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 the University of Sydney

12 December, 2014

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 a loss function of the differences between the modelled and actual observations. Under this approach, 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. Unlike the error minimization approach, all information about dynamic misspecifications is channeled to the parameter estimation residuals under exact calibration. In the context of option pricing, we illustrate that standard time series tests are powerful in detecting various kinds of dynamic misspecifications. Compared to the error minimization approach, exact calibration yields more reasonable model comparison result, and delivers more accurate hedging performance that is robust to both gradual and abrupt structural shifts of state variables.

*The authors would like to thank the Guest Co-editors, Michael McAleer, Shiqing Ling and Howell Tong, and the referees for their invaluable comments and suggestions, which lead to significant improvements to the paper. We also thank the conference organizers and participants at the Frontiers of Time Series Analysis and Related Fields at HKU, the 2014 Asian Meeting of the Econometric Society at the Academia Sinica, and the 24th New Zealand Econometrics Study Group Meeting at the University of Waikato Management School, as well as all the seminar participants at the University of Hong Kong, the University of Sydney, and the University of Technology, Sydney. All errors are our own.

†Correspondence email: raj15@cornell.edu
‡Correspondence email: simon.kwok@sydney.edu.au

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 rare but severe market crash. 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 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* (*errmin calibration* hereafter), 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). Throughout this paper, parameters are broadly defined as any unknown or unobserved quantities in the model. In particular, latent state variables such as stochastic volatilities are regarded as parameters.

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

In this paper, we view standard time series tests as an indispensable tool for detecting *dynamic model misspecification* for calibrated models. Under the conventional GMM framework, tests of moment restrictions and overidentifying restrictions are commonly employed as model checking

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

devices (e.g., Pan, 2002). They are residual-based tests that check whether the residuals are close enough to zero as suggested by the moment condition. However, these moment tests are not designed to pick up different degrees of serial dependence in the residuals sequence. We illustrate that time series tests exhibit power against different types of dynamic model misspecification when applied to residuals in observation space under the errmin approach, or to the residuals in the parameter space under the exact approach. There may be a slight difference in power between the two methods due to a nonlinearity in the transformation from the observation to the parameter space. The situation is further complicated by the existence of observation noise and parameter estimation error in the residuals that mix with model error. The analysis in Section 6 reveals that, under exact calibration, it is possible to disentangle dynamic model error (in the form of systematic and predictable dynamics) from the constant parameter estimation error and unpredictable observation noise in the residuals. A similar separation does not carry over to residuals from errmin calibration because the estimation error varies across different observations.

More importantly, we study how model misspecification affects the secondary usage of an exact calibrated model. In particular, exact calibration delivers more accurate in-sample estimators for functions of parameters (e.g., hedge ratios) than the errmin approach under a misspecified model. The reason is that the errmin approach averages over a global set of observations and assumes the same model structure over the whole sample. In contrast, the exact approach yields parameter estimates by exact-matching to a small number of observations and, hence, is robust to potential model misspecification.

An outline for this paper is as follows. Section 2 provides a literature review. Section 3 discusses calibration in the context of option pricing. Section 4 decomposes the calibration estimation errors excluding observation noise. The acceptable error structure (properties) of a "good" model, for misspecification testing, is developed in section 5. Section 6 characterizes the properties of the parameter estimates in the presence of observation noise. The prime example of calibration in option pricing using the Black-Scholes model is described in Section 7. Section 8 discusses a collection of standard statistical tests useful for testing model misspecification. Section 9 investigates the small sample properties of the suggested statistics through simulations. Section 10 applies the methodology to S&P 500 index options to test and compare different models. Section 11 concludes the paper. The proofs of theorems 7 and 8 are collected in an online appendix.

2 Literature Review

By far, one of the most widely adopted methods in the literature for calibrating option models is the generalized method of moments (GMM) (Hansen, 1982) – nonlinear least squares being a popular special case. The objective is to minimize a pre-specified discrepancy measure or loss function, in many cases taking the form of a weighted average of sample moments – for this reason we refer to this method as the error minimizing approach (errmin). The loss function is sometimes mediated with the addition of regularization terms for internal consistency or parameter stability (Andersen et al. (2012), for matching the realized volatility under the physical measure and the stochastic volatility from the model under the risk neutral measure) or parameter stability (e.g.,

regularization of ill-posed problem. See the discussion in Cont and Tankov (2004)). The inputs to the loss function may be pricing errors (measured in the actual price space or the implied volatility space) or the discrepancy between the modeled and real transition densities under the physical measure (Gagliardini et al. (2011)). Different loss functions represent different optimality criteria and they have different purposes. For instance, the optimal loss function for pricing is different from that for hedging. (see Christoffersen and Jacobs (2004) for an in-depth study. See also Detlefsen and Härdle (2007). Moment restrictions are motivated by financial theories (e.g., no arbitrage principle), and the parameter space can be finite or infinite dimensional.[2]

How can we study the goodness-of-fit of a proposed parametric model? For the above mentioned applications, calibration is done by averaging a large set of observations (e.g., time series, cross sections, or panels). These methods deliver accurate calibration results provided the sample is large enough and the model represents the true data generating process (DGP). The parameter estimate is consistent for the true parameter, which is kept constant over the entire sample. A common approach is to study parameter stability using Chow-type tests (e.g., Andersen et al., 2012), which check the equality of parameter estimates across two disjoint samples. There are drawbacks to Chow-type tests, however.[3] Most notably, even though calibrated parameters are stable over the sample, this does not imply that the errmin model yields a good fit to the data. The reason is that the errmin approach minimizes the average of individual errors, and the parameter is assumed to be a constant over the entire sample. Any modeling error is contained in the residuals (i.e., individual errors evaluated at the calibrated parameters).

One may test the moment conditions under a GMM framework as in Pan (2002). Her residual-based test is useful for detecting static misspecification by checking whether the moment conditions are compatible with the data at each sample point, but it is not powerful against dynamic model misspecification, which is revealed as time dependent residuals. For more complete model diagnostics, we need to test both the individual size and time series properties of the residuals. We emphasize the importance of uncovering dynamic residual patterns in a systematic manner. As illustrated by the simulations in Section 8, standard time series tests applied to the residuals of errmin calibrated models are useful in revealing dynamic model misspecification.

Exact calibration is an alternative method by which a modeler can recover a model's parameters. Under this approach, the parameters are obtained by exactly equating the modeled and observed quantities. The system is exactly identified in the sense that the number of unknown parameters is equal to the number of equalities. It is motivated by the familiar option pricing application which obtains the Black-Scholes implied volatility from an observed option price. Provided that there is a unique solution (which is the case for the Black-Scholes model), exact calibration gives zero residuals in the observation space and yields parameter estimates via a one-to-one transformation from the observations. In this case, any model misspecification is fully reflected in the estimated para-

[2] Aït-Sahalia and Lo (1998) calibrate state-price densities using option prices. Gagliardini et al. (2011) consider a semi-parametric framework which involves a parametric model of the stock return dynamics and a non-parametric estimator of the transition density of historical stock returns.

[3] The test results can be sensitive to the prespecified cutoff points. Moreover, these tests are usually less powerful against continuous structural shifts and breaks near the boundaries of the sample period. (e.g., see Chen and Hong, 2011).

meters of an exact-calibrated model, and manifested as different types of dynamic trends and/or patterns in the parameter estimates. Under the null hypothesis of a correct model specification, the model parameter is a constant. Model misspecification is exhibited as non-constant parameters under exact calibration, or non-zero residuals under errmin calibration. In practice, however, this is complicated by the presence of observation noise.

Observation noise arises from the micro-structure considerations of option trading (e.g., bid-ask spreads, which relates to the option's liquidity) and/or data issues (e.g., the non-synchronization of the option's and underlying's prices, recording error). Observation noise exhibits peculiar dynamics that reflect the stylized facts specific to option contracts, such as periodicity (due to cycling maturity dates), autocorrelations, and heterogeneity with respect to moneyness. This observation noise can affect the accuracy of parameter estimation and statistical inference. Under the errmin approach, however, it is possible to minimize such effects in large samples. With weakened assumptions that partially address the option features (e.g., an exogenous and weakly dependent error structure), the errmin parameters can be shown to be asymptotically consistent.[4]

Observation noise also plays a crucial role in exact calibration. Observation noise introduces randomness into the calibrated parameters under the null, and allows for formal statistical inference as shown in Section 6. Unlike the standard method for treating observation noise as a nuisance to be averaged out, we regard any exactly calibrated model as misspecified if it transfers trends or predictable parts of the observation noise (e.g., periodicity, autocorrelations and non-i.i.d. dynamics) to the parameter space. This perspective leads to more stringent standards on what constitutes a correctly specified model and more model rejections. Nevertheless, it motivates the explicit modeling of key option characteristics that affect pricing and secondary usages. This is largely in line with Bates (2000) who explicitly builds into the likelihood function an AR pricing error structure with a heterogeneous common factor that differs across time and option classes.

From a statistical point of view, errmin calibration is a fully parametric estimation method over the entire sample with a small number of parameters, while exact calibration is a non-parametric estimation method without smoothing – the number of parameters equals the number of exact-match combinations. There exist intermediate cases. Bakshi, Cao and Chen (1997) estimate a parametric model by errmin calibration over each cross section. In Gagliardini et al. (2011), exact-matching of model and market prices for options at each time point is possible through an extended method of moments. Compared to exact calibration or errmin calibration over separate cross sections, these parameter estimates are much more stable over time, thus validating the discretized stochastic volatility model used therein. This is because their estimation is mediated by the historical sequence of stock returns used for estimating the non-parametric transition density, which absorbs most of the dynamic misspecification that would otherwise exist if a purely parametric set-up has been used. Rather than relying on non-parametric estimation as a shock absorber, we are concerned with the problems of testing and calibrating a parametric dynamic model using only option data.

[4] Pan (2002) shows that the GMM estimator is consistent under a time series setting with large T, small N asymptotics. The same consistency result holds for the extended method of moments estimator in Gagliardini et al. (2011). On the other hand, Andersen et al. (2012) prove the consistency of the penalized least squares estimator under a panel setting with large N, small T asymptotics.

3 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 DGP, $\mathcal{M}_{it}(\vartheta) \equiv \mathcal{M}(\vartheta; \mathbf{z}_{it})$, where ϑ, possibly infinitely dimensional, is the true unknown parameter vector invariant over t. When there is no observation error, we have

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

for $i = 1, \ldots, n; t = 1, \ldots, T$. We will discuss the case with observation error in Section 6.

3.1 Error Minimization Calibration

One popular way of estimating the model is *error minimization calibration* (errmin). Under this approach, the modeler first chooses a parametric option pricing model $M(\boldsymbol{\theta})$ indexed by a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$, where $d < \infty$, 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 $\boldsymbol{\theta}$.

A crucial feature of the errmin approach is that the parameter vector $\boldsymbol{\theta}$ is constant across the entire sample. With a constant parameter, the model governs all the cross sectional variation and time series dynamics of the option prices. Latent state variables such as stochastic volatility may appear in the model, but we assume that their dynamics are fully characterized by $\boldsymbol{\theta}$.[5]

Given $\boldsymbol{\theta}$, the *pricing error* matrix $\mathbf{e} \equiv \mathbf{e}(\boldsymbol{\theta}) = (e_{it}(\boldsymbol{\theta}))$ is 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. An example is $L_2(\mathbf{e}(\boldsymbol{\theta}); \mathbf{D}) = \sum_{i=1}^{n} \sum_{t=1}^{T} w_{it} e_{it}^2$, with the weights w_{it} summing to one. In this case we interpret $M_{it}(\theta)$ as the *conditional mean model* of the observed market price m_t using the property of the weighted least squares estimation.

