THE UNIVERSITY OF SYDNEY

Economics Working Paper Series

2013 - 08

Specification Tests of Calibrated Option Pricing Models

Robert Jarrow & Simon Kwok

May 2013

Specification Tests of Calibrated Option Pricing Models

Robert Jarrow and Simon Kwok
Cornell University and University of Sydney

May 31, 2013

Abstract

In spite of the popularity of model calibration in finance, empirical researchers have put more emphasis on model estimation than on the equally important goodness-of-fit problem. This is due partly to the ignorance of modelers, and more to the ability of existing statistical tests to detect specification errors. In practice, models are often calibrated by minimizing the sum of squared difference between the modelled and actual observations. It is challenging to disentangle model error from estimation error in the residual series. To circumvent the difficulty, we study an alternative way of estimating the model by exact calibration. We argue that standard time series tests based on the exact approach can better reveal model misspecifications than the error minimizing approach. In the context of option pricing, we illustrate the usefulness of exact calibration in detecting model misspecification. Under heteroskedastic observation error structure, our simulation results shows that the Black-Scholes model calibrated by exact approach delivers more accurate hedging performance than that calibrated by error minimization.

1 Introduction

Calibration is a method to obtain the parameters of a parametric model. Calibration is essential when it is difficult, if not impossible, to estimate the model's parameters directly using historical time series data. For example, in equity option pricing calibration can be used to estimate the probability of a severe market crash, a rare event. Directly estimating the probability of a market crash using time series data is problematic due to structural shifts in the economy and the sparsity of market crashes in time series data. Other examples include estimating the probability of a firm's or country's debt defaulting using credit default swaps (CDS) or the probability of a default contagion occurring using collateralized debt obligations (CDOs). Even when it is not essential, calibration is often used for convenience. Common examples include estimating a stock return's volatility using options or the expected inflation rate using Treasury inflation protected securities (TIPs).

The problem with using calibrated parameters is that the calibration procedure embeds any model misspecification into the estimated parameter. This generates two difficulties: (1) if the purpose of the calibration is to test the model, then calibration may lead to an inappropriate acceptance (and usage) of the model, and (2) if the purpose is to use the calibrated parameters for a secondary reason, the parameter estimates may be significantly

biased. Both of these difficulties were present in the use of calibrated parameters by the financial industry prior to the 2007 credit crisis, and there is significant evidence that this misuse was a contributing cause of the crisis (see Jarrow (2011)).

There are two methods for estimating calibrated parameters: exact and error minimizing. Under *exact calibration*, the model parameters are the exact solution to an equation that matches the observed value to its theoretical model counterpart. Under *error minimizing calibration*, the model parameters are the error minimizing solution to a set of equations matching the observed values to their model counterparts. Of course, if the model is properly specified, then error minimizing calibration is equivalent to exact calibration. The two approaches only differ when the model is misspecified (this is the strictest sense of correct specification).

Although calibration is widely used in both practice and academics[1], the econometric foundations for calibration is lacking and formal statistical tests for the goodness-of-fit of a calibrated model unavailable. The purpose of this paper is to fill this gap by providing a set of statistical procedures for testing a calibrated model's validity, and consequently for testing the validity of the calibrated parameters themselves. For the reasons previously discussed, without loss of generality, we focus on exact calibration. For clarity of presentation, we also focus on the use of calibration for option pricing models, although the statistical procedures can be more widely applied.

An outline for this paper is as follows. Section 2 discusses calibration in the context of option pricing. Section 3 decomposes the calibration estimation errors excluding observation noise. The acceptable error structure (properties) of a "good" model, for misspecification testing, is contained in section 4. Section 5 adds observation noise and characterizes the properties of the parameter estimates. The prime example of option pricing calibration using the Black-Scholes Model is contained in section 6. Section 7 discusses a collection of standard statistical tests useful for testing model misspecification. Section 8 presents simulations to investigate the small sample properties of the suggested statistics, section 9 applies the methodology to S&P 500 index options to test the Black-Scholes option pricing model, and section 10 concludes the paper.

2 Calibration

We observe a panel of option prices over a cross section of n options and a sample period $[0, T]$. Let m_{it} $(i = 1, \ldots, n,\ t = 1, \ldots, T)$ be the observed price of the i^{th} option at time t, with strike price K_{it} and time to maturity τ_{it}. Suppose that all the options are written on a common underlying stock with price S_t and dividend rate q_t. Let r_t be the risk-free interest rate. We collect all of these observables into the vector $\mathbf{z}_{it} = (K_{it}, \tau_{it}, S_t, q_t, r_t)$. The resultant data set consists of option prices and observables $\mathbf{D} = \{(m_{it}, \mathbf{z}_{it}) : i = 1, \ldots, n; t = 1, \ldots, T\}$.

Suppose the option prices come from the data generating process (DGP), $\mathcal{M}_{it}(\vartheta) \equiv \mathcal{M}(\vartheta; \mathbf{z}_{it})$, where ϑ is the true unknown parameter vector invariant over t. When there is no observation error, we have

$$\mathcal{M}_{it}(\vartheta) \equiv m_{it} \tag{1}$$

[1] For example, Bates (2000) and Pan (2000).

for $i = 1, \ldots, n; t = 1, \ldots, T$. We will discuss the case with observation error in Section 5.

2.1 SSE Calibration

One popular way of estimating the model is *sum of squared error (SSE) minimization calibration*. In SSE calibration the modeler first chooses a parametric option pricing model $M(\boldsymbol{\theta}, \mathbf{z})$ indexed by a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$, so that $M_{it}(\boldsymbol{\theta}) \equiv M(\boldsymbol{\theta}, \mathbf{z}_{it})$ is the theoretical price of the i^{th} option at time t for vector $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_0$ be the "pseudo-true" value of the parameter to be estimated.

Given $\boldsymbol{\theta}$, the *pricing error* matrix $\mathbf{e} \equiv \mathbf{e}(\boldsymbol{\theta}) = (e_{it}(\boldsymbol{\theta}))$ captures the difference between all theoretical and observed option prices, with individual elements

$$e_{it}(\boldsymbol{\theta}) = m_{it} - M_{it}(\boldsymbol{\theta}) \tag{2}$$

for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

The modeler chooses a norm $L(\mathbf{e}(\boldsymbol{\theta}); \mathbf{D})$, a loss function, that takes the pricing errors as inputs and provides an aggregate measure of loss caused by the pricing errors deviating from zero. This calibration is called the *sum of squared error* approach if the loss function $L_2(\mathbf{e}(\boldsymbol{\theta}); \mathbf{D}) = \sum_{i=1}^{n} \sum_{t=1}^{T} w_{it} e_{it}^2$, with weights $w_{it} = m_{it}^{-2}$ is used.

The modeler minimizes the loss function and obtains the solution

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\mathbf{e}(\boldsymbol{\theta}); \mathbf{D}).$$

If the model specification is perfect, i.e. the proposed model M coincides the DGP $\mathcal{M}$, then the estimated pricing residuals will be identically zero, i.e. $\hat{\mathbf{e}} = \mathbf{e}(\widehat{\boldsymbol{\theta}}) \equiv 0$. In this case, it is easy to determine if the model is accepted. However, because models are approximations of a complex reality, it is rarely the case that M and $\mathcal{M}$ coincide, and an alternative criteria for accepting a model is required. It is reasonable to accept a model with non-zero pricing errors as a "good approximation" if the model only differs from the DGP in a random and unpredictable fashion. Such a "good approximation" can be characterized by requiring that the residuals $\hat{\mathbf{e}} = \mathbf{e}(\widehat{\boldsymbol{\theta}})$ satisfy a set of properties $\mathbf{P}$ which capture the random and unpredictable behavior. A collection of such properties $\mathbf{P}$ is discussed in section 3 below. Given these properties, specification tests can be performed and the model accepted or rejected (see Section 7).

2.2 Exact Calibration

An alternative method for estimating an option pricing model is exact calibration. Under this approach, the modeler matches the observed option price to the theoretical counterpart $M_{it}(\boldsymbol{\theta}) \equiv M(\boldsymbol{\theta}; \mathbf{z}_{it})$. For all $i \in \{i_1, \ldots, i_d\}$ and $t = 1, \ldots, T$, the following equality holds:

$$m_{it} = M(\boldsymbol{\theta}_t; \mathbf{z}_{it}).$$

Using vector notation,

$$\mathbf{m}_t = \mathbf{M}(\boldsymbol{\theta}_t; \mathbf{Z}_t),$$

where $\mathbf{m}_t = (m_{i_1 t}, \ldots, m_{i_d t})'$ and $\mathbf{Z}_t = \left(\mathbf{z}'_{i_1 t}, \ldots, \mathbf{z}'_{i_d t}\right)$. Recall that d is the dimension of $\boldsymbol{\theta}_t$.

A crucial feature of exact calibration is that it requires d equations to match the market prices to the theoretical prices to solve for the d unknown parameters. A unique solution $\hat{\boldsymbol{\theta}}_t$ exists if the function is invertible:

$$\hat{\boldsymbol{\theta}}_t = \mathbf{M}^{-1}(\mathbf{m}_t; \mathbf{Z}_t).$$

In the sequel, we assume that the model has a scalar parameter, i.e. $d = 1$.[2] For notational convenience, we drop the subscript i.

To ensure a unique solution $\hat{\theta}_t = M^{-1}(m_t; \mathbf{z}_t)$ for each $t = 1, \ldots, T$, we need to impose one of the following monotonicity assumptions.

Assumption SM: The option pricing function $M(\theta; \mathbf{z}_t)$ is strictly monotone in θ.

Assumption LSM: The option pricing function $M(\theta; \mathbf{z}_t)$ is strictly monotone in θ in a neighborhood $\Theta_0 = (\theta_0 - \epsilon, \theta_0 + \epsilon)$ of the hypothetical parameter θ_0 under model M.

An example is the Black-Scholes (1973) model $BS(\theta_t; \mathbf{z}_t)$ where θ_t corresponds to the stock's volatility, a scalar. The Black-Scholes model satisfies the strict monotonicity assumption for the volatility parameter. The unknown volatility θ_t can be found by solving the equation

$$m_t = BS(\theta_t; \mathbf{z}_t)$$

for each t. The solution

$$\hat{\theta}_t \equiv IV_t = BS^{-1}(m_t; \mathbf{z}_t) \tag{3}$$

is commonly known as the *implied volatility*.

In SSE calibration, as discussed in the previous section, the procedure for testing a model is based on examining the properties of the pricing errors. For exact calibration, however, there are no pricing errors, i.e. we *always* have $M(\hat{\theta}_t; \mathbf{z}_t) = \mathcal{M}(\vartheta; \mathbf{z}_t) = m_t$ for all $t = 1, \ldots, T$. Consequently, an alternative procedure for accepting or rejecting a model needs to be developed.

If the model specification is perfect, i.e., the proposed model M coincides the DGP $\mathcal{M}$, then all the solutions from exact calibration are identical to the true parameter: $\hat{\theta}_t = \vartheta$ for all $t = 1, ..., T$. In this case, it is easy to determine whether the model should be accepted. The model should be accepted if the estimated parameter is a constant for all t. Of course, because all models are approximations of a complex reality, no model will be accepted using this criteria. Hence, we seek an alternative criteria for determining whether the model is a "good approximation."

As before, it is reasonable to accept an exact-calibrated model as a "good approximation" if the model only differs from the DGP in a random and unpredictable fashion. Such a "good approximation" can be characterized by requiring that the *parameter estimation error*,

$$\varepsilon_t = \hat{\theta}_t - \vartheta \tag{4}$$

[2]The case of $k > 1$ is dealt with in a separate paper. Both SSE and exact calibration can handle multiple parameter calibration.

satisfies a set of properties **P** which capture the random and unpredictable behavior of the pricing model. A collection of such properties **P** is discussed in section 3 below. Given these properties, specification tests can be performed and the model accepted or rejected (see Section 7).

Because the true parameter ϑ of the DGP $\mathcal{M}$ is unknown, we need to test for properties **P** using the parameter estimation error instead, i.e.

$$\hat{\varepsilon}_t = \hat{\theta}_t - \bar{\theta}, \tag{5}$$

where

$$\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t.$$

We will rely on a battery of statistical tests on the estimated errors $\hat{\varepsilon}_t$ to study whether the randomness is systematic or not (see Section 7).

3 The Error Decomposition

To develop the properties **P** used to accept or reject a pricing model, it is necessary to study a decomposition of the pricing errors. This decomposition is discussed for both SSE and exact calibration.

3.1 SSE Calibration

Assume that the DGP is $\mathcal{M}_t(\vartheta)$ with the true parameter ϑ. The proposed model is $M_t(\theta)$ with the *pseudo-true parameter* θ_0. Under the SSE calibration approach, there are three types of errors:

1. **Measurement error**: the difference between the estimated model price and the market price,
$$\hat{e}_t = M_t(\widehat{\theta}) - m_t. \tag{6}$$

2. **Model error**: the difference between the true (unknown) DGP and model M evaluated at the pseudo-true parameter θ_0,
$$e_t = \mathcal{M}_t(\vartheta) - M_t(\theta_0).$$

3. **Estimation error**: the price difference due to the discrepancy between the pseudo-true parameter θ_0 and the estimated parameter $\widehat{\theta}$, both associated with model M,
$$u_t = M_t(\widehat{\theta}) - M_t(\theta_0).$$
In general, $\widehat{\theta} - \theta_0$ is not zero in a finite sample.

Summing the three errors yields $\mathcal{M}_t(\vartheta) - m_t$, which is zero by the definition of DGP. The error decomposition under SSE approach is illustrated in Figure 1.

