On the Number of State Variables in Options Pricing

Gang Li, Chu Zhang
Hong Kong Baptist University, Kowloon Tong, Hong Kong; and Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong {garyli@hkbu.edu.hk, czhang@ust.hk}

In this paper, we investigate the methodological issue of determining the number of state variables required for options pricing. After showing the inadequacy of the principal component analysis approach, which is commonly used in the literature, we adopt a nonparametric regression technique with nonlinear principal components extracted from the implied volatilities of various moneyness and maturities as proxies for the transformed state variables. The methodology is applied to the prices of S&P 500 index options from the period 1996–2005. We find that, in addition to the index value itself, two state variables, approximated by the first two nonlinear principal components, are adequate for pricing the index options and fitting the data in both time series and cross sections.

Key words: options pricing; state variables; nonparametric method; nonlinear principal component analysis
History: Received October 22, 2009; accepted May 31, 2010, by Wei Xiong, finance. Published online in Articles in Advance September 3, 2010.

1. Introduction
The recent literature on modeling options prices has made tremendous progress in tackling the problems of volatility smile and volatility smirk with respect to the well-known Black-Scholes (1973) model. There are two major classes of new models: the stochastic volatility models in continuous time framework and the generalized autoregressive conditional heteroskedasticity (GARCH) models in discrete time framework. Empirical studies confirm that these models have made advances in resolving the pricing issue. However, existing models are still inadequate, to varying degrees, in fitting the observed options prices. For example, a typical problem with stochastic volatility models is, as observed by Bakshi et al. (1997), that the models require highly implausible parameters of the volatility-return correlation and the volatility-of-variability. The existing stochastic volatility models with jumps are still incapable of matching long-term options prices, as shown by Bates (2000).

There can be two sources of model misspecification. One is the omitted state variables, or factors. Recent research has shown the benefit of increasing the number of state variables. For example, Chernov et al. (2003) study the dynamics of the Dow Jones index and find evidence that favors two-factor models, which separate the tail effect from the volatility persistence effect, over one-factor models. Christoffersen et al. (2008) propose a GARCH options pricing model with long-run and short-run volatility factors that outperforms the one-factor options pricing model of Heston and Nandi (2000), especially for pricing long maturity options. Christoffersen et al. (2009) find that two-factor stochastic volatility models provide more flexibility in modeling the time variation in the smirk and the volatility term structure than do single-factor stochastic volatility models. It is not clear, however, if having more state variables is better. In the extant literature, state variables beyond the price of the underlying security itself are mostly unobserved, and each additional state variable unavoidably introduces a slew of parameters; making the identification and estimation of the unobserved state variables and parameters a very complicated procedure. As a result, no consensus has been reached on how many state variables are ideal for options pricing and what these state variables should be.

The other source of model misspecification is the functional form of the process for the state variables, including the specification of risk premiums associated with the state variables. Bates (2000) finds strong evidence against the square-root diffusion process driving instantaneous volatility, because it cannot...
account for the large and typically positive volatility shocks implied in the prices of long-term options. Jones (2003) finds that the square-root stochastic volatility model is incapable of generating realistic return behavior and concludes that the stochastic volatility models in the constant elasticity of variance class or with a time-varying leverage effect are more consistent with the underlying asset and options data. Chernov et al. (2003) consider 10 models with various numbers of state variables and functional forms, but, in conclusion, they are unwilling to declare an overall winner.

The two sources of model misspecification—i.e., the uncertain number of state variables and the functional form of the underlying process of the state variables—are obviously related. With omitted state variables, the functional form can never be correct. With an incorrect specification of the functional form, determining the correct number of state variables is difficult and the result is unreliable.

In this paper, we address the issue of how to determine the number of state variables for options pricing. Although the goal is rather limited, we view the correct specification of the number of state variables as the top priority in modeling options prices; without this, a correct functional form for the state variables is out of the question. Because an incorrect functional form will also confound the determination of the number of state variables, we resolve the issue by using a nonparametric approach with state variables approximated by nonlinear principal components (NPCs) extracted from the implied volatilities. The nonparametric approach avoids misspecification of the relationship between options prices and the state variables. A feature that sets our approach apart from others in the literature is our use of NPCs as the observed proxies for state variables. More specifically, we fit the options prices nonparametrically as functions of these NPCs. Using the fitted options price functions, we then test how many NPCs are sufficient to capture the dynamics of the options prices.

Our approach of using NPCs as proxies for state variables requires some explanation. The NPCs we use are obtained from the prices of European calls and puts. This approach disqualifies the resulting nonparametric function as a structural options pricing model even if the functional form is known, because options prices are needed as input to fit the function. For our purpose of determining the number of state variables, however, this is not a problem. The approach is analogous to the one in the empirical studies of stock markets in which systematic factors of mimicking portfolio returns are constructed using returns on individual stocks, and these factors are then used to determine the alphas and betas of individual stocks.

We apply the methodology to the analysis of S&P 500 index options, using weekly data from January 1996 to December 2005. We extract NPCs from the weekly average implied volatilities of various moneyness and maturities. We find that the first two NPCs are sufficient to capture the dynamics of the state variables in the time series and cross sections of the options prices. Although the third NPC is also statistically significant, its contribution to options pricing is economically insignificant, because the differences between the fitted options prices from the model with three NPCs and those from the model with two NPCs are smaller than the typical transaction costs in trading options. The result suggests that, for S&P 500 index options, there will be little gain from extending the number of state variables beyond two.

The rest of this paper is organized as follows. Section 2 briefly presents the methodology. Section 3 describes the S&P 500 index options price data for the application and discusses the problems of using linear principal components (PCs) for options pricing. Section 4 discusses the nonparametric testing method and reports the results of determining the number of required state variables for S&P 500 index options. Section 5 concludes the paper.

2. Methodology

2.1. The Empirical Model and Nonparametric Estimation

Let \( S_t \) be the price of an underlying asset on which options are traded. Suppose the stochastic process of \( S_t \) is governed by a \( k \) vector of unobserved state variables, \( x_t = (x_{t1}, \ldots, x_{tk}) \), which follows a Markov process. The value at \( t \) of a European option expiring at \( t+T \) can be written as \( \psi_t = e^{-rT}E_t^Q[\xi(S_{t+T})] \), where \( \xi(S_{t+T}) \) is the option payoff at expiration date \( t+T \), \( E_t \) denotes the expected value conditioned on time-\( t \) information, \( Q \) denotes the risk-neutral measure, and \( r \) is the risk-free rate. Because \( x_t \) is Markov, the option price is a function of \( x_t \), plus other parameters. In particular, prices of calls and puts are functions of \( (x_t, S_t, K, T, r, \delta) \), where \( K \) is the strike price and \( \delta \) is the dividend yield. We restrict our attention to the case in which, given \( x_t \), the prices of European call and put options are homogeneous functions of degree one of \( S_t \) and \( K \). As a result, options prices are functions of \( K/S_t \) or, equivalently, of \( K/F_t \), where \( F_t \) is the \( t+T \) futures price of the underlying security. An options pricing function can be formally written as \( \psi_t = f(x_t, K/F_t, T) \).

In the empirical work below, we use the implied volatility of an option from the Black-Scholes formula instead of the option price itself as the dependent
variable. Suppressing the dependence of observations on time $t$, the function we fit is of the form

$$\sigma_t = \sigma(x_t, \ldots, x_{t_k}, K_t, T_t) + \epsilon_t, \quad (1)$$

where $\sigma_t$ is the implied volatility, $K_t/F_t$ is the moneyness, $T_t$ is the time-to-maturity of option $i$, and $\epsilon_t$ is an error term. Given other parameters, the Black-Scholes price of an option is a strictly increasing function of its implied volatility, so using options prices is equivalent to using their implied volatilities.