The modeler minimizes the loss function and obtains the solution

$$\widehat{\boldsymbol{\theta}} \equiv \widehat{\boldsymbol{\theta}}(n, T) \equiv \arg\max_{\boldsymbol{\theta}} L(\mathbf{e}(\boldsymbol{\theta}); \mathbf{D}).$$

By standard asymptotic arguments, the solution $\widehat{\boldsymbol{\theta}}$ converges to a constant vector $\boldsymbol{\theta}_0$ as the

[5] For example, the European option price under the Heston's stochastic volatility model (see the formulation in the appendix) is a function of five parameters: $\kappa, \theta, V_0, \sigma_V, \rho$.

sample size increases. We refer to $\boldsymbol{\theta}_0$ as the "pseudo-true" parameter vector.[6]

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: $\hat{\mathbf{e}} \equiv \mathbf{e}(\widehat{\boldsymbol{\theta}}) \equiv 0$. In this case, it is easy to determine if the model is rejected or not. Since models are approximations of a complex reality, it is rarely the case that M and $\mathcal{M}$ coincide. Alternative goodness-of-fit criteria are required.

One natural criterion is the closeness of $\hat{\mathbf{e}}$ to zero, which can be easily measured based on a studentized sample mean of the residuals. We can then carry out a goodness-of-fit test of the model using this test statistic.[7] Such a test accepts a model if the pricing errors are close to zero on average, but it generally ignores potential temporal and cross sectional dependence among the residuals. This motivates a different set of criteria.

An alternative set of criteria is based on the degree of randomness and non-predictability of $\hat{\mathbf{e}}$. In other words, we 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 $\hat{\mathbf{e}}$ exhibit random and unpredictable behavior over time. A collection of such properties $\mathbf{P}$ is discussed in Section 5 below. Based on these properties, model specification tests can be performed (see Section 8).

3.2 Exact Calibration

An alternative method for estimating an option pricing model is exact calibration. The modeler first proposes a model $M_{it}(\boldsymbol{\theta}) \equiv M(\boldsymbol{\theta}; \mathbf{z}_{it})$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a d-dimensional real-valued vector of parameters. Under this calibration approach, the observed option prices are matched to the modeled prices. For any fixed $t = 1, \ldots, T$, the system of equations $m_{it} = M_{it}(\boldsymbol{\theta})$ holds for all $i \in I = \{i_1, \ldots, i_d\}$. With a slight abuse of notation, the system can be expressed in vector form: $\mathbf{m}_{It} = \mathbf{M}_{It}(\boldsymbol{\theta})$,where $\mathbf{m}_t = (m_{i_1 t}, \ldots, m_{i_d t})'$ and $\mathbf{z}_t = (\mathbf{z}'_{i_1 t}, \ldots, \mathbf{z}'_{i_d t})$.

A crucial feature of exact calibration is that it requires d nonlinear equations to match the market prices to the theoretical prices to solve for the d unknown parameters. Unlike errmin calibration, the parameter vector $\boldsymbol{\theta}$ varies over different matchings for different t and I. If the function $\mathbf{M}_{It}(\cdot)$ is invertible, a unique solution $\hat{\boldsymbol{\theta}}_{It} = \mathbf{M}_{It}^{-1}(\mathbf{m}_{It})$ exists. In the sequel, we assume that the model has a scalar parameter, i.e., $d = 1$.[8] 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 the following monotonicity assumption[9].

[6] Gagliardini et al. (2011) prove the asymptotic normality of the more general extended method of moments estimator of the model parameter as $T \to \infty$, holding the cross sectional size n fixed. On the other hand, Andersen et al. (2012) deal with the asymptotics of a GMM estimator as $n \to \infty$ for a fixed time horizon T. Different regularity conditions on the pricing errors are necessary.

[7] For example, under the GMM framework, Pan (2002) tests the moment conditions $\mathbf{e}(\boldsymbol{\theta}_0) = 0$ jointly using a heteroskedasticity-robust residuals.

[8] The case of $d > 1$ raises more complex issues such as the existence and uniqueness of solutions and is dealt with in a separate paper (Jarrow and Kwok, 2014). Both the errmin and exact calibration can handle multiple parameter calibration.

[9] A weaker version is the locally strict monotonicity assumption: the option pricing function $M(\theta)$ is strictly monotone in θ in a neighborhood of the hypothetical parameter θ_0 under model M.

Assumption SM. The option pricing function $M(\theta)$ is strictly monotone in θ.

An example is the Black-Scholes (1973) model $BS(\theta; \mathbf{z})$, where θ corresponds to the stock's volatility, a scalar. The Black-Scholes model satisfies the strict monotonicity assumption on the volatility parameter. The unknown volatility θ can be found by solving the equation $m_t = BS(\theta; \mathbf{z}_t)$ for each t. The solution $\hat{\theta}_t \equiv IV_t = BS^{-1}(m_t; \mathbf{z}_t)$ is commonly known as the *implied volatility*.

Under errmin 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 $M_t(\hat{\theta}_t) = \mathcal{M}_t(\vartheta) = m_t$ for all $t = 1, \ldots, T$. If the model specification is perfect, i.e., the proposed model M coincides the DGP $\mathcal{M}$, then all the solutions $\hat{\theta}_t$ from exact calibration are identical to the *pseudo-true parameter* θ_0 for all $t = 1, ..., T$. In this case, it is easy to determine whether the model should be accepted. The model is accepted if the estimated parameter is a constant for all t. Since all models are approximations of a complex reality, all model will be rejected 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*,

$$\varepsilon_t = \hat{\theta}_t - \theta_0, \tag{3}$$

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 5 below. Given these properties, dynamic and static specification tests can be performed which leads to acceptance or rejection of the model (see Section 8).

Because the pseudo-true parameter θ_0 is unknown, we need to test for properties **P** using the *parameter estimation residual* instead:

$$\hat{\varepsilon}_t = \hat{\theta}_t - \bar{\theta}, \tag{4}$$

where $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t$.We will rely on a battery of statistical tests on the residuals $\hat{\varepsilon}_t$ to study whether the randomness is systematic or not (see Section 8).

4 The Error Decomposition and its Dynamic Properties

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 errmin and exact calibration.

4.1 Errmin 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 errmin calibration approach, there are three types of errors:

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

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

3. **Estimation error**: $u_t = M_t(\widehat{\theta}) - M_t(\theta_0)$, 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. In general, u_t 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. Hence we have $\hat{e}_t = e_t + u_t$. The error decomposition under errmin approach is illustrated in Figure 1 (a).

To test an errmin 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.

4.2 Exact Calibration

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

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

1. **Measurement error**: $M_t(\widehat{\theta}_t) - \mathcal{M}_t(\vartheta) = M_t(\widehat{\theta}_t) - m_t \equiv 0$.

2. **Model error**: $M_t(\theta_0) - \mathcal{M}_t(\vartheta)$.

3. **Estimation error**: $M_t(\widehat{\theta}_t) - M_t(\theta_0)$.

In this setting, the model and estimation errors are identical because the measurement error is zero. The error decomposition under the exact approach is illustrated in Figure 1 (b).

This approach offers an alternative method for detecting model misspecifications that avoids the confounding influence of estimation error in finite samples. Indeed, the equality of model and estimation errors suggests that model misspecifications can be detected from the dynamic properties of $\widehat{\theta}_t$[10]; therefore, misspecification tests are carried out in the *parameter space*.

In practice, the estimation error $\varepsilon_t = \hat{\theta}_t - \theta_0$ is unobservable due to the unknown pseudo-true parameter θ_0. We substitute it by the sample mean $\bar{\theta}$ of $\hat{\theta}_t$.Using $\bar{\theta}$ as a proxy for the pseudo-true parameter θ_0 introduces a bias $\varsigma \equiv \bar{\theta} - \theta_0$. The parameter estimation residual $\hat{\varepsilon}_t = \hat{\theta}_t - \bar{\theta}$ under exact calibration can be decomposed into parameter estimation error and bias $\hat{\varepsilon}_t = \varepsilon_t + \varsigma$. The

[10] Since $d = 1$, there is no cross-sectional model misspecification; all model misspecification is dynamic in nature. Under a multivariate set-up ($d > 1$) in which there are multiple ways of exact-calibrating the model, both dynamic and cross-sectional misspecifications can be detected from the panel of parameters $\hat{\theta}_{It}$. This is discussed in Jarrow and Kwok (2014).

bias ς is constant over t, so the dynamic properties of the estimation error ε_t are fully reflected in the parameter residuals $\hat{\varepsilon}_t$. If the model is dynamically misspecified, such information is revealed in the dynamics of $\hat{\varepsilon}_t$, which we discuss in the next section.

5 Properties P

To judge whether a model is a "good approximation", it is necessary to have a benchmark that allows for the presence of observation noise. The benchmark consists of a set of assumptions on the errors that we refer to as properties **P**. For errmin calibration, these properties apply to the model errors, $e_t = m_t - M_t(\theta_0)$. For exact calibration, these properties apply to the parameter estimation errors, $\varepsilon_t = \hat{\theta}_t - \theta_0$. For brevity, we present the properties **P** only under exact calibration. The same properties apply under errmin calibration by substituting e_t for ε_t in the various properties.

Let the probability space be $(\Omega, \mathbb{P}, \mathcal{F})$, where $\mathcal{F} = (\mathcal{F}_t)_{t=1,...,T}$ is the natural filtration of $\{m_t, \mathbf{z}_t\}$, i.e., $\mathcal{F}_t$ is the sigma algebra generated by $\{(m_s, \mathbf{z}_s) : s = 1, \ldots, t\}$. To allow for statistical inference, we require that the sequence of exact solutions $\hat{\theta}_t$ exists and is ergodic. Under this assumption, we introduce the properties **P**.

Properties P: for all $t = 1, ..., T$, and all $0 < s < t$,
WN1. ε_t is a white noise: $E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma^2$ and $Cov(\varepsilon_s, \varepsilon_t) = 0$.
MDS1. ε_t is a martingale difference sequence: $E(\varepsilon_t | \mathcal{F}_s) = 0$, $\mathbb{P}$-a.s..
WN2. $\varepsilon_t^2 - \sigma^2$ is a white noise: $E(\varepsilon_t^2) = \sigma^2, Var(\varepsilon_t^2) = \varkappa$ and $Cov(\varepsilon_t^2, \varepsilon_s^2) = 0$.
MDS2. $\varepsilon_t^2 - \sigma^2$ is a martingale difference sequence: $E(\varepsilon_t^2 - \sigma^2 | \mathcal{F}_s) = 0$, $\mathbb{P}$-a.s..
IID. $\varepsilon_{t_1}, \varepsilon_{t_2}, \ldots, \varepsilon_{t_b}$ are i.i.d. for all $t_1 \neq \cdots \neq t_b$ and all integers $b > 1$.

Some remarks are in order. **MDS1** implies that the parameter estimation errors are unpredictable based on the past information. **MDS1** implies that ε_t is uncorrelated with $f(\varepsilon_s)$ for any measurable function $f(\cdot)$ and $0 < s < t$:

$$Cov(\varepsilon_t, f(\varepsilon_s)) = 0. \tag{5}$$

Similarly, **MDS2** implies the squared estimation error is unpredictable. Furthermore, property **MDS2** is equivalent to saying that ε_t is (conditional) homoskedastic: for all $0 < s < t$, $E(\varepsilon_t^2 | \mathcal{F}_s) = \sigma^2$, $\mathbb{P}$-a.s.. It implies that ε_t^2 is uncorrelated with any measurable function of the past errors, i.e., for any measurable function $f(\cdot)$ and $0 < s < t$:

$$Cov(\varepsilon_t^2, f(\varepsilon_s)) = 0. \tag{6}$$

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

$$Cov(f(\varepsilon_t), g(\varepsilon_s)) = 0, \tag{7}$$

provided that the covariance exists.

The following relations among properties **P** are valid.