Hence, under the SSE approach, the model error plus the estimation error is equal to the measurement error, i.e.

$$\begin{aligned}\hat{e}_t &= M_t(\widehat{\theta}) - m_t \\ &= M_t(\theta_0) - m_t + M_t(\widehat{\theta}) - M_t(\theta_0) \\ &= e_t + u_t. \end{aligned} \tag{7}$$

To test a SSE calibrated model, one tests the properties $\mathbf{P}$ of the measurement errors $\hat{e}_t$ for $t = 1, ..., T$. One often ignores the estimation error component u_t in these tests by assuming certain asymptotic conditions hold. All model specification tests that are based on the measurement errors confound both the model and estimation errors.

We can understand the properties of the estimation error by performing a first-order Taylor expansion on the measurement error:

$$\begin{aligned}\hat{e}_i &= M_t(\widehat{\theta}) - m_t \\ &= M_t(\theta_0) + \left(\widehat{\theta} - \theta_0\right) M_t'(\theta_0) - m_t + o_p\left(\widehat{\theta} - \theta_0\right) \\ &= e_t + \left(\widehat{\theta} - \theta_0\right) M_t'(\theta_0) + o_p\left(\widehat{\theta} - \theta_0\right). \end{aligned}$$

Substituting back into (7), we see that the estimation error is given by

$$u_t = \left(\widehat{\theta} - \theta_0\right) M_t'(\theta_0) + o_p\left(\widehat{\theta} - \theta_0\right).$$

Under the null of a true model (and other regularity conditions), the estimator $\widehat{\theta}$ of the SSE approach (e.g. OLS estimator, GMM estimator) follows an asymptotic normal distribution with convergence rate $n^{-1/2}$, and so

$$u_t = O_p\left(\frac{1}{\sqrt{n}}\right) M_t'(\theta_0).$$

Even with a correct model specification, the estimation error u_t may exhibit heteroskedasticity (for time series) or heterogeneity (for cross section). Being contaminated by estimation error, the measurement error $\hat{e}_t$ may also exhibit heteroskedasticity/heterogeneity. The intuition is analogous to that of the linear regression model. The SSE approach is based on minimizing the sum of squared estimation errors, so that $M_t(\theta)$ is a *conditional mean model* of the observed market price m_t. The model M estimated under the SSE approach does not impose higher-order restrictions on the market prices.

3.2 Exact Calibration

Let us turn to the exact calibration approach. Conceptually, the exact approach only differs from the SSE approach in the "space" used to test the model.

Assume that the DGP is $\mathcal{M}_t(\vartheta)$ with the true parameter ϑ. The proposed model is $M_t(\theta)$ with a fixed *pseudo-true parameter* θ_0. We can also decompose an exact-calibrated model's errors in price space.

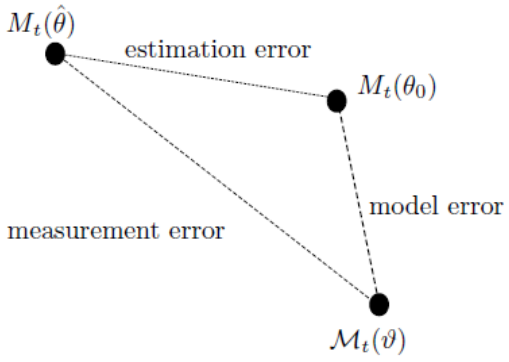

Figure 1: Error decomposition under the SSE approach.

1. **Measurement error**:

$$M_t(\widehat{\theta}_t) - \mathcal{M}_t(\vartheta) = M_t(\widehat{\theta}_t) - m_t \equiv 0, \tag{8}$$

   by (1).

2. **Model error**:

$$\mathcal{M}_t(\vartheta) - M_t(\theta_0). \tag{9}$$

3. **Estimation error**:

$$M_t(\theta_0) - M_t(\widehat{\theta}_t). \tag{10}$$

Note that for exact calibration, the model and estimation errors are identical because there is zero measurement error. The error decomposition under the exact approach is illustrated in Figure 2.

This approach offers an alternative method for detecting model misspecifications that avoids the confounding influence of estimation error. From (10), it is clear that model misspecifications are tested using the properties of $\widehat{\theta}_t$; therefore, misspecification tests are carried out in the *parameter space*.

4 Properties P

This section discusses the properties **P** that either a SSE-calibrated or an exact-calibrated model should satisfy if it is a "good approximation." For SSE calibration, these properties apply to the model errors, $e_t = \mathcal{M}_t(\vartheta) - M_t(\theta_0)$. For exact calibration, these properties apply to the parameter estimation errors, $\varepsilon_t = \hat{\theta}_t - \vartheta$. For brevity, we present the properties **P** only for exact calibration. The identical properties apply for SSE calibration substituting e_t for ε_t in the various properties.

Let the probability space be $(\Omega, \mathbb{P}, \mathcal{F})$, where $\mathcal{F} = (\mathcal{F}_t)_{t=1,\dots,T}$ is the natural filtration of $\{m_t, \mathbf{z}_t\}$, i.e. $\mathcal{F}_t$ is the sigma algebra generated by $\{(m_s, \mathbf{z}_s) : s = 1, \dots, t\}$. To study the moments of the parameter estimation errors, we need to impose the following assumption.

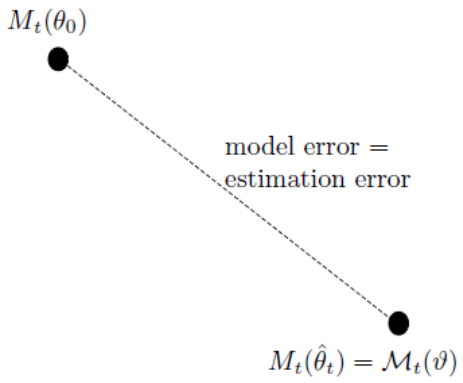

Figure 2: Error decomposition under the exact approach.

Assumption E&S: The sequence of zeros $\hat{\theta}_t$ is ergodic and second-order stationary, so that $\theta_0 \equiv E(\hat{\theta}_t)$ and $Var(\hat{\theta}_t)$ exist.

Assumption E&S implies that $E(\varepsilon_t) = 0$ and $\sigma^2 \equiv Var(\varepsilon_t) < \infty$.

With assumption E&S, we introduce the following properties **P**. The five properties are denoted **WN1**, **WN2**, **MDS1**, **MDS2**, and **IID.**

(WN1): ε_t is a white noise, i.e. for all $t = 1, ..., T$, and for all $s \neq t$,

$$E(\varepsilon_t) = 0,$$
$$Var(\varepsilon_t) = \sigma^2, \tag{11}$$
$$Cov(\varepsilon_s, \varepsilon_t) = 0. \tag{12}$$

(MDS1): ε_t is a martingale difference sequence, i.e. for all $t = 1, ..., T$, and for all $0 < s < t$, we have, $\mathbb{P}$-a.s.,

$$E(\varepsilon_t | \mathcal{F}_s) = 0. \tag{13}$$

Intuitively, MDS1 implies that the estimation error is unpredictable based on the past information given in $\mathcal{F}_s$. Note that this property imposes no restriction on the second moment of ε_t (which may even not exist).

MDS1 implies that ε_t is uncorrelated with $f(\varepsilon_s)$ for any measurable function $f(\cdot)$ and $0 < s < t$:

$$Cov(\varepsilon_t, f(\varepsilon_s)) = 0. \tag{14}$$

Additional properties are:

(WN2): $\varepsilon_t^2 - \sigma^2$ is a white noise, i.e. for all $t = 1, ..., T$, and for all $0 < s < t$,

$$\begin{aligned} E\left(\varepsilon_t^2\right) &= \sigma^2, \\ Var\left(\varepsilon_t^2\right) &= \varkappa, \\ Cov\left(\varepsilon_t^2, \varepsilon_s^2\right) &= 0. \end{aligned} \tag{15}$$

(MDS2): $\varepsilon_t^2 - \sigma^2$ is a martingale difference sequence, i.e. for all $t = 1, ..., T$, and for all $0 < s < t$,

$$E\left(\varepsilon_t^2 - \sigma^2 | \mathcal{F}_s\right) = 0.$$

Similar to MDS1, MDS2 implies the squared estimation error is unpredictable. Moving σ^2 to the right side of this expression, we see that property MDS2 is equivalent to saying that ε_t is (conditional) homoskedastic: for all $0 < s < t$, $\mathbb{P}$-a.s.,

$$E\left(\varepsilon_t^2 | \mathcal{F}_s\right) = \sigma^2.$$

Similar to the previous discussion on MDS1, MDS2 imposes no fourth order moment condition on ε_t (as it may even not exist).

Property MDS2 implies that ε_t^2 is uncorrelated with any measurable function of the past error, i.e. for any measurable function $f(\cdot)$ and $0 < s < t$:

$$Cov\left(\varepsilon_t^2, f(\varepsilon_s)\right) = 0. \tag{16}$$

(IID): ε_t is serially independent and identically distributed, i.e. for all $t_1 \neq t_2 \neq \cdots, \neq t_b$ and all integers $b > 1$,

$$\varepsilon_{t_1}, \varepsilon_{t_2}, \ldots, \varepsilon_{t_b} \text{ are i.i.d..}$$

IID implies pairwise independence: for all $t = 1, ..., T$, $0 < s < t$, and for any measurable functions $f(\cdot)$ and $g(\cdot)$,

$$Cov\left(f(\varepsilon_t), g(\varepsilon_s)\right) = 0, \tag{17}$$

provided that the covariance exists.

We can prove the following relations among the above five properties.

Lemma 1a: If $Var\left(\varepsilon_t\right)$ exists and is constant, then MDS1 $\Longrightarrow$ WN1.
Lemma 1b: If $Var\left(\varepsilon_t^2\right)$ exists and is constant, then MDS2 $\Longrightarrow$ WN2.
Lemma 2a: If $E\left(\varepsilon_t | \mathcal{F}_s\right)$ exists $\mathbb{P}$-a.s., then IID $\Longrightarrow$ MDS1.
Lemma 2b: If $E\left(\varepsilon_t^2 | \mathcal{F}_s\right)$ exists $\mathbb{P}$-a.s., then IID $\Longrightarrow$ MDS2.
Lemma 3a: (17) $\Longrightarrow$ (14) $\Longrightarrow$ (12).
Lemma 3b: (17) $\Longrightarrow$ (16) $\Longrightarrow$ (15).
Lemma 4: MDS2 $\Longrightarrow$ (11).

According to Lemmas 1a and 2a, given the existence of first conditional moment, and the constancy of the second unconditional moment, IID is the strongest assumption on ε_t,

followed by MDS1 and then WN1. Similarly, according to Lemmas 1b and 2b, given the existence of the second conditional moment, and the constancy of the fourth unconditional moment, IID is the strongest null hypothesis on ε_t, followed by MDS2 and then WN2.

Lemmas 3a and 3b exhibit links among WN, MDS and IID. In terms of the pairwise dependence measure, IID is the strongest concept, followed by MDS1/2 and then WN1/2.

Lemma 4 says that conditional homoskedasticity implies unconditional homoskedasticity.

These properties form the basis for the time series tests considered in subsequent sections.

5 Calibration with Observation Noise

In practice, we observe the true market price m_t with observation noise υ_t, i.e. we observe $m_t + \upsilon_t$. This section explores the impact that observation noise has on both SSE and exact calibration.

For analysis, we assume that the observation noise υ_t satisfies some exogeneity conditions.

Assumption SE (strong exogeneity) For all $t = 1, \ldots, T$, $\mathbb{P}$-a.s.,

$$E\left(\upsilon_t | Z\right) = 0.$$

Assumption WE (weak exogeneity) For all $t = 1, \ldots, T$, $\mathbb{P}$-a.s.,

$$E\left(\upsilon_t | Z_t\right) = 0.$$

Considering exogeneity, all of the moments and distributional results should be interpreted as conditional statements.

For inference purposes, we need two additional assumptions on the noise's covariance structure.

Assumption CV$_1$ (conditional homoskedasticity and independence) υ_t are independent with $Var(\upsilon_t | Z_t) = \sigma^2$ $\mathbb{P}$-a.s. for all t.

Assumption CV$_2$ (conditional heteroskedasticity and independence) υ_t are independent with $Var(\upsilon_t | Z_t) = \sigma_t^2$ $\mathbb{P}$-a.s. for all t.

The null and alternative hypotheses to be tested are:

$$\begin{aligned} &\mathbf{H}_0 : M_t\left(\theta_0\right) = m_t \text{ for all } t \in \{1, \ldots, T\}; \\ &\mathbf{H}_1 : M_t\left(\theta_0\right) \neq m_t \text{ for some } t \in \{1, \ldots, T\}. \end{aligned}$$

In order to test the model, we calibrate the model $M_t\left(\theta\right)$ using the noisy price observations $m_t + \upsilon_t$.

For later analyses, the model $M_t\left(\theta\right)$ has to obey some smoothness conditions.

Assumption D: The model $M_t(\theta) \equiv M_t(\theta; z_t)$ is continuously differentiable in θ and S_t, so that

$$\nabla_t(\theta) = \frac{\partial M_t(\theta)}{\partial \theta}$$

and

$$\nabla_{St}(\theta) = \frac{\partial^2 M_t(\theta)}{\partial \theta \partial S}$$

exist. Denote $\nabla_t = \nabla_t(\theta_0)$ and $\nabla_{St} = \nabla_{St}(\theta_0)$.

5.1 SSE Calibration

Under the SSE approach, the estimation error (in the price space) is:

$$e_t(\theta) = M_t(\theta) - (m_t + \upsilon_t).$$

Simplifying the loss function (letting the weights be unity), the SSE parameter estimator $\hat{\theta}$ solves the minimization problem

$$\inf_{\theta} \sum_{t=1}^{T} e_t^2(\theta). \tag{18}$$

Since $M_t(\theta)$ is nonlinear, expression (18) is a standard nonlinear least squares problem. Given the exogenous variables, $M_t(\theta)$ is the conditional mean of the market price. The estimation error has a zero conditional mean, i.e. $E(e_t(\theta)) = E(\upsilon_t) = 0$. Since the only random component is the observation noise υ_t, the conditional variance of the estimation error is $Var(e_t(\theta)) = Var(\upsilon_t)$.

