pi\hat{\sigma}_\alpha^2}} e^{-\frac{(d_\alpha - \hat{\mu}_\alpha)^2}{2\hat{\sigma}_\alpha^2}} \qquad (A-1)$$

n data points d_α are drawn independently from distributions with (possibly varying) actual means (μ_α) and (possibly varying) actual variances (σ_α). Assume that all models are predicting the mean correctly ($\hat{\mu}_\alpha^i = \mu_\alpha$), while the variances predicted by the models are given by

$$\hat{\sigma}_\alpha^i = \lambda_\alpha^i \sigma_\alpha, \quad \lambda_\alpha^i \geq 0 \qquad (A-2)$$

The likelihood of model i with parameters $\vec{\lambda}_i$ is

$$l(\vec{d} \mid \vec{\lambda}_i) = \prod_\alpha N_{d_\alpha}(\mu_\alpha, \lambda_\alpha^i \sigma_\alpha) \qquad (A-3)$$

Treat the data outcome $\vec{d}$ as a random variable $\vec{x}$. Then this likelihood itself is a random variable, depending on the parameter $\vec{\lambda}$. We seek the probability that the random variable with parameter $\vec{\lambda}$ is greater than the random variable with parameter $\vec{\lambda} = \vec{1}$, i.e., we are asking for the frequency with which the 'worse' model is chosen over the *actual* model. This probability is given by

$$P_{\vec{\lambda}} = \mathrm{Prob}[l_{\vec{\lambda}} \geq l_{\vec{1}}] = \int_{\{l_{\vec{\lambda}} \geq l_{\vec{1}}\}} \mathrm{d}^n\mathrm{x}\, l(\vec{x} \mid \vec{\lambda} = \vec{1}) \qquad (A-4)$$

The boundary condition for the integral is non trivial. The condition $l_{\vec{\lambda}} \geq l_{\vec{1}}$ implies $log(l_{\vec{\lambda}}) \geq log(l_{\vec{1}})$ due to the monotonicity of the log function. Thus the boundary condition reduces to

$$\sum_\alpha \frac{(x_\alpha - \mu_\alpha)^2}{2\sigma_\alpha^2}[1 - \frac{1}{\lambda_\alpha^2}] \geq \sum_\alpha log(\lambda_\alpha) \qquad (A-5)$$

The boundary condition is symmetric in i, the constraint on the λ_i's is independent of i and the likelihood function itself is symmetric in i so we expect the condition for $P_{\vec{\lambda}}$ to be maximized to be symmetric in i. With this as motivation, we consider the class of models for which $\vec{\lambda} = \lambda\vec{1}$, i.e., all λ_i's constant.

The boundary condition now reduces to

$$\begin{cases} \sum\limits_{\alpha} \frac{(x_\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2} \leq \beta n, & 0<\lambda<1 \\ \sum\limits_{\alpha} \frac{(x_\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2} \geq \beta n, & 1<\lambda \end{cases} \qquad \beta = \frac{\lambda^2 \, log\, \lambda}{\lambda^2-1}, \quad \beta>0 \tag{A-6}$$

$$0<\lambda<1 \Rightarrow 0<\beta<\frac{1}{2}, \qquad \lambda \to 0,1 \Rightarrow \beta \to 0, \frac{1}{2} \tag{A-7}$$

The integral can now be reduced to one dimensional form by changing variables to $u_\alpha = \frac{(x_\alpha-\mu_\alpha)}{\sigma_\alpha}$ and then transforming to spherical coordinates with the aid of the following result derivable by elementary methods

$$\int_{r<R} \mathrm{d}^n\mathrm{x}\, f(r) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \int_0^R \mathrm{dr}\; r^{n-1} f(r) \tag{A-8}$$

Thus with some manipulation one finally gets

$$P_{\vec{\lambda}} = P_\lambda = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})} \int_0^{\beta n} \mathrm{dy}\;\; y^{\frac{n}{2}-1} e^{-y} & 0<\beta<\frac{1}{2} \\ \frac{1}{\Gamma(\frac{n}{2})} \int_{\beta n}^{\infty} \mathrm{dy}\;\; y^{\frac{n}{2}-1} e^{-y} & \frac{1}{2}<\beta \end{cases} \tag{A-9}$$