Lemma 1a: If $Var(\varepsilon_t)$ exists and is constant, then **MDS1** $\Longrightarrow$ **WN1**.
Lemma 1b: If $Var(\varepsilon_t^2)$ exists and is constant, then **MDS2** $\Longrightarrow$ **WN2**.
Lemma 2a: If $E(\varepsilon_t|\mathcal{F}_s)$ exists $\mathbb{P}$-a.s., then **IID** $\Longrightarrow$ **MDS1**.
Lemma 2b: If $E(\varepsilon_t^2|\mathcal{F}_s)$ exists $\mathbb{P}$-a.s., then **IID** $\Longrightarrow$ **MDS2**.
Lemma 3a: **IID** $\Longrightarrow Cov(\varepsilon_t, f(\varepsilon_s)) = 0 \Longrightarrow Cov(\varepsilon_s, \varepsilon_t) = 0$.
Lemma 3b: **IID** $\Longrightarrow Cov(\varepsilon_t^2, f(\varepsilon_s)) = 0 \Longrightarrow Cov(\varepsilon_t^2, \varepsilon_s^2) = 0$.
Lemma 4: **MDS2** $\Longrightarrow Var(\varepsilon_t) = \sigma^2$.

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** and **MDS2**, and then by **WN1** and **WN2**. Lemma 4 says that conditional homoskedasticity implies unconditional homoskedasticity. These properties form the basis of the time series tests considered in Section 8.

6 Calibration with Observation Noise

In practice, the observed market price m_t is contaminated by observation noise υ_t, so that $m_t = \mathring{m}_t + \upsilon_t$, where $\mathring{m}_t = \mathcal{M}_t(\vartheta)$ is the true but unobserved price. This section explores the impact of the observation noise on the accuracy of both the errmin and exact calibration.

Given a proposed model $\{M(\theta; \mathbf{z}) : \theta \in \Theta\}$, where Θ is the admissible set of parameters, the null and alternative hypotheses to be tested are:

$$\mathbf{H}_0 : \ M(\theta; \mathbf{z}_t) = \mathring{m}_t \text{ for some } \theta = \theta_0 \text{ and for all } t \in \{1, \ldots, T\}$$
$$\mathbf{H}_1 : M(\theta; \mathbf{z}_t) \neq \mathring{m}_t \text{ for all } \theta \in \Theta, \text{ for some } t \in \{1, \ldots, T\}.$$

In other words, the model is correctly specified under $\mathbf{H}_0$ in that the modeled prices match the true prices. In order to test the model, we exact-calibrate the model using the noisy observations m_t.

For identification and inference purpose, we impose a subset of the following assumptions on the noise and the model.

Assumption SE. For all $t = 1, \ldots, T$, $\mathbb{P}$-a.s., $E(\upsilon_t|\mathbf{z}) = 0$.
Assumption WE.[11] For all $t = 1, \ldots, T$, $\mathbb{P}$-a.s., $E(\upsilon_t|\mathbf{z}_t) = 0$.
Assumption CV$_1$. υ_t satisfies $Var(\upsilon_t|\mathbf{z}_t) = \sigma^2$ and $Cov(\upsilon_s, \upsilon_t|\mathbf{z}_s, \mathbf{z}_t) = 0$, $\mathbb{P}$-a.s,. for all $t \neq s$.

[11] Alternatively, we may impose stronger exogeneity conditions by requiring $E(\upsilon_t|\mathcal{F}_{t-1}, \mathbf{z}) = 0$ or $E(\upsilon_t|\mathcal{F}_{t-1}, \mathbf{z}_t) = 0$, which imply Assumptions **SE** and **WE** respectively.

Assumption CV$_2$. υ_t satisfies $Var(\upsilon_t|\mathbf{z}_t) = \sigma_t^2$ and $Cov(\upsilon_s, \upsilon_t|\mathbf{z}_s, \mathbf{z}_t) = 0$, $\mathbb{P}$-a.s., for all $t \neq s$.
Assumption LC. The model $M_t(\theta)$ is strictly convex in θ in a neighborhood around θ_0.
Assumption D. The model $M_t(\theta)$ is continuously differentiable in θ and S_t, so that $\nabla_t(\theta) = \frac{\partial M_t(\theta)}{\partial \theta}$and $\nabla_{St}(\theta) = \frac{\partial^2 M_t(\theta)}{\partial \theta \partial S_t}$ exist. Denote $\nabla_t = \nabla_t(\theta_0)$ and $\nabla_{St} = \nabla_{St}(\theta_0)$.

6.1 Errmin Calibration with Observation Noise

In this subsection, we study the finite sample and asymptotic properties of the errmin estimator when the model is correctly specified and the option prices are contaminated by observation noise. To fix the idea, we focus on the calibration problem with an L_2 loss function with equal weights. We calibrate the model to the contaminated prices m_t by minimizing $\sum_{t=1}^T e_t^2(\theta)$, where $e_t(\theta) = M_t(\theta) - m_t$ is the pricing error in the presence of observation noise. This is a standard nonlinear least squares problem with the first order condition $\sum_{t=1}^T \nabla_t(\hat{\theta}) e_t(\hat{\theta}) = 0$.

We note that the residual is decomposed into three components: estimation error $M_t(\hat{\theta}) - M_t(\theta_0)$, model error $M_t(\theta_0) - \mathring{m}_t$, and observation noise υ_t:

$$e_t(\hat{\theta}) = M_t(\hat{\theta}) - (\mathring{m}_t + \upsilon_t) = \left[M_t(\hat{\theta}) - M_t(\theta_0)\right] + [M_t(\theta_0) - \mathring{m}_t] - \upsilon_t.$$

Under $\mathbf{H}_0$, there exists a parameter $\theta_0 \in \Theta$ such that $M_t(\theta_0) = \mathring{m}_t$ holds for all t, so there is no model error. The first order condition becomes $\sum_{t=1}^T \nabla_t(\hat{\theta}) \nabla_t(\tilde{\theta})'(\hat{\theta} - \theta_0) = \sum_{t=1}^T \nabla_t(\hat{\theta}) \upsilon_t$ for some $\tilde{\theta}$ lying between $\hat{\theta}$ and θ_0. To derive the large sample properties of $\hat{\theta}$, we need to solve for $\hat{\theta} - \theta_0$, which is possible with the following assumption.

Assumption PL$_a^{errmin}$. 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 $a = 1$) or $\hat{S}_{\sigma^2\nabla\nabla}(\theta) := \frac{1}{T}\sum_{t=1}^T \sigma_t^2 \nabla_t(\theta) \nabla_t(\theta)'$ (for $a = 2$), so that their probability limits $S_{\nabla\nabla}$ and $S_{\sigma^2\nabla\nabla}$ exist in a neighborhood Θ_0 of θ_0.[12] Furthermore, $S_{\nabla\nabla}$ is invertible.

Theorem 1 *Under Assumption* **SE**, **C** *and* $\mathbf{H}_0$, $\hat{\theta}$ *is biased in finite samples, 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 samples, i.e.,* $\hat{\theta} \xrightarrow{p} \theta_0$ *as* $T \to \infty$.

Theorem 3 *Under Assumption* **WE**, **D**, **CV**$_a$, **PL**$_a^{errmin}$ *(*$a = 1, 2$*) and* $\mathbf{H}_0$, $\hat{\theta}$ *is asymptotically normally distributed:*

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V_a^{errmin}) \quad \text{as } T \to \infty,$$

with asymptotic variance

$$V_a^{errmin} = \begin{cases} \sigma^2 S_{\nabla\nabla}^{-1} & \text{for } a = 1, \\ S_{\nabla\nabla}^{-1} S_{\sigma^2\nabla\nabla} S_{\nabla\nabla}^{-1} & \text{for } a = 2. \end{cases}.$$

[12] $\text{plim}_{T\to\infty} \sup_{\theta\in\Theta_0} \left[\hat{S}_{\nabla\nabla}(\theta) - S_{\nabla\nabla}\right] = 0$ and $\text{plim}_{T\to\infty} \sup_{\theta\in\Theta_0} \left[\hat{S}_{\sigma^2\nabla\nabla}(\theta) - S_{\sigma^2\nabla\nabla}\right] = 0$.

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

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

$$\sqrt{T}\left[M_t(\hat{\theta}) - M_t\left(\theta_0\right)\right] \xrightarrow{d} N(0, \nabla_t V_a^{errmin} \nabla_t') \quad \text{as } T \to \infty.$$

The statement for the delta hedge ratio is obtained by replacing $M_t(\cdot)$ by $\frac{\partial M_t(\cdot)}{\partial S}$, and ∇_t by ∇_{St}.

6.2 Exact Calibration with Observation Noise

In contrast to errmin calibration, exact calibration finds a solution $\hat{\theta}_t$ that satisfies

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

for each t. The solution $\hat{\theta}_t$ varies across observations and gives $e_t(\hat{\theta}_t) \equiv 0$ for all t, so the errors do not appear in the observation space. The errors are instead transferred to the parameter space through exact calibration. Denote the estimation error (in the parameter space) by $\varepsilon_t = \hat{\theta}_t - \theta_0$. We can estimate the unknown parameter θ_0 with the sample mean of the calibrated parameters,

$$\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T} \hat{\theta}_t = \theta_0 + \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t. \tag{8}$$

By defining the residual as $\hat{\varepsilon}_t = \hat{\theta}_t - \bar{\theta}$, the errors can be decomposed as follows:

$$\hat{\varepsilon}_t = [\hat{\theta}_t - M_t^{-1}(\hat{m}_t)] + \left[M_t^{-1}(\hat{m}_t) - \theta_0\right] + \left[\theta_0 - \bar{\theta}\right].$$

The three parts correspond to observation noise, model error, and estimation error $\varsigma \equiv \theta_0 - \bar{\theta}$, all in the parameter space. Since the estimation error ς remains constant over t, all dynamic model misspecifications are reflected in the observation noise and model error. By a first-order Taylor's expansion, the first term can be expressed in terms of υ_t:

$$\hat{\theta}_t - M_t^{-1}(\hat{m}_t) = M_t^{-1}(m_t) - M_t^{-1}(m_t - \upsilon_t) = \frac{\partial M_t^{-1}\left(\tilde{m}_t\right)}{\partial m}\upsilon_t,$$

for some $\tilde{m}_t$ lying between m_t and $m_t - \upsilon_t$. The equation links the observation space to the parameter space.

Under $\mathbf{H}_0$, the model error is zero, and the residual becomes

$$\hat{\varepsilon}_t = (\hat{\theta}_t - \theta_0) + \left(\theta_0 - \bar{\theta}\right) = \varepsilon_t + \varsigma. \tag{9}$$

This sets the stage for studying the finite and large sample properties of $\bar{\theta}$ under $\mathbf{H}_0$.

Theorem 4 *Under Assumption* **SE**, **C** *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 5 *Under Assumption* **WE** *and* $\mathbf{H}_0$, $\bar{\theta}$ *is consistent in large sample, i.e.* $\bar{\theta} \xrightarrow{p} \theta_0$ *as* $T \to \infty$.

Averaging expression (9) over $t = 1, \ldots, T$, we have

$$\sqrt{T}\left(\bar{\theta} - \theta_0\right) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \varepsilon_t = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\partial M_t^{-1}\left(\tilde{m}_t\right)}{\partial m} \upsilon_t.$$

We can derive the asymptotic distribution with the help of the following theorem.

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 $a = 1$) or $\hat{S}_{\sigma^2(\nabla\nabla)^{-1}}(\theta) := \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2\left[\nabla_t(\theta)\nabla_t(\theta)'\right]^{-1}$ (for $a = 2$), so that their probability limits $S_{(\nabla\nabla)^{-1}}$ and $S_{\sigma^2(\nabla\nabla)^{-1}}$ exist in a neighborhood Θ_0 of θ_0.[13]

Theorem 6 *Under Assumption* **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{exact}$ *($a = 1, 2$) and* $\mathbf{H}_0$,

$$\sqrt{T}(\bar{\theta} - \theta_0) \xrightarrow{d} N(0, V_a^{exact})$$

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

$$V_a^{exact} = \begin{cases} \sigma^2 S_{(\nabla\nabla)^{-1}} & \text{for } a = 1, \\ S_{\sigma^2(\nabla\nabla)^{-1}} & \text{for } a = 2. \end{cases}.$$

Using the delta method, we obtain the asymptotic distribution of $M_t(\bar{\theta})$.

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

$$\sqrt{T}\left[M_t(\bar{\theta}) - M_t\left(\theta_0\right)\right] \xrightarrow{d} N(0, \nabla_t V_a^{exact} \nabla_t') \quad \text{as } T \to \infty.$$

The statement for the delta hedge ratio is obtained by replacing $M_t(\cdot)$ *by* $\frac{\partial M_t(\cdot)}{\partial S}$, *and* ∇_t *by* ∇_{St}.