Assumption $\mathbf{PL}_a^{SSE}$: A uniform weak law of large numbers applies to $\hat{S}_{\nabla\nabla}(\theta) := \frac{1}{T}\sum_{t=1}^{T} \nabla_t(\theta)\nabla_t(\theta)'$ (for $a = 1$) and $\hat{S}_{\sigma^2\nabla\nabla}(\theta) := \frac{1}{T}\sum_{t=1}^{T} \sigma_t^2 \nabla_t(\theta)\nabla_t(\theta)'$ (for $a = 2$), so that their probability limits exist in a neighborhood Θ_0 of θ_0:

$$S_{\nabla\nabla} := \operatorname*{plim}_{T\to\infty} \sup_{\theta\in\Theta_0} \hat{S}_{\nabla\nabla}(\theta),$$
$$S_{\sigma^2\nabla\nabla} := \operatorname*{plim}_{T\to\infty} \sup_{\theta\in\Theta_0} \hat{S}_{\sigma^2\nabla\nabla}(\theta).$$

Furthermore, $S_{\nabla\nabla}$ is invertible.

Given this assumption, the SSE estimator $\hat{\theta}$ has the following characteristics (see e.g., Davidson and MacKinnon, 2004):

Theorem 1 *Under* ***Assumption*** **SE** *and* $\mathbf{H}_0$, $\hat{\theta}$ *is biased in finite sample, i.e.* $E(\hat{\theta}) \neq \theta_0$.

Proof. By Jensen's inequality and nonlinearity of $M_t(\theta)$. ■

Theorem 2 *Under* ***Assumption*** **WE** *and* $\mathbf{H}_0$, $\hat{\theta}$ *is consistent in large sample, i.e.* $\hat{\theta} \xrightarrow{p} \theta_0$ *as* $T \to \infty$.

Theorem 3 *Under* ***Assumption*** **WE**, **D**, $\mathbf{CV}_1$, $\mathbf{PL}_1^{SSE}$ *and* $\mathbf{H}_0$,

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N(0, V_{SSE,1})$$

as $T \to \infty$, *with asymptotic variance*

$$V_{SSE,1} = \sigma^2 S_{\nabla\nabla}^{-1}.$$

Theorem 4 *Under* ***Assumption*** **WE**, **D**, $\mathbf{CV}_2$, $\mathbf{PL}_2^{SSE}$ *and* $\mathbf{H}_0$,

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N(0, V_{SSE,2})$$

as $T \to \infty$, *with asymptotic variance*

$$V_{SSE,2} = S_{\nabla\nabla}^{-1} S_{\sigma^2\nabla\nabla} S_{\nabla\nabla}^{-1}.$$

Using the delta method, we can derive the asymptotic distribution of $M_t\left(\hat{\theta}\right)$.

Corollary 1 *Under* ***Assumption*** **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{SSE}$ *(*$a = 1, 2$*) and* $\mathbf{H}_0$, *we have, for any fixed* t,

$$\sqrt{T}\left[M_t\left(\hat{\theta}\right) - M_t\left(\theta_0\right)\right] \xrightarrow{d} N(0, \nabla_t V_{SSE,a} \nabla_t')$$

as $T \to \infty$.

Corollary 2 *Under* ***Assumption*** **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{SSE}$ *(*$a = 1, 2$*) and* $\mathbf{H}_0$, *we have, for any fixed* t,

$$\sqrt{T}\left[\frac{\partial M_t\left(\hat{\theta}\right)}{\partial S} - \nabla_{St}\right] \xrightarrow{d} N(0, \nabla_{St} V_{SSE,a} \nabla_{St}')$$

as $T \to \infty$.

For the estimators, one uses $\hat{\theta}$ for the parameter, and $M_t(\hat{\theta})$ for the model price.

5.2 Exact Calibration

Exact calibration finds $\hat{\theta}_t$ such that

$$M_t(\hat{\theta}_t) = m_t + \upsilon_t$$

for each t. This corresponds to $e_t(\hat{\theta}_t) \equiv 0$ for all t.

We can still regard $M_t(\hat{\theta}_t)$ as a conditional mean model for the true price m_t (so that υ_t has a zero conditional mean). The key difference from SSE calibration is that the parameter estimate $\hat{\theta}_t$ varies across observations.

Define the estimation error (in the parameter space) as:

$$\varepsilon_t = \hat{\theta}_t - \theta_0.$$

We can estimate the unknown parameter θ_0 with the sample mean of calibrated parameters, i.e.

$$\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t = \theta_0 + \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t. \tag{19}$$

We now study the finite and large sample properties of $\bar{\theta}$.

Theorem 5 *Under **Assumption** SE and* $\mathbf{H}_0$*,* $\bar{\theta}$ *is biased in finite sample, i.e.* $E(\bar{\theta}) \neq \theta_0$.

Proof. By Jensen's inequality and nonlinearity of $M_t(\theta)$. ■

Theorem 6 *Under **Assumption** WE and* $\mathbf{H}_0$*,* $\bar{\theta}$ *is consistent in large sample, i.e.* $\bar{\theta} \xrightarrow{p} \theta_0$ *as* $T \to \infty$.

By a first order Taylor's expansion, we can link the errors in the parameter and price spaces. Under H_0,

$$\begin{aligned}\varepsilon_t &= \hat{\theta}_t - \theta_0 \\ &= M_t^{-1}(m_t + \upsilon_t) - M_t^{-1}(m_t) \\ &= \frac{\partial M_t^{-1}(\tilde{m}_t)}{\partial m}\upsilon_t,\end{aligned} \tag{20}$$

where $\tilde{m}_t$ lies between m_t and $m_t + \upsilon_t$ (i.e. $|\tilde{m}_t - m_t| < |\tilde{m}_t - (m_t + \upsilon_t)|$). This relationship is useful for comparing the asymptotic distributions of the SSE and exact estimators.

Averaging expression (20) over $t = 1, \ldots, n$, we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\varepsilon_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\partial M_t^{-1}(\tilde{m}_t)}{\partial m}\upsilon_t.$$

But the term on the left side is just $\sqrt{T}\left(\bar{\theta} - \theta_0\right)$, by expression (19). Note that for some $\tilde{\theta}_t$ lying between θ_0 and $\hat{\theta}_t$, we have $\frac{\partial M_t^{-1}(\tilde{m}_t)}{\partial m} = \left(\frac{\partial M_t(\tilde{\theta}_t)}{\partial \theta}\right)^{-1}$. Using the central limit theorem, we see that $\bar{\theta}$ is asymptotically normal, as given in the following theorems.

Assumption $\mathbf{PL}_a^{Exact}$: A uniform weak law of large numbers applies to $\hat{S}_{(\nabla\nabla)^{-1}}(\theta) := \frac{1}{T}\sum_{t=1}^{T}\left[\nabla_t(\theta)\nabla_t(\theta)'\right]^{-1}$ (for $a = 1$) and $\hat{S}_{\sigma^2(\nabla\nabla)^{-1}}(\theta) := \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2\left[\nabla_t(\theta)\nabla_t(\theta)'\right]^{-1}$ (for $a = 2$), so that their probability limits exist in a neighborhood Θ_0 of θ_0:

$$\begin{aligned}S_{(\nabla\nabla)^{-1}} &:= \operatorname*{plim}_{T\to\infty}\sup_{\theta\in\Theta_0}\hat{S}_{(\nabla\nabla)^{-1}}(\theta), \\ S_{\sigma^2(\nabla\nabla)^{-1}} &:= \operatorname*{plim}_{T\to\infty}\sup_{\theta\in\Theta_0}\hat{S}_{\sigma^2(\nabla\nabla)^{-1}}(\theta).\end{aligned}$$

Theorem 7 *Under **Assumption** WE,* $\mathbf{D}$*,* $\mathbf{CV}_1$*,* $\mathbf{PL}_1^{Exact}$ *and* $\mathbf{H}_0$*,*

$$\sqrt{T}\left(\bar{\theta} - \theta_0\right) \xrightarrow{d} N(0, V_{Exact,1})$$

as $T \to \infty$*, with asymptotic variance* $V_{Exact,1} = \sigma^2 S_{(\nabla\nabla)^{-1}}$.

Theorem 8 *Under* ***Assumption*** **WE**, **D**, $\mathbf{CV}_2$, $\mathbf{PL}_2^{Exact}$ *and* $\mathbf{H}_0$,

$$\sqrt{T}\left(\bar{\theta}-\theta_0\right) \xrightarrow{d} N(0, V_{Exact,2})$$

as $T \to \infty$, *with asymptotic variance* $V_{Exact,2} = S_{\sigma^2(\nabla\nabla)^{-1}}$.

Finally, using the delta method, we can derive the asymptotic distribution of $M_t\left(\bar{\theta}\right)$.

Corollary 3 *Under* ***Assumption*** **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{Exact}$ *(*$a = 1, 2$*) and* $\mathbf{H}_0$, *we have*

$$\sqrt{T}\left[M_t\left(\bar{\theta}\right) - M_t\left(\theta_0\right)\right] \xrightarrow{d} N(0, \nabla_t V_{Exact,a} \nabla_t')$$

as $T \to \infty$.

Corollary 4 *Under* ***Assumption*** **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{Exact}$ *(*$a = 1, 2$*) and* $\mathbf{H}_0$, *we have, for any fixed* t,

$$\sqrt{T}\left[\frac{\partial M_t\left(\bar{\theta}\right)}{\partial S} - \nabla_{St}\right] \xrightarrow{d} N(0, \nabla_{St} V_{Exact,a} \nabla_{St}')$$

as $T \to \infty$.

For the estimators, one uses $\bar{\theta} = \frac{\sum_{t=1}^{n} \hat{\theta}_t}{n}$ for the parameter. For the model price, there are three possible estimators:

$$\begin{aligned}
\widehat{M_t\left(\theta\right)} &= M_t\left(\hat{\theta}_t\right), \\
\overline{M_t\left(\theta\right)} &= M_t\left(\bar{\theta}\right), \\
\overline{\widehat{M_t\left(\theta\right)}} &= \frac{1}{n}\sum_{k=1}^{n} M_t\left(\hat{\theta}_k\right).
\end{aligned}$$

We discuss all of these estimators below.

5.3 A Comparison

In this section, we want to study two comparisons, using the different SSE and exact-calibration estimators for the model's price.

Theorem 9 *Suppose* ***Assumptions*** **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{SSE}$, $\mathbf{PL}_a^{Exact}$ *(*$a = 1, 2$*) and* $\mathbf{H}_0$ *hold.*

(i) With conditional homoskedastic noise ($a = 1$*),* $\widehat{M_t\left(\theta\right)}$ *is asymptotically more efficient than* $\overline{M_t\left(\theta\right)}$.

(ii) With conditional heteroskedastic noise ($a = 2$*), it is possible that* $\overline{M_t\left(\theta\right)}$ *is asymptotically more efficient than* $\widehat{M_t\left(\theta\right)}$.

Proof. *We note that both estimators are biased in finite samples, but consistent in large samples.* ■

Proof. Under **Assumption CV**$_1$, $\widehat{M_t(\theta)}$ is asymptotically more efficient than $\overline{M_t(\theta)}$, because

$$\left(\frac{1}{T}\sum_{t=1}^{T}\nabla_t\nabla_t'\right)^{-1} \leq \frac{1}{T}\sum_{t=1}^{T}\left(\nabla_t\nabla_t'\right)^{-1},$$

by Jensen's inequality, so that $V_{SSE,1} < V_{Exact,1}$ in general, when ∇_t are not identically the same for all t.

However, under **Assumption CV**$_2$, it is possible that $\overline{M_t(\theta)}$ is *asymptotically more efficient* than $\widehat{M_t(\theta)}$. Here is a simple example: $T = 2$, $\nabla_1 = \sigma_1 = 1$, $\nabla_2 = \sigma_2 = 0.2$. Then, we see that $V_{SSE,2} = 1.852 > V_{Exact,2} = 1$. ■

Theorem 10 *Under **Assumptions** **WE**, **D**, **CV**$_a$, **PL**$_a^{SSE}$, **PL**$_a^{Exact}$ ($a = 1, 2$) and* $\mathbf{H}_0$, $\overline{\widehat{M_t(\theta)}}$ *is more efficient than* $\widehat{M_t(\theta)}$. *and is as efficient as* $\overline{M_t(\theta)}$ *asymptotically.*

Proof. Let us first compare $\overline{\widehat{M_t(\theta)}}$ and $\widehat{M_t(\theta)}$ under $\mathbf{H}_0$.

For $t = k$, $M_t(\hat{\theta}_t) = m_t + \upsilon_t$, so

$$E[M_t(\hat{\theta}_t)] = m_t,$$

which also shows that $\widehat{M_t(\theta)}$ is unbiased under $\mathbf{H}_0$.