When n is even we can get an exact answer. When n is odd, and large, we can get an answer in terms of the asymptotic form for the error function. We pursue the case n even, leaving the rest to numerical analysis. We also restrict ourselves to the interesting domain $0<\lambda<1$. The case $1<\lambda$ follows an identical analysis. We rewrite (A-9) as

$$\frac{(-1)^{m-1}}{\Gamma(m)} \left[\left(\frac{\partial}{\partial q}\right)^{m-1} \int_0^{2\beta m} dy\;\; e^{-qy} \right]_{q=1} \tag{A-10}$$

where $n=2m$. Doing the integral followed by the derivatives and noting that $\Gamma(m)=(m-1)!$ for positive integer m yields after some algebra

$$P_\lambda = 1 - e^{-\beta n} \sum_{k=0}^{\frac{n}{2}-1} \frac{(\beta n)^k}{k!} \tag{A-11}$$

We use the following lemma to discover an asymptotic form for this as $n \to \infty$

Lemma:

$$\lim_{n\to\infty} \left[e^{-n} \sum_{k=0}^{[f(n)]} \frac{n^k}{k!} \right] = \int_{-\infty}^{\alpha} \mathrm{du}\, \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} \qquad \alpha = \lim_{n\to\infty} \left[\sqrt{n} \left(\frac{f(n)}{n} - 1 \right) \right] \tag{A-12}$$

The asymptotic form of P_λ has $\alpha = \lim\limits_{n\to\infty} \left[\sqrt{\frac{n}{\beta}} \left((\frac{1}{2}-\beta) - \frac{1}{n} \right) \right]$. Thus for $\alpha<0$ ($\Rightarrow \mathrm{P}_\lambda > \frac{1}{2}$), $\beta(n)$ must approach $\frac{1}{2}$ faster than $\frac{1}{n}$. We can make this more precise. Suppose that $\lambda = 1-g(n)$ and $\beta = \frac{1}{2} - \eta(n)$. Then expanding to second order using the definition of $\beta(\lambda)$ one finds

$$g(n) = 2\eta(n) - \frac{2}{3}\eta^2(n) + \vartheta(\eta^3) \qquad \eta(n) = \frac{g(n)}{2} + \frac{g^2(n)}{12} + \vartheta(g^3) \tag{A-13}$$

Expanding the expression for P_λ in $\eta(n)$ we can get an asymptotic form as $\beta \rightarrow \frac{1}{2}$. We can then answer the question: what are the values of β that give $P_\lambda > \frac{1}{2}$. Tedious but elementary algebra yields

$$P_\lambda \overset{\eta\to 0}{\sim} 1 - e^{-\frac{n}{2}} \sum_{k=0}^{\frac{n}{2}-1} \frac{1}{k!}(\frac{n}{2})^k - \epsilon(n)[\eta(n) + \eta^2(n) + \vartheta(n\eta^3)] \qquad \epsilon(n) = \frac{ne^{-\frac{n}{2}}(\frac{n}{2})^{\frac{n}{2}-1}}{(\frac{n}{2}-1)!} \qquad (A-14)$$

Using Stirling's approximation for the factorial function, we can now get an expression for that $\lambda_{min}(n)$ for which $P_\lambda = \frac{1}{2}$. Thus $P_\lambda > \frac{1}{2}$ for $\lambda_{min}(n) < \lambda < 1$

$$\lambda_{min}(n) \overset{n\to\infty}{\sim} 1 - \frac{1}{n}\left[0.6 - \frac{0.3}{n}\right] \qquad (A-15)$$

As $\lambda \rightarrow 1^-$, $P_\lambda(n)$ tends toward its maximum value ($P_{max}(n)$) for given n

$$P_\lambda \xrightarrow{\lambda\to 1^-} P_{max}(n) = \frac{1}{2} + \frac{1}{\sqrt{12\pi n}}\left[1 - \frac{1}{36n}\right] \qquad (A-16)$$

Note the fact that for n fairly large, $\lambda_{min}(n)$ is significantly less than 1. All models that lie in the region between $\lambda_{min}(n)$ and 1 are more likely to be chosen than the actual model, so the question arises as to whether one can compensate for this systematic underestimation of the variance when the maximum likelihood scheme is used. This is the subject of Appendix B.