6.3 A Comparison

After carrying out the calibration, it is a common practice to use the calibrated model by substituting the estimated parameter into the function of interest. How do the errmin and exact calibration estimators facilitate pricing and secondary usage? In this subsection, we study and compare the asymptotic efficiency of the pricing function and hedging delta, which are in the form of some functions of the estimated parameters from errmin and exact calibration.

As we will show, under the ideal setting of correct model specification with weakly exogenous and conditionally homoskedastic noise, the errmin estimator is asymptotically more efficient than the counterpart from exact calibration. This is not surprising given the optimality of the errmin estimation. However, there exists circumstances in which the exact calibration gives more efficient

[13] $\operatorname{plim}_{T\to\infty} \sup_{\theta\in\Theta_0}\left[\hat{S}_{(\nabla\nabla)^{-1}}(\theta) - S_{(\nabla\nabla)^{-1}}\right] = 0$ and $\operatorname{plim}_{T\to\infty} \sup_{\theta\in\Theta_0}\left[\hat{S}_{\sigma^2(\nabla\nabla)^{-1}}(\theta) - S_{\sigma^2(\nabla\nabla)^{-1}}\right] = 0$.

estimation than the errmin calibration even when the model is correctly specified. This occurs when the observation noise is conditional heteroskedastic, for instance.

To set the stage, let $M_t(\hat{\theta})$ be the errmin-calibrated price obtained by substituting the errmin parameter estimate $\hat{\theta}$ into the pricing formula. On the other hand, we obtain the exact-calibrated price $M_t(\bar{\theta})$ by substituting the sample mean $\bar{\theta}$ of the parameter estimates from exact calibration, defined in (8). Since both $\hat{\theta}$ and $\bar{\theta}$ are a function of all the observations in the sample, this allows for a fair comparison. The result is formally stated in the following theorem.

Theorem 7 *Suppose Assumptions* **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{errmin}$, $\mathbf{PL}_a^{exact}$ *($a = 1, 2$) and* $\mathbf{H}_0$ *hold.*

(i) With conditional homoskedastic noise ($a = 1$), $M_t(\hat{\theta})$ is asymptotically more efficient than $M_t(\bar{\theta})$.

(ii) With conditional heteroskedastic noise ($a = 2$), it is possible that $M_t(\bar{\theta})$ is asymptotically more efficient than $M_t(\hat{\theta})$.

The above result still holds if we replace the pricing function $M_t(\cdot)$ by the hedging delta $\frac{\partial M_t(\cdot)}{\partial S}$.

Since exact calibration yields different parameter estimates from different observations, there are several ways to compute an estimate for the option price and the hedging delta. One of them is the exact-calibrated price $M_t(\bar{\theta})$ using the simple average of exact parameter estimates over the sample, as analysed above. Alternatively, we may consider $M_t(\hat{\theta}_t)$ or $\frac{1}{T}\sum_{k=1}^{T} M_t(\hat{\theta}_k)$ as two other option price estimates at time t. The former one directly uses the exact parameter estimate at time t (and thus equals the noisily observed option price), while the latter is the simple average of the price estimates obtained from all the exact parameter estimates.

The following theorem compares the three estimates of prices and hedging deltas from the exact parameter estimates. It states that, when the prices are contaminated by weakly exogenous noise and when they are modelled correctly, it is desirable to pool in other sampled observations when computing the estimate. In particular, the two estimates, $M_t(\bar{\theta})$ and $\frac{1}{T}\sum_{k=1}^{T} M_t(\hat{\theta}_k)$, are equally efficient in large samples. The naïve estimate $M_t(\hat{\theta}_t)$ is not good because it is a noisy estimator of the true price $\mathring{m}_t$ at time t even though it matches the observed price m_t exactly.

Theorem 8 *Under Assumptions* **WE**, **D**, $\mathbf{CV}_a$, $\mathbf{PL}_a^{exact}$ *($a = 1, 2$) and* $\mathbf{H}_0$, $\frac{1}{T}\sum_{k=1}^{T} M_t(\hat{\theta}_k)$ *is more efficient than $M_t(\hat{\theta}_t)$ and is as efficient as $M_t(\bar{\theta})$ asymptotically. The result still holds if we replace the pricing function $M_t(\cdot)$ by the hedging delta $\frac{\partial M_t(\cdot)}{\partial S}$.*

It is important to note that the finite sample and asymptotic analyses on the exact and errmin calibration are valid under correct model specification. In particular, the results may break down if the model is far from the true DGP, as is often the case in practice.

7 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 different properties **P** are needed to capture different aspects of dynamic model misspecification.

Let us suppose that the true DGP of the underlying stock price time series is the Merton's jump diffusion model, and the misspecified model to be tested is the Black-Scholes model. 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 with the price S_t of the underlying asset. In terms of partial derivative, we have

$$\frac{\partial IV_t}{\partial S_t} = \frac{\partial IV_t}{\partial C_t}\frac{\partial C_t}{\partial S_t} = \frac{1}{\frac{\partial C_t}{\partial IV_t}}\frac{\partial C_t}{\partial S_t}.$$

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

The Black-Scholes vega is given by $\frac{\partial C_t}{\partial IV_t} = S_t e^{-qt}\phi(d_1)\sqrt{\tau}$, 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 rate, dividend yield, moneyness, maturity, and model parameters, which are all assumed to be fixed over the in-sample period. It follows that $\frac{\partial IV_t}{\partial S_t}$ is inversely proportional to S_t. After discretization, we obtain

$$\Delta IV_t \approx c\frac{(\Delta S_t)}{S_t}. \tag{10}$$

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, IV_t displays heteroskedasticy as S_t evolves over time. The phenomenon is more pronounced under the jump diffusion model as the level of S_t changes dramatically at jump times, as illustrated by a simulated path in Figure 4.

The relation (10) 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 (10), 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 (10). This shows that different properties **P** are necessary to identify model misspecifications under the different calibration methods. Hence, when testing for model misspecification using either the errmin or exact calibration, it is important to test for all five properties **P**. Testing for only a subset (e.g., first moment properties) may not capture all possible model misspecifications (e.g., second moment properties).

8 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 errmin calibration.

8.1 Time Series Tests

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

First, let us consider testing of property **WN1** based on the parameter estimation error $\hat{\varepsilon}_t$, as defined in expression (4). 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 $Cov(\varepsilon_t, \varepsilon_s) = 0$, requires that any two parameter estimation errors are uncorrelated. A comprehensive statistical test for property **WN1** thus relies on both the sample variances and covariances of $\hat{\varepsilon}_t$ to reveal any potential model misspecification. Assuming unconditional homoskedasticity of ε_t, the Ljung-Box test (LB) is powerful against a departure from zero covariances up to a pre-specified maximum lag. Hong and Lee (2005) generalize the LB test by allowing for heteroskedasticity and avoiding the need to pre-specify the maximum lag. Their test (HL11) is powerful against all pairwise correlations of unknown form.

Next, we turn to the problem of testing property **MDS1**. A consistent test for a zero conditional mean 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 **MDS1**. 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. Hong (1999) gets around this conundrum by proposing a consistent test of the generalized spectral density that checks for the pairwise implication (5) 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 $\hat{\varepsilon}_t^2$. In parallel to the HL11 test, Hong and Lee (2005) provide a generalized version of the ML test (HL22). Similarly, the implication (6) of property **MDS2** can be checked by applying the Hong-Lee martingale test (HL20) to $\hat{\varepsilon}_t^2$. The implication 7 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 would lose any power by focusing on testing the implications expressions (5), (6), and (7) instead of the original properties **MDS1**, **MDS2** and **IID**. In theory, the Hong-Lee tests are designed to detect only all pairwise dependences of $\hat{\varepsilon}_t$; however, pairwise uncorrelatedness or independence is weaker than joint independence. Nevertheless, pairwise tests are sufficient for

detecting model misspecification for some common stochastic processes.[14]

8.2 Model Comparison Tests

The time series tests in the previous subsection provide a way to test a given model against the *absolute* criteria as given by properties **P**. While all models are approximations to the reality, it is sometimes more useful to have a set of statistical procedures that allow for a *relative* comparison between a number of competing models. Model comparisons can be either carried out by making use of various metrics that measures different aspects of departure from the ideal model (the DGP), or by formal statistical tests. We will discuss the following model comparison tools that are useful for our option price modelling purposes: (i) the sum-of-squared residuals, the Aikaiki information criteria, and chi-squared test; (ii) variance decomposition; (iii) time series test statistics; and (iv) orthogonality tests. In this subsection on model comparison, we accommodate to multivariate models and panel datasets by allowing for $d > 1$ and $n > 1$, where d is the model dimension and n is the cross sectional sample size.

(i) The *sum-of-squared residuals* (SSR). The SSR is defined on the pricing residuals $\hat{e}_{it}$ under the errmin approach: $SSR^{errmin} = \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{e}_{it}$. Under the exact approach, it is defined on the parameter estimation residuals $\hat{\varepsilon}_{it}^{(k)}$ for each parameter vector component: for parameter component $\theta^{(k)}$, $SSR^{exact(k)} = \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{\varepsilon}_{it}^{(k)}$ for $k = 1, \ldots, d$. Through summing over time and cross sections, the SSR measures the overall goodness-of-fit of a calibrated model. The closer is the SSR towards zero, the closer the model approximates the observables.

The SSR from the errmin approach can be readily used for model comparison. Using the quasi-likelihood methodology (by assuming i.i.d. Gaussian errors), the *Aikaiki information criteria* (AIC) can be computed in terms of the errmin SSR (apart from a common constant) as $AIC = 2d + nT \log\left(SSR^{errmin}/nT\right)$. Furthermore, the errmin SSR is useful for testing two nested models. To test model 0 against model 1, where model 0 is nested in model 1, we can apply the chi-squared test. It follows, under the null hypothesis, that the test statistic $\frac{nT}{2} \log\left(SSR_1^{errmin}/SSR_0^{errmin}\right)$ is asymptotically chi-squared distributed with $d_1 - d_0$ degrees of freedom, where d_k is the dimension of model k. With minor modification, the test is applicable to the testing of non-nested models (Pesaran and Deaton, 1978. See McAleer (1995) for a survey).

(ii) Variance decomposition. While the SSR provides a measure of the overall fit to the panel data, it ignores possible abnormal variation of pricing errors over time and cross-sectionally. One way to remedy this shortcoming is to decompose the variation of the pricing residuals from the errmin calibration into the variation only over time, the variation over only cross-sectionally, and the idiosyncratic variation. Denotng the cross-sectional mean at time t by $\bar{e}_t$, and the temporal mean for cross-section i by $\bar{e}_i$, the decomposition is presented below:

$$SSR^{errmin} = n \sum_{t=1}^{T} \bar{e}_t^2 + T \sum_{i=1}^{n} \bar{e}_i^2 + \sum_{i=1}^{n} \sum_{t=1}^{T} (\hat{e}_{it} - \bar{e}_i - \bar{e}_t)^2. \tag{11}$$

[14] 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 (5), (6) and (7).

The smaller is the variation, the more capable the model is in capturing the corresponding dynamics (cross-sectional or time) of the option panel.

Alternatively, we can apply a decomposition on the variation of parameter estimates from exact calibration. The decomposition for the component-k parameter can be defined analogously by replacing $\hat{e}_{it}$ by $\hat{\varepsilon}_{it}^{(k)}$ in (11). This approach allows for a finer and more revealing model specification analysis, as model misspecification may show up in the form of cross-sectional and/or time variations in certain parameters but not others. This would facilitate model diagnostics.