For $t \neq k$,

$$\begin{aligned} M_t(\hat{\theta}_k) &= M_t\left(M_k^{-1}(m_k + \upsilon_k)\right) \\ &= M_t\left(\theta_0 + \frac{\partial M_k^{-1}(\tilde{m}_k)}{\partial m}\upsilon_k\right). \end{aligned} \tag{21}$$

Using the delta method and the Central Limit Theorem, as $T \to \infty$,

$$\sqrt{T}\overline{\widehat{M_t(\theta)}} \xrightarrow{d} N\left(M_t(\theta_0), \sigma^2\nabla_t \operatorname*{plim}_{n\to\infty}\left[\frac{1}{T}\sum_{k=1}^{T}\left(\nabla_k\nabla_k'\right)^{-1}\right]\nabla_t'\right). \tag{22}$$

Therefore, for large T, the finite sample variance can be approximated by

$$Var\left(\overline{\widehat{M_t(\theta)}}\right) \approx \frac{\sigma^2}{T}\nabla_t\left[\frac{1}{T}\sum_{k=1}^{T}\left(\nabla_t\nabla_t'\right)^{-1}\right]\nabla_t',$$

which is smaller than the variance of $\widehat{M_t(\theta)}$

$$\begin{aligned} Var\left(\widehat{M_t(\theta)}\right) &= Var\left(M_t\left(\hat{\theta}_t\right)\right) \\ &= Var\left(m_t + \upsilon_t\right) \\ &= \sigma^2. \end{aligned}$$

Furthermore, a comparison between (22) and the result of Corollary 3 (for $a = 1$) shows that $\overline{\widehat{M_t(\theta)}} = \frac{1}{T}\sum_{k=1}^{T} M_t(\hat{\theta}_t)$ is as efficient as $M_t\left(\bar{\theta}\right)$ asymptotically. ■

6 An Example: the Black-Scholes Model

This section illustrates the previous results using the Black-Scholes option pricing model. In this example, we show that when using SSE versus exact calibration, different properties **P** are needed to capture typical model misspecifications.

Let us suppose that the true DGP of the underlying stock price time series is Merton's jump diffusion process, and the model to be tested is the Black-Scholes model. Of course, the Black-Scholes option model is misspecified. We assume that the modeler does not know the true DGP, and proceeds by using exact calibration with the Black-Scholes model for at-the-money (near-the-money, in practice) call prices C_t with time to maturity τ. Exact calibration gives an estimate of the implied volatility for each call price, IV_t.

First, let us consider how the exact-calibrated implied volatility IV_t changes as the price of the underlying, S_t changes, i.e.

$$\begin{aligned}\frac{\partial IV_t}{\partial S_t} &= \frac{\partial IV_t}{\partial C_t}\frac{\partial C_t}{\partial S_t} \\ &= \frac{1}{\frac{\partial C_t}{\partial IV_t}}\frac{\partial C_t}{\partial S_t}.\end{aligned}$$

The first term after the second equal sign is the reciprocal of the option's *vega* under the misspecified Black-Scholes model, while the second term is the *delta* of the call option under the DGP, Merton's jump diffusion model.

The Black-Scholes vega is given by

$$\frac{\partial C_t}{\partial IV_t} = S_t e^{-qt}\phi\left(d_1\right)\sqrt{\tau},$$

while the option's delta under the jump diffusion model is given by

$$\frac{\partial C_t}{\partial S_t} = e^{-qt}\Pi_1,$$

where both d_1 and Π_1 are functions of the risk-free interest rate, the dividend yield q, the option moneyness, the time to maturity τ, and their respective model parameters, which are all assumed to be fixed over the in-sample period. As a result,

$$\frac{\partial IV_t}{\partial S_t} = \frac{\Pi_1}{S_t\phi\left(d_1\right)\sqrt{\tau}},$$

which is inversely proportional to S_t. Discretizing, we have that for some constant c,

$$\Delta IV_t \approx c\frac{(\Delta S_t)}{S_t}. \tag{23}$$

This implies that for a change in S_t, the modeler expects a larger change in IV_t, the lower the level of S_t. Hence, as the level of S_t changes across time, IV_t will exhibit heteroskedasticy. The phenomenon is more pronounced under the jump diffusion model as the level of S_t changes dramatically at jump times. We illustrate this by a simulation given in Figure 11.

The relation (23) is also useful for analyzing the effect of errors in the price space on the errors in the parameter space. Suppose the underlying prices (or call prices, if the option delta is a constant over the sample period) contain i.i.d. observation noise. Then, by expression (23), we see that the i.i.d. error in the price space is translated to heteroskedastic error with respect to IV_t in the parameter space. Suppose instead the call price errors are proportional to S_t (or equivalently C_t, with fixed moneyness). Then, such heteroskedastic error in the price space is translated to i.i.d. error in the parameter space, again by expression (23). This shows that different properties **P** will be needed to identify model misspecifications under the different calibration methods. Hence, when testing for model misspecification using either the SSE or exact calibration, it is important to test for all five properties **P**. Indeed, testing for only a subset (e.g. first moment properties) may not capture all possible model misspecifications (e.g. second moment properties).

7 Statistical Tests

This section discusses the various test statistics that can be used to identify model misspecification with a calibrated model. Both time series tests and comparison tests are considered. These tests are illustrated for exact calibration, although similar tests can be used for SSE calibration.

7.1 Time Series Tests

In this section, we consider a collection of known time series tests suitable for testing the properties **P** discussed in the previous sections.

First, let us consider testing of property WN1 based on the parameter estimation error $\hat{\varepsilon}_t$, as defined in expression (5). The sample counterpart of the first condition in WN1: $E(\varepsilon_t) = 0$, is automatically satisfied by the construction of $\hat{\varepsilon}_t$.

The second and third conditions of property WN1 are much more useful for detecting model misspecifications. The second property is unconditional homoskedasticity: $Var(\varepsilon_t) = \sigma^2$, which requires that the variance of the parameter estimation error remains constant over time. However, this condition is guaranteed if property MDS2 is in place, due to Lemma 4.

The third property of zero covariances (12): $Cov(\varepsilon_t, \varepsilon_s) = 0$, requires that any two parameter estimation errors are uncorrelated. A comprehensive statistical test for property WN1 would thus rely on both the sample variances and covariances of $\hat{\varepsilon}_t$ revealing any potential model misspecification. This point will be discussed further in the next section.

Assuming unconditional homoskedasticity of ε_t, the Ljung-Box test (LB) is powerful against a departure from zero covariances (12).

Next, we turn to the problem of testing property MDS1. A consistent test for a zero conditional mean (13) would require taking into account all linear and nonlinear dependences of the current parameter estimation errors ε_t on all elements in the information set $\mathcal{F}_{t-1}$ simultaneously. There exists nonparametric tests in the literature that consistently check for (13). Some drawbacks include low power compared to parametric tests, complex and intensive computation involved in obtaining the value of test statistics and the optimal bandwidth, and the curse of dimensionality. A reasonable solution, provided

Tests	Conditions to Test	Equations
LB: Ljung-Box test on $\hat{\varepsilon}_t$	$Cov(\varepsilon_t, \varepsilon_{t-j}) = 0$	(12)
HL10: Hong-Lee martingale test on $\hat{\varepsilon}_t$	$Cov(\varepsilon_t, f(\varepsilon_{t-j})) = 0$	(14)
ML: McLeod-Li test on $\hat{\varepsilon}_t^2$	$Cov\left(\varepsilon_t^2, \varepsilon_{t-j}^2\right) = 0$	(15)
HL20: Hong-Lee martingale test on $\hat{\varepsilon}_t^2$	$Cov\left(\varepsilon_t^2, f(\varepsilon_{t-j})\right) = 0$	(16)
HL00: Hong-Lee serial indep test on $\hat{\varepsilon}_t$	$Cov(f(\varepsilon_t), g(\varepsilon_{t-j})) = 0$	(17)

Table 1: A summary of time series tests for different conditions in properties **P**.

by Hong (1999), gets around this conundrum by proposing a consistent test of the generalized spectral density that checks for the pairwise implication (14) of MDS1. Hong and Lee (2005) improve the generalized spectral test given conditional heteroskedasticity of unknown form. We apply the Hong-Lee martingale test (HL10) to $\hat{\varepsilon}_t$.

In parallel to the tests of WN1, we rely on the test proposed by McLeod and Li (1983) (ML) to test for property WN2. It is essentially the Ljung-Box test applied to the squared estimation errors $\hat{\varepsilon}_t^2$.

Similarly, the implication (16) of property MDS2 can be checked by applying the Hong-Lee martingale test (HL20) to $\hat{\varepsilon}_t^2$, and the implication (17) of property IID can be checked by applying the Hong-Lee test for serial independence (HL00) to $\hat{\varepsilon}_t$.

One may question if we lose any power by focusing on testing the implications expressions (14), (16), and (17) instead of the original properties MDS1, MDS2 and IID. In theory, the Hong-Lee tests only capture all pairwise dependences in the time series of $\hat{\varepsilon}_t$, and pairwise uncorrelatedness or independence is weaker than joint independence. Nevertheless, pairwise tests are sufficient for detecting model misspecification for certain classes of stochastic processes. For instance, Pierre (1971) shows that pairwise independence is equivalent to joint independence for infinitely divisible processes. Since many of the time series processes in financial modeling (e.g. Lévy processes) belong to such classes, we restrict our attention to testing expressions (14), (16) and (17).

The time series tests discussed above are summarized in Table 1.

7.2 Model Comparison Tests

The time series tests in the previous section provide a way to test a given model against the *absolute* criteria as given by properties **P**. In this section, we consider testing procedures that allow for a *relative* comparison between two competing models.

As a starting point, let us focus on testing property WN1. As discussed in the previous section, any departure from WN1 would be reflected in either time-varying variances or non-zero covariances (or both) of the parameter estimation error, $\hat{\varepsilon}_t$. Therefore, a natural candidate for a model comparison test is the variance ratio test. It compares two models by comparing the long-run sample variances of their $\hat{\varepsilon}_t$, which are essentially linear combinations of the sample variances and covariances.

Let us consider this test. Assume that the sample lengths of the two models are the same. From model k ($k = 1, 2$), we obtain the parameter estimation errors $\hat{\varepsilon}_t^{(k)}$ and we

compute the heteroskedasticity-autocorrelation consistent (HAC) estimator, defined as

$$\hat{V}_T^{(k)} = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\varepsilon}_t^{(k)}\right)^2 + \frac{1}{n}\sum_{j=1}^{p}\sum_{t=j+1}^{T} k\left(\frac{j}{p}\right)\hat{\varepsilon}_t^{(k)}\hat{\varepsilon}_{t-j}^{(k)}, \tag{24}$$

where $k\left(\frac{j}{p}\right)$ is a kernel function with bandwidth p.

Without loss of generality, we suppose that model 1 is "less misspecified" than model 2 under the null hypothesis, and vice versa under the alternative hypothesis. Denoting $V^{(k)}$ the population counterpart of $\hat{V}_T^{(k)}$, we have

$$\mathbf{H}_0 : V^{(1)} \leq V^{(2)},$$
$$\mathbf{H}_0 : V^{(1)} > V^{(2)}.$$

It is well known that a variance ratio test is not consistent for testing **(WN1)**, as $\hat{V}_T^{(k)}$ can be small even if **(WN1)** is false. This is possible when some of the covariances are negative, and they cancel other positive covariances and the variance. This problem can be resolved by modifying $\hat{V}_T^{(k)}$ and squaring each summand.

Furthermore, the estimation errors $\hat{\varepsilon}_t^{(k)}$ from two different models may be of different scales. This can be resolved by computing the long-run variances of the normalized estimation errors,

$$\hat{\eta}_t^{(k)} \equiv \hat{\varepsilon}_t^{(k)} / s_T^{(k)}, \tag{25}$$

where $s_T^{(k)}$ is the first term in (24). The first term of the resulting HAC estimator thus becomes one and can be ignored.

Consequently, to address the above two problems, we use the following long-run squared covariance estimator:

$$V_T^{2(k)} = \frac{1}{T^2}\sum_{j=1}^{p}\sum_{t=j+1}^{T} K^2\left(\frac{j}{p}\right)\left(\hat{\eta}_t^{(k)}\hat{\eta}_{t-j}^{(k)}\right)^2 \tag{26}$$

for some kernel function $K(\cdot)$. This is the generalized version of Ljung-Box and Hong-Lee tests for serial correlations before normalization. In other words, the Ljung-Box and Hong-Lee tests in the previous section provide a metric that measures how far the normalized residual sequence deviates from white noise.

As an alternative to a variance ratio test, more generally, any two test statistics, $S_T^{(1)}$ (model 1) and $S_T^{(2)}$ (model 2), can be employed for a comparison between a pair of models, either or both of which may be misspecified. If (1) $S_T^{(k)}$ is asymptotically $N(0,1)$ distributed for $k = 1, 2$, and (2) their covariance converges to zero in the limit, i.e. $\lim_{T\to\infty} Cov(S_T^{(1)}, S_T^{(2)}) = 0$, then, as $T \to \infty$, we have

$$z_T = \frac{S_T^{(1)} - S_T^{(2)}}{\sqrt{2}} \xrightarrow{d} N(0,1).$$

We may set $S_T^{(k)}$ to be the normalized Ljung-Box, McLeod-Li, or any of the Hong-Lee test statistics. They are generically denoted $S_{jT}^{(k)}$, $j = 1, 2, \ldots, 5$.Using the Hong-Lee test

statistic $S_{5T}^{(k)}$ of serial independence, the z_T test can be regarded as an omnibus model comparison test that incorporates all aspects of model misspecification.

If we are concerned more about one particular aspects of model departure from the null, then we can compare the weighted sum of squared statistics between the two models,

$$z_T^w = \frac{W_T^{(1)} - W_T^{(2)}}{\sqrt{2\sum_{j=1}^5 w_j^2}},$$

where $W_T^{(k)} = \sum_{j=1}^5 w_j \left(S_{jT}^{(k)}\right)^2$. The weights, which are required to sum to one, reflect our *a priori* view of the relative importance attached to the different kinds of model misspecifications. As $T \to \infty$, we have

$$z_T^w \xrightarrow{d} N(0,1).$$

An alternative to comparing two test statistics, Vuong (1989) proposes a likelihood ratio test for two arbitrary models, and derives the limiting distribution of the test statistic when the two models are non-nested, nested and overlapping. The test assumes the existence of smooth (twice continuously differentiable) likelihood function. A difficulty is that the likelihood function is most likely unknown. In this case, we can impose a distributional assumption on the estimation errors ε_t (which can be checked for goodness of fit). For example, one could assume:

Assumption N: ε_t is i.i.d. $N(0, \sigma^2)$ for all $i = 1, \ldots T$.