Theoretically, prices of European options are functions of $S_t$ and $x_t$, so for any set of more than $k + 1$ options, their prices are functionally dependent. Practically, however, the observed options prices differ from their theoretical ones. The differences, $\epsilon_t$, can be measurement errors from the nonsynchronicity of options prices and their underlying and microstructural errors due to the bid-ask bounce and the discrete tick size. Thus, inferring the number of state variables from observed options prices becomes a matter of statistical inference.

We use factors extracted from the implied volatilities to proxy for the state variables $x_t$ and use these factors as independent variables to fit the function $\sigma$ nonparametrically. Having volatility measures on both sides of the equation eases the difficulty in nonparametric fitting and makes the typical homoskedasticity assumption about $\epsilon_t$ more appropriate. Because of put-call parity, the implied volatilities of puts and calls corresponding to the same strike price and maturity are the same. We use out-of-the-money calls and puts only because out-of-the-money options are more liquid.

2.2. Nonlinear Principal Components and Nonparametric Tests

In the existing literature, principal component analysis is used to extract factors and identify the number of factors in the implied volatilities. Fengler et al. (2003), Mixon (2002), Skiadopoulos et al. (1999), and Bondarenko (2007) use PCs to examine the issue for options prices. Cao and Huang (2007) use PCs to examine options returns. There are two major problems with PCs, however. First, there is no unique, universally accepted method to determine the systematic factors when using PCs. Although more than a dozen heuristic stopping rules exist, they often give different results. Second, linear PC analysis assumes that the variables being studied are linearly related to each other. The true relationships among implied volatilities, however, are nonlinear. In the next section, we demonstrate the two problems using data on S&P 500 index options.

In this study, to resolve these issues, we use NPCs as proxies for state variables and conduct a rigorous nonparametric test on the number of state variables in options prices. The NPCs take care of the second problem of the PC approach, and the nonparametric test looks after the first problem. They are briefly explained below.

For a given random $m$ vector $w_t = (w_{1t}, \ldots, w_{mt})$, where the relationships among the components of $w_t$ might be nonlinear, let $\Phi(w_t) = \phi_t = (\phi_{1t}, \ldots, \phi_{mt})$ be a mapping from the $m$-dimensional space of $w_t$ to an $n$-dimensional space of $\phi_t$, where $n$ can be much greater than $m$ and the components of $\phi_t$ consist of linear and nonlinear transformations of $w_t$. The NPCs of $w_t$ associated with the mapping $\Phi$ are simply the PCs of $\phi_t$. By choosing an appropriate mapping $\Phi$, the NPCs capture the most important variations in the data $w_t$ without missing potential nonlinearity. The detailed description of the procedure can be found in Scholkopf et al. (1999).3

We divide S&P 500 index options into groups according to their maturity and moneyness, calculate the average implied volatility for each group, and extract NPCs from these average implied volatilities. We then use the first few NPCs to estimate (1). Let $M_k$ be the model that uses the first $k$ NPCs for $k = 0, 1, 2, 3, 4$. $M_0$ stands for the model that uses no additional state variables. For each model, we calculate the corresponding root mean squared errors, $\text{RMSE}_k$. For $k = 1, 2, 3, 4$, we calculate the partial $R^2_k$, defined as $R^2_k = 1 - \text{RMSE}_k^2/\text{RMSE}_1^2$, which provides a measure of the incremental contribution of the $k$th state variable, given the explanatory power of the first $k-1$ state variables. For the null hypothesis that the $k$th state variable contributes no additional explanatory power, we run bootstrapping tests with data generated from $M_{k-1} + \epsilon_t$ to obtain an empirical distribution of the $R^2_k$ under the null. We also use simulations to verify that our nonparametric test has better performance than the stopping rules in the PC analysis.

3. Data and Preliminary Analysis

3.1. Data Description

We use weekly data of the S&P 500 index call and put options from 1996 to 2005. The options written on the S&P 500 index are the most actively traded European-style contracts, and the S&P 500 index options and the S&P 500 futures options have been the focus of recent empirical options studies. The options data are provided by OptionMetrics. OptionMetrics includes daily closing options prices,

3 In earlier versions of the paper, we also used variance swap prices and average implied volatilities for various maturities to proxy for state variables. The results are qualitatively the same. We appreciate one referee’s suggestion of adopting NPCs as proxies for state variables.
open interests, and trading volume for all U.S. index options and equity options since 1996. For fitting options prices, we use weekly observations on Wednesdays to reduce the burden of nonparametric fitting. Wednesdays are chosen because the number of holidays that fall on Wednesdays is the least out of the five weekdays. If Wednesday is a holiday, we choose Thursday. If both are holidays, we choose Tuesday, and so on. Options with a value of less than $\frac{1}{2}$, or with a time-to-expiration of less than seven days, are excluded for liquidity reasons. Daily interest rate data are obtained from the U.S. Treasury Department’s website (http://www.ustreas.gov/offices/domestic-finance/debt-management/interest-rate/yield.shtml).

To calculate the implied volatilities, we use the forward price of the underlying security inferred from options prices instead of the spot price of the underlying security. By doing this, we reduce the potential errors introduced by estimating the future dividends and the risk-free rate. We use the call-put options pair with the strike price for which the difference between the call and put options prices is the smallest and apply the put-call parity to infer the forward price, $F = (c(K^*) - p(K^*))e^{rt} + K^*$, where $K^* = \arg \min |c(K) - p(K)|$, and $c(K)$ and $p(K)$ are the call and put prices with strike price $K$, respectively.

Table 1 reports the summary statistics of the implied volatilities from the out-of-the-money options, i.e., put options with $K/F \leq 1.00$ and call options with $K/F > 1.00$. The total sample size is 105,574.

We divide the entire sample into six moneyness and six maturity bins, creating 36 intersecting groups in all. Because there are more short-term than long-term options contracts, to balance between missing values and maturity coverage in each group, we set the cutoff points in maturity dimension to 39 days, 69 days, 189 days, 279 days, and 369 days, which correspond to 1-month, 2-month, 6-month, 9-month, and 12-month options, respectively. Table 1 shows the number of options contracts and the average implied volatility in each moneyness and maturity group, and panel B shows the average implied volatility in each moneyness and maturity group.

Table 2 Variance Decomposition of PCs and NPCs

<table>
<thead>
<tr>
<th>Panel A: Variance decomposition of Ps</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_p$</td>
<td>0.0813</td>
<td>0.0048</td>
<td>0.0013</td>
<td>0.0006</td>
<td>0.0003</td>
<td>0.0003</td>
</tr>
<tr>
<td>$V_p/TVP$</td>
<td>0.9080</td>
<td>0.0536</td>
<td>0.0141</td>
<td>0.0067</td>
<td>0.0038</td>
<td>0.0029</td>
</tr>
</tbody>
</table>

Panel B: Variance decomposition of $P^*$

<table>
<thead>
<tr>
<th>$P^*$</th>
<th>0.7228</th>
<th>0.0318</th>
<th>0.0075</th>
<th>0.0038</th>
<th>0.0028</th>
<th>0.0020</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_p/P^*$</td>
<td>0.9289</td>
<td>0.0409</td>
<td>0.0097</td>
<td>0.0049</td>
<td>0.0036</td>
<td>0.0026</td>
</tr>
</tbody>
</table>

Notes. Panel A reports the variance explained by the first six principal components ($V_p$), and the proportion of total variation in the 36 average implied volatilities explained by these principal components ($V_p/TVP$). Panel B reports the variance explained by the first six nonlinear principal components ($V_p^*$), and the proportion of total variation in all the linear and quadratic combinations of the 36 average implied volatilities explained by these nonlinear principal components ($V_p^*/TVP^*$).
groups, the implied volatility exhibits a smile pattern, whereas for the long-maturity groups, the implied volatility exhibits a smirk pattern. On average, the implied volatility of short-maturity options is slightly higher than that of long-maturity options.

