Errors in Implied Volatility Estimation

LUDGER HENTSCHEL

Simon School, University of Rochester, Rochester, NY 14627

September 26, 2003

Forthcoming in
Journal of Financial and Quantitative Analysis

Comments Welcome

ABSTRACT: Estimating implied volatility by inverting the Black–Scholes formula is subject to considerable error when option characteristics are observed with plausible errors. Especially for options away from the money, large changes in volatility produce small changes in option prices. Conversely, small errors in option prices and other option characteristics produce large errors in implied volatilities. In the presence of small measurement errors, unobserved truncation of option prices that violate lower bounds for absence of arbitrage can also lead to systematic volatility "smiles." The paper proposes feasible GLS estimators that reduce the noise and bias in implied volatility estimates.

KEYWORDS: Black–Scholes, generalized least squares, implied volatility, options, truncation, VIX, volatility, volatility index, volatility smile, volatility smirk.

JEL CLASSIFICATIONS: G13, C24.

This paper is a revised version of "Caveat Invertor: Errors in Implied Volatility Estimation," Chapter 3 of my Ph.D. dissertation (Hentschel 1994). I am grateful to James Bodurtha, Jr., John Campbell, Chris Jones, Shakeeb Khan, John Long, Jr., Will Melick, John Montgomery, Matt Pritsker, Bill Schwert, Cliff Smith, Jr., Charlie Thomas, Jerry Warner, and two anonymous referees for helpful conversations and suggestions. Part of this work was completed at the Board of Governors of the Federal Reserve System. The John M. Olin Foundation provided generous financial support for this research.

Contents

1 Introduction

In the frictionless market of Black and Scholes (1973), we observe all prices without error and every option price can be inverted to find the unique implied volatility consistent with the observed prices. In reality, however, we observe prices with errors stemming from finite quote precision, bid–ask spreads, nonsynchronous observations, and other measurement errors. This seemingly innocuous difference turns implied volatility calculations into implied volatility estimations and raises questions about the precision of the estimates. Implied volatility estimates are imprecise when large changes in volatility produce small changes in option prices, because, conversely, small errors in option prices or other option characteristics produce large errors in implied volatilities. This is especially true for options far from the money.

This paper shows that many implied volatility estimates are noisy and biased. 95% confidence intervals for implied volatility from an at-the-money stock option with 20 days to maturity are on the order of plus or minus 6 percentage points. Implied volatilities can also show pronounced "smiles" or "smirks," even if all of the Black–Scholes assumptions hold but we measure prices with plausible errors. The bias arises because the lower absence of arbitrage bounds on option prices systematically eliminate low implied volatilities. Option prices typically are far from the upper absence of arbitrage boundary, so that the truncation is one-sided. If dealers post prices that obey absence of arbitrage, then this truncation is not directly observed but nonetheless biases implied volatilities. Since the truncation bias is most severe for options away from the money, observed implied volatilities show smiles or smirks.

To gauge and reduce the noise and bias in implied volatilities, this paper shows how to (i) construct confidence intervals for implied volatility, (ii) construct efficient estimators of implied volatilities, and (iii) assess and reduce the bias in implied volatility estimates. Although the paper focuses on stock options and stock index options under the Black–Scholes assumptions, the methods proposed here translate directly to other options and option pricing models. The only requirement is the ability to calculate option price derivatives with respect to the options' characteristics.

Researchers as early as Latané and Rendleman (1976), Chiras and Manaster (1978), and Beckers (1981) tried to form weighted average implied volatilities that give low weights to less informative implied volatilities and high weights to more informative implied volatilities.[1] The common theme of these weighting schemes is that they pay exclusive attention to the sensitivity of volatility to option price errors. Other errors are ignored. This paper derives efficient implied volatility estimators when all option characteristics are measured with errors. Variance decompositions show that the error from the underlying asset

[1] Beckers (1981) chooses implied volatility to minimize the weighted sum of squared option pricing errors. Whaley (1982) minimizes the equally-weighted sum of squared pricing errors over implied volatility estimates since he is primarily concerned with option prices, not implied volatility.

price accounts for the majority of the variance in implied volatility estimates from in-the-money options.

Recently, exchanges have begun to provide implied volatility indices, such as the Chicago Board Options Exchange Market Volatility index, VIX, based on implied volatilities from eight S&P 100 options. The paper shows that the VIX-methodology produces a fairly efficient estimate of implied volatility—even though the methods proposed here produce slightly more precise volatility estimators. 95% confidence intervals for the VIX are on the order of plus or minus 25 basis points. The three main sources of the VIX's precision are the focus on near-the-money options, the low weights assigned to implied volatility from options near expiration, and the averaging of implied volatilities from puts and calls to cancel errors in the underlying asset price.

Implied volatilities appear to be neither unbiased nor efficient forecasts of future volatility. Franks and Schwartz (1991), Lyons (1988), Day and Lewis (1988), Chiras and Manaster (1978), Schmalensee and Trippi (1978), Canina and Figlewski (1993), Fleming (1998), and Christensen and Prabhala (1999) document this. The standard test of regressing realized volatility on past implied volatility produces a positive intercept and a slope coefficient that is significantly less than one.[2] Measurement error and upward bias in implied volatilities are broadly consistent with this finding.

Implied volatility patterns like smiles or smirks appear to violate the Black–Scholes model, and in particular the assumption that the underlying asset returns are generated by a lognormal diffusion. Recent work by Andersen, Benzoni, and Lund (2002), Bates (1996), Jones (2001), or Pan (2002), for example, has focused on augmenting the return process to include stochastic volatility or jumps in order to replicate the observed patterns in implied volatilities. This paper points out that plausible measurement errors can induce smiles or smirks in implied volatilities even when the Black–Scholes model is true. Although the biases in implied volatilities stem from biases in observed prices, the non-linear amplification of the price biases means that small biases in some option prices lead to large biases in the associated implied volatilities. As a result, one has to be careful when using implied volatilities to assess the performance of option pricing models (see Rubinstein 1985, or Bakshi, Cao, and Chen 1997 in addition to the papers cited above).[3] The methods of this paper are designed to improve the statistical assessment of the observed implied volatilities and can be adapted to any option pricing model. Whether the primary cause of the observed patterns in implied volatilities is stochastic volatility, jumps, or measurement errors is an

[2] Some of these results must be interpreted with care. Christensen, Strunk Hansen, and Prabhala (2001) show that OLS regression estimates are inconsistent and standard inference is inappropriate when the overlap between successive volatilities is large relative to the sample size and the overlap varies over time.

[3] Christoffersen and Jacobs (2001) make a related point when they show that the two-step procedure of first fitting an implied volatility function and then using the fitted implied volatilities to compute theoretical option prices can be very different from the one-step procedure of directly fitting a theoretical option pricing model.

empirical question this paper does not answer, since the answer requires detailed option price data.

For any option pricing model, there may be specification errors that lead to "model errors" in implied volatility estimates. Ledoit and Santa-Clara (2000) argue that, as at-the-money options converge to zero time to expiration, Black–Scholes implied volatility converges to true volatility, even if the Black–Scholes model is wrong. Unfortunately, measurement errors limit the power of this asymptotic argument. In the presence of measurement errors, at-the-money options near expiration provide extremely noisy volatility estimates. Moreover, near expiration even options with small differences between the strike price and underlying asset price are far from the money. The paper shows that implied volatilities from these options can have substantial upward bias.

Section 2 briefly explains how implied volatility is typically calculated and how measurement errors in the inputs are propagated to implied volatility. Section 3 describes what the sources of error are, and just how sensitive implied volatility can be to these errors. Section 4 shows how to form efficient estimates of implied volatilities. Section 5, shows how measurement errors in prices can lead to biases in implied volatility. Section 6 concludes.

2 Precision of Implied Volatility Calculations

Although implied volatility has to be computed numerically, this section derives analytical results for errors in implied volatility conditional on errors in the underlying prices.[4] To establish notation, I first restate the well-known Black–Scholes (1973) and Merton (1973) option pricing formula.

2.1 The Black–Scholes option pricing formula

Black and Scholes (1973) derive the value of European options under the assumption that the price of the underlying asset follows a geometric Brownian motion, that agents can continuously trade any fraction of a non-dividend-paying asset without transaction costs, and that agents can borrow and lend at a constant riskless interest rate. Merton (1973) extends the Black–Scholes model to include proportional dividends on the underlying asset. This model shows that the price, C, of a European call option with strike price K and t years to expiration is

$$C(S,K,r,\delta,t,\sigma) = Se^{-\delta t}\Phi(z) - Ke^{-rt}\Phi(z-\sigma\sqrt{t}), \tag{1}$$

where

$$z = \frac{\ln\left(Se^{-\delta t}/(Ke^{-rt})\right)}{\sigma\sqrt{t}} + \frac{1}{2}\sigma\sqrt{t}. \tag{2}$$

S is the price of the underlying asset, such as a stock or a stock index; r is the riskless interest rate; δ is the dividend rate on the underlying asset; $\Phi(\cdot)$ is the

[4] The standard numerical solution for implied volatility uses the Newton–Raphson procedure described in Manaster and Koehler (1982) to find the implied volatility that equates theoretical and observed option prices.

standard normal cumulative density function; and σ is the standard deviation of the annual log return of the underlying asset. By put call parity, the price of a European put, P, is $P = C - Se^{-\delta t} + Ke^{-rt}$, which can be expanded to yield formulas similar to equations (1) and (2).

2.2 Error sensitivity of implied volatility calculations

If the Black–Scholes model is correct and I observe prices without errors, there is a unique implied volatility $\tilde{\sigma}$ that equates the theoretical option price at the observed underlying prices $\tilde{S}$, $\tilde{K}$, $\tilde{r}$, $\tilde{\delta}$, and $\tilde{t}$ to the observed option price $\tilde{C}$,

$$C(\tilde{S}, \tilde{K}, \tilde{r}, \tilde{\delta}, \tilde{t}, \tilde{\sigma}) = \tilde{C}. \tag{3}$$

The problem is that observed prices typically are contaminated by measurement errors. If the observed prices contain measurement errors, so that $\tilde{C} = C + dC$ for example, then the observed implied volatility, $\tilde{\sigma}$, is equal to the true volatility, σ, plus an error, $d\sigma$,

$$C(S + dS, K + dK, r + dr, \delta + d\delta, t + dt, \sigma + d\sigma) = C + dC, \tag{4}$$

where dC, dS, dK, dr, $d\delta$, and dt are measurement errors in prices.

For small measurement errors dx_i, totally differentiating the definition of implied volatility in equation (3) allows me to find the volatility error

$$\begin{aligned} d\sigma &= \left(\frac{\partial C}{\partial \sigma}\right)^{-1} \left\{ dC - \frac{\partial C}{\partial S}\, dS - \frac{\partial C}{\partial K}\, dK - \frac{\partial C}{\partial r}\, dr - \frac{\partial C}{\partial \delta}\, d\delta - \frac{\partial C}{\partial t}\, dt \right\} \\ &= \frac{\partial \sigma}{\partial \boldsymbol{x}'}\, d\boldsymbol{x}, \end{aligned} \tag{5}$$

where $\boldsymbol{x} = (C, S, K, r, \delta, t)'$ is the column vector of underlying prices. Given the monotonicity of the call price in volatility, $(\partial C/\partial \sigma)^{-1} = \partial \sigma / \partial C$. Equation (5) also defines the remaining partial derivatives $\partial \sigma / \partial x_i = -(\partial C/\partial x_i)(\partial C/\partial \sigma)^{-1}$. The relevant partial derivatives of the option price are

$$\frac{\partial C}{\partial \sigma} = Se^{-\delta t}\sqrt{t}\, \phi(z) > 0 \tag{6}$$

$$\frac{\partial C}{\partial S} = e^{-\delta t}\Phi(z) > 0 \tag{7}$$

$$\frac{\partial C}{\partial K} = -e^{-rt}\Phi(z - \sigma\sqrt{t}) < 0 \tag{8}$$

$$\frac{\partial C}{\partial r} = tKe^{-rt}\Phi(z - \sigma\sqrt{t}) > 0 \tag{9}$$

$$\frac{\partial C}{\partial \delta} = -tSe^{-\delta t}\Phi(z) < 0 \tag{10}$$

$$\frac{\partial C}{\partial t} = -\delta Se^{-\delta t}\Phi(z) + rKe^{-rt}\Phi(z - \sigma\sqrt{t}) + \frac{\sigma Se^{-\delta t}}{2\sqrt{t}}\, \phi(z) \gtrless 0, \tag{11}$$

where $\phi(\cdot)$ is the standard normal probability density function.

Equations (5) and (6) characterize the main problem. For options that are deep in or out of the money, $|z|$ is large and $\phi(z)$ is close to zero. Hence, $\partial\sigma/\partial C = (\partial C/\partial\sigma)^{-1}$ is very large for options far from the money and small errors in any of the inputs produce large errors in the implied volatility.