Under Assumption N, the log-likelihood function of $\varepsilon_t^{(k)}$ is

$$L^{(k)}(\sigma^{(k)}) = -T\log\left(\sigma^{(k)}\right) - \frac{1}{2\left(\sigma^{(k)}\right)^2}\sum_{t=1}^T \left(\hat{\varepsilon}_t^{(k)}\right)^2 \tag{27}$$

$$= \sum_{t=1}^T \left[-\log\left(\sigma^{(k)}\right) - \frac{\left(\hat{\varepsilon}_t^{(k)}\right)^2}{2\left(\sigma^{(k)}\right)^2}\right] \equiv \sum_{t=1}^T \ell_t^{(k)}. \tag{28}$$

In this case, maximizing the log-likelihood is equivalent to minimizing the sum of squares of the feasible estimation error $\hat{\varepsilon}_t^{(k)}$,

$$SSE_T^{(k)} = \sum_{t=1}^T \left(\hat{\varepsilon}_t^{(k)}\right)^2.$$

Even if **Assumption N** does not hold, one can still use quasi-likelihood methods and obtain a consistent estimation. In practice, one obtains the concentrated log-likelihood function by substituting the sample variances $\hat{\sigma}^{(k)}$ for the population variances $\sigma^{(k)}$ in (27)

$$\hat{L}_T^{(k)} = L_T^{(k)}(\hat{\sigma}^{(k)}) = -T\log\left(\hat{\sigma}^{(k)}\right) - \frac{1}{2\left(\hat{\sigma}^{(k)}\right)^2}\sum_{t=1}^T \left(\hat{\varepsilon}_t^{(k)}\right)^2$$

$$= -T\log\left(\hat{\sigma}^{(k)}\right) - \frac{T}{2},$$

and then compute the quasi-log-likelihood ratio,

$$\hat{L}_T^{(1)} - \hat{L}_T^{(2)} = T \log\left(\frac{\hat{\sigma}^{(2)}}{\hat{\sigma}^{(1)}}\right).$$

Suppose that model k has $d^{(k)}$ parameters, $k = 1, 2$. Assume that the models 1 and 2 are non-nested. Vuong (1989) proposes the likelihood ratio test as follows:

$$LR_T = \frac{\hat{L}_T^{(1)} - \hat{L}_T^{(2)} - \frac{d^{(1)}-d^{(2)}}{2}\log T}{\sqrt{T}\omega_T},$$

where ω_T^2 equals the long-run sample variance computed using the squared t^{th} contributions of the log-likelihood ratio, $\hat{\ell}_t$, obtained by substituting $\hat{\sigma}^{(k)}$ for $\sigma^{(k)}$ in (28). Specifically,

$$\omega_T^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\ell}_t^{(k)}\right)^2 + \frac{1}{T}\sum_{j=1}^{p}\sum_{t=j+1}^{T} k\left(\frac{j}{p}\right)\hat{\ell}_t^{(k)}\hat{\ell}_{t-j}^{(k)}.$$

Under the null hypothesis that model 1 is better than model 2, LR_T converges in distribution to a standard normal distribution as $T \to \infty$. Vuong's result assumes i.i.d. observations. One can generalize this approach to serially correlated and heteroskedastic estimation errors by replacing $\hat{\boldsymbol{\varepsilon}}^{(k)\prime}\hat{\boldsymbol{\varepsilon}}^{(k)} = \sum_{t=1}^{T}\left(\hat{\varepsilon}_t^{(k)}\right)^2$ with $\hat{\boldsymbol{\varepsilon}}^{(k)\prime}\left(\hat{\mathbf{V}}^{(k)}\right)^{-1}\hat{\boldsymbol{\varepsilon}}^{(k)}$, where $\hat{\mathbf{V}}^{(k)}$ is a covariance matrix estimator of $\boldsymbol{\varepsilon}^{(k)}$ (see Rivers and Vuong, 2002).

8 Simulation Studies

In this section, using simulations, we study the finite sample performance of the various statistical tests discussed in the previous section in detecting model misspecifications. The observations obtained herein prove useful when employing these same statistical tests with actual market prices.

8.1 Exact vs SSE calibration

In this section, we study the finite sample performance of the five time series tests as discussed in section 7. They include the Ljung-Box test of serial correlations (LB), the McLeod-Li test of heteroskedasticity (ML), the Hong-Lee martingale test of residuals (HL10), the Hong-Lee martingale test for squared residuals (HL20), and the Hong-Lee test of serial independence (HL00) under different DGP specifications.

The data generating processes are Merton's jump-diffusion model (1976), Heston's stochastic volatility model (1993), and Heston-Nandi's GARCH(1,1) model (2000). We generate 1000 sample paths of the underlying price process (with initial price $S_0 = 1000$) for each simulation experiment. Each sample path is of length $T = 63$. From the generated sample paths of the underlying, the theoretical prices of (near) at-the-money (ATM) calls are obtained. The strike price granularity *mesh* is specified so that the strike of the ATM call option is equal to the underlying price rounded to the nearest *mesh* unit. The observed

Exp	**DGP**	**Parameters**	*mesh*
1	Merton	$\sigma = 0.1, \lambda \in [0, 10], \mu_J = 0.1, \sigma_J = 0.8$	25
2	Merton	$\sigma = 0.1, \lambda = 2, \mu_J = 0.2, \sigma_J \in [0, 0.6]$	25
3	Heston	$\kappa \in [1, 1000], V_0 = 0.5, \bar{V} = \frac{V_0}{\kappa}, \sigma_V = 1.5, \rho = 0$	25
4	Heston	$\kappa = 100, V_0 = 0.5, \bar{V} = \frac{V_0}{\kappa}, \sigma_V \in [0.01, 3], \rho = 0$	25
5	Heston	$\kappa \in 50, V_0 = 0.5, \bar{V} = \frac{V_0}{\kappa}, \sigma_V = 0.8, \rho \in [0, 0.99]$	25
6	GARCH	$\omega = 5 \times 10^{-6}, \beta_1 \in [0, 0.9], \alpha_1 = 3 \times 10^{-7}, \gamma_1 = 400, \lambda = 0$	5
7	GARCH	$\omega = 1 \times 10^{-6}, \beta_1 = 0.3, \alpha_1 \in [0, 0.9] \times 10^{-5}, \gamma_1 = 0, \lambda = 0$	5
8	GARCH	$\omega = 1 \times 10^{-6}, \beta_1 = 0.5, \alpha_1 \in 1 \times 10^{-6}, \gamma_1 \in [0, 600], \lambda = 0$	5

Table 2: The data generating processes used in various simulation experiments.

ATM call price is the theoretical price contaminated with i.i.d. $N(0, 0.1)$ noise. We fix the interest rate and dividend yield to be $r = 0.05$ and $q = 0$, respectively.

Next, we calibrate the Black-Scholes model using both exact and SSE calibration and carry out the tests. Since the Black-Scholes model is misspecified, we expect the five statistical tests to detect the model misspecification. Comparisons of the five tests' power are done among themselves and between the two competing calibration methods. For all power studies in this section, the smoothing parameter[3] is set to 1, and the nominal rejection rate is 10%. Table 2 summarizes the simulation experiments, including the DGP and the parameters.

In experiments 1-2, the DGP is the Merton model. We vary the jump rate λ (experiment 1) and jump size standard deviation σ_J (experiment 2), respectively, holding all other parameters constant.

The power curves for the five tests are shown in Figure 3. The left panel shows the power curves for exact calibration, while the right panel corresponds to SSE calibration. Here are the key observations.

1. The tests on the squared residuals (ML and HL20), and the test of serial independence of the residuals (HL00) are relatively more powerful than LB and HL10 tests for detecting jumps. The power of the former tests generally increases with the jump intensity λ and decreases with the volatility σ.

2. The ML, HL20 and HL00 tests applied to the residuals from exact calibration are more powerful than the same tests applied to the residuals from SSE calibration.

We note from the first observation that the powers of the five tests increase differently as the jump component of the DGP begins to dominate the diffusion component (Merton, 1976b).

The second observation shows that the residuals in the parameter space are better than the residuals in the price space for detecting model misspecification. The reason relates to the strike price granularity (or *mesh* size, to be discussed in section 8.2). The

[3]The smoothing parameter for Ljung-Box and McLeod-Li test statistics is the number of residual autocorrelations, and that for Hong-Lee test statistics is the bandwidth of the nonparametric kernel estimator for the corresponding generalized spectral density.

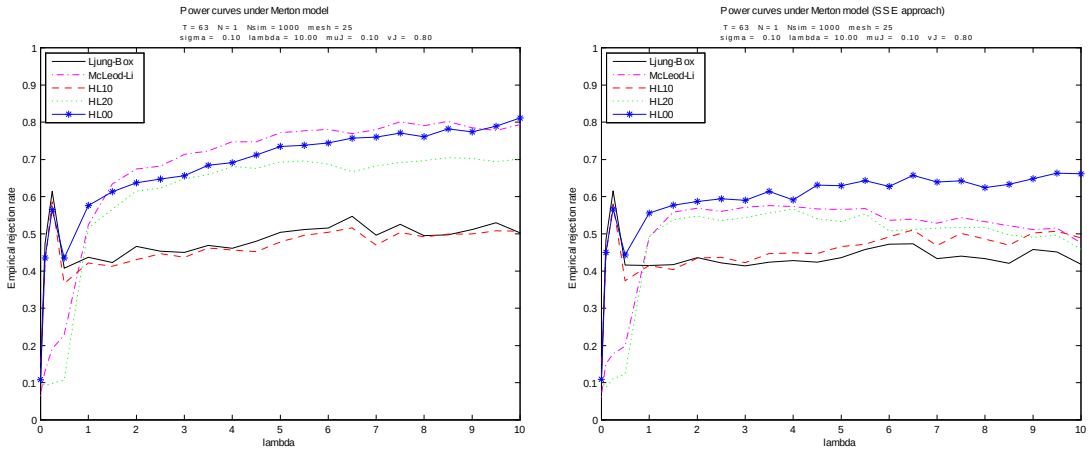

Figure 3: The power of five model misspecification tests over jump rate λ under exact (left panel) and SSE (right panel) calibration. DGP = Merton; Misspecified model = Black-Scholes.

log-moneyness of a near-the-money option, defined as $\log(S_t/K_t)$ (where K_t is set to S_t rounded to the nearest $mesh$ unit), has different magnitudes for different levels of the underlying's price S_t. Since the underlying index level jumps in the Merton model, the log-moneyness exhibits heteroskedasticity. The exact calibration residuals $\hat{\varepsilon}_t = IV_t - \overline{IV}$, which are a function of log-moneyness, thus exhibit heteroskedasticity in the parameter space (see Figure 5). This is captured by the second-order tests on the squared residuals (ML and HL20) as well as the omnibus HL00 test. However, no such heteroskedasticity is displayed for the SSE calibrated residuals in the price space (see Figure 6).

In experiments 3-5, the Heston model is set as the DGP. We consider three types of departures from the null by increasing the logarithm of the speed of the volatility mean reversion, $\log_{10}(\kappa)$, the volatility of volatilities, σ_V, and the correlation ρ between the Brownian motions that drive the underlying's price and stochastic volatility processes. All other parameters are kept constant. The power curves are displayed in Figures 7-9, and the power surface over $\log_{10}(\kappa)$ and σ_V for the HL00 test is displayed in Figure 10. Here are the key observations.

1. The tests on the residuals (LB, HL10), and the test of serial independence of the residuals (HL00) are relatively more powerful than the tests on the squared residuals (ML and HL20). The powers of all five tests generally increase with κ, σ_V, and ρ.

2. The model misspecification tests are comparable in terms of power under both exact and SSE calibration.

In experiments 6-8, the DGP is the Heston-Nandi GARCH(1,1) model. We consider three kinds of departure from the null of a constant volatility (Black-Scholes model): increasing the persistence of volatility (GARCH parameter β_1), the shock to volatility (ARCH parameter α_1), and the degree of volatility asymmetry (γ_1). The underlying prices and observed call prices are generated in the same manner as in the first two experiments.

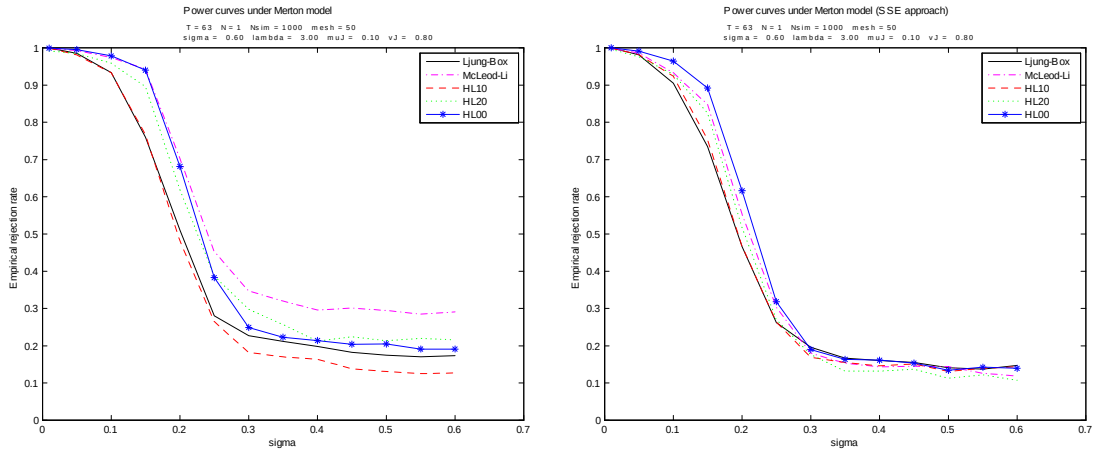

Figure 4: The power of five model misspecification tests over volatility σ under exact (left panel) and SSE (right panel) calibration. DGP = Merton; Misspecified model = Black-Scholes.