From the 36 average implied volatilities, we calculate the PCs. The variance decomposition of the first six PCs is listed in panel A of Table 2. The first four PCs explain 90.8%, 5.36%, 1.41%, and 0.67% of the total variation in the 36 average implied volatilities, respectively.

The eigenvectors associated with the first four PCs are plotted in Figure 1. The interpretation of the PCs can be made from the patterns of eigenvectors. In the upper-left panel, the coefficients of the first eigenvector are all positive with similar magnitude, suggesting that the first PC captures the variation in the overall level of the average implied volatilities. In the upper-right panel for the second PC, the coefficients corresponding to the average implied volatilities of short maturities and large $K/F$ values are positive, whereas those of long maturities and small $K/F$ values are negative, indicating that the second PC captures the variation in the slopes along the maturity and moneyness dimensions. The third PC mainly captures the variation in the average implied volatility of the deep out-of-the-money put group; the fourth one captures the variation in the curvature along the moneyness dimension for the short-maturity average implied volatilities.

The 36 average implied volatilities, $(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{36})$, are mapped to a vector of dimension 702, $(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_{36}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_{35}, \tilde{\sigma}_1, \tilde{\sigma}_2, \ldots, \tilde{\sigma}_{35}, \tilde{\sigma}_{36})$. This is a natural extension of the linear PCs, because it captures the nonlinearity in the implied volatilities using quadratic terms. The NPCs are extracted by applying the PC analysis on the 702 dimensional vector.$^4$

$^4$ In many applications, the dimension of the mapped vector is greater than the number of observations. In that case, a computational technique known as “kernel principal component analysis” can be used, as explained in Scholkopf et al. (1999).
Panel B of Table 2 shows that the first NPC explains 92.89% of the total variation in all the linear and quadratic combinations of the 36 average implied volatilities, which is greater than that of the first PC. The second, third, and fourth NPCs explain 4.09%, 0.97%, and 0.49% of the total variation, respectively. The time-series plots of the first four NPCs are shown in Figure 2. The first NPC exhibits a familiar pattern of the average volatility over the sample period. The second and third NPCs are persistent, and the fourth one has a few spikes.

Figure 3 shows the scatter plots of the first four pairs of NPCs and PCs. The straight line in each panel represents the linear regression fit between the pair of NPC and PC. As seen in the figure, the NPCs and PCs are closely related in general, so NPCs can be interpreted similarly as PCs. Nevertheless, there are significant differences between the two. In particular, the upper-left panel shows a clear nonlinear relationship between the first NPC and the first PC. The nonlinear relationship is also apparent between the third NPC and the third PC.

Let $P_i$ be the $i$th PC and $P_i^*$ be the $i$th NPC. We examine how much the first few $P_i$s can explain a $P_j$, a $P_j^2$, or an $P_j^{*2}$, where $j = 1, 2, 3, 4$. Similarly, we examine how much the first few $P_j$'s can explain a $P_i$, a $P_i^2$, or a $P_i^{*2}$ for $j = 1, 2, 3, 4$. We run regressions

$$P_j^* = b_0 + \sum_{i=1}^{k} b_i P_i + \epsilon_i \quad k = 1, 2, 3, 4 \quad (2)$$

and replace the $P_j^*$ on the left-hand side with a $P_j^2$ or a $P_j^{*2}$ for $j = 1, 2, 3, 4$. We then run regressions

$$P_j = b_0 + \sum_{i=1}^{k} b_i P_i^* + \epsilon_i \quad k = 1, 2, 3, 4, \quad (3)$$

and replace the $P_j$ on the left-hand side with a $P_j^2$ or a $P_j^{*2}$ for $j = 1, 2, 3, 4$. The $R^2$s of these regressions are reported in Table 3. The results show that
the goodness-of-fit of $P_i$s by $P_i^*$'s is as good as the
goodness-of-fit of $P_i^*$'s by $P_i$s. This indicates that PCs
and NPCs can be interpreted similarly in general. However, the goodness-of-fit of the quadratic terms
by $P_i^*$ is much greater than that by $P_i$s. This indicates
that NPCs capture the higher-order variations in the
implied volatility surface better than do PCs.

3.2. Problems with Principal Component Analysis
In this subsection, we empirically show the two prob-
lems involved with using PC analysis to determine
the number of state variables in options pricing,
namely, the inconsistency of stopping rules and the
nonlinearity in implied volatilities. Because of these
two problems, the conclusion drawn from PCs cannot
be relied on.

To illustrate the first problem, we adopt five re-
presentative methods out of more than a dozen
and show that they generate different numbers of
nontrivial PCs for the average implied volatilities.

The detailed descriptions of these methods can be
found in, for example, Jackson (1993) and Peres-
Neto et al. (2005). The PC analysis is known to
have lack of the invariance to scale transformations.
Most methods are based on the correlation matrix,
rather than on the variance matrix.\textsuperscript{5} Let $\lambda_1 \geq \cdots \geq \lambda_n$
be the ordered eigenvalues of the correlation matrix
with $(1/n) \sum \lambda_i = 1$. The simplest method is the
Kaiser-Guttman (KG) method, which determines
the number of factors by $k^* = \max \{k : \lambda_k > 1\}$. The next
method, known as the broken stick (BS) method,
takes a benchmark in which the total variance of
$n$ random variables is randomly divided among
the $n$ variables. In such a case, the proportion of
total variance being explained by the $k$th PC is

\textsuperscript{5}This is not a serious issue for implied volatilities across money-
ness and maturities because they have natural common units and
their variances are more or less similar, so the essential difference
between the variance matrix and the correlation matrix is small.
Table 3 Cross-Explanatory Power of PCs and NPCs

<table>
<thead>
<tr>
<th>Panel A: Explanatory power of PCs</th>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_i^1$</td>
<td></td>
<td>0.9874</td>
<td>0.9875</td>
<td>0.9882</td>
<td>0.9882</td>
</tr>
<tr>
<td>$P_i^2$</td>
<td></td>
<td>0.9817</td>
<td>0.9817</td>
<td>0.9838</td>
<td>0.9838</td>
</tr>
<tr>
<td>$P_i^3$</td>
<td></td>
<td>0.9616</td>
<td>0.9705</td>
<td>0.9376</td>
<td>0.9376</td>
</tr>
<tr>
<td>$P_i^4$</td>
<td></td>
<td>0.1133</td>
<td>0.1452</td>
<td>0.1922</td>
<td>0.1923</td>
</tr>
<tr>
<td>$P_i^5$</td>
<td></td>
<td>0.1843</td>
<td>0.1893</td>
<td>0.3882</td>
<td>0.3882</td>
</tr>
<tr>
<td>$P_i^6$</td>
<td></td>
<td>0.1010</td>
<td>0.1039</td>
<td>0.6310</td>
<td>0.6310</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Explanatory power of NPCs</th>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_i^1$</td>
<td></td>
<td>0.9874</td>
<td>0.9876</td>
<td>0.9929</td>
<td>0.9893</td>
</tr>
<tr>
<td>$P_i^2$</td>
<td></td>
<td>0.9815</td>
<td>0.9816</td>
<td>0.9828</td>
<td>0.9828</td>
</tr>
<tr>
<td>$P_i^3$</td>
<td></td>
<td>0.9609</td>
<td>0.9734</td>
<td>0.9346</td>
<td>0.9346</td>
</tr>
<tr>
<td>$P_i^4$</td>
<td></td>
<td>0.2433</td>
<td>0.3374</td>
<td>0.3494</td>
<td>0.3494</td>
</tr>
<tr>
<td>$P_i^5$</td>
<td></td>
<td>0.1978</td>
<td>0.2153</td>
<td>0.4465</td>
<td>0.4465</td>
</tr>
<tr>
<td>$P_i^6$</td>
<td></td>
<td>0.1026</td>
<td>0.1105</td>
<td>0.6657</td>
<td>0.6657</td>
</tr>
<tr>
<td>$P_i^7$</td>
<td></td>
<td>0.3704</td>
<td>0.3704</td>
<td>0.3883</td>
<td>0.3883</td>
</tr>
<tr>
<td>$P_i^8$</td>
<td></td>
<td>0.2390</td>
<td>0.2347</td>
<td>0.4782</td>
<td>0.4782</td>
</tr>
<tr>
<td>$P_i^9$</td>
<td></td>
<td>0.1817</td>
<td>0.2170</td>
<td>0.6285</td>
<td>0.6285</td>
</tr>
</tbody>
</table>

Notes. This table reports the $R^2$s of the regressions of $P_i^j$s, $P_j^k$s, and $P_j^k$s on $P_i^j=(1, P_1, \ldots, P_n)$ in panel A and the $R^2$s of the regressions of $P_i^j$s, $P_j^k$s, and $P_j^k$s on $P_i^j=(1, \ldots, P_n)$ in panel B, where $P_i^j$ and $P_j^k$ indicate the $j$th principal component and nonlinear principal component, respectively. The first column of panels A and B indicates the dependent variable in the regressions.