APPENDIX B

In this appendix we derive the probability distribution used in the development of the λ correction factor.A similar analysis as that which led up to (A-6) yields the following conditions

$$l_{\lambda_1} \leq l_\lambda \Rightarrow \begin{cases} \sum_i \frac{(x_i-\mu_i)^2}{2\sigma_i^2} \leq n\beta(\lambda,\lambda_1), & \lambda < \lambda_1 \\ \sum_i \frac{(x_i-\mu_i)^2}{2\sigma_i^2} \geq n\beta(\lambda,\lambda_1), & \lambda > \lambda_1 \end{cases} \qquad \beta(\lambda,\lambda_1) = \frac{\lambda^2\, log\left(\frac{\lambda}{\lambda_1}\right)}{\left[\frac{\lambda^2}{\lambda_1^2}-1\right]}, \quad \beta > 0 \qquad (B-1)$$

What is needed is that $l_{\lambda+\frac{d\lambda}{2}} \leq l_\lambda$ and that $l_{\lambda-\frac{d\lambda}{2}} \leq l_\lambda$. From (B-1) this is equivalent to the condition that the data satisfy

$$n\beta(\lambda,\lambda-\tfrac{d\lambda}{2}) \leq \sum_\alpha \frac{(x_\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2} \leq n\beta(\lambda,\lambda+\tfrac{d\lambda}{2}) \qquad (B-2)$$

So to get the probability that this is true we simply need to integrate the probability density for the x_i's over the region where this condition is true. Thus we want

$$\int_{n\beta(\lambda,\lambda-\frac{d\lambda}{2})\leq\sum_\alpha \frac{(x_\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2}\leq n\beta(\lambda,\lambda+\frac{d\lambda}{2})} \mathrm{d}^n\mathrm{x}\ \prod_\alpha N_{x_\alpha}(\mu_\alpha,\sigma_\alpha) \qquad (B-3)$$

This is the probability that the data fall within the range required to ensure that $l_{\lambda+\frac{d\lambda}{2}} \leq l_\lambda$ and that $l_{\lambda-\frac{d\lambda}{2}} \leq l_\lambda$. But this precisely the probability that our algorithm stops at λ. Using (A-8) we can reduce this integral into one dimensional form as was done for (A-9). Thus the probability that $\lambda' \in [\lambda, \lambda+\frac{d\lambda}{2}]$ is given by

$$P(\lambda)\frac{d\lambda}{2} = \frac{1}{\Gamma(\frac{n}{2})}\int_{n\beta(\lambda,\lambda-\frac{d\lambda}{2})}^{n\beta(\lambda,\lambda+\frac{d\lambda}{2})} \mathrm{dy}\ y^{\frac{n}{2}-1}e^{-y} \qquad (B-4)$$

The idea now is to expand the limits in dλ. To this end we require the expansion $\beta(\lambda,\lambda+\delta) = \frac{\lambda^2}{2}\left[1+\frac{\delta}{\lambda}\right] + \vartheta(\delta^2)$. We use this to expand the limits of integration and then using the substitution $nx = y - n\frac{\lambda^2}{2}$ the integral reduces to

$$P(\lambda)d\lambda = \frac{2n^{\frac{n}{2}}e^{-n\frac{\lambda^2}{2}}}{\Gamma(\frac{n}{2})}\left(\frac{\lambda^2}{2}\right)^{\frac{n}{2}-1}\int_{-\frac{\lambda}{4}d\lambda}^{\frac{\lambda}{4}d\lambda} \mathrm{dx}\ \left[1+\frac{2x}{\lambda^2}\right]^{\frac{n}{2}-1} e^{-nx} \qquad (B-5)$$

Expanding the integrand as a taylor series in x and doing the integral gives a result as a taylor series in $d\lambda$. Taking the limit as $d\lambda \to 0$ one finally gets

$$P(\lambda)d\lambda = \frac{n^{\frac{n}{2}}e^{-n\frac{\lambda^2}{2}}}{\Gamma(\frac{n}{2})}\left(\frac{\lambda^2}{2}\right)^{\frac{n}{2}-1} \lambda d\lambda \qquad (B-6)$$