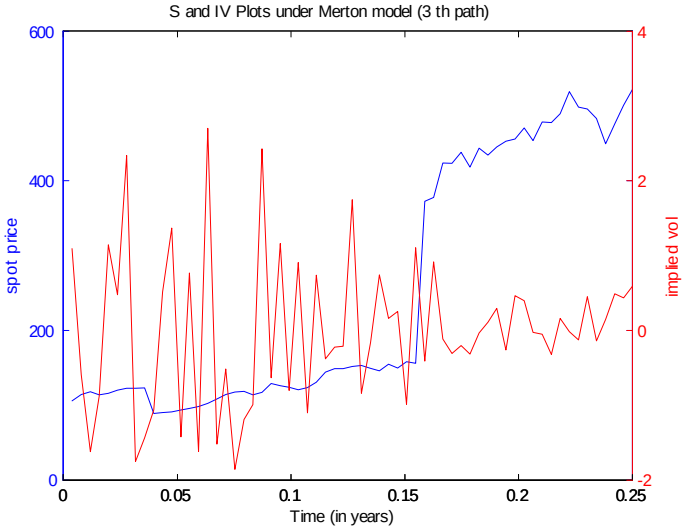

Figure 5: Under the exact approach, the IV residuals (in red and in the parameter space) exhibits heteroskedasticity as the underlying index (in blue) jumps over time, as in the Merton model.

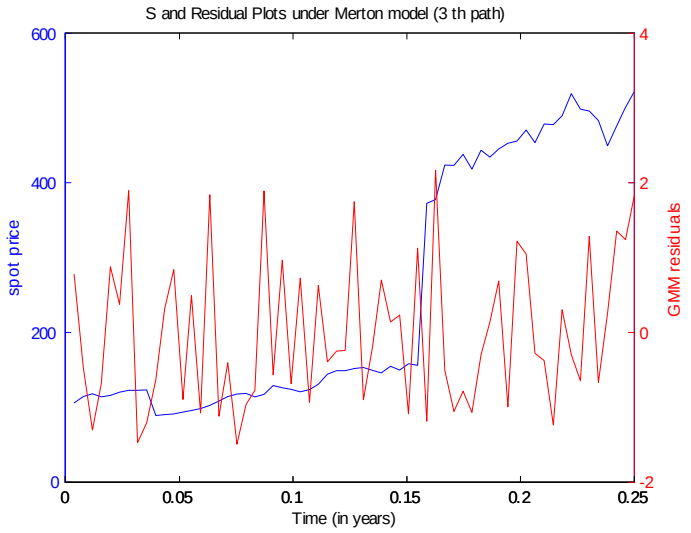

Figure 6: Under the SSE approach, the pricing residuals (in red and in the price space) do not exhibit heteroskedasticity as the spot price (in blue) jumps over time, as in the Merton model.

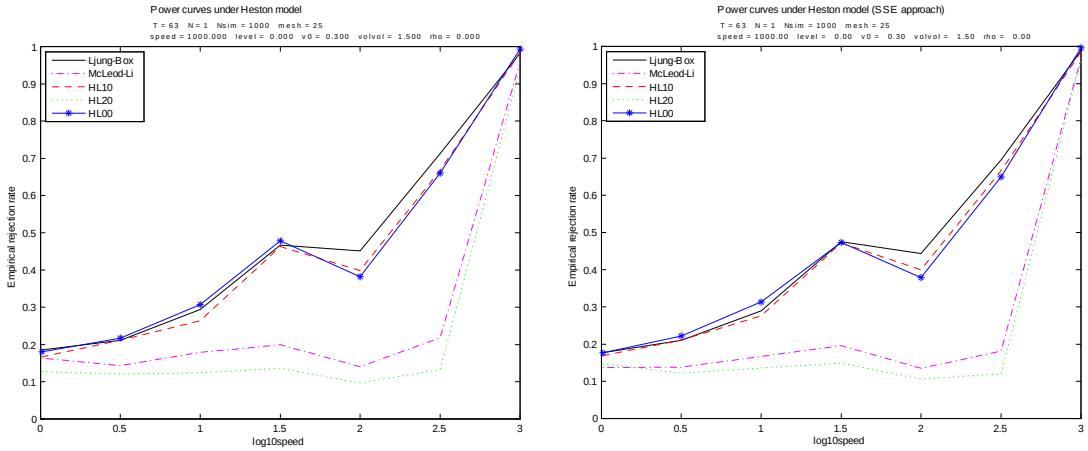

Figure 7: The power of five model misspecification tests over log speed $\log_{10}(\kappa)$ under exact (left panel) and SSE (right panel) calibration. DGP = Heston; Misspecified model = Black-Scholes.

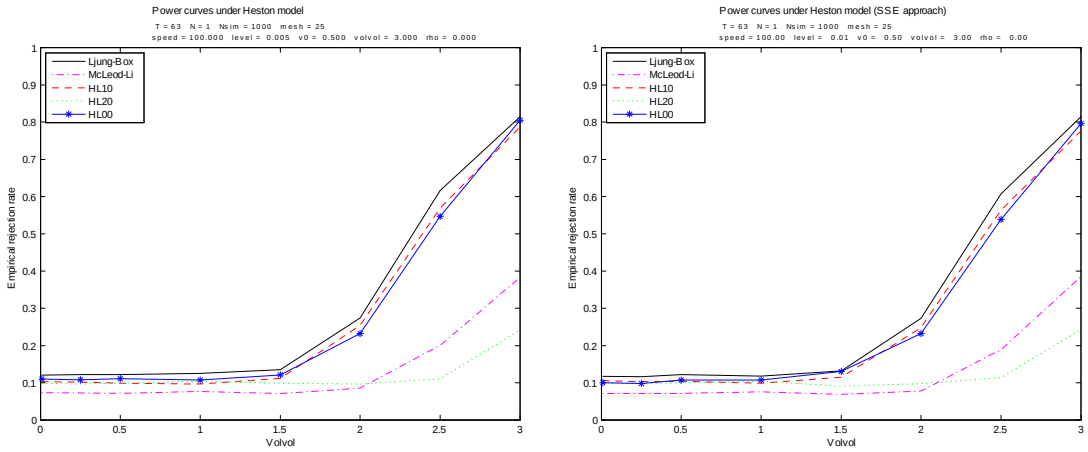

Figure 8: The power of five model misspecification tests over volatility of volatilities σ_v under exact (left panel) and SSE (right panel) calibration. DGP = Heston; Misspecified model = Black-Scholes.

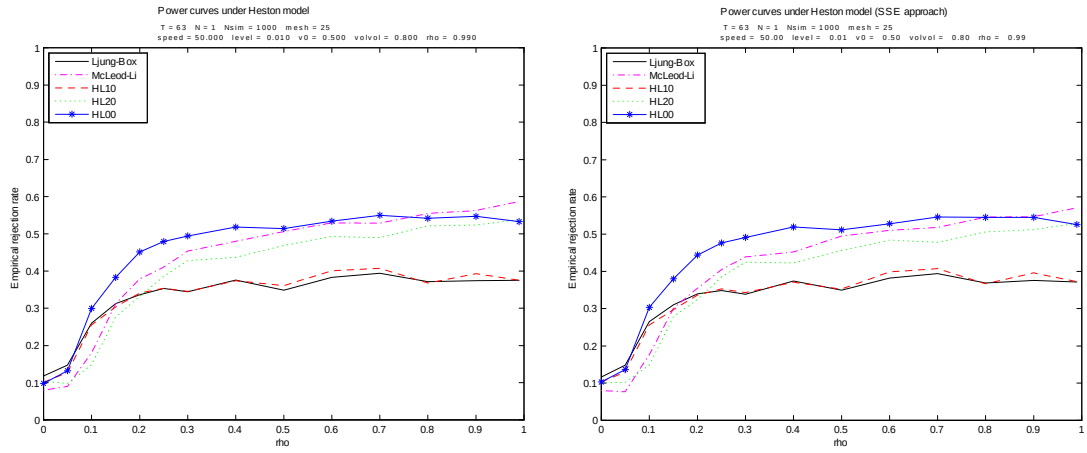

Figure 9: The power of five model misspecification tests over correlation ρ under exact (left panel) and SSE (right panel) calibration. DGP = Heston; Misspecified model = Black-Scholes.

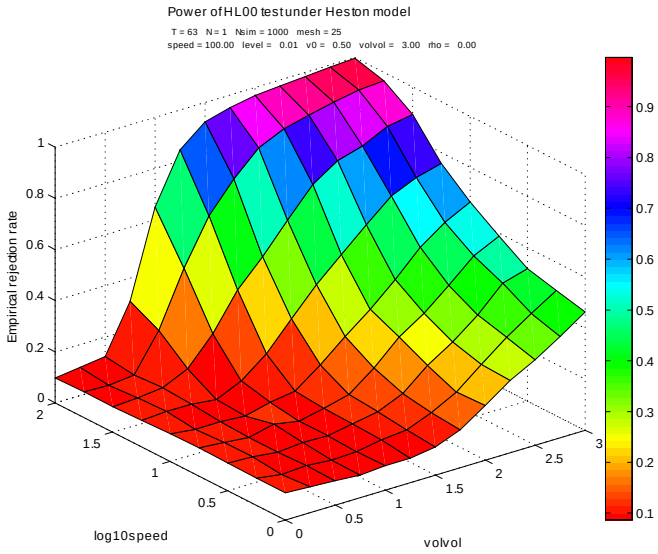

Figure 10: The power surface of Hong-Lee test of serial independence over $\log_{10}(\kappa)$ and σ_V under Heston model (Exp 3-4), Nominal rate = 0.01.

The power results are shown in Figures 11-13. The key observations are summarized below.

1. The test performances between the exact and SSE calibration are similar.

2. In general, the LB, HL10, and HL00 tests are more powerful than the ML and HL20 tests as the DGP departs from the null.

8.2 Strike Price Granularity

In this section, we study the finite sample performance of the time series tests as we vary the call options' strike price granularity.

The motivation for incoporating strike price granularity is related to detecting model misspecification. In practice, options of different strike prices are available. Strike prices for options on the same underlying are separated by a fixed distance, denoted the *mesh* size, that depends on the underlying's price, the option's moneyness, the time-to-maturity, the liquidity, among many other factors. Because of strike price granularity, we cannot obtain a perfect at-the-money option almost surely; but we can always obtain a near-the-money option with a strike price equal to the current underlying's price rounded to the nearest *mesh* unit (e.g. if the $mesh = 5$ and if the current underlying's price is 97, then the strike price is 95). The mesh size measures the step size that the call moves away from the at-the-money option as the underlying's price changes over time. A non-zero *mesh* size mimics the discrete nature of actual strike prices, but more importantly it controls how any model misspecification is revealed in the time series properties of $\hat{\varepsilon}_t$ and $\hat{e}_t$ as

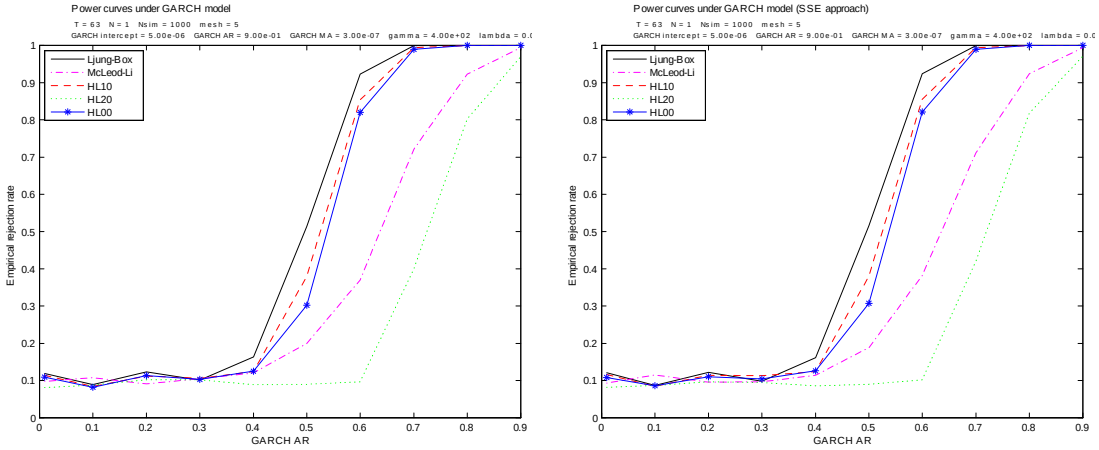

Figure 11: The power of five model misspecification tests over the AR parameter β_1 under exact (left panel) and SSE (right panel) calibration. DGP = Heston-Nandi; Misspecified model = Black-Scholes.