$\Delta \lambda_i^{nb}$ is the $m$th percentile of the empirical distribution of the $k$th eigenvalue from the 100 randomized samples, $\lambda_i^{nb}$, and their differences, $\Delta \lambda_i^{nb}$, and $nb_i=\sum_{i=1}^{n}(1/i)$.

Table 4 Tests of the Number of Nontrivial Principal Components

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_k$</td>
<td>32.7587</td>
<td>2.0125</td>
<td>0.3731</td>
<td>0.2274</td>
<td>0.1267</td>
<td>0.1078</td>
</tr>
<tr>
<td>$\Delta \lambda_k$</td>
<td>30.7462</td>
<td>1.6393</td>
<td>0.1458</td>
<td>0.1006</td>
<td>0.0189</td>
<td>0.0210</td>
</tr>
<tr>
<td>$\lambda_k^{50}$</td>
<td>30.0622</td>
<td>1.7607</td>
<td>0.3324</td>
<td>0.1420</td>
<td>0.1145</td>
<td>0.0940</td>
</tr>
<tr>
<td>$\lambda_k^{95}$</td>
<td>35.6238</td>
<td>2.2452</td>
<td>0.4312</td>
<td>0.3335</td>
<td>0.1446</td>
<td>0.1172</td>
</tr>
<tr>
<td>$\Delta \lambda_k^{50}$</td>
<td>1.0666</td>
<td>1.5289</td>
<td>1.4646</td>
<td>1.4256</td>
<td>1.3867</td>
<td>1.3522</td>
</tr>
<tr>
<td>$\Delta \lambda_k^{95}$</td>
<td>0.1113</td>
<td>0.0909</td>
<td>0.0846</td>
<td>0.0746</td>
<td>0.0733</td>
<td>0.0653</td>
</tr>
</tbody>
</table>

Method KG BS BE RE RED

Notes. This table reports the test results of five methods in determining the number of nontrivial PCs of the 36 average implied volatilities. The five methods are the KG method, the BS method, the BE method, the RED method, and the BE method. The test results are based on the first six eigenvalues, $\lambda_i$, and the eigenvalue differences, $\Delta \lambda_i$, of the correlation matrix of the original data, the 5th and 95th percentiles of the first six eigenvalues from the correlation matrices of 100 bootstrapped samples, $\lambda_i^{50}$ and $\lambda_i^{95}$, the 95th percentile of the first six eigenvalues from the correlation matrices of 100 randomized samples, $\lambda_i^{nb}$, and their differences, $\Delta \lambda_i^{nb}$, and $nb_i=\sum_{i=1}^{n}(1/i)$.

According to the randomization method based on eigenvalues (RE), $k^* = \max\{k: \lambda_k > \lambda_k^{95}\}$, where $\lambda_k^{m}$ is the $m$th percentile of the empirical distribution of the $k$th eigenvalue from the 100 randomized samples. According to the randomization method based on the eigenvalue differences (RED), $k^* = \max\{k: \Delta \lambda_k > \Delta \lambda_k^{95}\}$, where $\Delta \lambda_k = \lambda_k - \lambda_{k+1}$ and $\Delta \lambda_i^{95}$ is the $m$th percentile of the empirical distribution of $\lambda_k - \lambda_{k+1}$. Of the five methods, KG, BS, and RE are based on the value of each eigenvalue relative to certain benchmarks, and BE and RED are based on the value of each eigenvalue relative to the next eigenvalue.

We apply the five methods to the 36 average implied volatilities. Table 4 shows the first six eigenvalues, $\lambda_i$, and the eigenvalue differences, $\Delta \lambda_i$, of the correlation matrix of the original data, the 5th and 95th percentiles of the first six eigenvalues from the correlation matrices of 100 bootstrapped samples, $\lambda_i^{50}$ and $\lambda_i^{95}$, the 95th percentile of the first six eigenvalues from the correlation matrices of 100 randomized samples, $\lambda_i^{nb}$, and their differences, $\Delta \lambda_i^{nb}$. Also reported are $nb_i$s for $k=1, \ldots, 6$, which do not depend on the data. The KG method gives $k^* = 2$ because $\Delta \lambda_2 = 2.0125 > 1$, and $\lambda_3 = 0.3731 \leq 1$. The BS method gives $k^* = 1$ because $\lambda_1 = 32.7587 > 36b_1 = 4.1746$, whereas $\lambda_1 = 2.0125 < 36b_2 = 3.1746$. The BE method gives $k^* = 2$ because $(\lambda_5^{50} - \lambda_5^{95}) = (1.7607, 2.2452)$ does not overlap with $(\lambda_5^{50} - \lambda_5^{95}) = (0.3324, 0.4312)$, but the latter overlaps with $(\lambda_4^{50} - \lambda_4^{95}) = (0.1420, 0.3335)$. The RE method gives $k^* = 2$ because $\lambda_2 = 2.0125 > \lambda_5^{95} = 1.5289$, whereas $\lambda_3 = 0.3731 < \lambda_5^{95} = 1.4646$. The RED method gives $k^* = 4$ because $\Delta \lambda_4 = 0.1006 > \lambda_4^{95} = 0.0746$, and $\Delta \lambda_5 = 0.0189 < \lambda_4^{95} = 0.0733$. The five methods give three different answers.

\[b_i = (1/n) \sum_{i=1}^{n}(1/i)^6\]

The BS method determines $k^*$ by $k^* = \max\{k: \lambda_k > \lambda_k^{95}\}$, i.e., $k^*$ is chosen such that the 90% confidence interval of $\lambda_k$ does not overlap with that of $\lambda_{k+1}$. The last two methods are randomization methods in which the observations of each variable are resampled independently across variables, so that for the reshuffled data, the variances of the variables remain the same, but the covariances among the variables are randomized. The eigenvalues are calculated for the reshuffled data. The procedure is repeated, say, 100 times.

\[\Delta \lambda_i^{nb} = \sum_{i=1}^{n}(1/i)\]

For a line of unit length divided randomly by $n-1$ points, the expected length of the $k$th longest interval is $(1/n) \sum_{i=1}^{n}(1/i)$.
To illustrate the second problem related to nonlinearity, we use a Ramsey (1969) type of the regression equation specification error test (RESET) to examine the linear relationships among the implied volatilities. For a given $k$, we run a linear regression to estimate

$$
\tilde{\sigma}_j = \alpha_{j,0} + \sum_{i=1}^{k} \alpha_{j,i} P_i + \eta_i,
$$

where $\tilde{\sigma}_j$ is the weekly observations of one of the 36 average implied volatilities and $P_i$ is the $i$th PC extracted from the average implied volatilities. The fitted value of this regression is denoted as $\hat{\sigma}_j$. We then estimate

$$
\hat{\sigma}_j = \beta_{j,0} + \sum_{i=1}^{k} \beta_{j,i} P_i + \sum_{i=1}^{q} \gamma_i, \hat{\sigma}_j^{1+i} + \xi_j
$$

and test the hypothesis that $(\gamma_{j,1}, \ldots, \gamma_{j,q}) = 0$. Under the assumption of linear relationships among all the average implied volatilities, their PCs, which are their linear transformations, capture the important variations in the average implied volatilities, but their higher-order power transformations should be useless.