In order to facilitate comparison across exact-calibrated multivariate models, parameter variations need to be aggregrated over the parameters using some weighting schemes. Three different weights (equal weight, the reciprocal of the sample variance, and the reciprocal of the fourth central sample moment) are adopted for aggregation. The last two weighting schemes give heavier weights to more precisely calibrated parameters, and are preferred to the first scheme.[15]

(iii) Time series tests. The time series test statistics in the previous subsections can be treated as a model comparison device. The Ljung-Box and Hong-Lee (1,1) test statistics, when applied to normalized estimation residuals, are nonparmetric estimators of the long run variance of the residuals and can be regarded as a metric that measures how far the normalized residual sequence deviates from white noise. More generally, all the time series test statistics in the previous subsection can be viewed as measures of deviations of the model from different levels of "good approximations" as represented by the set of properties **P**. The time series tests are either applied to the pricing residuals under errmin calibration, or to each of the parameter's estimation residuals under exact calibration. Similar to the SSR and variance decomposition, comparison of exact-calibrated models can be achieved by aggregating the statistics over all parameter components.

(iv) Orthogonality tests. One shortcoming of the time series tests is that it ignores misspecifications along the cross-sectional dimension, which could result in a volatility smile, for example. To address this point, orthogonality tests can be applied directly to the panel of parameters (under exact) and pricing residuals (under errmin). We run panel regressions of the panel of parameters (under exact) or pricing residuals (under errmin) on a chosen set of variables that capture cross-section and time variations. The variables include the option's moneyness, moneyness squared and time to maturity, a linear time trend, and lagged parameters/pricing residuals. Random effects in time and cross section can be incorporated in the panel regressions. If the option pricing model is correct, the regressions will not have any significant explanatory power for the variations of the parameters (under exact) or pricing residuals (under errmin); hence the regression R-squared is close to zero, and the F test on the variable coefficients is insignificant.

To focus on just the cross-sectional or dynamic misspecifications, we carry out separate static and dynamic tests, whereby regressions are run on only the set of cross-sectional variables and on only the set of time variables, respectively. To compare exact-calibrated multivariate models, we apply the equal weighting scheme when averaging the adjusted R-squared and F statistics associated with regressions of different parameter components.

[15] In particular, the last weighting scheme enjoys optimality property. It can be shown that the variance of the measures is minimized when the weights are the reciprocal of the fourth central sample moments of the associated parameter estimation residuals.

9 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.

9.1 Exact vs errmin calibration

In this section, we study the finite sample performance of the five time series tests as discussed in Section 8. 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 (1976) jump-diffusion model, Heston's (1993) stochastic volatility model, and Heston and Nandi's (2000) GARCH(1,1) model. 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 underlier, 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 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 at $r = 0.05$ and $q = 0$, respectively.

Next, we calibrate the Black-Scholes model using both exact and errmin 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[16] is set to 1, and the nominal rejection rate is 10%. Table 1 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 2. The left panel shows the power curves for exact calibration, while the right panel corresponds to errmin calibration. First, we note that the powers of the five tests increase differently as the jump component of the DGP begins to dominate the diffusion component (as the jump intensity λ increases and the volatility σ decreases. See Merton, 1976b). The ML, HL20, and HL00 tests are more powerful than the LB and HL10 tests for detecting jumps. Second, 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 errmin calibration. This supports the discussion in Section 7. We can see from a simulation of the Merton model that heteroskedasticity is detectable in the time series of the exact parameter estimates but

[16] 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.

not in the errmin calibrated residuals in the observation space (Figure 4).

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. The power curves under exact calibration are displayed in Figure 5, and the power surface over $\log_{10}(\kappa)$ and σ_V for the HL00 test is displayed in Figure 6. We observe that the LB, HL10, and HL00 tests are more powerful than the ML and HL20 tests. The powers of all five tests generally increase with κ, σ_V, and ρ. Although not shown in the plots, the tests display similar power under errmin calibration.

Similar observations are made when the DGP is the Heston-Nandi GARCH(1,1) model (experiments 6-8). As shown in Figure 7, the test power under exact calibration increases with the persistence of volatility (β_1), the shock to volatility (α_1), and the degree of volatility asymmetry (γ_1). The tests display similar power under errmin calibration.

9.2 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 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.

Under exact calibration, we obtain the time series of implied volatilities (IV), and then apply the statistical tests on the demeaned IVs $\hat{\varepsilon}_t$, $t = 1, \ldots, T$. Under errmin 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 $\hat{e}_t$, $t = 1, \ldots, T$, 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 8. The same figure also shows the power curve associated with other DGPs, including the Heston SV model (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 (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. The power curves under the errmin approach are very similar and not displayed here.

9.3 Hedge Ratio Accuracy

In practice, one is not only interested in the pricing accuracy but also hedging performance of a calibrated model. This section compares the accuracy of the estimated hedge ratio obtained from both errmin 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 noise 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 errmin 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 errmin approach, we obtain the constant parameter $\hat{\theta} = \widehat{IV}$ as the solution as in Section 3.1. 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 compute the Black-Scholes hedge ratios.

To measure the accuracy of the estimated hedge ratios in comparison to the true ones, 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 errmin hedge ratio has a smaller RMSE than the exact hedge ratio using $\overline{IV}$. The latter becomes more accurate if the standard deviation of the heteroskedastic noise is $0.01 S_t$. This is consistent with the asymptotic result that estimators from exact calibration can be more efficient than those from errmin calibration under heteroskedastic noise, as suggested in Corollaries 1 and 2, coupled with Theorem 7, (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.

We also compare the accuracy of hedge ratios under the two calibration approaches when the

option pricing model is misspecified. Model misspecification arises when (i) the DGP comes from a different model family and/or (ii) the model structure varies over the sample period. Since exact calibration allows model parameters to vary over different observations, we expect that hedge ratios from an exact-calibrated model are more accurate than those from an errmin-calibrated model. While the model to be calibrated remains to be the Black-Scholes model, we consider four different DGP's: the Merton jump-diffusion model, the Heston SV model, the Heston-Nandi GARCH(1,1) model, and the geometric Brownian motion with a volatility break in the middle of the sample period. All four DGP's belong to a more general model family than the Black-Scholes model, but only the last one involves an abrupt change in the model parameter.

The results are displayed in Table 2. Under all DGP's, the hedge ratios computed with $\overline{IV}$ from exact calibration come close to those computed with $\widehat{IV}$ from errmin calibration in terms of accuracy. This is justified by the way we constructed $\overline{IV}$ and $\widehat{IV}$. Recall that $\overline{IV}$ is the sample average over the exact-calibrated parameters directly, while $\widehat{IV}$ is the parameter that minimizes the sum of squared pricing errors in errmin calibration. Both $\overline{IV}$ and $\widehat{IV}$ are obtained from an averaging process, except that the former acts on the parameter space and the latter acts on the observation space.

The advantage of using exact-calibrated IV_t directly when computing hedge ratios is manifested when the way the model is misspecified changes over the sample period. This takes the form of drifting volatility in the Heston-Nandi GARCH(1,1) model, and an abrupt change in the Black-Scholes volatility parameter in the form of a break in the middle of the sample period. We observe that the hedge ratios computed with exact-calibrated IV_t beat those computed with $\overline{IV}$ and $\widehat{IV}$ in terms of accuracy when the volatility drift is persistent ($\beta_1 = 0.95$) under the Heston-Nandi GARCH(1,1) model, or when the volatility break is relatively big ($\sigma_1 = 0.3$, $\sigma_2 = 0.8$) under geometric Brownian motion with volatility break.

This is however not the case when the DGP follows the Merton or Heston model. Indeed, if the parameters are kept constant over the sample period, the volatility dynamics of the underlying remain stable over time.[17] Provided that we fix the moneyness and time-to-maturity of the options used for calibration, the time series dynamics of Black-Scholes implied volatilities IV_t will remain stable too. We thus expect that the hedge ratios that directly use IV_t from exact calibration will not have an advantage over those computed with $\overline{IV}$ and $\widehat{IV}$ in terms of accuracy. This is confirmed by the simulation results when the DGP's are the Merton and Heston models.

10 An Empirical Study

In this section, we calibrate some well-known option pricing models using market data. The purpose is to provide an illustration of the misspecification tests and model comparison techniques applied to models calibrated under both the errmin and exact approaches.

The data sample consists of the S&P 500 index and its associated European call option prices

[17] Provided the parameters remain constant over the sample period, the degree of drifting for the diffusion component is fixed under the Merton model, and the stochastic volatility exhibits mean-reversion under the Heston model. Both cases give rise to stable volatility dynamics over the sample period.

from Option Metrics over the period January 1 – December 31, 2010 ($T = 252$ business days).

The sampling scheme is designed to minimize the effect of market micro-structure irregularities while ensuring liquidity. Specifically, we consider for each cross-section three groups of call options with three nearest time-to-maturities that are at least eight days. The calls in each group have different strike prices but share the same maturity date, which rolls forward in time to the next nearest date once it drops below eight days. The sample sizes are 32, 24 and 16 for the three groups in ascending order of maturities. An inspection of the data set reveals that the common granularity for the strikes of the S&P 500 index calls is $mesh = 25$.

There are two additional inputs for the option pricing 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&P 500 index, we divide the monthly dividend by the 21-day moving average of the S&P 500 index.[18]

We consider the Black-Scholes, Heston, Merton, and Bates models for our data set. The models are first calibrated by errmin and exact approaches[19], and then ranked using the model comparison devices as discussed in section 8.2.

Table 3 reports the SSR of the four models estimated from the common data set described earlier. Since the models are nested, it is straightforward to compare the models based on the SSR. Under the assumption that the errors are i.i.d. Gaussian, the test statistic can be expressed in terms of the log-SSR ratio, and is asymptotically chi-squared distributed with degrees of freedom equal to the difference of the number of parameters between the two models. We conclude from a sequence of pairwise tests that the BS, Merton, and Heston models are rejected in favor of the Bates model. To allow for a fair comparison that takes into account the number of parameters, the Akaike information criterion (AIC) is computed (except for a common constant) under the Gaussian error assumption. Under this criterion, the Bates model is the most preferred, followed by the Heston, Merton, and BS models.

To prepare for exact calibration, we tailor the data set to the models by setting the sample size to be a multiple of the number of parameters in the model, where the multiple is the number of non-overlapping matchings between the observed and the modelled sets of call prices for exact calibration. Table 4 contains model-specific details on the grouping of observations under exact and errmin calibration.[20] The sample sizes for different models are slightly different as a result.

The results are presented in Table 5, which displays the variance decompositions of errmin pricing residuals and estimated parameters from the exact calibration. Using the reciprocals of sample variance or sample kurtosis as weights, the variance decomposition of parameter estimates from the exact calibration reveals that the Bates model outperforms the Heston model in capturing both cross-sectional and time variations. The Heston model in turn beats the Merton and BS models, but the comparisons between the last two models are mixed. Under the errmin calibration,

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

[19] Exact calibration on multivariate models are discussed in more details in another paper (Jarrow and Kwok, 2014).

[20] There are clearly more than one way to group the observations for exact calibration of multivariate models. The effect of groupings on the identifiability of model parameters is related to the information contents of the observations in the group. This non-trivial issue is studied in Jarrow and Kwok (2014).

the Bates model has the smallest pricing residuals variance cross-sectionally and over the whole panel, but the residual variance over time achieves the minimum for the BS model. We point out that under the errmin model calibration, the model parameters are assumed to remain constant over the entire panel; the parameter estimates and hence the pricing residuals are thus susceptible to the effect of estimation error and model misspecification error in finite samples. The model comparison result is likely to be affected accordingly and has to be interpreted with care.

More importantly, we note that the above model comparison based on the variance decomposition ignores properties **P**. Indeed, dynamic misspecifications in the form of serial correlations, heteroskedasticity, etc., are not readily reflected in the sample variance, which only measures the overall extent of fluctuation around the mean. To conduct a more careful model specification analysis, we apply the time series tests in Section 8.1 to the parameter estimates from exact calibration and the pricing residuals from errmin calibration. The time series test statistics serve as natural norms of deviation from the various benchmarks as represented by properties **P**. Since the tests under consideration are univariate by nature, it is necessary to reduce the time series to one dimension before testing by averaging the parameters (under exact) and pricing residuals (under errmin) over each cross section at each point in time. After the tests, the statistics associated with different parameters under exact calibration are aggregated by a weighted sum, where the weights are either equal or the inverted sample variance of the parameter estimates.