Equations (5)–(11) allow us to find the total error in implied volatility conditional on knowing the individual component errors. By definition, the component errors are unknown however, and we have to confine ourselves to computing the standard deviation or variance of $d\sigma$. If all of the component errors are measurement errors, they have zero means and are mutually independent. The mutual independence implies that the covariances of the component errors vanish. Therefore,

$$V(d\sigma) = \left(\frac{\partial\sigma}{\partial C}\right)^2 \left\{ V(dC) + \left(\frac{\partial C}{\partial S}\right)^2 V(dS) + \left(\frac{\partial C}{\partial K}\right)^2 V(dK) + \left(\frac{\partial C}{\partial r}\right)^2 V(dr) + \left(\frac{\partial C}{\partial \delta}\right)^2 V(d\delta) + \left(\frac{\partial C}{\partial t}\right)^2 V(dt) \right\}. \tag{12}$$

This variance is easily computed from the partial derivatives in equations (6)–(11) and the variances of the component errors. For future convenience, it is also possible to rewrite equation (12) in vector notation as

$$V(d\sigma) = \frac{\partial\sigma}{\partial \boldsymbol{x}'} \boldsymbol{\Lambda} \frac{\partial\sigma}{\partial \boldsymbol{x}}, \tag{12a$'$}$$

where

$$\boldsymbol{\Lambda} = E[d\boldsymbol{x}\, d\boldsymbol{x}']. \tag{12b$'$}$$

Due to the mutual independence of the measurement errors dx_i, their covariance matrix, $\boldsymbol{\Lambda}$, is a diagonal matrix with the variances of the underlying price errors on the diagonal, $\text{diag}(\boldsymbol{\Lambda}) = (V(dC), V(dS), \ldots, V(dt))'$.[5]

For implied volatility from a European put option, the variance can be found by substituting the measurement error of the put option price, dP, for dC and using the appropriate partial derivatives of the put price instead of the call price derivatives. For American put options, this is more difficult since the possibility of optimal early exercise prevents analytical pricing formulas and derivatives.

3 Error Sources and Magnitudes

A considerable amount of work has been done to improve the precision of implied volatility estimates and investigate their behavior.[6] Only scant attention has been paid to the *level* of accuracy of these estimates. This section uses

[5] Equation (12) can be viewed as an application of the "delta method." The delta method is used to find the standard error of an estimate that is a nonlinear function of other estimates for which standard errors are available. See Gallant (1997), for example.

[6] Examples are Latané and Rendleman (1976), Chiras and Manaster (1978), Schmalensee and Trippi (1978), Patell and Wolfson (1979), Beckers (1981), Day and Lewis (1988), Harvey and Whaley (1991), or Harvey and Whaley (1992).

TABLE 1
Assumptions for Price Levels and Bid–Ask Spreads

Underlying Asset	Levels				Bid–Ask Spreads				
	S	σ	r	δ	S	K	r	δ	t
Stock	25	25%	5%	2.5%	.25	0	.02%	.02%	8 hours
Stock Index	500	15%	5%	2.5%	2.00	0	.02%	.02%	8 hours

Assumptions for price levels, volatility levels, and bid–ask spreads used in the paper.

equation (12) to construct confidence intervals for implied volatility. To accomplish this, we need to compute the partial derivatives $\partial\sigma/\partial x_i$ and specify the measurement error variances in $\boldsymbol{\Lambda}$.

The partial derivatives of the option price depend on price levels. I make assumptions about price levels that are indicative of realistic prices. The results in this paper are not sensitive to moderate variations in the price levels. I choose a stock price of \$25. During the period 1980–2000, the month-end closing prices of New York Stock Exchange (NYSE) stocks recorded on the Center for Research in Securities Prices (CRSP) database had mean and median values of \$20.56 and \$34.35, respectively. I use a stock index level of 500, which is representative of S&P 100 and S&P 500 levels in the early 1990s. Major U.S. stock indices have annual volatility in the neighborhood of $\sigma = 15\%$. Large corporations, whose shares support stock options, have annual equity return volatility around $\sigma = 25\%$. I assume that the interest rate $r = 5\%$. Finally, I assume that the dividend yield $\delta = 2.5\%$. During 1980–2000, the dividend yield on the CRSP value-weighted index averaged 3.9% but has fallen steadily. Although these rates vary over time and across firms, the results of this paper are not sensitive to plausible variations in interest or dividend rates. Table 1 summarizes these price levels for the underlying variables.

If the Black–Scholes model is true, measurement errors in the inputs, $\boldsymbol{x} = (C, S, K, r, \delta, t)'$, are the only possible source of error in implied volatility. Apart from clerical errors, the strike prices of ordinary options are measured perfectly. The other five inputs, however, are all subject to systematic measurement errors. For each of the market-determined inputs, C, S, r, and δ, there conceptually exists a unique market-clearing price at any time. These are the prices we should use to compute implied volatility. Unfortunately, these "true" prices are typically obscured by market structures like minimum tick sizes and bid–ask spreads. In addition, non-synchronous prices can introduce errors even if we observe the "true" prices at certain times. In practice, there are other measurement difficulties in addition to finite quote precision, bid–ask spreads, and non-synchronous quotes. Due to these omitted errors, the reported implied volatility errors are almost certainly conservative estimates.[7]

[7] Transaction costs beyond bid–ask spreads can add noise to price observations. Phillips and Smith (1980) and Figlewski (1988) show that brokerage fees and other administrative costs may be substantial relative to transaction value and that the fees vary widely by customer type and transaction size. Due to the difficulty of assessing the applicable transaction costs, I ignore them.

3.1 Finite quote precision

Unlike the idealized Black–Scholes market, the actual markets for options, underlying securities and interest rates are not characterized by infinitely small price increments. Tick sizes are part of the contract specifications for most exchange-traded options and other securities. Tick sizes constitute a source of measurement error, since the true price is not constrained to move in discrete steps. Even though prices are recorded in discrete increments, traders can achieve much more finely gradated average prices by splitting trades across adjacent prices.[8]

If the true price is rounded to quote precision, then the true price is within plus or minus one half tick size of the quoted price. If the typical price variation is larger than the tick size, then the observed price equals the true price plus a uniformly distributed error term with zero mean and a range of plus or minus one half of the tick size. To keep matters simple, the remainder of the paper only discusses normally distributed measurement errors based on bid–ask spreads.[9] The main results are qualitatively similar if finite quote precision is the only source of measurement error.

3.2 Bid–Ask spreads

In practice, we frequently observe "bid" and "ask" quotes. We typically average these quotes to obtain a single price. In most markets, the bid–ask spread is bigger than the tick size and obscures the true price beyond the uncertainty caused by the finite quote precision.

Ho and Stoll (1981) show that, depending on the inventory of the dealer, the true price is closer to the bid or the ask price at any particular point in time. Over time, however, one would expect the bid–ask spread to be centered over the true price. Nonetheless, the true price does not have to always lie between the bid and ask quotes. Vijh (1990) documents that just over 4% of all option trades in his sample of Chicago Board Options Exchange (CBOE) options occur outside the last quoted spread. In Vijh's sample, transactions outside the spread are evenly split between prices below the bid and above the ask. The transactions in the underlying NYSE stocks are similarly distributed in Vijh's sample.

---

[8] Prior to June 1997, stocks on the NYSE generally traded in increments of $1/8$ of a dollar per share. In June 1997, the tick size declined to $1/16$ of a dollar. On January 29, 2001, the tick size declined to 1¢. Large equity indices, such as the S&P 100, are quoted in increments of 0.01, even though they are not directly traded. The Chicago Board Options Exchange used to quote equity and equity index options in $1/8$ of a dollar per share for option prices above \$3 and in $1/16$ of a dollar per share for option prices below \$3. As part of decimalization, these tick sizes were lowered to 10¢ and 5¢, respectively. The discount on U.S. Treasury bills is quoted in cents per \$100 of face value, which is roughly equivalent to quoting interest rates up to one basis point. Dividends on common stocks are announced in cents, although fractions of a cent as small as $1/8$ are not uncommon.

[9] Olds (1952) shows how to compute the joint density of a convolution of uniform densities with different supports. While this is a straightforward computation, it makes the exposition in this paper cumbersome.

There is no obvious distribution for the errors stemming from bid–ask quotes. For convenience, I postulate that the errors have a normal distribution with a zero mean and a standard deviation proportional to the magnitude of the bid–ask spread. The data presented in Vijh (1990) can be broadly characterized as stemming from a normal distribution with a zero mean and a standard deviation of one quarter of the bid–ask spread.[10]

Table 1 also shows my bid–ask spread assumptions for the underlying prices. Vijh (1990) reports that roughly half of all NYSE stocks with CBOE-traded options are quoted with a bid–ask spread of $1/4$ in his sample. I assume that K is observed without error. The typical bid–ask spread of U.S. Treasury bills is 0.02% discount, but widens to 0.1% discount, or more, for bills with less than about thirty days to maturity.

There are no quoted bid and ask prices for the dividend rates on individual stocks or stock indices. Although dividends are announced ahead of their payment date, there is considerable uncertainty about the level of dividends ahead of the announcement. Although this is almost certainly an understatement of the error in dividend rates, I assume that dividend rates are measured with an error equivalent to the bid–ask spread of the interest rates. I also assign a "spread" to time to expiration, t. Time to expiration is typically measured in whole days, regardless of when quotes are collected, or when the option expires.

Based on stylized facts, I assume that the bid–ask spreads for in-the-money options are higher than the spreads for out-of-the-money options. The assumed spreads fall in steps of $1/16$. Figure 1 shows the spreads for call options. For stock options, I assume that the bid–ask spread varies from $3/8$ for in-the-money-options to $1/8$ for out-of-the-money options. For the maturities I consider here, the in-the-money stock options at $K/S = 0.8$ have a theoretical price of $C \approx \$5$. For index options, I assume that the bid–ask spread varies from $3/4$ for in-the-money-options to $1/8$ for out-of-the-money options. The in-the-money index options at $K/S = 0.9$ have a theoretical price of $C \approx \$50$. Empirically, options appear to have higher bid–ask spreads than what I assume here.[11]

3.3 Non-synchronous prices

In addition to the bid–ask spreads on the traded securities, table 1 also shows a "bid–ask spread" for the non-traded index level. For the index level, a large error typically comes from using closing prices for the options and index that are measured fifteen minutes apart. This time difference can be reduced by using transaction prices, but such careful alignment of prices is not typical. Even when option prices and published index levels are perfectly synchronous, large indexes often contain stale component prices. (See Scholes and Williams 1977.)

[10] When the standard deviation of the errors is equal to one quarter of the bid–ask spread, roughly 2 percent of all true prices are above the ask and below the bid quotes, respectively.

[11] Phillips and Smith (1980) report that there are many options with values below fifty cents for which the bid–ask spread exceeds 100% of the option value.

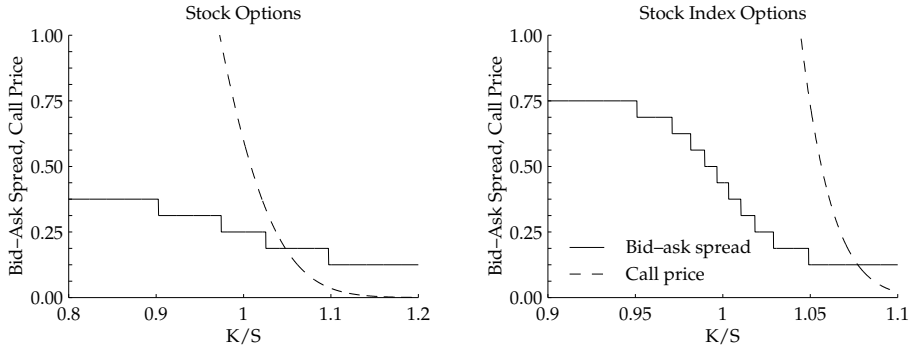

FIGURE 1: BID–ASK SPREADS FOR CALL OPTION PRICES

The solid lines in the figure show the assumed bid–ask spreads for call options on a stock or a stock index. For comparison, the dashed line shows the Black–Scholes (1973) call option prices for options with 20 days to maturity. For the stock options in the left panel, $S = 25$, $r = 5\%$, $\delta = 2.5\%$, and $\sigma = 25\%$. For the stock index options in the right panel, $S = 500$, $r = 5\%$, $\delta = 2.5\%$, and $\sigma = 15\%$.

If returns follow an i.i.d. random walk in calendar time with an annual volatility of 15%, then the standard deviation of half-hour returns is 0.113%. Based on this volatility, I construct a "bid–ask spread" for the assumed index level of 500, so that the volatility is one quarter of the spread. This constructed "bid–ask spread" is equal to 2.26. I round this value down to 2.0, which corresponds to a 23 minute alignment error.

3.4 Effects on implied volatility

Figure 2 shows 95% confidence intervals for implied volatility estimates based on equation (12) and input errors with variances equal to one quarter of the bid–ask spreads summarized in table 1 and figure 1. The left panel shows confidence intervals for implied volatility from a call option on a stock; the right panel shows confidence intervals for implied volatility from a call option on a stock index. The plotted range of moneyness may seem extreme. In practice, however, listed options cover a larger range of moneyness than is shown in the figure.[12] The small jumps in the confidence intervals, most noticeable in the left panel, stem from the discrete changes in the bid–ask spread for the option price.

For at-the-money or slightly out-of-the-money stock options with ninety days until maturity, the 95% confidence interval is just over ± 2.8 percentage points wide. From there, the interval widens dramatically as the option moves in or out of the money or approaches maturity. With twenty days until maturity, the 95% confidence intervals are at least ± 5.9 percentage points wide. Stock options with 20 days to maturity that are more than 5% in or out of the money contain

[12] For stock options, CBOE rules space strike prices every $2\frac{1}{2}$ points when the strike price is between \$5 and \$25, every 5 points when the strike price is between \$25 and \$200, and every 10 points when the strike price is above \$200. According to these rules, the plotted range only covers 3 CBOE-listed stock options at $K = \{22.5, 25, 30\}$. For S&P 100 and S&P 500 index options, CBOE rules space strike prices every 5 points for near expiration months. Accordingly, the plotted range for stock index options covers 21 strike prices at $K = \{450, 460, \ldots, 550\}$.