Table 5 reports the RESET results for the 36 average implied volatilities and their first four PCs. The choice of the first four PCs is based on the results in Table 4. We consider $q = 1$, 2, and 3, where the maximum of $q = 3$ is suggested by Ramsey (1969). The numbers in the table are the frequency of rejection of the hypothesis $(\gamma_{j,1}, \ldots, \gamma_{j,q}) = 0$ under various significance levels. The results show that the rejection rates are very high, indicating the linearity assumption that underlies the PC analysis is false for the 36 average implied volatilities.

4. Number of State Variables: A Nonparametric Approach

In this section, we conduct a rigorous analysis of the number of state variables using a nonparametric approach. We fit the implied volatility surface using the NPCs and adopt a bootstrap procedure to test the number of state variables in the time series and cross sections of the options prices. We also compare the performance of the bootstrap test with the PC stopping rules using simulation.

4.1. Nonparametric Estimation

The function $\sigma$ is fitted nonparametrically with the local linear regression. The local linear regression has advantages over the classical Nadaraya-Watson kernel regression used in earlier papers, such as Ait-Sahalia and Lo (1998). The Nadaraya-Watson estimator is biased for the points at the boundary region, while the boundary effect is absent for the local linear regression. This is useful especially when dealing with multidimensional situations, as the boundary can be quite substantial in terms of the number of data points involved. The superior boundary behavior of the local linear regression is important for our case, as we have at most six regressors. Recent applications of local linear/polynomial regression to options pricing include Ait-Sahalia and Duarte (2003) and Li and Zhao (2009).

Let $(P_{1i}^*, \ldots, P_{ki}^*)$ be the first $k$ NPCs for $k = 0, \ldots, 4$, where the case of $k = 0$ stands for an empty set. The implied volatility $\sigma$ is fitted as a function of $(P_{1i}^*, \ldots, P_{ki}^*, K/F, T)$. The local linear regression estimator of $\sigma(P_{1i}^*, \ldots, P_{ki}^*, K/F, T)$ is the first element of $(\alpha, \beta_1, \ldots, \beta_{k+2})$ that minimizes

$$
\begin{align*}
&\sum_{i=1}^{N} \left[ \sigma_i - \alpha - \sum_{j=1}^{k} \beta_j (P_{ji}^* - P_{ji}^0) \\
&- \beta_{k+1} (K_i/F_i - K/F) - \beta_{k+2} (T_i - T) \right]^2 \\
&\cdot \prod_{j=1}^{k} \frac{1}{h_{P_j}^0} \left( \frac{P_{ji}^* - P_{ji}^0}{h_{P_j}^0} \right) \frac{1}{h_{K/F}} \left( \frac{K_i/F_i - K/F}{h_{K/F}} \right) \frac{1}{h_T} \left( \frac{T_i - T}{h_T} \right),
\end{align*}
$$

where $G(\cdot)$ is a kernel function, $h_w$ is the bandwidth for an explanatory variable $w$, and $N$ is the number of observations. We choose the second order Gaussian kernel $G(z) = (1/\sqrt{2\pi})e^{-z^2/2}$, which is commonly used in the literature. It is important to choose

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Nonlinearity Among Implied Volatilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q = 1$</td>
<td>$q = 2$</td>
</tr>
<tr>
<td>$k$</td>
<td>$10%$</td>
</tr>
<tr>
<td>1</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>0.67</td>
</tr>
<tr>
<td>4</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Notes. This table presents the results of the RESET for the weekly average implied volatilities, $\tilde{\sigma}_j$, which are first regressed on their principal components, $P_i$, up to the first four:

$$
\tilde{\sigma}_j = \alpha_{j,0} + \sum_{i=1}^{k} \alpha_{j,i} P_i + \eta_i.
$$

They are then reestimated in

$$
\hat{\sigma}_j = \beta_{j,0} + \sum_{i=1}^{k} \beta_{j,i} P_i + \sum_{i=1}^{q} \gamma_i, \hat{\sigma}_j^{1+i} + \xi_j,
$$

where $\hat{\sigma}_j$ is the fitted value from the previous regression. The RESET examines the hypothesis that $(\gamma_{j,1}, \ldots, \gamma_{j,q}) = 0$. For a given $k$ and a given $q$, the rejection rates out of the total of 36 models with different average implied volatilities as dependent variables at 10%, 5%, and 1% significance levels are reported.
a proper bandwidth in the nonparametric studies. We adopt an optimal plug-in bandwidth selection rule that has been shown to exhibit superior asymptotic and practical performance to the cross-validation method. Because the results of nonparametric analysis may sometimes vary with the choice of bandwidth, we also consider an oversmoothed bandwidth, $h_{rs}$, which is 125% of the optimal bandwidth, and an undersmoothed bandwidth, $h_{us}$, which is 80% of the optimal bandwidth, for robustness checks.

The models with various numbers of NPCs as proxies for state variables are labeled as $M_k(P_1, \ldots, P_t)$, where $k$ is the number of state variables (in addition to the price of the underlying security). We calculate the root mean squared error for the model $M_k$, RMSE$_k$, to gauge its absolute performance and calculate partial $R^2_k$, defined as

$$R^2_k = 1 - \frac{\text{RMSE}_{k}^2}{\text{RMSE}_{k-1}^2}, \quad (7)$$

to gauge the improvement on $M_{k-1}$ from the $k$th state variable. Similar to the adjusted $R^2$ in a linear regression, the partial $R^2$ used here penalizes more explanatory variables with a larger optimal bandwidth, which may give a larger RMSE for the model with more explanatory variables. As a result, $R^2_k$ is not guaranteed to be positive.

Panel A of Table 6 reports the RMSEs for models $M_0$ to $M_4$ and the corresponding $R^2$s. The results show that the RMSE drops from 0.0457 for $M_0$ to 0.0125 for $M_4$ when $P_1$ is included in the model. The $R^2$ is 0.9255. When the second NPC, $P_2$, enters the model as the second state variable, the RMSE further drops to 0.0066 in $M_4$. The second state variable explains an additional 72.16% of the variation that $M_4$ fails to explain. When $P_3$ enters the model as the third state variable, the RMSE decreases only marginally. The $R^2$ for the third state variable is 0.0645. Finally, the $R^2$ for $M_4$ is $-0.1182$, which suggests that the nonparametric fit is actually worse when the fourth NPC is included in the model. The negative $R^2$ indicates that the improvement in fit with the additional variable is dominated by the penalty imposed for a larger bandwidth. Overall, the RMSE statistics suggest that the NPCs beyond the first two do not improve the fit significantly.

Christoffersen et al. (2009) estimate parametric stochastic volatility models using S&P 500 index options from 1990 to 2004. They report that the RMSEs for a one-factor and a two-factor stochastic volatility model are 0.0199 and 0.0151, respectively. These numbers can be compared with the RMSEs of $M_4$ and $M_2$ reported here. Because we use a nonparametric function for fitting, the RMSEs are naturally smaller. Nevertheless, in both papers, the importance of the second factor is demonstrated.

These results suggest that the models with only one state variable in addition to the underlying price are inadequate, even though jump components can be added to the underlying security price and stochastic volatility process with jump intensity and jump size determined by the same state variable. The results also suggest that improvements to the existing two-factor models in the literature can be achieved by considering more flexible functions that govern the evolution of the underlying price and state variables, but not by adding more state variables.