Table 6 reports the realized time series test statistics across the four models estimated under both calibration approaches.[21] Under the equal weighting scheme, we observe that all the tests uequivocally rank the Bates model as the best one, and that the Heston model beats the Merton and BS models. The test results imply that, in terms of the deviation from properties P, the Bates model has the least dynamic misspecification as exhibited by the parameters from exact calibration. In contrast, the tests on errmin residuals vote the BS model to be the best model. The test results could be affected by the errmin parameter estimates.

The dynamic fluctuation of these parameters, aggregated cross-sectionally, from the four models can be visualized in Figure 9. A comparison between the plots of Heston and Bates model parameters reveals that the presence of the jump component stablizes the parameters in the stochastic volatility component over time ($a, b, V_0, \sigma_V, \rho$, where $b = \kappa$ and $a = \kappa\theta$. See the appendix for model formulation). Similarly, allowing the volatility to be stochastic narrows down the range of fluctuations of the diffusion parameter (compare the plots of σ in the BS and Merton, and the plots of σ in Merton and V_0 in the Heston and Bates models).

Table 7 presents the results of the orthogonality tests, including the panel regression orthogonality tests (with and without random effects), static tests, and dynamic tests. Under exact calibration, all the tests on the parameters unanimously point to one conclusion: the Bates model is the best, followed in turn by the Heston, Merton, and BS models. As for the tests on the pricing residual panel under errmin calibration, the panel regression and static tests favor the Heston model, but the results are mixed under the dynamic tests.

In order to study the effect of calibration on the secondary usage, we compare the hedging

[21] In our empirical analysis, the maximum lags and the tuning parameters of the time series tests, M, are set to be five. The results remain qualitatively identical for $M = 2$ and 10.

performance under different calibration schemes. At any point in time, we form a dynamic hedge portfolio by longing one unit of call option priced at m_t, and shorting ϕ_t units of the S&P 500 portfolio priced at S_t, where ϕ_t is the hedge ratio determined by either the delta hedging or the minimum variance hedging method. The latter method is more appropriate for models that exhibit stochastic volatility and jumps. The hedge ratios can be computed using the model parameters under different models (Bakshi, Cao and Chen, 1997). Three calibration schemes are considered to estimate the hedge ratio ϕ_t at time t: using the group-specific parameter estimate $\hat{\theta}_{It}$ from exact calibration, using their simple average $\bar{\theta}$ over the sample, and using the errmin estimate $\hat{\theta}$. From the estimated hedge ratio $\hat{\phi}_t$, the hedge portfolio value at time t is obtained as $m_t - \hat{\phi}_t S_t$. The standard deviations and the cumulated value of the hedge portfolio value over the sample period are reported in Table 8.

One caveat is that it is not very enlightening to compare the hedging performance across models as all of them are misspecified to different extents, which means that the hedge portfolio was formed based on an incorrect hedge ratio. Nevertheless, the results show that, except for the Merton model, the minimum variance hedge portfolio has a smaller cumulated value with smaller fluctuation if we adopt the exact calibration (both using $\hat{\theta}_{It}$ and $\bar{\theta}$), which has slight advantage over errmin calibration. The advantage is the biggest for the Bates model, which happens to be the most favorable model out of the four.

11 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 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 noise, our simulation results show that the Black-Scholes model calibrated by the exact approach delivers more accurate hedging performance than that obtained by error minimization calibration. In the empirical study, we found from a series of model comparison and test procedures that the Bates model generally outperforms the Heston, Merton, and Black-Scholes models in the overall fit and in explaining both the time series and cross section variations of the S&P 500 index options. Compared to the error minimization approach, exact calibration often delivers richer model diagnostic information and more reasonable model comparison results, as illustrated in Section 10.

Due to the fundamental nature of the exact calibration methodology, this paper points to an array of future research directions, both theoretical and empirical. On the theoretical side, exact

calibration of a multivariate model creates interesting issues to be resolved, and is the focus of another paper by the same authors. On the empirical side, it would be informative to re-estimate the 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 applications is a fruitful area for future research.

References

[1] Aït-Sahalia, Yacine and Andrew Lo (1998), "Non-Parametric Estimation of State-Price Densities Implicit in Financial Asset Prices," *Journal of Finance*, 53, 2, pp.499–547.

[2] Andersen, Torben, G., Nicola Fusari, and Viktor Todorov (2012), "Parametric Inference and Dynamic State Recovery from Option Panels," NBER Working Paper Series.

[3] Bakshi, Gurdip, Charles Cao, and Zhiwu Chen (1997), "Empirical Performance of Alternative Option Pricing Models," *Journal of Finance*, **52**, 5, pp.2003–2049.

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

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

[6] Chen, Bin, and Yongmiao Hong (2012), "Testing for Smooth Structural Changes in Time Series Models Via Nonparametric Regression," *Econometrica*, **80**, 3, pp.1157–1183.

[7] Christoffersen, Peter, F., and Kris Jacobs (2004), "The Importance of the Loss Function in Option Valuation," *Journal of Financial Economics*, **72**, 2, pp.291–318.

[8] Cont, Rama, and Peter Tankov (2004), "Nonparametric calibration of jump-diffusion option pricing models," *Journal of Computational Finance*, **7**, 3, pp.1–49.

[9] Detlefsen, K., and Wolfgang K. Härdle (2007), "Calibration Risk for Exotic Options," *Journal of Derivatives*, **14**, 4, pp.47–63.

[10] Gagliardini, Patrick., Christian Gourieroux, and Eric Renault (2011), "Efficient Derivative Pricing by the Extended Method of Moments," *Econometrica*, **79**, 4, pp.1181–1232.

[11] Hansen, Lars Peter (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, **50**, 4, pp.1029–1054.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] Jarrow, Robert, and Simon Kwok (2014), "Specification Test of Multivariate Calibrated Financial Models," Working Paper.

[18] McAleer, Michael (1995), "The Significance of Testing Empirical Non-nested Models," *Journal of Econometrics*, **67**, pp.150–171.

[19] 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.

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

[21] 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.

[22] 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.

[23] Pesaran, M. H., and A. S. Deaton (1978), "Testing Non-Nested Nonlinear Regression Models," *Econometrica*, **46**, 3, pp.677–694.

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

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

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

11.1 Figures

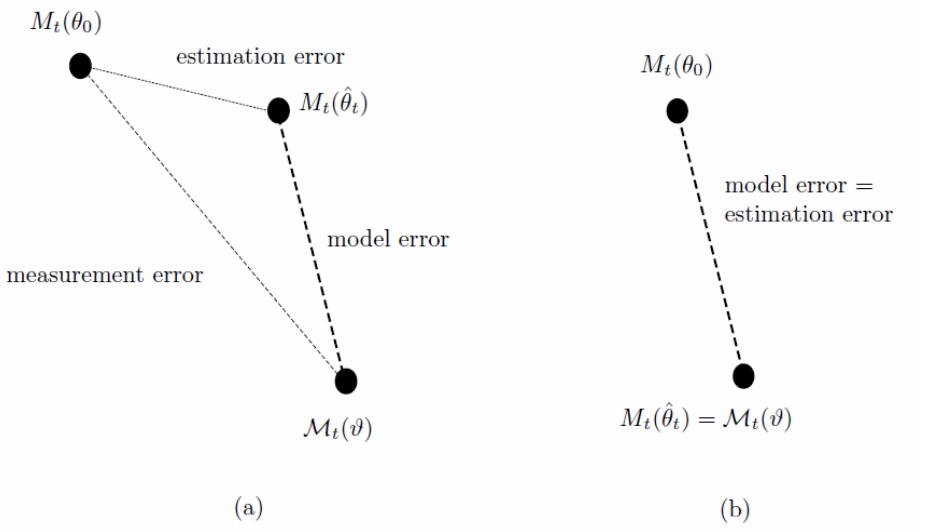

Figure 1: Error decomposition under the (a) errmin and (b) exact approaches.

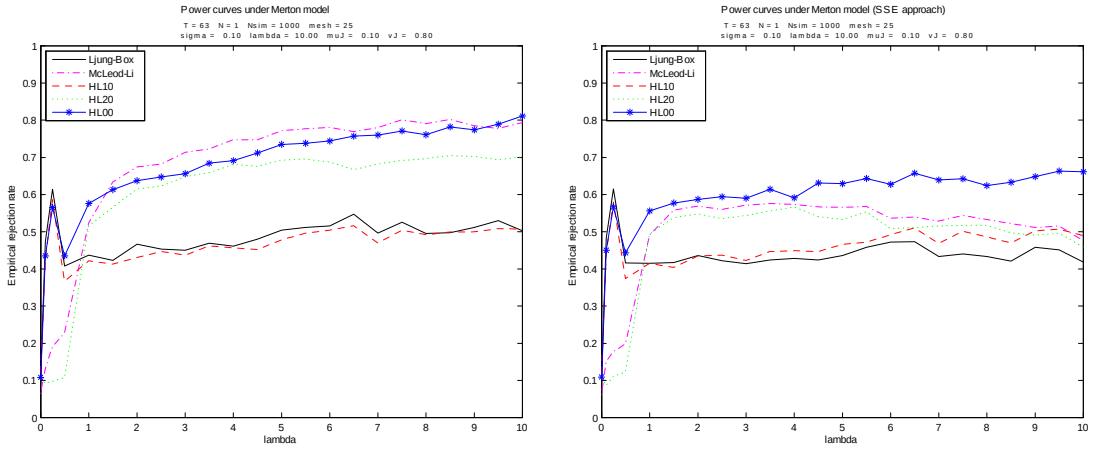

Figure 2: The power of five model misspecification tests over jump rate λ under exact (left panel) and errmin (right panel) calibration. DGP = Merton; Misspecified model = Black-Scholes.

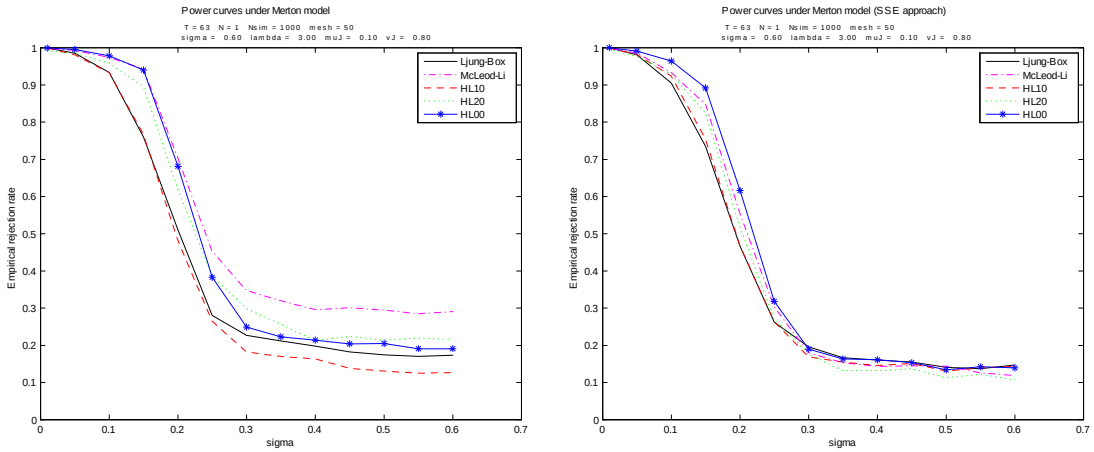

Figure 3: The power of five model misspecification tests over volatility σ under exact (left panel) and errmin (right panel) calibration. DGP = Merton; Misspecified model = Black-Scholes.