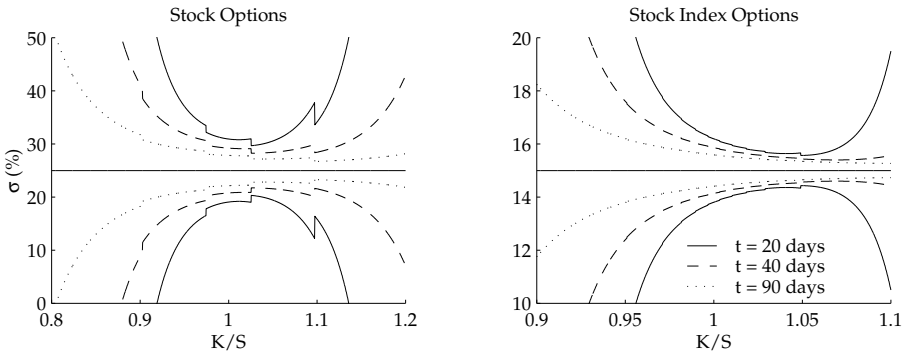

FIGURE 2: 95% CONFIDENCE INTERVALS FOR IMPLIED VOLATILITY

The figure shows 95% confidence intervals for the implied volatility from individual call options with three different times to maturity, t. The left panel shows results for call options on a stock. The right panel shows results for call options on a stock index. The confidence intervals were constructed based on mutually independent, normally distributed measurement errors with standard deviations of one quarter of the bid–ask spreads shown in table 1 and figure 1.

virtually no information about volatility. Based on the assumed measurement errors, stock index options deliver more precise implied volatility estimates. (Note the different scales for the two panels.) Nonetheless, even for index options, the precision of implied volatility is low for options near expiration or away from the money. Jackwerth and Rubinstein (1996) observe these features in their S&P 500 options data: They find large volatility of implied volatilities for options away from the money and rising volatility of implied volatility as the options move farther from the money.

The widths of the confidence intervals in figure 2 depend on the level of volatility. All else equal, higher volatility implies that an option with $K/S \neq 1$ has a higher probability of reaching $K/S = 1$ by maturity. In this sense, higher volatility makes these options closer to the money. As a result, the confidence bands have less curvature at higher volatility. At low volatility, nearly all options are far from the money and contain little information about implied volatility. For options at $K/S = 1$, however, the confidence intervals are nearly independent of the level of volatility. For example, the width of the 95% confidence intervals for stock index options with 20 days to maturity is ±1.250 percentage points at $\sigma = 5\%$, ±1.193 at $\sigma = 15\%$, ±1.193 at $\sigma = 25\%$, ±1.201 at $\sigma = 35\%$, and ±1.221 percentage points at $\sigma = 45\%$.

Figure 2 demonstrates the limitations of two widely used rules of thumb. First, the figure shows that options near expiration do not provide more precise implied volatility estimates than similar options farther from expiration. In general, there is a trade-off between the lower bid–ask spread for options near expiration and the inherently more precise volatility estimates from longer-dated options. I assume that the at-the-money index options in figure 2 have a bid–ask spread of $^7/_{16}$. The 95% confidence intervals for the at-the-money index options are ±1.193 percentage points with 20 days to maturity and ±0.589 percentage points with 90 days to maturity. To fall to the level of precision of the 20-day at-the-

money index option, the 90-day option would have to have a bid–ask spread of 2-1/8, all else equal. That is almost five times the spread I assume.

Second, the figure shows that it is not generally true that at-the-money options provide the most precise estimates of implied volatility when inputs besides the option price are measured with error. It is important to recognize that the asymmetry in the precision of implied volatility estimates is *not* a result of the pattern in bid–ask spreads shown in figure 1. For in-the-money options, errors in the underlying asset price contribute most of the uncertainty in implied volatility. The larger confidence intervals for implied volatility from in-the-money options stem from the increasing importance of errors from the underlying asset price. For out-of-the-money options, errors in the option price contribute most of the uncertainty in implied volatility.

Based on the assumptions in table 1 and figure 1, variance decompositions of the implied volatility errors show that, for most options, price errors from options and underlying assets dominate the total error variance. Only for options with maturities of a year or more do errors from interest and dividend rates contribute materially to the overall error variance.[13] For the magnitudes shown in table 1 and figure 1, errors in the time to expiration never make substantive contributions to the overall error in implied volatility.

It is interesting to compare the confidence intervals in figure 2 to confidence intervals for historical volatility estimates. If returns follow a random walk in calendar time, a sample of 50 stock returns with annual standard deviation $\sigma = 25\%$ produces 95% confidence intervals of ± 5 percentage points.[14] A sample of 200 stock index returns with annual standard deviation $\sigma = 15\%$ produces 95% confidence intervals of ± 1.5 percentage points.

The same analysis used to derive confidence intervals for implied volatilities is useful in deriving confidence intervals for option prices and option price derivatives based on implied volatility estimates. These confidence intervals include the direct contributions of errors in the underlying prices and the indirect contribution of these errors through errors in implied volatility. This analysis reveals that, for options that are five or ten percent out of the money and have less than a month to expiration, 95% confidence intervals for prices quickly reach zero and twice the true price. Even for at-the-money calls with 20 days to maturity, however, confidence intervals include prices between 75% and 125% of the true price. For an option near expiration, the errors in the underlying prices and implied volatility also result in very noisy estimates of the option's delta.

[13] For long-term options, however, dividend rates may also be far less certain than is implied by the numbers in table 1. When the dividend rate has a "bid–ask spread" of 0.25%, the dividend rate accounts for less than 3% of the overall variability of the errors in the implied volatilities for options with 90 days or less to maturity. For stock index options with 3 years to maturity, however, the same errors in the dividend rate account for 75% of the variance of implied volatility errors. The corresponding variance share for stock options is 15%.

[14] For a sample of N i.i.d. returns, the annualized sample standard deviation of returns sampled h times per year, $\sqrt{h}\,\sigma_h$, has variance equal to $\sigma^2/(2N)$ regardless of sampling frequency h.

4 Efficient Implied Volatility Estimation

The previous section discussed the precision of an implied volatility estimate from a single option when prices are observed with measurement errors. In practice, we frequently observe a vector of implied volatilities, $\tilde{\boldsymbol{\sigma}} = \{\tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_n\}'$, from several options on the same underlying asset. This section shows how to exploit information in volatility estimates from multiple options to form efficient estimates of implied volatility.

The main issues are that the n implied volatilities, $\tilde{\boldsymbol{\sigma}}$, contain n errors, $d\boldsymbol{\sigma}$, that are heteroskedastic and correlated with each other. The heteroskedasticity stems directly from the fact that some implied volatilities are more sensitive to errors in underlying prices than others. The correlations of the errors in implied volatility arise because we typically try to synchronize all prices. If the observations of S, r, δ, and t are identical across all options in a cross-section, so are their measurement errors dS, dr, $d\delta$, and dt.[15] These common errors induce correlations in the errors in implied volatility, $d\sigma_i$.

With mutually independent measurement errors on perfectly synchronized prices, the covariance matrix for implied volatilities from a cross-section of options is a straightforward generalization of the scalar case. Now $V(d\boldsymbol{\sigma}) = E[d\boldsymbol{\sigma}\, d\boldsymbol{\sigma}'] = \boldsymbol{\Sigma}$ is a matrix and

$$\boldsymbol{\Sigma} = \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{x}'} \boldsymbol{\Lambda} \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{x}}. \tag{13}$$

$\boldsymbol{\Lambda} = E[d\boldsymbol{x} d\boldsymbol{x}']$ is still the covariance matrix of the underlying errors. But now $\mathrm{diag}(\boldsymbol{\Lambda}) = (V(dC_1), \ldots, V(dC_n), V(dS), \ldots, V(dt))'$ because there are n option prices in $\boldsymbol{x} = (C_1, \ldots, C_n, S, \ldots, t)'$. $\partial \boldsymbol{\sigma} / \partial \boldsymbol{x}'$ is the Jacobian matrix of implied volatility derivatives, $(\partial \sigma_i / \partial C_i)(\partial C_i / \partial x_j)$.

If implied volatilities from European put options with prices P_l are part of the cross-section, their variances and covariances can be found from equation (13) as well. This merely involves an expansion of the underlying price vector $\boldsymbol{x}$ to include the put prices, as well as an expansion of the covariance matrix $\boldsymbol{\Lambda}$ to include their measurement error variances.

Following historical developments, and to build intuition, I now present a sequence of implied volatility estimators in order of increasing complexity.

4.1 Estimation across strike prices

If we restrict all options with a common maturity date to have the same implied volatility, regardless of strike price, there exists a weighted average $\hat{\sigma} = \boldsymbol{\omega}' \tilde{\boldsymbol{\sigma}}$ of the observed implied volatilities that is minimum variance among all such unbiased weighted averages. By standard least squares arguments, this efficient

[15] Even if not all the measurements are simultaneous, it is common practice to use the same set of measurements on S, r, δ, and t for options whose prices were observed at slightly different times. If strike prices are observed without error, the errors resulting from this procedure have the same covariance matrix as errors from synchronous prices.

estimator is the GLS weighted average obtained by regressing implied volatilities on a constant,

$$\tilde{\boldsymbol{\sigma}} = \boldsymbol{\iota}\sigma + d\boldsymbol{\sigma}, \tag{14}$$

where $\boldsymbol{\iota}$ is an n-vector of ones. The GLS estimate is

$$\hat{\sigma} = \boldsymbol{\omega}'\tilde{\boldsymbol{\sigma}} = (\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\sigma}}. \tag{15}$$

In practice, implied volatility and the weights have to be found by iteration because the partial derivatives, $\partial\boldsymbol{\sigma}/\partial\boldsymbol{x}$, in equation (13) depend on the level of volatility. All weighting methods that rely on option sensitivities share this feature. Since the iterations usually converge quickly and the partial derivatives are easy to compute, the iterations are not burdensome.

Using the covariances in equation (13) and the assumptions in table 1 and figure 1, one can work out theoretical confidence intervals for different weighted averages of implied volatilities. These calculations reveal three general conclusions: (i) an equally weighted average of implied volatilities produces extremely noisy volatility estimates; (ii) the Chiras–Manaster (1978) method performs very poorly when deep-out-of-the-money options are included because it assigns large weights to these options; and (iii) the Chiras–Manaster (1978) and Latané–Rendleman (1976) methods perform relatively poorly when measurement errors other than those for option prices make material contributions to errors in implied volatility.[16]

The weights derived in equation (15) are by construction optimal for any given covariance matrix $\boldsymbol{\Sigma}$. Unfortunately, we don't know or observe the true covariance matrix. Therefore, the optimal weights involve an estimate of the true covariance matrix, and are themselves subject to estimation error.

Equation (15) is homogeneous of degree zero in $\boldsymbol{\Sigma}$, and the GLS weights are independent of the scale of the covariance matrix. Unfortunately, when the relative error is not the same for all variance and covariance estimates, the weights, $\boldsymbol{\omega}$, are suboptimal. This fact notwithstanding, there are no known superior weights.

In simulations, the GLS weighting method performs very well, even when the individual measurement error variances used in the construction of the covariance matrix are between 0.01 and 100 times the true variances. The weights are not sensitive to substantial variation in measurement error variances. Only when the error variances used in the construction of the covariance matrix approach zero, do the weights change materially. In this extreme case, the weight vector tries to exploit the covariances by using very large positive and negative weights.

[16] The Latané–Rendleman (1976) weights are similar to the GLS weights when option price errors are the most important errors. Unlike the Latané–Rendleman weights, some of the GLS weights may be negative. In principle, negative weights may generate a negative estimate of implied volatility. This is unlikely, however, since the weight ω_i for $\tilde{\sigma}_i$ generally declines with the variance of $d\sigma_i$. Nonnegative weights can be found through standard iterative quadratic programming algorithms but prevent analytical solutions. (See chapter 14 in Luenberger 1984, for example.)

These large weights can produce noisy volatility estimates when the true errors are large.

In this paper, my strategy for developing feasible GLS estimators for implied volatility is first to make plausible assumptions about bid–ask spreads and second to make plausible assumptions about how bid–ask spreads translate in measurement error variances, $\text{diag}(\boldsymbol{\Lambda}) = \boldsymbol{\lambda}$. For n options with synchronous price measurements, $\boldsymbol{\lambda}$ contains $n + 4$ variances if strike prices are observed without error. We then can generate the full $n \times n$ GLS covariance matrix $\boldsymbol{\Sigma}$ from the $n+4$ variances in $\boldsymbol{\lambda}$ and the known option pricing model derivatives $\partial\boldsymbol{\sigma}/\partial\boldsymbol{x}'$ according to equation (13). The experiments described above verify that the GLS estimators are not sensitive to changes in $\boldsymbol{\lambda}$. Based on the framework developed here, however, there are at least two other feasible GLS estimators. If we observe bid–ask spreads, we can dispense with the first assumptions and use the observed bid–ask spreads. Alternatively, we can try to estimate $\boldsymbol{\lambda}$ without either assumption about how the variances were generated. If we are willing to place restrictions on how $\boldsymbol{\lambda}$ depends on option characteristics or how $\boldsymbol{\lambda}$ changes over time, we can estimate $\boldsymbol{\lambda}$ as part of the GLS procedure. Even for moderate n, this is much easier than estimating $\boldsymbol{\Sigma}$ directly.