Table 6 Testing the Number of State Variables

<table>
<thead>
<tr>
<th>Unrestricted model</th>
<th>RMSE</th>
<th>Restricted model</th>
<th>$R^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Optimal bandwidth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_0$</td>
<td>0.0457</td>
<td>$M_0$</td>
<td>0.9255</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_1$</td>
<td>0.0125</td>
<td>$M_1$</td>
<td>0.7216</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_2$</td>
<td>0.0066</td>
<td>$M_2$</td>
<td>0.0645</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_3$</td>
<td>0.0064</td>
<td>$M_3$</td>
<td>$-0.1182$</td>
<td>0.18</td>
</tr>
<tr>
<td>$M_4$</td>
<td>0.0067</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panel B: Oversmoothed unrestricted model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_1^*$</td>
<td>0.0127</td>
<td>$M_1^*$</td>
<td>0.9233</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_2^*$</td>
<td>0.0068</td>
<td>$M_2^*$</td>
<td>0.7024</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_3^*$</td>
<td>0.0065</td>
<td>$M_3^*$</td>
<td>0.0346</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_4^*$</td>
<td>0.0068</td>
<td>$M_4^*$</td>
<td>$-0.1305$</td>
<td>0.15</td>
</tr>
<tr>
<td>Panel C: Undersmoothed unrestricted model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M_1^*$</td>
<td>0.0123</td>
<td>$M_1^*$</td>
<td>0.9279</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_2^*$</td>
<td>0.0064</td>
<td>$M_2^*$</td>
<td>0.7415</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_3^*$</td>
<td>0.0063</td>
<td>$M_3^*$</td>
<td>0.0968</td>
<td>0.00</td>
</tr>
<tr>
<td>$M_4^*$</td>
<td>0.0067</td>
<td>$M_4^*$</td>
<td>$-0.0985$</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Notes. This table reports the test results on the number of state variables. $M_0$ is the model with the underlying price as the only state variable; $M_1$, $M_2$, $M_3$, and $M_4$ represent the models with one, two, three, and four additional state variables, respectively. Panel A is based on the optimal bandwidth. Panel B is based on the oversmoothed unrestricted models, indicated by a superscript *. Panel C is based on the undersmoothed unrestricted models, indicated by a superscript -. Column one indicates the more general models. Column two reports the RMSE. Column three indicates the restricted models, which are tested against the more general models. Column four is the $R^2$, defined as $1 - \text{RMSE}^2/\text{RMSE}^2_{*}$, which measures the improvement of the more general models $M_k$ over the restricted models $M_{k-1}$. The last column reports the bootstrap p-values.

The performance of each model is also illustrated graphically. Figure 4 shows the weekly average residual (the solid line) and the 5th to 95th percentile of the residual distribution (the shaded area) for each of the models. For the model $M_4$ in the upper-left panel, the average residual exhibits a persistent pattern and resembles the time-series plot of $P_1^t$. The 5th
and 95th percentiles of the distribution cover a rather wide range. The pattern of residuals indicates that there are omitted state variables in the model. There is a clear contrast between the upper-left panel and the upper-right panel, in which the residuals from the model $M_1$ are presented. In the upper-right panel, the variation in the average residual is much less pronounced, and the distribution of the residuals is much narrower than that in the upper-left panel. This suggests that the model with one state variable in addition to the underlying security price is significantly better than the model with just the underlying security price. The first NPC captures the variation of the average implied volatility across time. The lower-left panel shows the residuals of the model $M_2$. By comparing the upper-right panel and the lower-left panel, we can see that the overall fitting is further improved. The average residual from the model $M_2$ does not show any systematic deviation from zero. The residuals of the model $M_3$ are shown in the lower-right panel. Consistent with the RMSE statistic, the overall fitting only improves marginally for $M_3$.

The weekly RMSE of each model is plotted in Figure 5. Note the difference in scale for each panel. The average RMSE for $M_0$ is 0.0398 and the highest value is 0.1569, whereas the average implied volatility for the entire sample is merely 0.2288. The average RMSE for $M_1$ is 0.0109 and the majority of the weekly RMSEs are below 0.03. The average RMSE further drops to 0.0059 for $M_2$ and to 0.0056 for $M_3$. For these two models, the majority of the weekly RMSEs are below 0.02. Overall, the results suggest that a model with two state variables in addition to the underlying security price captures almost all the systematic variation in the options prices both cross-sectionally and over time.

The residuals across maturity for each model are shown in Figure 6. The residuals are divided into 50 bins across maturity, and each bin contains the same number of observations. The average residual, the 5th percentile, and the 95th percentile of
the residual distribution for each bin are plotted against the average maturity $T$ in years for that bin. For the model $M_0$ in the upper-left panel, the 90% confidence band covered by the 5th and 95th percentiles of the residual distribution is wide, implying large variations in implied volatilities for all maturities. The confidence band in the upper-right panel is much narrower, which indicates that the model $M_1$ with the first NPC as the additional state variable explains the variations in both short- and long-maturity implied volatilities. In the lower-left panel, the residuals from the model $M_2$ with the first two NPCs as additional state variables show that the fitting is further improved for all maturities. Interestingly, the improvement is mostly from long maturities, which suggests that the second NPC captures the variation mainly from long-maturity implied volatilities. The improvement of $M_3$ over $M_2$ is not significant along the maturity dimension.

The residuals across moneyness for each model are shown in Figure 7. Similar to the case of Figure 6, the residuals are divided evenly into 50 bins across moneyness. The average residual, the 5th percentile, and the 95th percentile of the residual distribution for each bin are plotted against the average $K/F$. For the model $M_0$ in the upper-left panel, the 90% confidence band of residuals is wide, especially for low strike options, i.e., out-of-the-money puts. The average residual does not exhibit the smile or smirk pattern because $K/F$ is controlled for in the model $M_0$. The 90% confidence band is much narrower for the model $M_1$ and further reduced for the model $M_2$. The model $M_3$ does not show significant improvement in fitting over $M_2$ along the moneyness dimension.

4.2. Nonparametric Testing

We use a bootstrap procedure to test the null hypothesis nonparametrically:

$$H_0: \sigma(P^r_1, \ldots, P^r_{k-1}, K/F, T) = \sigma(P^a_1, \ldots, P^a_k, K/F, T)$$

against the alternative

$$H_1: \sigma(P^r_1, \ldots, P^r_{k-1}, K/F, T) \neq \sigma(P^a_1, \ldots, P^a_k, K/F, T),$$

where $P^r$ and $P^a$ denote the realized and estimated implied volatilities, respectively.
where $\sigma(P_1^*, \ldots, P_{k-1}^*, K/F, T)$ is the restricted model and $\sigma(P_1^*, \ldots, P_k^*, K/F, T)$ is the unrestricted model. We start with $k = 1$ and increase the number of state variables until the null hypothesis cannot be rejected.

A few nonparametric tests of omitted variables in the literature do not rely on simulation, including the ones in Fan and Li (1996), Lavergne and Vuong (2000), Ait-Sahalia et al. (2001), and Christoffersen and Hahn (1999). These papers develop test statistics based on the fitted residuals from restricted and unrestricted models. Under the null hypothesis, the statistic is asymptotically distributed as a standard normal variate under certain regularity conditions. These tests work well when the number of explanatory variables is small, but the performance worsens when the number increases. In addition, the result is sensitive to the bandwidth choice for both the restricted and the unrestricted models in the nonparametric regression.