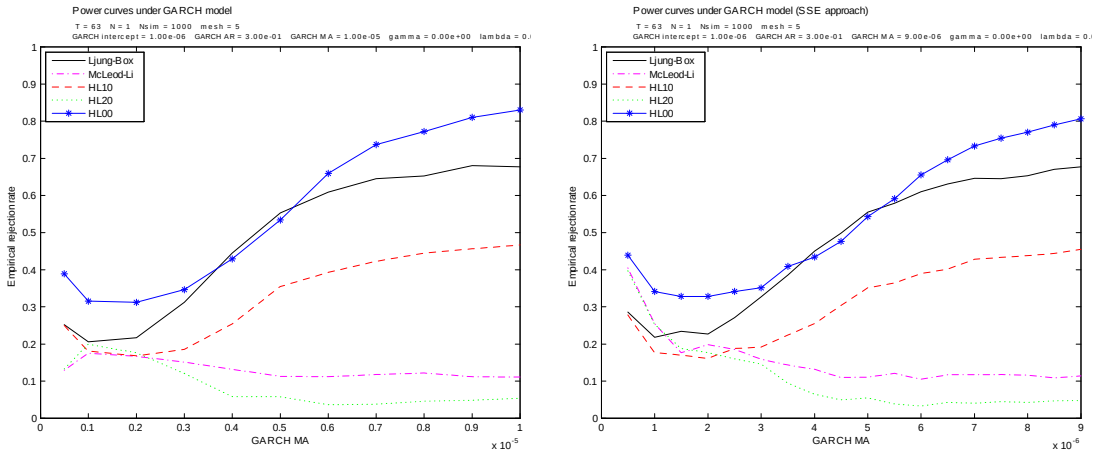

Figure 12: The power of five model misspecification tests over the MA parameter α_1 under exact (left panel) and SSE (right panel) calibration. DGP = Heston-Nandi; Misspecified model = Black-Scholes.

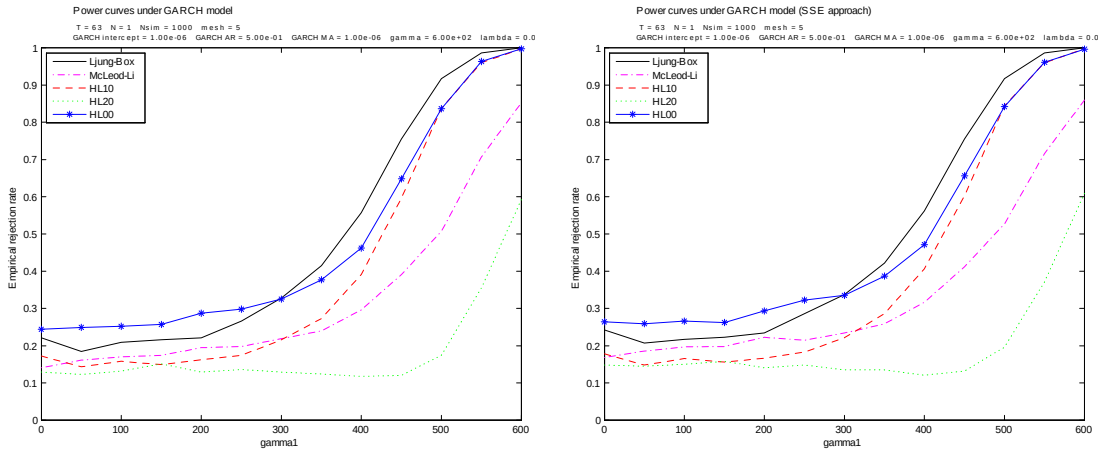

Figure 13: The power of five model misspecification tests over the volatility asymmetry parameter γ_1 under exact (left panel) and SSE (right panel) calibration. DGP = Heston-Nandi; Misspecified model = Black-Scholes.

the underlying's price evolves. This is important because if the DGP is fixed over the sample period, and if we consider perfect at-the-money calls (i.e. the *mesh* size is zero) with constant maturity, the corresponding IV sequence (and hence $\hat{\varepsilon}_t$) is a constant over time *even if the model is misspecified*. We expect that the residuals $\hat{\varepsilon}_t$ and $\hat{e}_t$ gradually lose properties **P** under the null hypothesis as the *mesh* size increases.

The simulation set-up under model misspecification is as follows. We assume that the underlying dynamics follow the Merton model, with parameters $\sigma = 0.1$, $\lambda = 2$, $\mu_J = -0.3$, and $\sigma_J = 0.1$ (the fixed DGP). The sample length T is set to 63 days. The number of simulation runs is set to 1000. The misspecified model to be calibrated is the Black-Scholes model.

With exact calibration, we obtain the time series of implied volatilities (IV) as in (3), and then apply the statistical tests on the demeaned IVs $\hat{\varepsilon}_t$, $t = 1, \ldots, T$. With SSE calibration, we minimize the sum of squared pricing error percentages with respect to the Black-Scholes volatility, and then obtain the time series of measurement errors $\hat{e}_t$, $t = 1, \ldots, T$, as in (6), on which we apply the tests.

The empirical rejection probabilities of the five statistical tests are plotted against the *mesh* size under the exact approach in Figure 14. The power curves under the SSE approach are very similar and not displayed here.

We also consider as DGPs the Heston SV model (Figures 15, with $\kappa = 50$, $V_0 = 0.3$, $\bar{V} = \frac{V_0}{\kappa}$, $\sigma_v = 1$, $\rho = 0$) and the Heston-Nandi (2000) GARCH(1,1) model (Figures 16, with $\omega = 1 \times 10^{-6}$, $\beta = 0.3$, $\alpha = 0.3 \times 10^{-8}$, $\gamma = 0$, $\lambda = 0.01$). In all cases, the powers of all of the tests generally increase with the *mesh* size.

8.3 Hedge Ratio Accuracy

In practice, one is not only interested in estimating the option's model price, but also the option model's hedge ratio. This section compares the accuracy of the estimated hedge

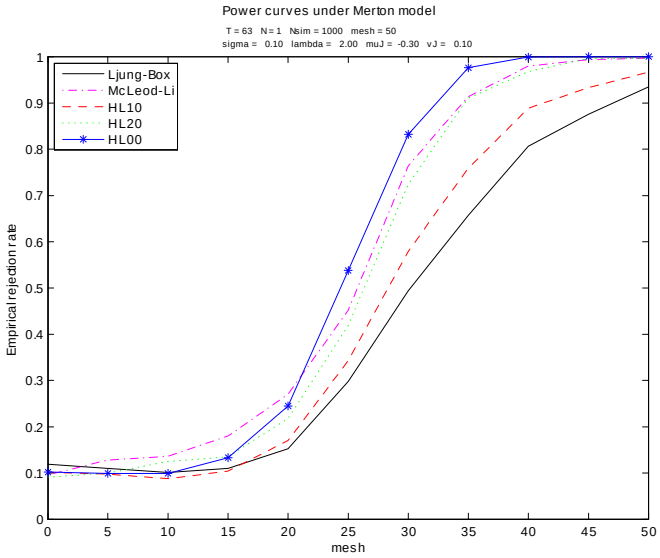

Figure 14: The plot of powers of five statistical tests against *mesh* under the exact approach. DGP = Merton model.

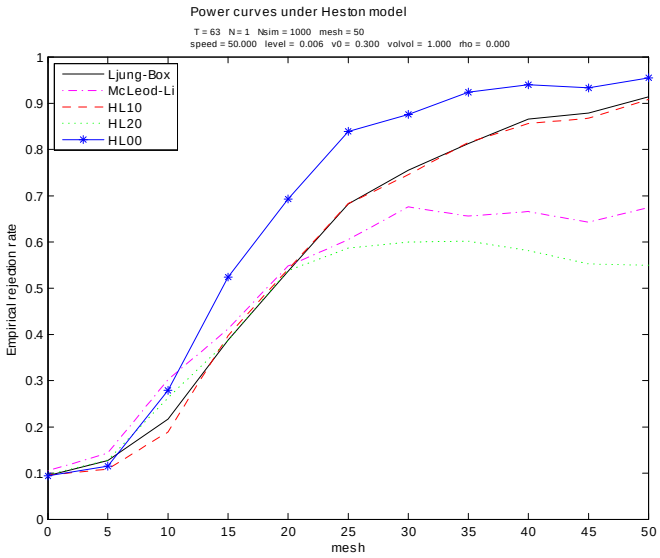

Figure 15: The plot of powers of five statistical tests against *mesh* under the exact approach. DGP = Heston model.

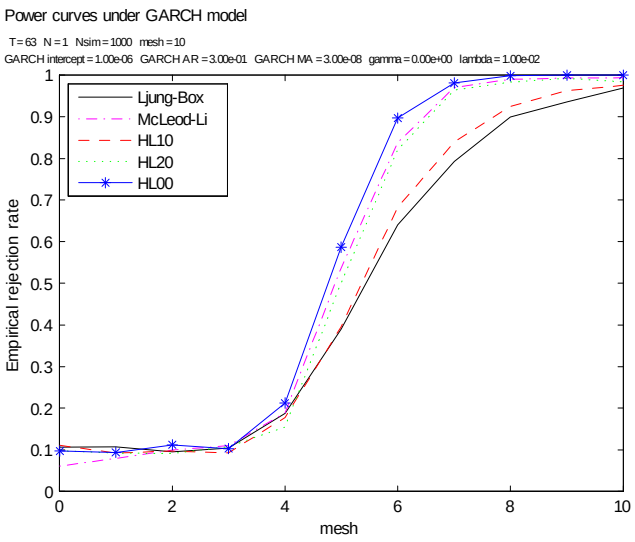

Figure 16: The plot of powers of five statistical tests against *mesh* under the exact approach. DGP = Heston-Nandi model.

ratio obtained from both SSE and exact calibration.

First, we assume that the true stock price DGP is a geometric Brownian motion with volatility $\sigma = 0.6$ and initial stock price $S_0 = 100$. We simulate 1000 sample paths. The sample length of each path is fixed at $T = 63$ days. The interest rate is set to be $r = 0.05$ and the dividend yield $q = 0$. Near-the-money call options are used for pricing, with $mesh = 5$ that controls the granularity of the strike prices K_t, and time-to-maturity $\tau = 63$ days.

Second, we assume that the call prices are contaminated with noise. Three types of noises are considered: (i) i.i.d. $N(0, 0.01^2)$ white noise; (ii) independent $N(0, 0.01^2 S_t^2)$ noise; and (iii) independent $N(0, 0.01 S_t)$ noise. The last two types of noise are heteroskedastic.

Next, we calibrate the (correctly specified) Black-Scholes model and obtain the implied volatility (IV) sequence IV_t using both the exact and SSE approaches. Under the exact approach, we obtain both the time-varying sequence of $\hat{\theta}_t = IV_t$ and its sample average over time $\bar{\theta} = \overline{IV} = \frac{1}{T}\sum_{t=1}^{T} IV_t$. Under the SSE approach, we obtain the constant parameter $\hat{\theta} = \widehat{IV}$ as the solution as in Section 2.1. Both $\overline{IV}$ and $\widehat{IV}$ are constants across time, providing a fair comparison of hedging performance under both approaches.

Using IV_t, $\overline{IV}$ and $\widehat{IV}$, we then compute the Black-Scholes hedge-ratio. For instance, with $\widehat{IV}$, the hedge-ratio is

$$\Delta_t = e^{-q\tau} N(d_1),$$

where

$$d_1 = \frac{\log(S_t/K_t) + (r - q + \widehat{IV}^2/2)\tau}{\widehat{IV}\sqrt{\tau}}.$$

RMSE under gBm	**Exact (with IV_t)**	**Exact (with $\overline{IV}$)**	**SSE (with $\widehat{IV}$)**
$N(0, 0.01^2)$ noise	0.4405×10^{-4}	0.564×10^{-5}	0.544×10^{-5}
$N(0, 0.01^2 S_t^2)$ noise	0.4001×10^{-2}	0.5165×10^{-3}	0.5284×10^{-3}
$N(0, 0.01^2 S_t)$ noise	0.4140×10^{-3}	0.5336×10^{-4}	0.5308×10^{-4}

Table 3: The RMSE results for geometric Brownian motion model, with σ=0.6.

To measure the accuracy of the estimated hedge-ratios in comparison to the true ones (which are known to us as $\sigma = 0.6$), we compute the root mean squared error (RMSE) by averaging the squared hedge-ratio errors over the sample period and over all simulation runs, and then taking the square root.

The finite sample results are displayed in Table 3, classified according to the types of noise that contaminates the call prices. In the cases of i.i.d. white noise and heteroskedastic noise with standard deviation $0.01\sqrt{S_t}$, the SSE hedge-ratios has a smaller RMSE than the exact hedge-ratios using $\overline{IV}$. The latter becomes more accurate if the standard deviation of the heteroskedastic noise is $0.01S_t$. This is consistent with the asymptotic result that estimators from exact calibration can be more efficient than those from SSE calibration under heteroskedastic noise, as suggested in Corollaries 2 and 4, coupled with Theorem 9, (ii).

It is not surprising that the exact hedge-ratios computed with IV_t are the least accurate because the exact solution is obtained from only a single observation. A comparison of a pair of realized sample paths of hedge-ratios under both approaches reveals that the exact hedge-ratios fluctuate more than the SSE counterparts around the true and constant Black-Scholes hedge-ratio (Figure 17).

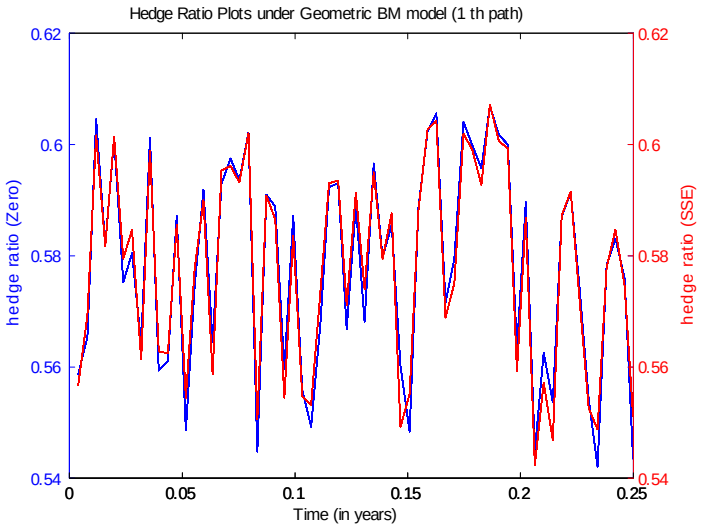

Figure 17: Realized sample paths of hedge-ratios under the exact (in blue) and SSE (in red) approaches when the true DGP is geometric Brownian motion. The call price noise is heteroskedastic.