The simple regression in equation (14) generalizes easily. We can specify a volatility smile for options with different strike prices but a common expiration date. For example, we can fit a polynomial in moneyness to implied volatilities for a single maturity. A quadratic specification of a volatility smile is

$$\begin{aligned}\tilde{\boldsymbol{\sigma}} &= \boldsymbol{\iota}\beta_0 + (\boldsymbol{K}/S - \boldsymbol{\iota})\beta_1 + (\boldsymbol{K}/S - \boldsymbol{\iota})^2\beta_2 + d\boldsymbol{\sigma} \\ &= \boldsymbol{X}\boldsymbol{\beta} + d\boldsymbol{\sigma}, \end{aligned} \tag{16}$$

where $\boldsymbol{K} = (K_1, K_2, \ldots, K_n)'$ is the column vector of strike prices, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, and $\boldsymbol{X} = \left[\boldsymbol{\iota}, (\boldsymbol{K}/S - \boldsymbol{\iota}), (\boldsymbol{K}/S - \boldsymbol{\iota})^2\right]$. In this case, the efficient fitted volatilities are based on the GLS estimate of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\sigma}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\sigma}}, \tag{17}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix in equation (13). Section 5 contains extensive simulation results for implied volatility estimates from regressions like equation (16). I defer the simulation results to section 5, so I can use the simulations to simultaneously investigate biases in implied volatilities.[17]

[17] The regressors in equation (16) contain measurement error through the error in the underlying asset price. For synchronized prices, dS is a single measurement error affecting all observations. In a particular regression, this error biases the intercept. Since $E[dS] = 0$ and $\text{Var}(dS)$ is small, however, the average bias is very small. The main effect of the measurement error dS is to introduce additional noise in the estimate of the intercept. The simulations in section 5 confirm this analysis.

4.2 Estimation across strike prices and maturities

Generally, we are unwilling to assume that volatility is constant over time and standard practice forms separate implied volatility estimates for each maturity. Even in this case, one can theoretically increase the precision of volatility estimates by estimating implied volatilities simultaneously if measurement errors are correlated across observed implied volatilities. For perfectly synchronous prices, the measurement errors for all of the underlying variables—except for the option prices—are identical.

If a sample contains options with q times to maturity, $t_1, \ldots, t_q$, then the efficient implied volatility estimate that restricts volatilities to be the same for all options with identical expiration dates consists of q volatilities, $\hat{\boldsymbol{\sigma}} = \{\hat{\sigma}_1, \ldots, \hat{\sigma}_q\}'$. If the sample contains n_i options for each maturity t_i, we can order them by maturity and stack their observed implied volatilities into a vector $\tilde{\boldsymbol{\sigma}}$. The resulting system of equations stacks q versions of equation (14) as

$$\tilde{\boldsymbol{\sigma}} = \begin{bmatrix} \boldsymbol{\iota}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \ddots & & \vdots \\ \vdots & & \ddots & \mathbf{0}_{n_{q-1}} \\ \mathbf{0}_{n_q} & \cdots & \mathbf{0}_{n_q} & \boldsymbol{\iota}_{n_q} \end{bmatrix} \begin{pmatrix} \sigma_1 \\ \vdots \\ \vdots \\ \sigma_q \end{pmatrix} + \begin{pmatrix} d\boldsymbol{\sigma}_1 \\ \vdots \\ \vdots \\ d\boldsymbol{\sigma}_q \end{pmatrix},$$

or

$$\tilde{\boldsymbol{\sigma}} = \boldsymbol{J}\boldsymbol{\sigma} + d\boldsymbol{\sigma}, \tag{18}$$

where $\boldsymbol{\iota}_{n_i}$ is an $n_i \times 1$ vector of ones, and $\mathbf{0}_{n_i}$ is an $n_i \times 1$ vector of zeros. Due to correlations across the measurement errors, $d\boldsymbol{\sigma}$, this is a seemingly unrelated regression equations (SURE) problem. The GLS–SURE estimator for the q implied volatilities is

$$\hat{\boldsymbol{\sigma}} = \boldsymbol{\Omega}'\tilde{\boldsymbol{\sigma}} = \left(\boldsymbol{J}'\boldsymbol{\Sigma}^{-1}\boldsymbol{J}\right)^{-1}\boldsymbol{J}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\sigma}}, \tag{19}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the errors $d\boldsymbol{\sigma}$. As before, $\boldsymbol{\Sigma}$ can be found by applying the logic of equation (13) to the errors $d\boldsymbol{\sigma}$.

This approach generalizes to cases where implied volatilities are a function of moneyness and time to expiration. Dumas, Fleming, and Whaley (1998) and Brandt and Wu (2001) contain recent examples of such volatility surfaces. For example, a quadratic volatility surface is obtained by specifying $\boldsymbol{X} = [\boldsymbol{\iota}, (\boldsymbol{K}/S - \boldsymbol{\iota}), \boldsymbol{t}, (\boldsymbol{K}/S - \boldsymbol{\iota})^2, \boldsymbol{t}^2, (\boldsymbol{K}/S - \boldsymbol{\iota}) \odot \boldsymbol{t}]$ in equation (17).[18]

Put–call parity restricts puts and calls with the same strike price and maturity to have the same implied volatility. We can impose put-call parity by fitting a single curve or surface to the implied volatilities from puts and calls with the same strike price and maturity in the data set. In this case, the implied volatility estimator can largely cancel the errors from the underlying asset prices

[18] The symbol $\odot$ stands for element-by-element multiplication of matrices. Exponents also apply on an element-by-element basis here.

and dividend rates.[19] Unfortunately, many authors preclude the cancellation of errors across puts and calls by using only the more liquid out-of-the-money options. Unless underlying asset prices and dividend rates are observed with high precision, this practice can result in a substantial loss of efficiency.

4.3 VIX Volatility Estimates

One popular volatility measure that averages implied volatilities from puts and calls and two separate maturities is the Market Volatility Index, VIX, of the CBOE. The VIX is an average of eight implied volatilities from near-the-money puts and calls for the S&P 100 option contract closest to expiration and the next-shortest maturity.[20] For each maturity and strike price, the VIX averages implied volatilities from puts and calls. Next, for each maturity, the VIX linearly interpolates the volatilities from the high and low strike prices to an at-the-money volatility. Finally, the VIX linearly interpolates these two volatilities to a single volatility for a 30-day maturity. For details on the construction of the VIX see Whaley (1993).[21]

The VIX employs several devices that mimic the GLS–SURE estimates in equation (19). Averaging volatilities across puts and calls with the same strike price and maturity partially cancels the effects of measurement error in the underlying asset price, since a measurement error in the underlying asset price has roughly opposite effects on put and call prices. Averaging these volatilities across strike prices near the money, further reduces noise. Finally, the noisy volatility estimates from options near expiration receive low weights when volatilities are interpolated to 30-day maturities.

Figure 3 directly compares the volatility weights employed by the VIX and GLS–SURE implied volatility estimators. For this comparison, the GLS–SURE volatility estimates in equation (19) use the same eight options as the VIX. To form a single implied volatility, I interpolate the two GLS–SURE volatility estimates to the one-month maturity employed by the VIX. This interpolation also can be thought of as a further transformation of the weights $\boldsymbol{\Omega}$ in equation (19). The figure shows the weights on all eight options in the cross-section for three different times to maturity and moneyness. The call options are identified as $C_{i,j}$. For $i = 1$, the call has the lower of the two strike prices; for $i = 2$, it has the higher of the two strike prices. For $j = 1$, the call has the shorter of the two times to maturity; for $j = 2$, it has the longer of the two times to maturity. Put options are similarly identified as $P_{i,j}$. Overall, the VIX and interpolated GLS–SURE weights are quite similar. The main difference is that the VIX ultimately focuses on two options, one call and one put, when these are at the money and have one month to maturity. In the same situation, the interpolated GLS–SURE estimator continues to place

[19] Put prices change with underlying asset prices as $(\partial P/\partial S) = (\partial C/\partial S)[1 - 1/\Phi(z)]$. For near-the-money options, $\Phi(z) \approx 0.5$, so that $(\partial P/\partial S) \approx -(\partial C/\partial S)$. A similar argument applies to option price changes due to the dividend rate, $(\partial P/\partial \delta) \approx -(\partial C/\partial \delta)$.

[20] The CBOE publishes a similar volatility index, VXN, for the NASDAQ–100 index.

[21] The S&P 100 options used in the construction of the VIX are American options. For simplicity, I assume that all options are European.

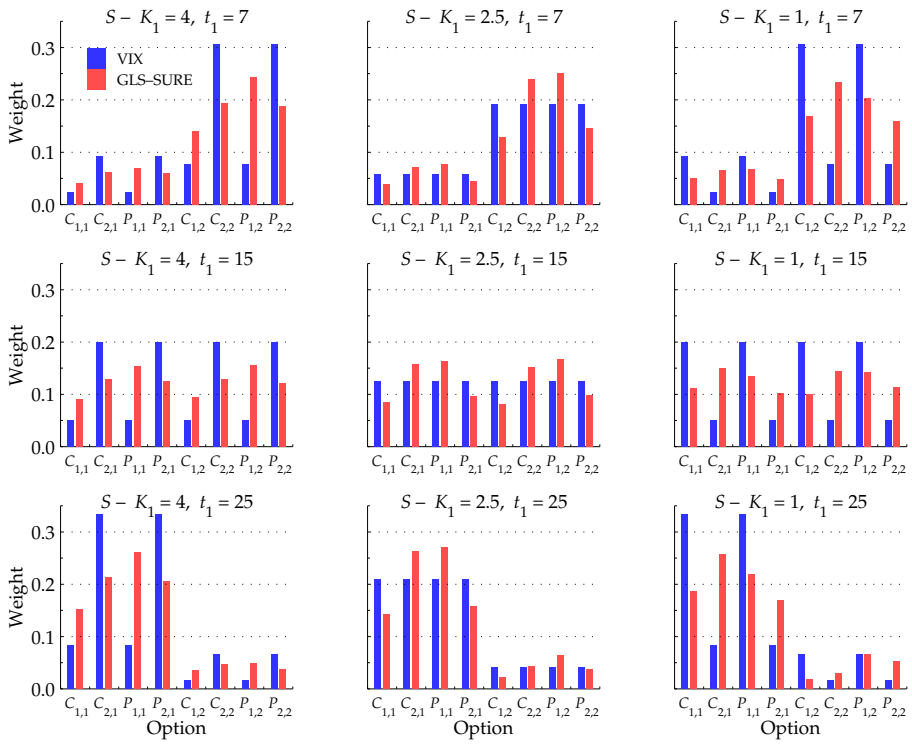

FIGURE 3: VIX AND INTERPOLATED GLS–SURE VOLATILITY WEIGHTS

The figure compares the interpolated GLS–SURE weights proposed here to the VIX weights suggested by Whaley (1993). Price levels are as given in table 1. Measurement errors for the option prices are assumed to be mutually independent and normally distributed with standard deviations of 1/4 of the bid–ask spreads in table 1. The other inputs are assumed to be measured with errors that are identical for all options but independent across inputs. These errors are also normally distributed with standard deviations of 1/4 of the bid–ask spreads in table 1 and figure 1.

All estimates use eight stock index options: four calls and four puts. The options are just in or just out of the money and expire on the nearest expiration date or the next one. The call options are identified as $C_{i,j}$. For $i = 1$, the call has the lower of the two strike prices; for $i = 2$, it has the higher of the two strike prices. For $j = 1$, the call has the shorter of the two times to maturity; for $j = 2$, it has the longer of the two times to maturity. Put options are similarly identified as $P_{i,j}$. The title in each panel lists moneyness for the lower strike price, $S - K_1$ and the shorter maturity in days, t_1. The higher strike price is $K_2 = K_1 + 5$; the longer maturity is $t_2 = t_1 + 30$ days.

positive weights on the two options away from the money but with one month to expiration. The options with two months to expiration receive no weight due to the interpolation, not because they contain no information about volatility. In many ways, the VIX implements equation (19) using an approximation of Σ.[22]

We can also use the GLS–SURE estimator of implied volatility to assess the precision of VIX volatility estimates. The left panel of figure 4 shows upper

[22] Figure 2 also suggests a possible improvement on the VIX estimator that is no harder to compute: Form the simple average of the implied volatilities from the four options with the same maturity, then interpolate these two average volatilities to one month.