To avoid potential problems arising from such tests, we use the so-called two-point wild bootstrap procedure for the nonparametric testing. Applications of the bootstrap approach in various nonparametric studies have demonstrated its good finite sample properties. See, for example, Li and Wang (1998) and Li and Racine (2007, Chap. 12). The bootstrap procedure is implemented as follows. We fit the implied volatility, $\sigma_i$, using the restricted model and calculate the residuals from the model as

$$
\epsilon_i = \sigma_i - \hat{\sigma}_i(P_1^*, \ldots, P_{k-1}^*, K_i/F_i, T_{i}).
$$

Using the residuals from the restricted model, we construct the bootstrapped samples under the null hypothesis,

$$
\sigma^*_i = \hat{\sigma}_i(P_1^*, \ldots, P_{k-1}^*, K_i/F_i, T_{i}) + \epsilon^*_i,
$$

where $\epsilon^*_i = (1 - \sqrt{5})/2 \epsilon_i$ with probability $(1 + \sqrt{5})/2\sqrt{5}$ and $\epsilon^*_i = (1 + \sqrt{5})/2 \epsilon_i$ with probability $(-1 + \sqrt{5})/2\sqrt{5}$. Such probabilities and weightings
lead to the following property for the new errors: $E^*(\epsilon^*_i) = 0$, $E^*(\epsilon^2_i) = \epsilon^2_i$, and $E^*(\epsilon^3_i) = \epsilon^3_i$, where $E^*$ indicates the expected value in the simulation. We calculate the $R^2_k$ from many bootstrapped samples and construct the empirical distribution of $R^2_k$ under the null hypothesis. By comparing the $R^2_k$ from the original sample with the empirical distribution of $R^2_k$ from the bootstrapped samples, we are able to make inferences about the validity of the null hypothesis. When the $R^2_k$ calculated from the original sample is higher than 95% of the empirical distribution of $R^2_k$ from the bootstrap samples, the bootstrap test rejects the null hypothesis at the 5% significance level and the bootstrap $p$-value is 5%.

The last column in Table 6 reports the $p$-values for the bootstrap test of the number of state variables in the options prices. The $p$-value is calculated from 100 bootstrapped samples. Not surprisingly, the model with one additional state variable significantly improves upon the model with the underlying security price as the only state variable. The bootstrap $p$-value suggests that the second state variable further improves the one-factor model. The bootstrap test also rejects the two-factor model in favor of the three-factor model. However, the average of the absolute difference in the fitted values between the two-factor and three-factor models is only 0.003, which accounts for at most a half bid-ask spread of options with various moneyness and maturities. Therefore, the improvement of the three-factor model over the two-factor model is not economically significant.

We also examine the robustness of the results of the bandwidth choice. We consider an over- and under-smoothed case. Specifically, we use 125% and 80% of the optimal bandwidth for the additional variable in the unrestricted models and test against the restricted models fitted using the optimal bandwidth. The results are reported in panel B of Table 6 for the oversmoothed case and in panel C of Table 6 for the undersmoothed case. As expected, the over-smoothed
(undersmoothed) unrestricted models give higher (lower) RMSEs, and lower (higher) $R^2$s, and $p$-values change only slightly. This is because the distributions of $R^2$s from the bootstrapped samples under the null are shifted to the left (right) for the oversmoothed (undersmoothed) case accordingly. The results suggest that the nonparametric test is insensitive to different bandwidth choices. The robustness of the test is also confirmed in the simulation analysis in the following subsection.

4.3. Simulation Analysis

In this subsection, we compare the performance of the stopping rules of PC analysis with that of the proposed nonparametric approach using simulated data. We simulate the following two-factor stochastic volatility process,

$$dS_t = rS_t dt + \sqrt{\sigma_{11}^2} S_t d\omega_{t1} + \sqrt{\sigma_{22}^2} S_t d\omega_{t2},$$

$$dx_{11} = \kappa_1 (\theta_1 - x_{11}) dt + \sigma_{11} \sqrt{\sigma_{11}^2} d\omega_{11},$$

$$dx_{22} = \kappa_2 (\theta_2 - x_{22}) dt + \sigma_{22} \sqrt{\sigma_{22}^2} d\omega_{22},$$

where $\omega_{t1}$ and $\omega_{t2}$ are uncorrelated, $\omega_{t1}$ has a correlation $\rho_1$ with $\omega_{11}^t$, $\omega_{t2}$ has a correlation $\rho_2$ with $\omega_{22}^t$, $\omega_{11}^t$, and $\omega_{22}^t$ are uncorrelated, and $\omega_{t1}$ and $\omega_{t2}^t$ are uncorrelated. We set the parameters for the volatility process to those estimated by Christoffersen et al. (2009), because they estimate such a two-factor stochastic volatility model using the S&P 500 index options data from a sample period similar to ours. In particular, $\kappa_1 = 0.179$, $\theta_1 = 0.007$, $\sigma_{11} = 3.667$, $\rho_1 = -0.957$, $\kappa_2 = 1.303$, $\theta_2 = 0.114$, $\sigma_{22} = 0.280$, and $\rho_2 = -0.834$.

The risk-free rate $r$ is assumed to be 0.05. To minimize the adverse effect of the discretization of the continuous-time model, we simulate daily data. The sample period is 10 years, which is the same as the actual sample. Then a sample of weekly data is selected, and for each week we calculate the option prices and implied volatilities for six different maturities of 1, 2, 6, 9, 12, and 18 months, and six different moneyness of 0.925, 0.955, 0.985, 1.015, 1.045, and 1.075, which correspond to the weekly average implied volatilities of the actual sample. We add random errors to implied volatilities to reflect the actual situation where implied volatilities are measured with errors. The random errors are drawn from the normal distribution with a mean of 0 and a standard deviation of 0.0151. We simulate 100 such samples.

The performance of stopping rules on the simulated data is shown in panel A of Table 7. The KG method identifies 68% of the simulated samples correctly, and the remaining 32% of the cases are underidentified as one factor. The BS method identifies only one factor from all the 100 simulated samples, whereas the true number of factors is two. The BE method tends to overidentify the number of factors, with 52% of the cases correctly identified. The RE method tends to underidentify the number of factors, with only 15% of the cases correctly identified. The RED method performs relatively better, identifying 84% of the cases correctly, with the remaining 16% overidentified as three factors. These stopping rules generally lead to incorrect conclusions and give inconsistent results in the simulated samples, the same as in the actual sample.

The nonparametric test is applied to the same simulated data sets used in the analysis of stopping rules. The $p$-values of the tests are calculated from 100 bootstrapped samples. The significance level of the tests is 5%. The performance of the nonparametric test is shown in panel B of Table 7. To gauge the sensitivity of the test performance to the bandwidth choice, results for different bandwidths are compared. The results show that the performance of the nonparametric test is quite good, with at least 94% of the cases correctly identified. In addition, the results are not sensitive to the bandwidth choice.

4.4. Out-of-Sample Analysis

We conduct an out-of-sample analysis in this subsection. For a given year in the 10-year sample period
Table 8  Out-of-Sample Performance