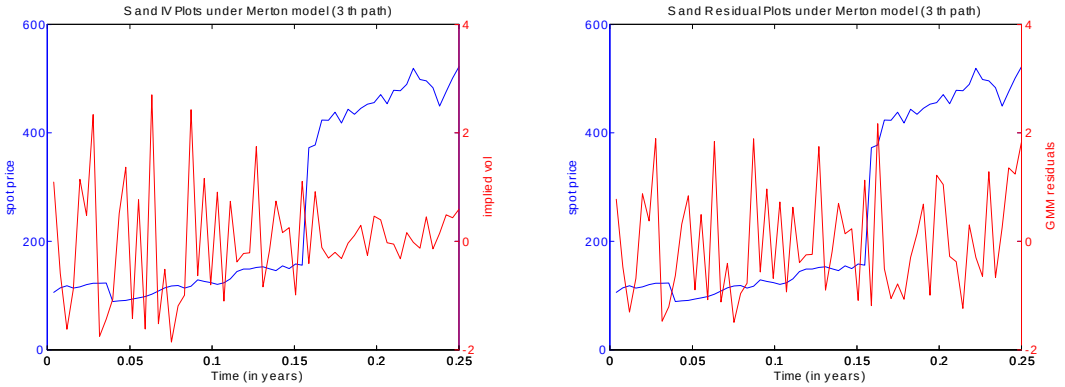

Figure 4: The residuals (in red and in the parameter space) exhibits heteroskedasticity under exact calibration but not under errmin calibration as the underlying index (in blue) jumps over time, as in the Merton model.

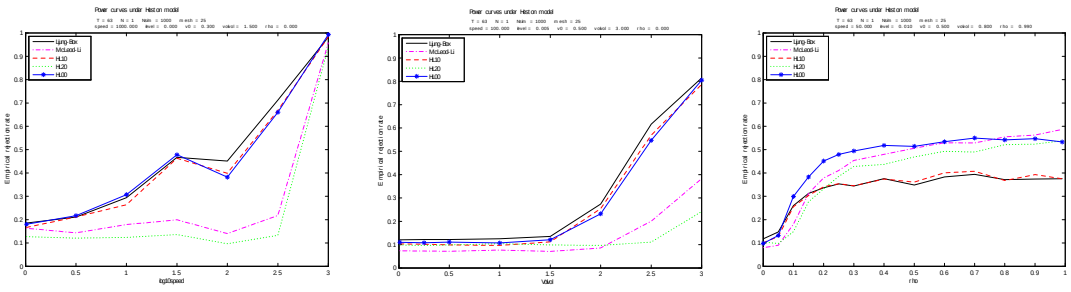

Figure 5: The power of five model misspecification tests over log speed $\log_{10}(\kappa)$, volatility of volatilities σ_V, and correlation ρ under exact calibration. DGP = Heston; Misspecified model = Black-Scholes.

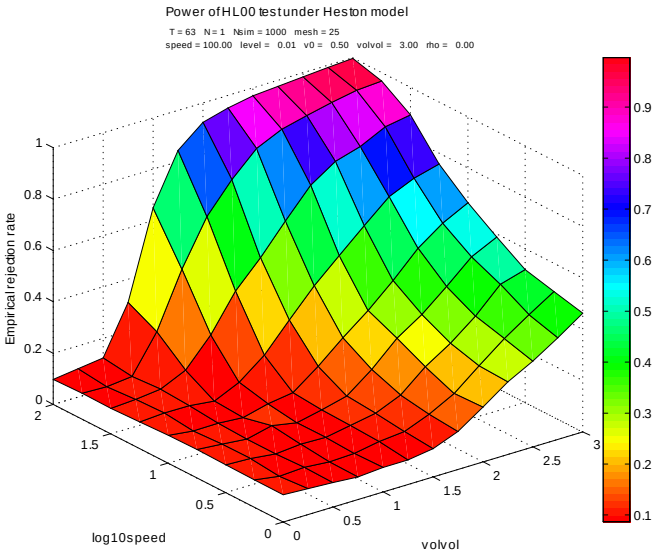

Figure 6: The power surface of Hong-Lee test of serial independence over $\log_{10}(\kappa)$ and σ_V under Heston model (Exp 3-4), Nominal rate = 0.01.

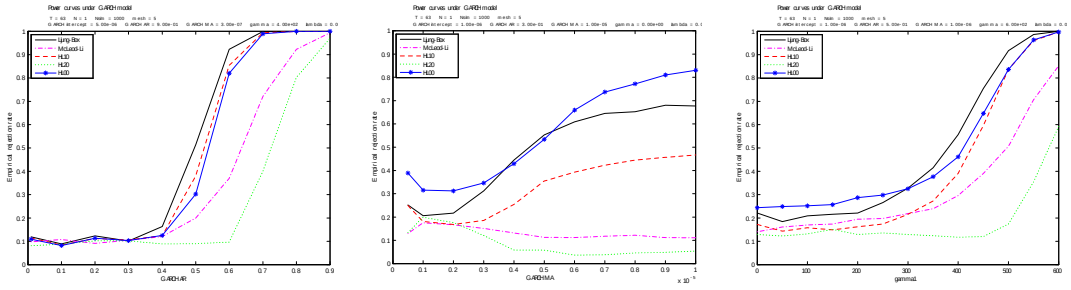

Figure 7: The power of five model misspecification tests over the persistence β_1, the shock α_1, and the asymmetry γ_1 of volatility under exact calibration. DGP = Heston-Nandi; Misspecified model = Black-Scholes.

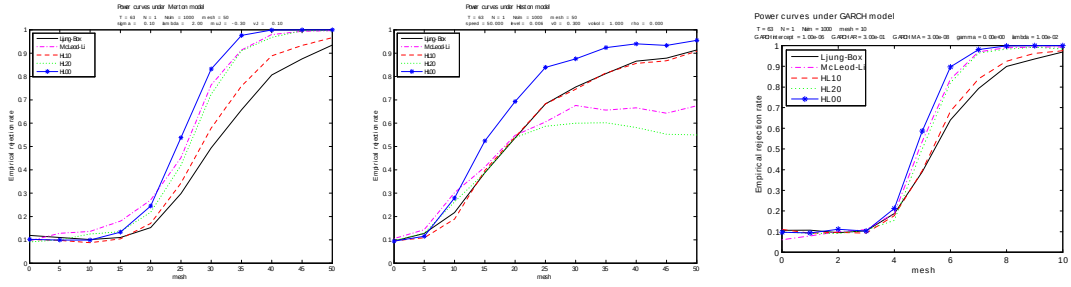

Figure 8: The plot of powers of five statistical tests against mesh under the exact approach. DGP = Merton, Heston and GARCH models.

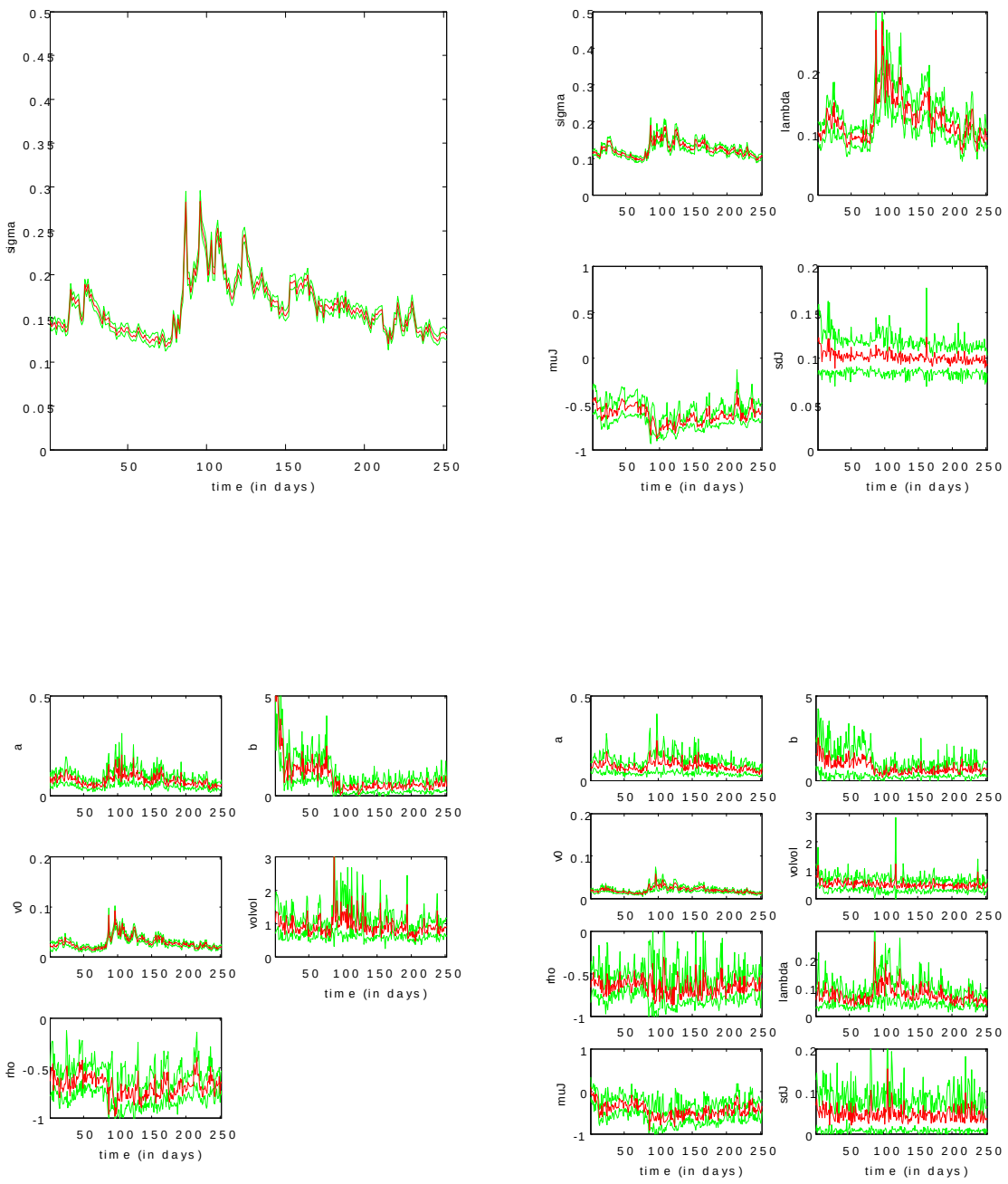

Figure 9: Time series plots of exact-calibrated parameter estimates of the Black-Scholes (upper left), Merton (upper right), Heston (lower left) and Bates (lower right) model, aggregated over cross-section (in red), and the confidence bound (in green).

11.2 Tables

Table 1: The data generating processes used in various simulation experiments.

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: Simulation results of delta hedging

		RMSE of hedge ratio		
DGP	noise	Exact	Exact mean	Errmin
BS	$N(0, 0.01^2)$	4.405×10^{-5}	5.640×10^{-6}	5.440×10^{-6}
	$N(0, 0.01^2 S_t{}^2)$	4.001×10^{-3}	5.165×10^{-4}	5.284×10^{-4}
	$N(0, 0.01^2 S_t)$	4.140×10^{-4}	5.336×10^{-5}	5.308×10^{-5}
Merton	$N(0, 0.01^2)$	7.433×10^{-2}	7.423×10^{-2}	7.425×10^{-2}
	$N(0, 0.01^2 S_t{}^2)$	7.447×10^{-2}	7.426×10^{-2}	7.429×10^{-2}
	$N(0, 0.01^2 S_t)$	7.433×10^{-2}	7.424×10^{-2}	7.425×10^{-2}
Heston	$N(0, 0.01^2)$	1.535×10^{-4}	4.473×10^{-4}	1.473×10^{-4}
	$N(0, 0.01^2 S_t{}^2)$	4.009×10^{-3}	5.239×10^{-4}	5.362×10^{-4}
	$N(0, 0.01^2 S_t)$	4.327×10^{-3}	1.558×10^{-4}	1.559×10^{-4}
GARCH	$N(0, 0.01^2)$	1.100×10^{-3}	7.025×10^{-3}	7.322×10^{-3}
	$N(0, 0.01^2 S_t{}^2)$	4.921×10^{-2}	7.055×10^{-3}	7.377×10^{-3}
	$N(0, 0.01^2 S_t)$	1.204×10^{-2}	7.026×10^{-3}	7.324×10^{-3}
BS with vol break	$N(0, 0.01^2)$	5.658×10^{-7}	1.986×10^{-5}	2.019×10^{-2}
	$N(0, 0.01^2 S_t{}^2)$	6.328×10^{-3}	1.987×10^{-2}	2.020×10^{-2}
	$N(0, 0.01^2 S_t)$	5.519×10^{-4}	1.986×10^{-2}	2.019×10^{-2}

The table shows the root mean squared errors of Black-Scholes delta computed from the implied volatilities associated with the observed call price (Exact), their simple average (Exact mean), and the volatility estimate from errmin calibration (Errmin). Three types of observations noises are considered: i.i.d. Gaussian with s.d. 0.01, independent Gaussian noises with s.d. $0.01S_t$, and independent Gaussian noises with s.d. $0.01S_t{}^{0.5}$. The DGP's are the Black-Scholes model (σ = 0.6), the Merton's jump diffusion model (σ = 0.6, λ = 15, μ_J = -0.3, σ_J = 0.6), the Heston's stochastic volatility model (κ = 10, ϑ = 0.05, V_0 = 0.5, σ_V = 0.4, ρ = 0), the Heston-Nandi GARCH(1,1) model (ω = 5×10^{-6}, β_1 = 0.95, α_1 = 5×10^{-4}, γ_1 = 0, λ = 0.3), and the Black-Scholes model with a volatility break (σ_1 = 0.3, σ_2 = 0.8).