RMSE under Merton	**Exact (with IV_t)**	**Exact (with $\overline{IV}$)**	**SSE (with $\widehat{IV}$)**
$N(0, 0.01^2)$ noise	0.07433	0.07423	0.07425
$N(0, 0.01^2 S_t^2)$ noise	0.07447	0.07426	0.07429
$N(0, 0.01^2 S_t)$ noise	0.07433	0.07424	0.07425

Table 4: The RMSE results for Merton jump-diffusion model, with $\sigma = 0.6$, $\lambda = 15$, $\mu_J = -0.3$ and $\sigma_J = 0.6$.

RMSE under Heston	**Exact (with IV_t)**	**Exact (with $\overline{IV}$)**	**SSE (with $\widehat{IV}$)**
$N(0, 0.01^2)$ noise	0.1535×10^{-3}	0.1473×10^{-3}	0.1473×10^{-3}
$N(0, 0.01^2 S_t^2)$ noise	0.4009×10^{-2}	0.5239×10^{-3}	0.5362×10^{-3}
$N(0, 0.01^2 S_t)$ noise	0.4327×10^{-2}	0.1558×10^{-3}	0.1559×10^{-3}

Table 5: The RMSE results for Heston SV model, with $\kappa = 10$, $\bar{V} = 0.05$, $V_0 = 0.5$, $\sigma_J = 0.4$ and $\rho = 0$.

We also compare the hedge-ratios under the two approaches when the model is misspecified. The set-up is the same as before, except that the option pricing calibration relies on the (misspecified) Black-Scholes model, and the hedge-ratios under the true DGP are now computed according to the correct DGP formula.

The results are displayed in Tables 4 and 5. The exact hedge-ratios computed with $\overline{IV}$ have a slightly smaller RMSE than the SSE hedge ratios under i.i.d. and heteroskedastic noise of both forms.

A pair of realized sample paths for the hedge-ratios under both calibration approaches under model misspecification (the true DGP is Merton) is illustrated in Figure 18. It shows essentially the same features as in the case of a correctly specified model.

9 An Empirical Study

In this section, we use exact calibration to test the validity of the Black-Scholes model using market data. The purpose of this section is two-fold. One, it provides an illustration of the misspecification tests for calibrated models as discussed in the previous section. Two, it presents a test of the Black-Scholes model using exact calibration. To our knowledge, all existing calibration based tests of the Black-Scholes model use the SSE approach. Of course, the accumulated evidence rejects the Black-Scholes model (see for example, Pan (2002)), which is what we expect to observe as well.

Our data, extracted from Option Metrics, consist of the S&P 500 index and its associated European call options spanning the time period January 1 – December 31, 2011 ($T = 253$ business days).

We consider calls with rolling maturity dates. Calls with large maturities usually have thin or even zero volume, while calls with very short maturities exhibit market microstructure irregularities. Thus, we only consider calls with a time to maturity >7 days and <180 days. In this set, we choose the call with the shortest time to maturity. An inspection of the data set reveals that the common granularity for the strikes of the S&P 500 index calls is $mesh = 25$.

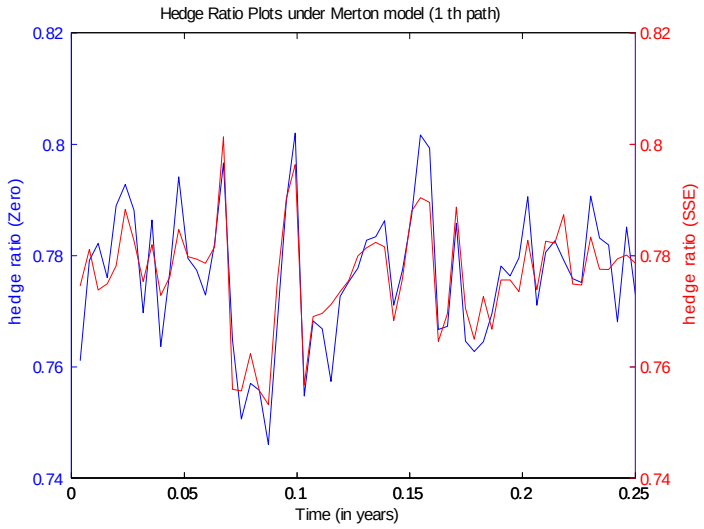

Figure 18: Realized sample paths of hedge-ratios under the exact (in blue) and SSE (in red) approaches when the true DGP is Merton model. The call price noise is heteroskedastic.

There are two additional inputs for the Black-Scholes model: the interest rate and dividend yield. For the interest rate, we employ the one-month Treasury bill rate released by the Federal Reserve Board. To construct the dividend yield for the S&P500 index, we divide the monthly dividend by the 21-day moving average of the S&P500 index.[4]

We first exact-calibrate the Black-Scholes model each day. The calibrated parameter is the implied volatility IV_t, so one call price is sufficient (one equation, one unknown) on a given day. We choose the at-the-money call option with a strike price closest to the current S&P 500 index.

Exact calibration yields a time series of IV_t, $t = 1, \ldots, T$. We then apply the various time series tests on the estimation error series $\hat{\varepsilon}_t = IV_t - \overline{IV}$. Table 6 shows that all the time series tests we considered strongly reject the Black-Scholes model. The time series of IV_t is plotted in Figure 19. The plot displays the serial correlation and trending behavior of the implied volatilities.

10 Conclusion

In spite of the popularity of model calibration in finance, empirical researchers have put more emphasis on model estimation than on the equally important goodness-of-fit problem. This is due partly to the ignorance of modelers, and more to the ability of existing statistical tests to detect specification errors. In practice, models are often calibrated by

[4]The monthly dividends of S&P500 are obtained from the website of Robert Shiller at www.econ.yale.edu/~shiller/data.htm.

gBm	IV_t		
	$M=1$	$M=5$	$M=10$
LB	197.2	731.0	1204.5
	0.0000	0.0000	0.0000
ML	82.77	183.6	306.8
	0.0000	0.0001	0.0001
HL10	50.40	54.66	52.39
	0.0000	0.0000	0.0000
HL20	8.99	10.93	9.88
	0.0000	0.0000	0.0000
HL00	75.16	70.53	64.90
	0.0000	0.0000	0.0000

Table 6: Time series tests are applied on a zero-calibrated geometric Brownian motion. Test statistic values and p values are displayed. S&P 500 index and the associated European call data span over 1 Jan - 31 Dec, 2011 (253 days). Time series tests considered include LB: Ljung-Box; ML: McLeod-Li; HL10: Hong-Lee test of martingale for residuals; HL20: Hong-Lee test of martingale for squared residuals; and HL00: Hong-Lee test of serial independence. M=smoothing parameter.

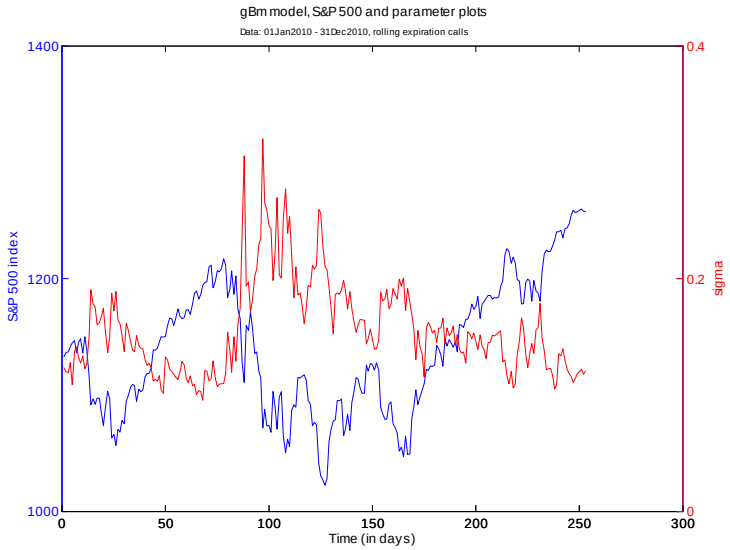

Figure 19: Time series plots of S&P 500 and IV_t under the Black-Scholes model, over 1 Jan - 31 Dec, 2011.

minimizing the sum of squared difference between the modelled and actual observations. It is challenging to disentangle model error from estimation error in the residual series. To circumvent the difficulty, we study an alternative way of estimating the model by exact calibration. We argue that standard time series tests based on the exact approach can better reveal model misspecifications than the error minimizing approach. In the context of option pricing, we illustrate the usefulness of exact calibration in detecting model misspecification. Under heteroskedastic observation error structures, our simulation results show that the Black-Scholes model calibrated by exact approach delivers more accurate hedging performance than that obtained by error minimization calibration.

Due to the fundamental and innovative nature of the exact calibration methodology, this paper points to an array of future research directions, both theoretical and empirical. On the theoretical side, even though the discussion was centered on exact calibration of a parametric model with a single parameter, the methodology is not confined to the univariate case. It turns out that exact calibration of a multivariate model creates interesting issues to be resolved, and is the focus of another working paper by the same authors. On the empirical side, it would be informative to re-estimate existing models in the option pricing literature by exact calibration, and see if they yield consistent or contradictory results. Furthermore, as mentioned in the introduction, exact calibration offers a methodology to estimate the unobservable attributes of many important financial structural models (e.g. risk aversion coefficient, probability of default) that would otherwise be impossible. Model misspecification testing for these other applications is a fruitful area for future research.

11 Appendix: Data generating processes

This appendix presents the different data generating processes of the underlying's price S_t. The interest rate r_t and dividend yield q_t are assumed to be exogenous.

1. Black-Scholes model
$$dS_t = (r_t - q_t)\, S_t dt + \sigma S_t dW_t$$
where W_t is the Brownian motion.

2. Merton's jump-diffusion model
$$dS_t = (r_t - q_t - \lambda\mu_J)\, S_t dt + \sigma S_t dW_t + J_t S_t dN_t$$
where N_t is a Poisson process with intensity λ. Given a jump occurs, the jump size J_t is distributed as $\log(1+J_t) \sim N\left(\log(1+\mu_J) - \frac{\sigma_J^2}{2}, \sigma_J^2\right)$.

3. Heston's stochastic volatility model
$$\begin{aligned} dS_t &= (r_t - q_t)\, S_t dt + \sqrt{V_t} S_t dW_t^{(1)} \\ dV_t &= \kappa(\theta - V_t)dt + \sigma_V \sqrt{V_t} dW_t^{(2)} \end{aligned}$$

where $W_t^{(1)}$ and $W_t^{(2)}$ are standard Brownian motion with $Corr(dW_t^{(1)}, dW_t^{(2)}) = \rho$.

4. GARCH(1,1) model of Heston and Nandi (2000)

$$\begin{aligned} \log S(t) &= \log S(t- \triangle) + (r_t - q_t) + \lambda h(t) + \sqrt{h(t)} W(t) \\ h(t) &= \omega + \beta_1 h(t- \triangle) + \alpha_1 \left(W(t- \triangle) - \gamma_1 \sqrt{h(t- \triangle)} \right)^2 \end{aligned}$$

where $\triangle$ is a one time step. Note that the model is expressed in discrete time form.

References

[1] Bates, David S. (2000), "Post-'87 Crash Fears in the S&P 500 Futures Option Market," *Journal of Econometrics*, **94**, pp.181–238.

[2] Black, Fischer, and Myron Scholes (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, **81**, pp.637–659.

[3] Davidson, Russell, and James, G. MacKinnon (2004), *Econometric Theory and Methods*, Oxford University Press.

[4] Heston, Steven L. (1993), "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bonds and Currency Options," *Review of Financial Studies*, **6**, 2, pp.327–343.

[5] Heston, Steven L. and Saikat Nandi (2000), "A Closed-Form GARCH Option Valuation Model," *Review of Financial Studies*, **13**, 3, pp.585–625.

[6] Hong, Yongmiao (1999), "Hypothesis Testing in Time Series via the Empirical Characteristic Function: A Generalized Spectral Density Approach," *Journal of the American Statistical Association*, **84**, pp.1201–1220.

[7] Hong, Yongmiao and Yoon-Jin Lee (2005), "Generalized Spectral Tests for Conditional Mean Models in Time Series with Conditional Heteroscedasticity of Unknown Form," *Review of Economic Studies*, **72**, pp.499–541.

[8] Jarrow, Robert (2011), "Risk Management Models: Construction, Testing, Usage," *Journal of Derivatives,* Summer, pp.1–10.

[9] McLeod, A. Ian and Wai Keung Li (1983), "Diagnostic Checking ARMA Time Series Models using Squared-Residual Autocorrelations," *Journal of Time Series Analysis*, **4**, pp.269–273.

[10] Merton, Robert (1976), "Option Pricing when the Underlying Stock Returns are Discontinuous," *Journal of Financial Economics,* **4**, pp.125–144.

[11] Merton, Robert (1976b), "The Impact on Option Pricing of Specification Error in the Underlying Stock Price Returns," *Journal of Finance,* **31**, 2, pp.333–350.

[12] Pan, Jun (2002), "The jump-Risk Premia Implicit in Options: Evidence from an Integrated Time-Series Study," *Journal of Financial Economics*, **63**, pp.3–50.

[13] Pierre, P.A. (1971), "Infinitely Divisible Distributions, Conditions for Independence, and Central Limit Theorem," *Journal of Mathematical Analysis and Applications*, **33**, pp.341–354.

[14] Rivers, Douglas, and Quang H. Vuong (2002), "Model Selection Tests for Nonlinear Dynamic Models," *Econometric Journal*, **5**, pp.1–39.

[15] Vuong, Quang H. (1989), "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, **57**,2, pp.307–333.