<table>
<thead>
<tr>
<th>Year</th>
<th>$M_0$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>0.0522</td>
<td>0.0121</td>
<td>0.0115</td>
<td>0.0110</td>
<td>0.9468</td>
<td>0.1073</td>
</tr>
<tr>
<td>1997</td>
<td>0.0349</td>
<td>0.0190</td>
<td>0.0148</td>
<td>0.0121</td>
<td>0.7040</td>
<td>0.3880</td>
</tr>
<tr>
<td>1998</td>
<td>0.0598</td>
<td>0.0282</td>
<td>0.0161</td>
<td>0.0170</td>
<td>0.7775</td>
<td>0.6760</td>
</tr>
<tr>
<td>1999</td>
<td>0.0561</td>
<td>0.0270</td>
<td>0.0170</td>
<td>0.0194</td>
<td>0.7682</td>
<td>0.6037</td>
</tr>
<tr>
<td>2000</td>
<td>0.0301</td>
<td>0.0146</td>
<td>0.0129</td>
<td>0.0140</td>
<td>0.7860</td>
<td>0.2262</td>
</tr>
<tr>
<td>2001</td>
<td>0.0341</td>
<td>0.0156</td>
<td>0.0108</td>
<td>0.0107</td>
<td>0.7891</td>
<td>0.5224</td>
</tr>
<tr>
<td>2002</td>
<td>0.0550</td>
<td>0.0224</td>
<td>0.0126</td>
<td>0.0146</td>
<td>0.8336</td>
<td>0.6850</td>
</tr>
<tr>
<td>2003</td>
<td>0.0363</td>
<td>0.0133</td>
<td>0.0112</td>
<td>0.0092</td>
<td>0.8660</td>
<td>0.2945</td>
</tr>
<tr>
<td>2004</td>
<td>0.0483</td>
<td>0.0125</td>
<td>0.0087</td>
<td>0.0077</td>
<td>0.9332</td>
<td>0.5165</td>
</tr>
<tr>
<td>2005</td>
<td>0.0743</td>
<td>0.0119</td>
<td>0.0091</td>
<td>0.0080</td>
<td>0.9743</td>
<td>0.4140</td>
</tr>
<tr>
<td>Overall</td>
<td>0.0502</td>
<td>0.0189</td>
<td>0.0128</td>
<td>0.0130</td>
<td>0.8583</td>
<td>0.5415</td>
</tr>
</tbody>
</table>

Notes. This table reports the out-of-sample RMSE and $R^2$ for different models, where $R^2$ for model $M_1$ is defined as $1 - \text{RMSE}_{M_1}^2/\text{RMSE}_{M_1}^2$. $M_1$, $M_2$, and $M_3$ represent the models with one, two, and three additional state variables, respectively. For each year, the implied volatilities are estimated using the relationship fitted with data for the other nine years.

and a model of a given number of state variables, we use the data from the other 9 years to fit a non-parametric function and use the fitted nonparametric function to price options for that given year. Such a procedure is rotated for each year in the sample period.

Table 8 reports the out-of-sample RMSEs and $R^2$s for each of the 10 years in 1996–2005, labeled as overall. The RMSEs for $M_0$ vary across the years, indicating that the effect of stochastic volatility on options pricing is time varying. The variations in RMSEs for $M_1$ and $M_2$ are considerably smaller, although the out-of-sample errors are 1.5 times and 2 times as large as the in-sample counterparts for $M_1$ and $M_2$, respectively. The pattern of $R^2$s is similar to that of the in-sample analysis, except that for certain years the incremental contribution of the third NFC is higher. The overall results are consistent with the in-sample analysis. The RMSE drops from 0.0502 for $M_0$ to 0.0189 for $M_1$; it further drops to 0.0128 for $M_2$, representing $R^2$s of 0.8583 for $M_1$ and 0.5415 for $M_2$, respectively. However, the out-of-sample fit of $M_3$ is worse than that of $M_2$ on average. The results indicate that options pricing models with two state variables in addition to the underlying security price are adequate for pricing options out-of-sample.

Table 9 provides further details of the out-of-sample analysis. Panels A, B, and C report the overall out-of-sample RMSEs and $R^2$s of the 10-year sample for each of $M_1$, $M_2$, and $M_3$ for moneyness and maturity groups. The $R^2$s for $M_1$ are greater than 0.7 for all the moneyness and maturity groups. The $R^2$s for $M_2$ are all greater than 0. However, the improvement of $M_2$ over $M_1$ is marginal for low-strike and short-to-medium maturity groups. The $R^2$s for $M_3$ are negative for the majority of the groups. The improvement of $M_3$ over $M_2$ is large only for the low-strike and short maturity group. The results show that the options pricing model with two state variables in addition to the underlying security price is a good choice.

Table 9  Out-of-Sample Performance for Moneyness and Maturity Groups

<table>
<thead>
<tr>
<th>RMSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 2m</td>
<td>3 – 9m</td>
</tr>
<tr>
<td>Panel A: $M_1$</td>
<td></td>
</tr>
<tr>
<td>$K/F \leq 0.97$</td>
<td>0.0210</td>
</tr>
<tr>
<td>$0.97 &lt; K/F \leq 1.03$</td>
<td>0.0195</td>
</tr>
<tr>
<td>$1.03 &lt; K/F$</td>
<td>0.0266</td>
</tr>
<tr>
<td>Subtotal</td>
<td>0.0220</td>
</tr>
<tr>
<td>Panel B: $M_2$</td>
<td></td>
</tr>
<tr>
<td>$K/F \leq 0.97$</td>
<td>0.0196</td>
</tr>
<tr>
<td>$0.97 &lt; K/F \leq 1.03$</td>
<td>0.0105</td>
</tr>
<tr>
<td>$1.03 &lt; K/F$</td>
<td>0.0133</td>
</tr>
<tr>
<td>Subtotal</td>
<td>0.0159</td>
</tr>
<tr>
<td>Panel C: $M_3$</td>
<td></td>
</tr>
<tr>
<td>$K/F \leq 0.97$</td>
<td>0.0162</td>
</tr>
<tr>
<td>$0.97 &lt; K/F \leq 1.03$</td>
<td>0.0121</td>
</tr>
<tr>
<td>$1.03 &lt; K/F$</td>
<td>0.0135</td>
</tr>
<tr>
<td>Subtotal</td>
<td>0.0145</td>
</tr>
</tbody>
</table>

Notes. This table reports the out-of-sample RMSE and $R^2$ for different models, where $R^2$ for model $M_1$ is defined as $1 - \text{RMSE}_{M_1}^2/\text{RMSE}_{M_1}^2$. $M_1$, $M_2$, and $M_3$ represent the models with one, two, and three state variables in addition to the underlying security price, respectively. For each year, the implied volatilities are estimated using the relationship fitted with data for the other nine years. Panels A–C report the overall RMSE and $R^2$ of the 10-year sample for $M_1$, $M_2$, and $M_3$ for different moneyness and maturity groups, respectively.
because introducing the third state variable would merely overfit the data in-sample without improving its out-of-sample performance.

5. Conclusion
This paper makes methodological contributions to the literature on determining the number of state variables required for options pricing. We show that PC analysis is not a reliable tool for this purpose because the various methods in PC analysis often give different results, and the average implied volatilities across different moneyness and maturity groups are non-linearly related. We use a nonparametric approach that avoids misspecifications together with the use of NPCs as proxies for the underlying state variables. The methodology is applied to the options on the S&P 500 index. The results indicate that the first two NPCs capture the dynamics of the state variables in the options prices. The model with the two state variables is able to eliminate most of the persistent pricing errors.

The nonparametric approach of the paper complements the recent growing literature on developing parametric multifactor options pricing models. Our results suggest that, for S&P 500 options, adding jumps to the one-factor model (in addition to the underlying price) with jump intensity and jump sizes depending on the same factor is not enough to resolve the problem in fitting options prices. On the other hand, extending the volatility process to higher dimensions than two is of little use either. Rather, improving on the specification of the two-factor volatility process is a promising direction in modeling options prices. Our methodology can be applied to options of other underlying securities to determine the number of state variables necessary for options pricing.

Acknowledgments
The authors thank Charles Cao, Jinchuan Duan, Nengjiu Ju, Tao Li, Sophie Ni, Liuren Wu, Yingzi Zhu, and conference/workshop participants at Quantitative Methods in Finance 2008, the Institute of Risk Management at National University of Singapore, Fudan University, China International Conference in Finance 2009, and Asia-Pacific Futures Research Symposium 2010 for their helpful comments on earlier versions. In particular, the authors thank two anonymous referees whose comments and suggestions helped improve the paper substantially. Chu Zhang acknowledges the financial support from the RGC Competitive Earmarked Research Grant HKUST646007. All remaining errors are the authors’ responsibility.

References


Li and Zhang: On the Number of State Variables in Options Pricing
Management Science 56(11), pp. 2058–2075, © 2010 INFORMS