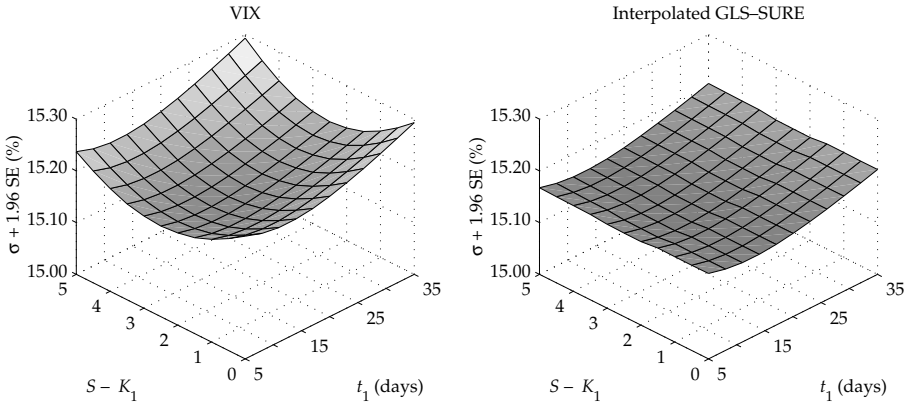

FIGURE 4: PRECISION OF VIX ESTIMATES, $\sigma = 15\%$

The left panel of the figure shows the upper 95% confidence bounds for the VIX implied volatility estimates. The lower confidence bounds are omitted for clarity. The upper and lower bounds are symmetric about the true volatility level of 15%. The right panel shows the upper 95% confidence bounds for interpolated efficient implied volatility estimates from two separate maturities using the same 8 options as the VIX. The figure shows confidence bounds for a range of shorter maturities, $t_1 = \{5, \ldots, 35\}$ calendar days, and lower strike prices, $K_1 = \{S, \ldots, S - 5\}$. The longer maturity is $t_2 = t_1 + 30$ days; the higher strike price is $K_2 = K_1 + 5$. All calculations are based on the price and measurement error assumptions in table 1 and figure 1.

95% confidence bounds for the VIX based on the price level and measurement error assumptions in table 1 and figure 1. The lower bounds are omitted for clarity but are symmetric about the assumed, constant volatility of 15%. Under the VIX choice of options, there are two maturities and two strike prices. The figure shows upper 95% confidence bounds for a range of short maturities, $t_1 = \{5, \ldots, 35\}$ calendar days, and lower strike prices, $K_1 = \{495, 496, \ldots, 500\}$. The longer maturity is $t_2 = t_1 + 30$ days and the higher strike price is $K_2 = K_1 + 5$.

Based on this analysis, one cannot rule out that VIX variations of 20 or 30 basis points are measurement error noise. For comparison, Fleming, Ostdiek, and Whaley (1995) report that the standard deviation of daily changes in the VIX averages 130 basis points between 1986 and 1992 when volatility changes during October 1987 and October 1989 are omitted. If the variance of volatility scales in calendar time, this translates to hourly volatility of volatility of 27 basis points. If the variance of volatility scales in trading time, this translates to hourly volatility of volatility of 46 basis points. With plausible measurement errors, the VIX seems to contain limited information about high-frequency variations in volatility.

The right panel of figure 4 shows the upper 95% confidence intervals for interpolated GLS–SURE estimates of implied volatility using the same eight options as the VIX. The two GLS–SURE estimates of implied volatility in $\hat{\boldsymbol{\sigma}}$ based on equation (19) are interpolated to a constant 30-day calendar maturity. The figure assumes that volatility is constant at 15%, but the GLS–SURE estimates obviously do not exploit this assumption. The GLS–SURE estimate of implied volatility has a slightly tighter confidence interval than the VIX, but the difference is modest. When either strike price is far from the money, then the interpolated GLS–SURE

estimates are slightly more efficient. This difference grows as the options move farther away from the money and may become important for volatility indexes for other assets with widely spaced strike prices.

Although exploiting the information across multiple maturities according to equation (19) reduces the noise in the implied volatility estimates, it is not commonly done. For the remainder of the paper, I focus on volatility estimates from a single maturity.

5 Bias in implied volatility

Thus far, I have assumed that errors in implied volatility have a mean of zero. This section shows that this assumption is not warranted. Nonetheless, the techniques designed to minimize noise in implied volatility estimates are also effective at reducing bias.

Bias in implied volatilities arises due to the systematic elimination of low implied volatilities. Implied volatilities are eliminated when option prices violate absence of arbitrage boundary conditions and there is no positive volatility that is consistent with the prices. This induces an upward bias because violations of the lower absence-of-arbitrage bounds is much more likely than violation of the upper absence-of-arbitrage bounds.[23] The bias is largest for deep-in- and deep-out-of-the-money options since their true prices are closest to the lower bounds and their implied volatilities are especially sensitive to price errors.[24]

The bias can arise in two distinct ways. First, researchers can mis-align prices to induce violations of the boundary conditions even though the actual prices satisfy absence of arbitrage. When a sample contains option prices that cannot be inverted into implied volatilities, the sample of implied volatilities is "censored." We observe the option characteristics but no associated implied volatility. Not carefully treating these observations causes bias. Closely matched transactions prices minimize this source of error.

A second, systematic, and unavoidable source of bias stems from the fact that dealers have strong incentives to post option prices that obey absence of arbitrage. For out-of-the-money options, the true price converges to zero as the option moves farther out of the money. The dealer, however, must set a non-negative bid price. The bid price typically converges to zero. The dealer also has to set an ask price that is an integer multiple of the tick size and above the bid price. Typically, the ask price converges to a tick or two. For deep-out-of-the-money options, the bid–ask average price is above the true price. Even transactions prices are biased upward, however, because no customer will sell

[23] For call options, the lower absence-of-arbitrage bound is $\max\{Se^{-\delta t} - Ke^{-rt}, 0\}$ and the upper absence-of-arbitrage bound is $Se^{-\delta t}$.

[24] Implied volatility is a nonlinear function of the underlying prices. Hence, implied volatility also suffers from a Jensen's inequality bias if the other variables are stochastic. The magnitude of the bias in implied volatility induced by option price errors should be similar to the magnitude of the bias in option prices induced by volatility errors. Simulation studies of Boyle and Ananthanarayanan (1977) and Butler and Schachter (1986) show that the latter bias is small. Jensen's inequality effects do not induce large biases in implied volatility.

at nonpositive prices, and all observed transactions for deep-out-of-the-money options occur at the ask price.

The upward bias in prices for deep-out-of-the-money options induces an upward bias in implied volatilities from these options. Yet, the resulting volatility smirk is statistically insignificant in the sense that the implied volatility is between the bid and ask volatilities. Because the bid volatility is zero, the interval of bid volatility to ask volatility includes all nonnegative volatility levels below the observed implied volatility for deep-out-of-the-money options. For deep-in-the-money options, the true price is also close to the lower absence of arbitrage boundary. Consequently, we would expect a similar upward bias in observed implied volatilities from deep-in-the-money options.

Although the true errors for option prices have a mean of zero, when a large negative error is drawn, the option price violates absence of arbitrage and is never posted. Posted option prices that satisfy absence of arbitrage form a "truncated" sample. This truncation only occurs at the lower boundary for options away from the money, because only there do true prices approach the absence of arbitrage boundaries. This truncation is not directly observable by researchers. The recorded prices obey absence of arbitrage and the option prices can be inverted into implied volatilities. Yet, option prices that would have led to low implied volatilities are systematically missing from the sample because they would have presented arbitrage opportunities. Even when all observed option prices can be inverted into implied volatilities, the implied volatilities may suffer from substantial truncation bias.

One can think of the posted option prices in one of two equivalent ways. Errors are drawn until they result in option prices that satisfy absence of arbitrage. Equivalently, errors for option prices are drawn from a distribution that is truncated so that the observed option prices obey absence of arbitrage. Either mechanism results in the same positive truncation bias in *observed* option prices.

I use simulations to show examples of the systematic biases of implied volatility estimates in the presence of measurement errors. The reported results focus on the unobserved price truncation and associated bias. I compute Black–Scholes call option prices for stock options and stock index options. The options use the price levels shown in table 1 and have 20 days to maturity. I simulate a cross-section of options with a range of strike prices. Then, I perturb all prices with normal errors that have zero means and standard deviations equal to one fourth the bid–ask spreads shown in table 1 and figure 1. This yields a different option price error for each strike price in the cross-section. Each of the other prices is constant within each cross-section, however, so that the underlying price error, for example, is the same within each cross-section. The option price errors are drawn from a normal distribution with zero mean and standard deviation equal to one quarter the bid–ask spread in figure 1 that is truncated so that the option price plus error obeys absence of arbitrage given the observed underlying prices. I then compute implied volatilities for the cross-section. I repeat these simulations 10,000 times. The repetitions are independent of each other.

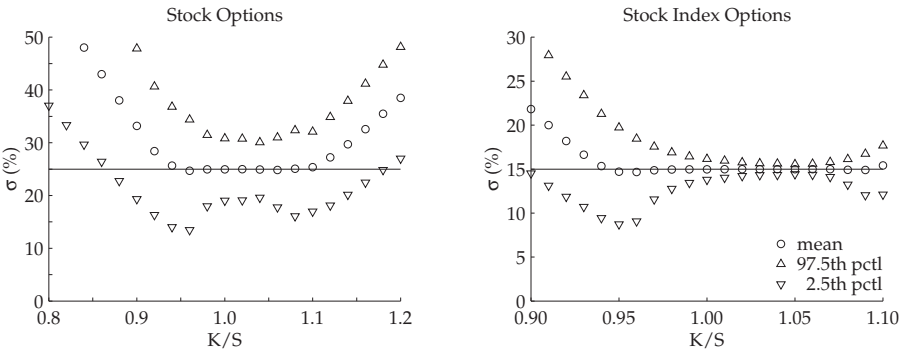

FIGURE 5: TRUNCATION BIAS IN IMPLIED VOLATILITY

The figure shows the means, 97.5th percentiles, and 2.5th percentiles for implied volatilities from simulated prices. The statistics are for options with a given moneyness. The left panel shows results for call options on a stock; the right panel shows results for call options on a stock index. For both cases, the samples are truncated: implied volatilities from options that violate boundary conditions are replaced. All options have 20 days to maturity.
The simulated prices obey all of the Black–Scholes assumptions, except that prices are perturbed by independent normal measurement errors. The measurement errors have zero means and standard deviations equal to $\frac{1}{4}$ of the bid–ask spreads in table 1 and figure 1. The price levels are also shown in table 1. Each statistic is based on a sample of 10,000 observations.

5.1 Bias by moneyness

Figure 5 plots statistics for simulated implied volatilities by moneyness. Circles mark the mean of implied volatility for a given moneyness. The empirical 95% confidence interval is marked with downward triangles for the 2.5th percentiles and upward triangles for the 97.5th percentiles of the simulated implied volatilities. The statistics are based on 10,000 implied volatilities for a given moneyness. The left panel shows implied volatilities for stock options. For these options, the true value of volatility is 25%. The right panel shows implied volatilities for stock index options. For these options, the true value of volatility is 15%.

The size of the biases shown in figure 5 depends on the level of volatility. At lower levels of volatility, options away from $K/S = 1$ are farther from the money and the size of the bias increases dramatically. An index option at $K/S = 0.95$, for example, provides nearly unbiased implied volatility when true volatility is 15%. If true volatility is 10%, however, the same option overestimates implied volatility by 2.2 percentage points on average. If true volatility is 5%, the upward bias rises to 7.5 percentage points. At higher levels of volatility, all options are closer to the money and the size of the biases falls.

The pattern of rising implied volatilities as the underlying options move deeper in or out of the money has been widely reported and is referred to as an "implied volatility smile." With measurement errors in prices, the existence of implied volatility smiles does not reject the Black–Scholes (1973) and Merton (1973) model. A striking feature of figure 5 is the fact that, in truncated samples, the range of implied volatilities excludes the true value for options away from the money. In the presence of realistic measurement errors, the bias in implied volatilities quickly becomes so severe that the finding of volatility

smiles is inevitable even when all other Black–Scholes assumptions hold.

Realistic measurement errors can also account for implied volatility smirks. For the stock index options in the right panel of figure 5, the errors result in a bias that is asymmetric around $K/S = 1$. In this example, the asymmetry is due to the fact that the error in the underlying asset price is more important for in-the-money options.[25] This feature is unlikely to fully explain the observed volatility smirks, however, since the smirk extracted from a matched set of put options tends to go the other way. Under the assumptions of this paper, imposing put-call parity on the implied volatilities from a matched set of puts and calls whose strike prices are symmetric about $K/S = 1$ produces a roughly symmetric volatility smile.

Although not shown in the figure, it is nonetheless true that the truncation bias is more severe for options near maturity and less severe for otherwise identical options far from maturity. This bias pattern is broadly consistent with Das and Sundaram's (1999) report that "the volatility smile in most markets is deepest at short maturities and flattens out monotonically as maturity increases" (p. 216). Bakshi, Cao, and Chen (1997) and Aït-Sahalia and Lo (1998) report this pattern for S&P 500 options. In large part, this pattern stems from the fact that volatility until expiration, $\sigma\sqrt{t}$, not volatility alone, σ, matters in determining how likely the spot price is to cross the strike price by the option's expiration date. Dumas, Fleming, and Whaley (1998) propose a time-adjusted measure of moneyness to control for this effect.

5.2 Bias in cross-sectional averages

Standard regression techniques for truncated samples should be useful in reducing the bias in average implied volatilities. Unfortunately, many of the techniques are limited by their reliance on asymptotic results that are unlikely to hold for the typically small cross-sections of options with the same maturity date.[26] Moreover, standard solutions to this problem, such as Tobit regressions or Heckman's (1976) two-step procedure, are known to be inconsistent under non-normality or heteroskedasticity of the errors. Under the GLS weighting scheme proposed here, volatility estimates from options that are away from the money are likely to receive low weights, which mitigates this bias but does not necessarily eliminate it.

Table 2 reports sample statistics for unweighted and weighted cross-sectional estimates of implied volatility across 10,000 replications. The statistics are the mean, median, standard deviation, root-mean-squared-error (RMSE), mean absolute deviation (MAD), minimum, and maximum for each estimator, as well as the correlations across estimates. The unweighted estimates are the cross-sectional mean, the cross-sectional median, implied volatility computed only from the at-

[25] For stock index options, the volatility smile in figure 5 is asymmetric about $K/S = 1$. Nonetheless, options that are far out of the money also produce implied volatilities with an upward bias.