Table 3: Pricing residual analysis

	Model			
	BS	Merton	Heston	Bates
d	1	4	5	8
SSE	616317.6	479787.9	474835.8	474334.6
AIC	63967.3	59429.7	59243.4	59230.3
ranking	4	3	2	1

SSE = sum of squared pricing residuals from errmin calibration. The *AIC* of a model is computed as (apart from a common constant) $2d + nT \log(SSE/nT)$, where $nT = 18144$ is the sample size, and d is the dimension of the model.

Table 4: Data structure of the panel of S&P 500 index calls

Sample period:	Jan 1 - Dec 31, 2010 (252 business days)
Mesh for strike prices:	25 (in units of index)
Time to maturities:	> 7 days from 3 nearest maturity groups

Data for errmin calibration

At each t, no. of obs. in the:	
1st nearest maturity group	32
2nd nearest maturity group	24
3rd nearest maturity group	16
No. of obs. in a x-section	72
Total no. of obs.	18,144

Data for exact calibration

	Model			
	BS	Merton	Heston	Bates
Model dimension	1	4	5	8
At each t, no. of obs. in the:				
1st nearest maturity group	30	32	35	32
2nd nearest maturity group	20	24	25	24
3rd nearest maturity group	10	16	15	16
No. of obs. in a x-section	60	72	75	72
Total no. of obs.	15,120	18,144	18,900	18,144

Table 5: Variance decomposition

		Variances				Ranking			
Exact calibration - variance decomposition of parameter estimates									
		BS	Merton	Heston	Bates	BS	Merton	Heston	Bates
Variance over x-section:									
	equal weight	0.0005	0.0073	0.0964	0.0626	1	2	4	3
weights	1/variance	0.0005	0.0009	0.0003	0.0002	3	4	2	1
	1/kurtosis	0.0005	0.0004	0.0001	0.00001	4	3	2	1
Variance over time:									
	equal weight	0.0010	0.0026	0.1715	0.0256	1	2	4	3
weights	1/variance	0.0010	0.0005	0.0004	0.0001	4	3	2	1
	1/kurtosis	0.0010	0.0004	0.0002	0.0001	4	3	2	1
Residue (idiosyncratic) variance:									
	equal weight	0.0001	0.0045	0.2420	0.0818	1	2	4	3
weights	1/variance	0.0001	0.0005	0.0006	0.0004	1	3	4	2
	1/kurtosis	0.0001	0.0002	0.0001	0.0001	3	4	2	1
Total variance:									
	equal weight	0.0016	0.0144	0.5099	0.1700	1	2	4	3
weights	1/variance	0.0016	0.0019	0.0012	0.0008	3	4	2	1
	1/kurtosis	0.0016	0.0010	0.0003	0.0001	4	3	2	1
Errmin calibration - variance decomposition of pricing residuals									
		BS	Merton	Heston	Bates	BS	Merton	Heston	Bates
Var over x-section		7.40	0.89	0.53	0.54	4	3	1	2
Var over time		21.42	21.11	19.62	21.21	4	2	1	3
Residue variance		4.24	4.45	4.77	4.39	1	3	4	2
Total variance		33.06	26.35	24.89	26.11	4	3	1	2

Table 6: Time series tests

	Test statistics				Ranking			
	BS	Merton	Heston	Bates	BS	Merton	Heston	Bates
Exact calibration								
LB	796.5	562.9	402.0	248.0	4	3	2	1
ML	205.5	164.0	111.1	49.2	4	3	2	1
HL(1,1)	79.3	73.9	45.3	35.1	4	3	2	1
HL(2,2)	14.1	11.4	4.6	3.5	4	3	2	1
HL(1,0)	62.4	52.1	36.5	29.0	4	3	2	1
HL(2,0)	13.7	12.6	3.3	2.8	4	3	2	1
HL(0,0)	79.8	62.4	50.8	33.3	4	3	2	1
Errmin calibration								
LB	836.0	855.6	852.4	853.8	1	4	2	3
ML	258.3	270.2	273.3	275.0	1	2	3	4
HL(1,1)	97.0	99.7	98.3	98.5	1	4	2	3
HL(2,2)	12.3	13.1	12.4	12.5	1	4	2	3
HL(1,0)	71.8	70.8	70.8	70.7	4	2	3	1
HL(2,0)	12.7	14.9	14.0	14.3	1	4	2	3
HL(0,0)	88.7	88.7	88.7	88.5	4	2	3	1

LB = Ljung-Box test, ML = McLeod-Li test, HL(i,j) = Hong-Lee test with orders (i,j). The maximum lag for LB and ML tests, and the bandwidth parameter for HL(i,j) tests are set to be 5.

Table 7: Orthogonality tests

		Statistics				Ranking			
		BS	Merton	Heston	Bates	BS	Merton	Heston	Bates
Exact calibration									
panel	Adj. R^2 (with RE)	0.97	0.76	0.50	0.41	4	3	2	1
regression	Adj. R^2 (no RE)	0.80	0.63	0.33	0.27	4	3	2	1
static	Adj. R^2	0.96	0.36	0.18	0.14	4	3	2	1
regression	F statistic	1063.42	10.36	2.96	2.75	4	3	2	1
dynamic	Adj. R^2	0.78	0.40	0.20	0.10	4	3	2	1
regression	F statistic	446.24	138.18	46.00	17.68	4	3	2	1
Errmin calibration									
panel	Adj. R^2 (with RE)	0.96	0.95	0.95	0.95	4	2	1	3
regression	Adj. R^2 (no RE)	0.85	0.83	0.82	0.83	4	2	1	3
static	Adj. R^2	0.82	0.78	0.77	0.82	4	2	1	3
regression	F statistic	176.73	43.45	38.31	45.42	4	2	1	3
dynamic	Adj. R^2	0.79	0.80	0.79	0.80	2	4	1	3
regression	F statistic	502.6	549.6	524.9	552.9	1	3	2	4

RE = random effects

Table 8: Hedging performance

	Std deviations of hedge portfolio's value			
	BS	Merton	Heston	Bates
Exact calibration				
delta hedging	11.8	15.1	12.6	14.5
min var hedging	11.8	13.0	11.6	13.1
Exact calibration (mean)				
delta hedging	11.8	15.0	12.5	14.4
min var hedging	11.8	13.0	11.6	13.2
Errmin calibration				
delta hedging	11.9	15.0	12.4	14.3
min var hedging	11.9	12.8	11.7	13.9

	Cumulated hedge portfolio's value			
	BS	Merton	Heston	Bates
Exact calibration				
delta hedging	0.251	0.286	0.234	0.235
min var hedging	0.251	0.142	0.160	0.148
Exact calibration (mean)				
delta hedging	0.238	0.277	0.223	0.224
min var hedging	0.238	0.094	0.156	0.126
Errmin calibration				
delta hedging	0.246	0.270	0.223	0.215
min var hedging	0.246	0.083	0.177	0.201

12 Appendix

12.1 Data generating processes

This appendix presents the different data generating processes of the underlier'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. Provided that a jump occurs, the jump size distribution is given by $\log(1+J_t) \sim N\left(\log(1+\mu_J) - \frac{\sigma_J^2}{2}, \sigma_J^2\right)$.

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) (expressed in discrete time):
$$\begin{aligned} \log S(t) &= \log S(t- \triangle) + (r_t - q_t) + \lambda h(t) + \sqrt{h(t)} W(t) \\ h(t) &= \omega + \beta_1 h(t- \triangle) + \alpha_1 \left(W(t- \triangle) - \gamma_1 \sqrt{h(t- \triangle)}\right)^2 \end{aligned}$$
where $\triangle$ represents the length of a unit time step.

12.2 Proof of Theorem 7

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

Under Assumption $\mathbf{CV}_1$, $M_t(\hat{\theta})$ is asymptotically more efficient than $M_t(\bar{\theta})$, 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 $V_1^{errmin} < V_1^{exact}$ in general, when ∇_t are not identically the same for all t.

However, under Assumption $\mathbf{CV}_2$, it is possible that $M_t(\bar{\theta})$ is *asymptotically more efficient* than $M_t(\hat{\theta})$. Here is an artificial but simple example: suppose the sample size is $2T$ and $d = 1$.

For $t = 1, \ldots, T$, we have $\nabla_t = \sigma_1 = 1$; for $t = T+1, \ldots, 2T$, we have $\nabla_2 = \sigma_2 = 0.2$. Then, we see that $V_2^{errmin} = 1.852 > V_2^{exact} = 1$.

The proof for the delta hedge ratio is obtained by replacing all ∇_t by ∇_{St} and all $M_t(\cdot)$ by $\frac{\partial M_t(\cdot)}{\partial S}$.

12.3 Proof of Theorem 8

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

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

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

Now, let us compare the asymptotic efficiency of the estimators when $a = 2$, which nests the case of $a = 1$. For $t \neq k$,

$$M_t(\hat{\theta}_k) = M_t\left(M_k^{-1}(\hat{m}_k + \upsilon_k)\right) = M_t\left(\theta_0 + \frac{\partial M_k^{-1}(\tilde{m}_k)}{\partial m}\upsilon_k\right).$$

for some $\tilde{m}_k$ lying between m_k and $\hat{m}_k$.

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

$$\frac{1}{\sqrt{T}}\sum_{k=1}^{T} M_t(\hat{\theta}_k) \xrightarrow{d} N\left(M_t(\theta_0), \nabla_t \operatorname*{plim}_{n\to\infty}\left[\frac{1}{T}\sum_{k=1}^{T}\sigma_k^2\left(\nabla_k\nabla_k'\right)^{-1}\right]\nabla_t'\right). \tag{12}$$

Therefore, for large T, its variance can be approximated by

$$Var\left(\frac{1}{T}\sum_{k=1}^{T} M_t(\hat{\theta}_k)\right) \approx \frac{1}{T}\nabla_t\left[\frac{1}{T}\sum_{k=1}^{T}\sigma_k^2\left(\nabla_k\nabla_k'\right)^{-1}\right]\nabla_t'.$$

By assumption $\mathbf{PL}_a^{exact}$, the limiting variance is finite, and so, for large enough T, the finite sample variance $Var\left(\frac{1}{T}\sum_{k=1}^{T} M_t(\hat{\theta}_k)\right)$ is smaller than the variance of $M_t(\hat{\theta}_t)$:

$$Var\left(M_t(\hat{\theta}_t)\right) = Var\left(\hat{m}_t + \upsilon_t\right) = \sigma_t^2.$$

Furthermore, a comparison between (12) and the result of Corollary 2 shows that $\frac{1}{T}\sum_{k=1}^{T} M_k(\hat{\theta}_k)$ is as efficient as $M_t\left(\bar{\theta}\right)$ asymptotically.

The proof for the delta hedge ratio follows from Corollary 2.