Thus we have found, using this method of model selection, the probability that a model with parameter λ is obtained. We are interested in quantities like the expected correction factor

$$\alpha_{correc} = \frac{1}{\lambda} \tag{$B-7$}$$

These quantities are easily calculated from $\mathrm{P}(\lambda)$ using the identity

$$\langle \lambda^m \rangle = \int_0^\infty \mathrm{d}\lambda\, \lambda^m \left\{ \frac{n^{\frac{n}{2}} e^{-n\frac{\lambda^2}{2}}}{\Gamma(\frac{n}{2})} \left(\frac{\lambda^2}{2}\right)^{\frac{n}{2}-1} \lambda \right\} = \left(\frac{2}{n}\right)^{\frac{m}{2}} \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \tag{$B-8$}$$

One finds by substituting the value $m = \pm 1$ that

$$\begin{aligned} \alpha_{correc} = \left\langle \frac{1}{\lambda} \right\rangle &= \sqrt{\left(\frac{n}{2}\right)} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \\ \frac{1}{\langle \lambda \rangle} &= \sqrt{\left(\frac{n}{2}\right)} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})} \end{aligned} \tag{$B-9$}$$

Thus we have done what we set out to achieve at the beginning of this appendix. Given that a model was selected (trained) using the maximum likelihood criterion, we have found a correction for the systematic underestimation of the variance. If one wishes, one could also calculate the variances of these correction factors.

REFERENCES

Black, F. and Scholes, M.S., ‘The Pricing of Options and Corporate Liabilities’, Journal of Political Economy, **3** (1973), 637-654.

Bollerslev, T., ‘Generalized Autoregressive Conditional Heteroscedasticity’, Journal of Econometrics, **31** (1986), 307-327.

Crouhy,M. and Galai,D., ‘Hedging With a Volatility Term Structure’, The Journal of Derivatives, **Spring** (1995), 45-52.

Engle, R.F., ‘Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation’, Econometrica, **50** (1982), 987-1008.

Fama, E.E., ‘The Behavior of Stock Market Prices’, Journal of Business, **38** (1965), 34-105.

French, K.R., 'Stock Returns and The Weekend Effect', Journal of Financial Economics, **8** (1980), 55-69.

French, K.R., Schwert, G.W. and Stambaugh, R.F., 'Expected Stock Returns and Volatility', Journal of Financial Economics, **19** (1987), 3-29.

Hull, John C., Options, Futures and other Derivative Securities 2nd Ed., Englewood Cliffs, New Jersey: Prentice Hall, 1993.

Hull, J. and White, A., 'The Pricing of Options on Assets with Stochastic Volatilities', Journal of Finance, **2** (1987), 281-300.

Ito, K., 'On Stochastic Differential Equations', Memoirs, American Mathematical Society, **4** (1951), 1-51.

Kat,H.M.,'Replicating Ordinary Call Options: A Stochastic Simulation Study', Presented at the 13th AMEX Options and Derivatives Colloquium, New York (1993).

Nelson, D.B., 'Conditional Heteroscedasticity in Asset Returns: A New Approach', Econometrica, **59** (1991), 347-370.

Poterba, J. and Summers, L., 'The Persistence of Volatility and Stock Market Fluctuations', American Economic Review, **76** (1986), 1142-1151.

Schwert, G.W., 'Why Does Stock Market Volatility Change Over Time ', Journal of Finance, **44** (1989), 1115-1153.

Shiller, Robert J., Market Volatility, Cambridge, Massachusetts: The MIT Press, 1993.

Valavanis, Stefan, Econometrics; an introduction to maximum likelihood methods, New York: McGraw-Hill, 1959.

Weigend, A.S., Nix, D.A., 'Learning Local Error Bars for Nonlinear Regression', Advances in Neural Information Processing Systems (NIPS): Proceedings of the 1994 Conference, Tesauro G., Touretzky D., Leen T., eds. **7** (1995), 489-496.