[26] See chapter 10 in Amemiya (1985) or chapter 6 in Maddala (1983) for details.

the-money option, and a Tobit estimate of the mean.[27] The weighted estimates are the Latané–Rendleman weighted average, the Beckers weighted average, the GLS weighted average based on equations (13) and (15), and a Tobit estimate that uses the diagonal of the covariance matrix in equation (13).[28] In the presence of a non-diagonal covariance matrix of the errors, maximum likelihood estimation of the Tobit regression requires numerical integration of n-dimensional multivariate normal cumulative distribution function, where n is the number of observations in the regression. Hajivassiliou and Ruud (1994) discuss fast methods for evaluating these integrals but the computations remain burdensome. To speed up the simulations, I ignore covariances in the Tobit estimators. Nonetheless, I provide results for Tobit estimators that account for the heteroskedasticity indicated by the diagonal terms of the covariance matrix in equation (13). I label these Tobit–HC.

The covariance and weighting matrices for the estimators whose weights depend on volatility (Tobit–HC, Latané–Rendleman, and Hentschel) are not based on the true volatility but iteratively compute implied volatility and weighting matrices until both converge. The GLS computations also randomly perturb the assumed distribution of the input errors. In each iteration, the covariance matrices are computed using bid–ask spreads drawn independently from a uniform distribution that has a mean and range equal to the true bid–ask spreads. For stock index options, for example, the GLS computations draw the bid–ask spread of the index level from a uniform distribution on the interval from 1 to 3, $U_{(1,3)}$, in each iteration. The spread used in generating the data is 2.

Many of the estimators in table 2 show severe upward bias. For the mean and median estimators, the bias stems directly from the truncation of the low volatility estimates. Although the Tobit-style estimator explicitly accounts for the truncation, heteroskedasticity biases this estimator.[29] The "at-the-money" estimator shows no bias. Since the VIX focuses on options that are essentially at the money, the bias of the VIX estimates is likely to be small.[30] Weighting the implied volatilities using the GLS matrix based on equation (13) turns out to sharply reduce the truncation bias. This is true for the GLS average, as well as the Tobit–HC maximum likelihood estimates that account for truncation and heteroskedasticity. Since the weighted Tobit estimators perform no better than the simple GLS mean, the additional computational burden of the Tobit estimators does not seem worthwhile in this application. When several near-the-money

[27] In case two options are equally close to the money, I designate the closest-to-the-money in-the-money option as the "at-the-money" option.

[28] Latané and Rendleman (1976) compute the average implied volatility by taking the square root of the average of the squares of the weighted volatilities. This induces an upward bias due to Jensen's inequality. I apply the Latané–Rendleman weights directly to the implied volatilities.

[29] The term Tobit regression is sometimes reserved for regressions with censored but not truncated samples. My Tobit estimators are full maximum likelihood estimators of the mean for truncated samples from an i.i.d. normal population.

[30] The absence of truncation bias and the relatively precise measurements suggest that measurement problems alone are unlikely to account for Fleming, Ostdiek, and Whaley's (1995) finding that VIX volatility is not an unbiased predictor of future volatility.

TABLE 2

The Precision of Cross-sectional Implied Volatility Estimates in Truncated Samples

Estimator	Mean	Median	SD	RMSE	MAD	Min	Max	Correlations							
Panel A: Call Options on a Stock, $\sigma = 25\%$, $S = 25$, $K = \{20.0, 22.5, 25.0, 27.5, 30.0\}$, $r = 5\%$, $\delta = 2.5\%$, $t = 20$ days															
Mean	36.060	36.000	3.396	11.570	11.060	25.010	48.194	1.000	0.629	0.382	0.952	0.471	0.629	0.380	0.442
Median	32.514	32.039	4.938	8.991	7.619	17.429	50.096		1.000	0.202	0.689	0.247	0.514	0.191	0.210
At-the-money	24.967	24.931	3.038	3.038	2.432	12.915	35.588			1.000	0.411	0.882	0.854	0.998	0.905
Tobit	35.598	35.570	3.364	11.119	10.611	−7.554	47.683				1.000	0.520	0.687	0.410	0.486
Tobit-HC	25.099	25.172	2.660	2.662	2.115	12.924	33.386					1.000	0.915	0.895	0.996
Latané–Rendelman	26.346	26.310	3.074	3.356	2.663	13.091	37.531						1.000	0.855	0.899
Beckers	24.912	24.908	3.044	3.045	2.445	12.899	35.380							1.000	0.917
Hentschel (GLS)	24.943	25.028	2.659	2.659	2.112	12.916	33.279								1.000
Panel B: Call Options on a Stock Index, $\sigma = 15\%$, $S = 500$, $K = \{450, 455, 460, \ldots, 550\}$, $r = 5\%$, $\delta = 2.5\%$, $t = 20$ days															
Mean	15.787	15.655	1.286	1.507	1.183	12.633	21.182	1.000	0.968	0.896	1.000	0.952	0.982	0.966	0.893
Median	15.123	15.076	0.570	0.584	0.463	13.214	17.823		1.000	0.908	0.968	0.962	0.974	0.971	0.920
At-the-money	14.986	14.991	0.616	0.616	0.491	12.539	17.823			1.000	0.896	0.910	0.925	0.939	0.888
Tobit	15.787	15.655	1.286	1.507	1.183	12.633	21.160				1.000	0.952	0.982	0.966	0.893
Tobit-HC	14.991	14.994	0.339	0.339	0.271	13.822	16.670					1.000	0.974	0.973	0.976
Latané–Rendelman	15.061	15.036	0.870	0.872	0.703	12.220	19.489						1.000	0.994	0.927
Beckers	15.003	15.001	0.647	0.647	0.518	12.612	18.404							1.000	0.941
Hentschel (GLS)	14.969	14.979	0.223	0.225	0.176	14.001	16.002								1.000

The table reports summary statistics for different cross-sectional estimators of implied volatility: mean; median; implied volatility from the single, closest-to-the-money option; a Tobit estimate of the mean that accounts for truncation; the same Tobit estimator using the diagonal of covariance matrix in equation (13) to also account for heteroskedasticity; a weighted average based on Latané and Rendelman (1976); an estimate based on Beckers (1981); and the GLS estimate of the mean in equation (15) using the covariance matrix in equation (13).

Call option prices are computed from the Black–Scholes formula. For each cross-section, all inputs into the implied volatility calculation are randomly perturbed using disturbances with zero means and standard deviations equal to one quarter of the bid–ask spreads in table 1 and figure 1. In each cross-section, the option price errors are independent but the other errors are identical. Option prices are constructed to obey absence of arbitrage, thereby creating unobserved truncation. Each simulation contains 10,000 independent cross-sections.

options are available, as in panel B, the GLS estimator has substantially lower variance than the "at-the-money" estimator. Nonetheless, when most of the options are well away from the money, as in panel A, the GLS estimator places such low weights on their implied volatilities that it effectively collapses to the "at-the-money" estimator. Although not designed to reduce the effects of truncation, the GLS estimator proposed in this paper performs best among these estimators.

Simulation for cross sections that contain call and put options show similar results. The main difference is that all estimates are more precise because the inclusion of call and put options allows more efficient cancellation of measurement errors in the underlying asset price.[31] The addition of the put options, however, does not materially reduce the bias of the cross-sectional implied volatility estimates. When put options are included, the RMSEs of the simple mean of the index volatilities falls from 1.507 to 0.705; for the at-the-money estimate, the RMSE drops from 0.616 to 0.173; for the GLS estimate, the RMSE falls from 0.225 to 0.058.

5.3 Bias in volatility function estimates

Instead of trying to construct a single efficient volatility estimate from a set of options with the same maturity, attention has increasingly turned to estimating a volatility function in which implied volatilities are smoothed across strike prices ("volatility smile") and time to maturity ("volatility surface"). (See Dumas, Fleming, and Whaley 1998 or Brandt and Wu 2001, for example.) Since the Black–Scholes assumptions specify a single volatility for each maturity, testing whether implied volatility is constant across strike prices can be thought of as a specification test of the Black–Scholes assumptions. More generally, there is interest in extracting implied volatilities for option pricing models that permit Black–Scholes implied volatilities to depend on moneyness.

To illustrate the effects of errors in implied volatilities on volatility functions, I investigate the performance of several parametric and non-parametric estimators for volatility smiles. The simulations assume that all of the Black–Scholes assumptions hold, but that prices are observed with the same errors as before. As before, this leads to unobserved truncation when all prices are required to satisfy absence of arbitrage. As shown in figure 4, this truncation can lead to pronounced biases in the observed implied volatilities. For each cross-section of implied volatilities, I now estimate a volatility smile to fit the observed implied volatilities. Naturally, many estimators fit the bias in the observed implied volatilities. Due to the small cross-sections for the stock options, I only perform these simulation for the stock index options.

The parametric estimates smooth the observed implied volatilities, $\tilde{\sigma}_i$, as a quadratic function of moneyness, $(K_i/S - 1)$,

$$\tilde{\sigma}_i = \beta_0 + \beta_1(K_i/S - 1) + \beta_2(K_i/S - 1)^2 + \epsilon_i. \tag{20}$$

[31] In these cross-sections, the at-the-money estimator averages the implied volatilities from the closest-to-the-money call and put options.

The quadratic form for the parametric regressions is a compromise between a desire for flexibility and the difficulty of estimating a large number of parameters from 21 observations at a time.[32] Specifying moneyness as $(K/S - 1)$ is attractive because it implies that the regressors are orthogonal when the included strike prices are symmetric about S, as is true here. I estimate these quadratic regressions using least squares or Tobit procedures. Both estimation methods are simulated in two different ways: ignoring heteroskedasticity and correlation for the errors, or accounting for these features using the covariance matrix in equation (13).

Least squares can use the full covariance matrix from equation (13) to specify generalized least squares, GLS. The Tobit–HC estimator accounts for heteroskedasticity using the diagonal of the covariance matrix but ignores correlations. As before, these estimators exploit the structure of the covariance matrix but not precise knowledge of the underlying error variances. In each iteration, the bid–ask spreads used in the construction of the covariance matrix are drawn independently from a uniform distribution that has a mean and range equal to the true bid–ask spreads.

Table 3 reports means, medians (in brackets), and standard deviations (in parentheses) of the simulation results. The at-the-money estimate of volatility, $\hat{\beta}_0$, is generally close to the true value of 15%. All of the homoskedastic estimators erroneously conclude there is a volatility smile, however, since the linear and quadratic coefficients are individually and jointly significant according to the standard t- and F-tests. The estimators that account for heteroskedasticity behave much better. The coefficient estimates are generally closer to their true values and the significance tests frequently cannot reject the hypothesis that $\hat{\beta}_1 = \hat{\beta}_2 = 0$. Nonetheless, in these small cross-sections, none of the estimators and tests are perfect. In theory, a Tobit estimator accounting for truncation, heteroskedasticity, and correlation should account for all of the features of the data and be well behaved. The Tobit–HC estimator ignores the correlations of the residuals to speed up the calculations. This omission is enough to bias the estimator toward finding volatility smiles when there are none. Although the GLS estimator accounts for the correlations, it ignores the effects of truncation. Since the observations that are most affected by truncation are also the ones that receive low weights under the GLS estimator, the GLS estimator performs best among these estimators.

Table 4 summarizes the empirical distributions of the t- and F-statistics. For each estimator, the table shows the following percentiles: 2.5, 97.5, 5, and 95. The table clearly shows that, for these samples, all of the tests are far from their nominal size. All of the significance tests overstate the true differences.

In addition to the parametric methods described above, I also investigate two non-parametric procedures: a standard locally-constant kernel regression,

[32] Experiments with higher-order specifications produce noisier parameter estimates and fitted values. This is true because the richer specifications add more parameters that are unnecessary under the constant volatility assumption.

TABLE 3
Volatility Smile Regressions

	$\beta_0 = 15\%$		$\beta_1 = 0$		$\beta_2 = 0$		$F(\hat{\beta}_1 = \hat{\beta}_2 = 0)$	
Estimator	$\hat{\beta}_0 \times 100$	t	$\hat{\beta}_1 \times 100$	t	$\hat{\beta}_2 \times 100$	t	F	P
OLS	14.500	−0.794	−19.630	−4.716	362.182	4.275	74.136	0.020
	[14.576]	[−0.911]	[−16.178]	[−3.135]	[342.926]	[3.382]	[21.091]	[0.000]
	(0.887)	(1.871)	(20.511)	(5.375)	(164.646)	(2.759)	(114.523)	(0.079)
OLS (White)	14.500	−1.077	−19.630	−4.243	362.182	4.299	66.379	0.019
	[14.576]	[−1.390]	[−16.178]	[−2.671]	[342.926]	[3.412]	[18.123]	[0.000]
	(0.887)	(2.406)	(20.511)	(5.123)	(164.646)	(2.952)	(132.137)	(0.081)
Tobit	14.500	−0.857	−19.630	−5.094	362.182	4.617	86.492	0.016
	[14.576]	[−0.984]	[−16.178]	[−3.386]	[342.926]	[3.653]	[24.606]	[0.000]
	(0.887)	(2.021)	(20.511)	(5.806)	(164.646)	(2.980)	(133.611)	(0.072)
Tobit–HC	15.029	−0.005	−1.652	−0.539	16.974	0.452	20.214	0.072
	[15.007]	[0.062]	[−0.236]	[−0.048]	[6.584]	[0.099]	[10.993]	[0.001]
	(0.694)	(5.752)	(18.729)	(3.882)	(188.186)	(2.882)	(26.500)	(0.176)
LS–HC	15.024	−0.629	−2.329	−0.686	30.534	0.566	28.700	0.070
	[15.010]	[0.136]	[−0.528]	[−0.136]	[10.828]	[0.164]	[15.044]	[0.000]
	(0.672)	(6.651)	(21.049)	(5.029)	(233.779)	(3.282)	(38.211)	(0.178)
GLS	14.904	−1.061	2.680	0.615	−30.879	−0.359	9.340	0.142
	[14.960]	[−0.292]	[1.376]	[0.260]	[−15.189]	[−0.195]	[4.819]	[0.019]
	(0.499)	(3.668)	(13.872)	(2.554)	(153.721)	(1.910)	(14.183)	(0.234)

The table reports simulation results for cross-sectional volatility smile regressions

$$\tilde{\sigma}_i = \beta_0 + \beta_1 (K_i/S - 1) + \beta_2 (K_i/S - 1)^2 + \epsilon_i. \tag{20}$$

The table shows simulation means for the scaled point estimates, $\hat{\beta}_j \times 100$, the t-statistics, t, for the tests that $\beta_0 = 15\%$, $\beta_1 = 0$, and $\beta_2 = 0$, F-statistics for the test $\hat{\beta}_1 = \hat{\beta}_2 = 0$, F, and P-values associated with the F-test, P. The numbers in square brackets are sample medians; the numbers in parentheses are sample standard deviations.

Implied volatilities, $\tilde{\sigma}_i$, are generated by simulating 10,000 independent cross-sections of Black–Scholes prices for call options on a stock index. The options have 20 days to maturity. In the simulations, $\sigma = 15\%$, $S = 500$, $K = \{450, 455, 460, \ldots, 550\}$, $r = 5\%$, $\delta = 2.5\%$. Each cross-section contains 21 call options. All prices are observed with errors based on table 1 and figure 1.

The regressions are estimated using one of the methods listed in the first column: ordinary least squares (OLS), ordinary least squares with heteroskedasticity-consistent standard errors based on White (1980) (OLS (White)), Tobit, Tobit taking into account the heteroskedasticity of the errors but ignoring their correlation (Tobit–HC), least squares taking into account the heteroskedasticity of the errors but ignoring their correlation (LS–HC), or generalized least squares (GLS).

and a locally-linear kernel regression.[33] The performance of the different nonparametric methods is similar and I only show results for the locally-linear regressions. Although the differences are small, the locally-linear regressions perform better in this application. For both locally-linear and locally-constant kernel regressions the choice of bandwidth is important. In these simulations, the plug-in methods summarized in Simonoff (1996) outperform cross-validation. I only report the locally-linear results based on the plug-in method.

[33] Given a kernel function $\theta(\cdot)$ and bandwidth h, a locally-constant regression finds y_0 to minimize $\sum_i [y_i - y_0]^2 \theta((x - x_i)/h)$ for all observations $y_i = f(x_i) + \epsilon_i$. The locally-linear regression generalizes the objective function to $\sum_i [y_i - y_0 - y_1(x - x_i)]^2 \theta((x - x_i)/h)$. When the true function, $f(\cdot)$, is not constant near the boundaries of the sample, locally-linear regressions are less biased than locally-constant regressions.

TABLE 4
Significance Tests in Volatility Smile Regressions

Estimator	$t(\beta_0 = 15\%)$ $\underline{t}$	$\overline{t}$	$t(\beta_1 = 0)$ $\underline{t}$	$\overline{t}$	$t(\beta_2 = 0)$ $\underline{t}$	$\overline{t}$	$F(\hat{\beta}_1 = \hat{\beta}_2 = 0)$ $\underline{F}$	$\overline{F}$
Panel A: 2.5% and 97.5%								
Nominal	−2.093	2.093	−2.093	2.093	−2.093	2.093	0.025	4.507
OLS	−3.997	3.261	−17.406	1.855	0.817	11.112	1.679	402.008
OLS (White)	−4.775	4.130	−17.065	1.898	0.794	12.099	1.685	422.448
Tobit	−4.317	3.522	−18.801	2.003	0.882	12.003	1.959	469.010
Tobit–HC	−11.455	10.341	−9.190	5.961	−4.400	6.895	0.354	92.646
LS–HC	−14.623	9.607	−11.180	7.602	−4.874	7.555	0.321	135.638
GLS	−9.810	4.006	−3.327	6.167	−4.334	2.916	0.168	47.790
Panel B: 5% and 95%								
Nominal	−1.729	1.729	−1.729	1.729	−1.729	1.729	0.051	3.522
OLS	−3.587	2.460	−15.218	1.393	1.179	9.820	2.629	314.544
OLS (White)	−4.377	3.173	−14.387	1.406	1.125	10.248	2.620	296.366
Tobit	−3.874	2.657	−16.438	1.505	1.274	10.606	3.067	366.968
Tobit–HC	−9.546	9.103	−7.616	5.013	−3.680	5.696	0.728	70.874
LS–HC	−12.311	8.482	−9.613	6.566	−4.106	6.454	0.696	102.079
GLS	−8.006	3.502	−2.861	5.234	−3.642	2.452	0.328	34.197

The table reports empirical critical values for tests in the volatility smile regressions

$$\tilde{\sigma}_i = \beta_0 + \beta_1 (K_i/S - 1) + \beta_2 (K_i/S - 1)^2 + \epsilon_i. \tag{20}$$

The table shows empirical 2.5, 97.5, 5, and 95% critical values for the t-statistics of the tests that $\beta_0 = 15\%$, $\beta_1 = 0$, and $\beta_2 = 0$. The table shows the same empirical crictical values for the F-test that $\hat{\beta}_1 = \hat{\beta}_2 = 0$. The results are based on 10,000 independent simulations. The nominal values are shown for comparison.

Implied volatilities, $\tilde{\sigma}_i$, are generated by simulating 10,000 independent cross-sections of Black–Scholes prices for call options on a stock index. The options have 20 days to maturity. In the simulations, $\sigma = 15\%$, $S = 500$, $K = \{450, 455, 460, \ldots, 550\}$, $r = 5\%$, $\delta = 2.5\%$. Each cross-section contains 21 call options. All prices are observed with errors based on table 1 and figure 1.

The regressions are estimated using one of the methods listed in the first column: ordinary least squares (OLS), ordinary least squares with heteroskedasticity-consistent standard errors based on White (1980) (OLS (White)), Tobit, Tobit taking into account the heteroskedasticity of the errors but ignoring their correlation (Tobit–HC), least squares taking into account the heteroskedasticity of the errors but ignoring their correlation (LS–HC), or generalized least squares (GLS).

The locally-linear non-parametric regressions can also be modified to locally-linear generalized least squared regressions. Unfortunately, for about 1% of all simulations this procedure produced volatility estimates in excess of 1,000%. Since there was no sharp cut-off between acceptable and obviously bad fits, I modify the locally-linear regressions to account for heteroskedasticity but ignore correlations. These regressions are well behaved. They are labeled "Locally-linear–HC."[34]

Figure 6 shows means and medians of the fitted values and empirical 95% confidence intervals for the six volatility smile estimators. The top three panels in the figure show results for estimators that assume homoskedastic and

[34] There is no known non-parametric estimator that accounts for truncation. See Khan and Lewbel (2002) for semi-parametric estimation of truncated regression models.

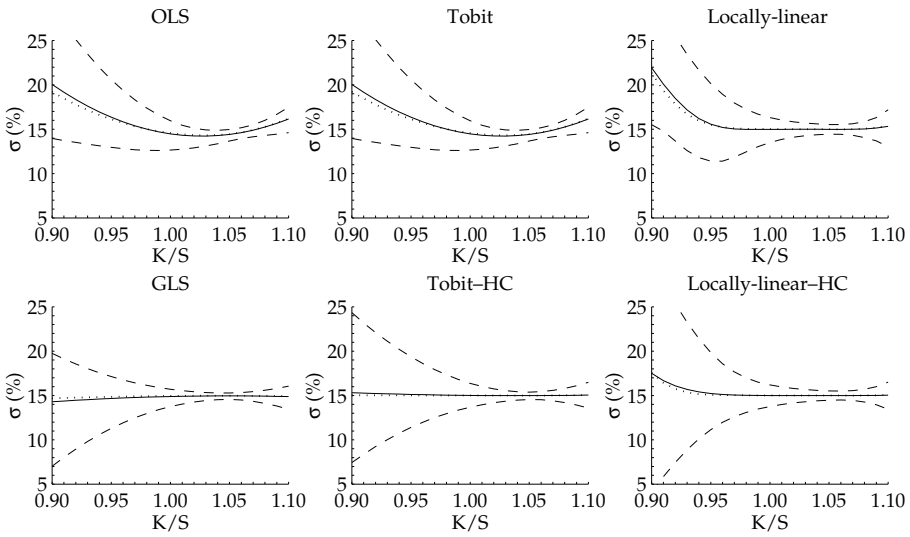

FIGURE 6: VOLATILITY SMILE ESTIMATES

The figure shows mean fitted values (solid lines), median fitted values (dotted lines), and empirical 95% confidence intervals (dashed lines) for volatility smile estimators when the Black–Scholes assumption are true but prices are observed with measurement errors. True volatility is constant at $\sigma = 15\%$. The parametric estimators, OLS, GLS, Tobit, Tobit–HC, assume that implied volatility is a quadratic function of moneyness, $(K_i/S - 1)$. The locally-linear non-parametric estimator uses plug-in bandwidth estimates. All options and underlying prices are observed with the errors described in table 1 and figure 1.

The estimators in the top panels (erroneously) assume that the implied volatilities are homoskedastic and uncorrelated. The estimators in the bottom panels make corrections for heteroskedasticity or correlations. The GLS estimator employs covariance matrix estimates based on equation (13). The Tobit-HC and locally-linear–HC estimators only use the diagonal elements of the covariance matrix.

Implied volatilities, $\tilde{\sigma}_i$, are generated by simulating 10,000 independent cross-sections of Black–Scholes prices for call options on a stock index. The options have 20 days to maturity. In the simulations, $S = 500$, $K = \{450, 455, 460, \ldots, 550\}$, $r = 5\%$, $\delta = 2.5\%$. Each cross-section contains 21 call options.

uncorrelated errors. All three estimators, OLS, Tobit, and the locally-linear non-parametric regression, show substantial bias and low precision for implied volatilities from options far from the money. The figure shows in striking fashion that these estimators almost surely find a volatility smile even though all of the Black–Scholes assumptions hold—except that prices are measured with small errors. In the bottom three panels, the estimators account for heteroskedasticity. GLS also accounts for correlation among the errors. This modification materially reduces the bias and noise of the estimators. On average, the two parametric estimators correctly fit a constant implied volatility function.

Results for cross-sections including calls and puts are generally similar. Ignoring heteroskedasticity and correlation in the errors still leads to noisy and biased fitted values. Nonetheless, the addition of the put options has two noticeable effects. First, the addition of the put options helps to cancel the errors from the underlying asset price and the fitted values are more precise. This also has the effect that the Tobit–HC estimator becomes the most efficient estimator,

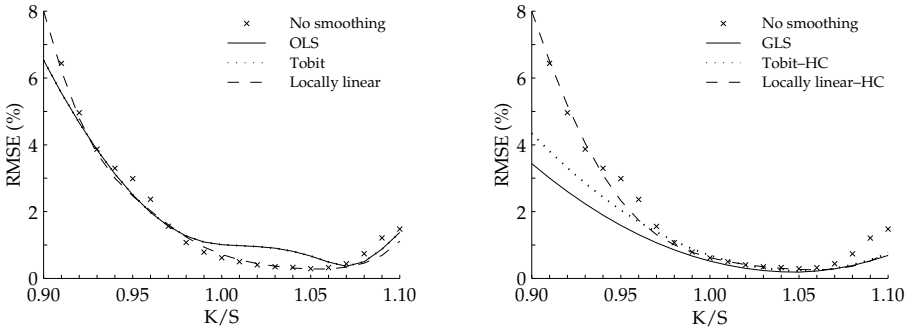

FIGURE 7: RMSEs FOR VOLATILITY SMILE ESTIMATES

The figure shows root-mean-squared errors (RMSE) for implied volatility estimators using call options on a stock index. The parametric estimators, OLS, GLS, Tobit, Tobit–HC, assume that implied volatility is a quadratic function of moneyness, $(K_i/S - 1)$. The locally-linear non-parametric estimator uses plug-in bandwidth estimates. All options and underlying prices are observed with the errors described in table 1 and figure 1.

The estimators in the left panel (erroneously) assume that the implied volatility estimates are homoskedastic and uncorrelated. (The OLS and Tobit estimators perform identically in this case and only the OLS line is visible.) The estimators in the right panel make corrections for heteroskedasticity or correlations. The GLS estimator employs the covariance matrix estimates based on equation (13). The Tobit-HC and locally-linear–HC estimators only use the diagonal elements of the covariance matrix. In both panels, the RMSE of the individual Black–Scholes implied volatilities is marked by an ×.

Implied volatilities, $\tilde{\sigma}_i$, are generated by simulating 10,000 independent cross-sections of Black–Scholes prices for call options on a stock index. The options have 20 days to maturity. In the simulations, $\sigma = 15\%$, $S = 500$, $K = \{450, 455, 460, \ldots, 550\}$, $r = 5\%$, $\delta = 2.5\%$. Each cross-sections contains 21 call options.

since the GLS estimator no longer gains a big advantage by exploiting the correlations stemming from the error in the underlying asset price. Second, the bias in the fitted values becomes more symmetric about $K/S = 1$. There is more of a smile rather than the smirk shown in figure 6.

Finally, figure 7 compares the root-mean-squared errors (RMSE) for several volatility smile estimators. There are three general findings in figure 7. First, the low-dimensional parametrically smoothed volatility smiles have lower RMSEs than the individual implied volatilities or the non-parametrically smoothed volatilities. Frequently these differences are large. Strikingly, a non-parametric smooth of implied volatilities from call options has basically the same RMSEs as the individual implied volatilities. This is a typical outcome for an estimator that overfits the data. Second, smoothing is most effective for the least reliable implied volatility estimates, those farthest from the money. Third, even for a low-order polynomial, estimating a volatility smile when volatility is constant is inefficient. Table 2 shows that the GLS mean of implied volatilities has a RMSE of 0.225%. The quadratic GLS estimates in the right panel of figure 7 have a RMSE of 0.518 for at-the-money options.[35] This is about twice as high as the RMSE of 0.225 for the constant GLS estimate reported in the bottom panel of table 2. Yet, for

[35] The minimum RMSE for the GLS estimates in figure 7 is 0.190%.

options away from the money, the RMSE of the quadratic estimates quickly rises to 10 times the RMSE of the constant estimates. Of course, when there really is a volatility smile, restricting volatility to be constant across strike prices produces inconsistent volatility estimates. Once again, adding puts to the cross-section of options leaves the qualitative conclusions unchanged. Nonetheless, the addition of puts roughly halves the RMSEs.

6 Conclusion

This paper shows that even if the Black–Scholes assumptions hold perfectly—except that we observe prices with measurement errors—estimates of implied volatility are subject to considerable noise and bias.

Measurement errors in option characteristics, especially in the option price and underlying asset price, directly translate into measurement errors in implied volatility. Implied volatilities from options away from the money are especially sensitive to such measurement errors. To gauge the magnitude of this noise, the paper constructs confidence intervals for implied volatilities. Based on conservative assumptions for measurement errors, the 95% confidence intervals for implied volatility from an at-the-money stock option with 20 days to maturity are on the order of plus or minus 6 percentage points.

If dealers post option prices that obey absence of arbitrage, observed implied volatilities are biased upward. This bias occurs because negative option price errors quickly lead to violations of absence of arbitrage for options away from the money. By posting only prices that obey absence of arbitrage, dealers truncate posted option prices and their associated implied volatilities. In simulations, stock options that have 20 days to expiration and are 10% in the money have average observed implied volatility of 33% when true volatility is 25%; stock options that have 20 days to expiration and are 20% in the money have average observed implied volatility of 58% when true volatility is 25%.

The paper constructs feasible GLS estimators of implied volatility. The estimators are designed to minimize the noise the implied volatility estimates but also sharply reduce the bias of the estimates. The methods apply to volatility estimators that model implied volatility as constant across moneyness, as volatility smiles across moneyness, or as volatility surfaces across moneyness and time to expiration. These GLS estimators outperform all of the previously proposed implied volatility estimators by considerable margins. The two principal reasons for this superior performance are that the GLS estimators proposed here recognize that the measurement error of the underlying asset price can make an important contribution to the total error in implied volatility and that the measurement errors of the option characteristics are highly correlated across options.

Given the potentially large errors, and their pattern, the errors may contribute to explanations of some of the reported patterns in implied volatility. The standard finding that regressing realized volatility on implied volatility yields a coefficient less than 1 is consistent with measurement errors in implied volatility.

The fact that other models of volatility, such as the models of generalized autoregressive heteroskedasticity (GARCH), introduced by Bollerslev (1987), have predictive power for volatility over and above observed implied volatility is also consistent with measurement errors in implied volatility.

The bias in implied volatility may contribute to explanations of the widely reported volatility smiles. The bias appears in individual implied volatilities and in standard estimates of volatility smiles or surfaces. The feasible GLS estimators proposed here largely eliminate these biases. The dramatic increase in the bias in implied volatility as options near expiration may play a role in the rise in measured implied volatility just prior to expiration reported in Day and Lewis (1988).

The abundant evidence on time-varying volatility in stock returns makes it unlikely that measurement errors are the only source of the observed deviations from Black–Scholes option prices and the associated patterns in implied volatilities. Nonetheless, it would be interesting to apply the proposed estimators to option data to see what fraction of the observed structure in implied volatility can be explained by noise and bias due to measurement errors. Ferri (2001) argues that accounting for tick sizes and the truncation bias due to absence of arbitrage is sufficient to explain the observed volatility smiles in foreign currency options.

References

Aït-Sahalia, Yacine and Andrew W. Lo, 1998, "Nonparametric estimation of state-price densities implicit in financial asset prices." *Journal of Finance* 53, 499–547.

Amemiya, Takeshi, 1985, *Advanced Econometrics.* Harvard University Press, Cambridge, MA.

Andersen, Torben G., Luca Benzoni, and Jesper Lund, 2002, "An empirical investigation of continuous-time equity return models." *Journal of Finance* 57, 1239–1284.

Bakshi, Gurdip, Charles Cao, and Zhiwu Chen, 1997, "Empirical performance of alternative option pricing models." *Journal of Finance* 52, 2003–2049.

Bates, David S., 1996, "Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options." *Review of Financial Studies* 9, 69–107.

Beckers, Stan, 1981, "Standard deviations implied in option prices as predictors of future stock price variability." *Journal of Banking and Finance* 5, 363–382.

Black, Fischer and Myron Scholes, 1973, "The pricing of options and corporate liabilities." *Journal of Political Economy* 81, 637–54.

Bollerslev, Tim, 1986, "Generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics* 31, 307–328.

Boyle, Phelim and A.L. Ananthanarayanan, 1977, "The impact of variance estimation in option valuation models." *Journal of Financial Economics* 5, 375–387.

Brandt, Michael W. and Tao Wu, 2001, "Cross-sectional tests of deterministic volatility functions." Unpublished paper, Wharton School, University of Pennsylvania, Philadelphia, PA.

BUTLER, J.S. AND BARRY SCHACHTER, 1986, "Unbiased estimation of the Black–Scholes formula." *Journal of Financial Economics* 15, 341–357.

CANINA, LINDA AND STEPHEN FIGLEWSKI, 1993, "The informational content of implied volatility." *Review of Financial Studies* 6, 659-681.

CHIRAS, DONALD P. AND STEVEN MANASTER, 1978, "The information content of option prices and a test for market efficiency." *Journal of Financial Economics* 6, 213–234.

CHRISTENSEN, BENT J., CHARLOTTE STRUNK HANSEN, AND NAGPURNANAND R. PRABHALA, 2001, "The telescoping overlap problem in options data." Unpublished paper, School of Economics and Management, University of Aarhus, Aarhus, Denmark.

CHRISTENSEN, BENT J. AND NAGPURNANAND R. PRABHALA, 1999, "The relation between implied volatility and realized volatility." *Journal of Financial Economics* 50, 125–150.

CHRISTOFFERSEN, PETER F., AND KRIS JACOBS, 2001, "The importance of the loss function in option pricing." Unpublished paper, McGill University, Montreal, Canada.

DAS, SANJIV R. AND RANGARAJAN K. SUNDARAM, 1999, "Of smiles and smirks: A term structure perspective." *Journal of Financial and Quantitative Analysis* 34, 211–239.

DAY, THEODORE E. AND CRAIG M. LEWIS, 1988, "The behavior of the volatility implicit in the prices of stock index options." *Journal of Financial Economics* 22, 103–122.

DUMAS, BERNARD, JEFF FLEMING, AND ROBERT E. WHALEY, 1998, "Implied volatility functions: Empirical tests." *Journal of Finance* 53, 2059–2106.

FERRI, ARTHUR, 2001, "The structure of implied volatility and the bid–ask spread in option markets." Unpublished paper, Fox School of Business, Temple University, Philadelphia, PA.

FIGLEWSKI, STEPHEN, 1989, "Options arbitrage in imperfect markets." *Journal of Finance* 44, 1289–1311.

FLEMING, JEFF, 1998, "The quality of market volatility forecasts implied by S&P 100 index option prices." *Journal of Empirical Finance* 5, 317–345.

FLEMING, JEFF, BARBARA OSTDIEK, AND ROBERT E. WHALEY, 1995, "Predicting Stock market volatility: A new measure." *Journal of Futures Markets* 15, 265–302.

FRANKS, JULIAN R. AND EDUARDO S. SCHWARTZ, 1991, "The stochastic behaviour of market variance implied in the prices of index options." *Economic Journal* 101, 1460–1475.

GALLANT, A. RONALD, 1997, *An Introduction to Econometric Theory.* Princeton University Press, Princeton, NJ.

HAJIVASSILIOU, VASSILIS A., AND PAUL A. RUUD, 1994, "Classical estimation methods for LDV models using simulation." *Handbook of Econometrics, Vol. IV.* Edited by Robert F. Engle and Daniel L. McFadden, Elsevier Science B.V., New York, NY.

HARVEY, CAMPBELL R. AND ROBERT E. WHALEY, 1991, "S&P 100 index option volatility." *Journal of Finance* 46, 1551–1561.

HARVEY, CAMPBELL R. AND ROBERT E. WHALEY, 1992, "Market volatility prediction and the efficiency of the S&P 100 index option market." *Journal of Financial Economics* 31, 43–73.

HECKMAN, JAMES, 1976, "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models." *Annals of Economic and Social Measurement* 5, 475–492.

HENTSCHEL, LUDGER, 1994, "Alternative models of asymmetric volatility in stock returns." Unpublished Ph.D. dissertation, Princeton University,Princeton, NJ.

HO, THOMAS AND HANS R. STOLL, 1981, "Optimal dealer pricing under transactions and return uncertainty." *Journal of Financial Economics* 9, 47–73.

JACKWERTH, JENS C. AND MARK RUBINSTEIN, 1996, "Recovering probability distributions from option prices." *Journal of Finance* 51, 1611–1631.

JONES, CHRISTOPHER S., 2001, "The dynamics of stochastic volatility: Evidence from underlying and option markets." Unpublished paper, Simon School, University of Rochester, Rochester, NY.

KHAN, SHAKEEB AND ARTHUR LEWBEL, 2002, "Weighted and two stage least squares estimation of semiparametric truncted regression models." Unpublished paper, University of Rochester, Rochester, NY.

LATANÉ, HENRY A. AND RICHARD J. RENDLEMAN, JR., 1976, "Standard deviations of stock price ratios implied in option prices." *Journal of Finance* 31, 369–381.

LEDOIT, OLIVIER, PEDRO SANTA-CLARA, AND SHU YAN, 2002, "Relative pricing of options with stochastic volatility." Unpublished paper, Anderson School of Management, UCLA, Los Angeles, CA.

LUENBERGER, DAVID G., 1984, *Linear and Nonlinear Programming.* 2nd ed., Addison–Wesley, Reading, MA.

LYONS, RICHARD K., 1988, "Tests of the foreign exchange risk premium using the expected second moment implied by option pricing." *Journal of International Money and Finance* 7, 91–108.

MADDALA, G.S., 1983, *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, New York, NY.

MANASTER, STEVEN AND GARY KOEHLER, 1982, "The calculation of implied variances from the Black–Scholes model: A note." *Journal of Finance* 37, 227–230.

MERTON, ROBERT C., 1973, "Theory of rational option pricing." *Bell Journal of Economics and Management Science* 1, 141–183.

OLDS, EDWIN G., 1952, "A note on the convolution of uniform distributions." *Annals of Mathematical Statistics* 23, 282–285.

PAN, JUN, 2002, "The jump-risk premia implicit in options: evidence from an integrated time-series study." *Journal of Financial Economics* 63, 3–50.

PATELL, JAMES M. AND MARK A. WOLFSON, 1979, "Anticipated information releases reflected in call option prices." *Journal of Accounting and Economics* 1, 117–140.

PHILLIPS, SUSAN M. AND CLIFFORD W. SMITH, JR., 1980, "Trading costs for listed options: implications for market efficiency." *Journal of Financial Economics* 8, 179–189.

RUBINSTEIN, MARK, 1985, "Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active CBOE option classes from August 23, 1976 through August 31, 1978." *Journal of Finance* 40, 455–480.

SCHMALENSEE, RICHARD AND ROBERT R. TRIPPI, 1978, "Common stock volatility expectations implied in option premia." *Journal of Finance* 33, 129–147.

SCHOLES, MYRON AND JOSEPH T. WILLIAMS, 1977, "Estimating betas from nonsynchronous data." *Journal of Financial Economics* 5, 309–327.

SIMONOFF, JEFFREY S., 1996, *Smoothing Methods in Statistics.* Springer-Verlag New York, New York, NY.

VIJH, ANAND M., 1990, "Liquidity of the CBOE equity options." *Journal of Finance* 45, 1157–1179.

WHALEY, ROBERT E., 1982, "Valuation of American call options on dividend-paying stocks: Empirical tests." *Journal of Financial Economics* 10, 29–58.

WHALEY, ROBERT E., 1993, "Derivatives on market volatility: Hedging tools long overdue." *Journal of Derivatives* 1, 71–84.

WHITE, HALBERT, 1980, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica* 48, 817–838.