Review of Derivatives Research, 3, 263–282 (1999)

The Dynamics of the S&P 500 Implied Volatility Surface

GEORGE SKIADOPOULOS* gskiadopoulos@hotmail.com
Athens Derivatives Exchange Research and Development Group. and Associate Research Fellow at the Financial Options Research Centre, Chrisanthemon 76, N. Iraklio 14121, Athens, Greece

STEWART HODGES s.d.hodges@warwick.ac.uk
Director of the Financial Options Research Centre, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK

LES CLEWLOW les_clewlow@compuserve.com
Associate Research Fellow at the Financial Options Research Centre

Abstract. This empirical study is motivated by the literature on "smile-consistent" arbitrage pricing with stochastic volatility. We investigate the number and shape of shocks that move implied volatility smiles and surfaces by applying Principal Components Analysis. Two components are identified under a variety of criteria. Subsequently, we develop a "Procrustes" type rotation in order to interpret the retained components. The results have implications for both option pricing and hedging and for the economics of option pricing.

Keywords: volatility smile, volatility surface, implied volatility, principal components analysis.

JEL classification: G13.

The developing literature on "smile consistent" no-arbitrage stochastic volatility models (Dupire, 1992, 1993; Derman and Kani, 1998; Ledoit and Santa-Clara, 1998) has been motivated by the need to price and hedge exotic options consistently with the prices of standard European options. The objective of this paper is to investigate the dynamics of implied volatilities, since this is a necessary prerequisite for the implementation of these models.

In recent years considerable interest has focused on the behavior of the implied volatilities of options contracts, derived from inverting Black–Scholes (1973) model. The empirical evidence shows (see Rubinstein, 1985, 1994; Derman and Kani, 1998) that implied volatil-

* We would like to thank participants at the 1998 meetings of the European and French Finance Associations, where an earlier version of this paper with the title "The Dynamics of Smiles," was presented, and especially the discussants Michael Rockinger and Norman Strong. We would also like to thank participants at the 1999 Derivatives Securities Conference in Boston University and Abhay Abhyankar, David Bates, Gordon Gemmill, Russell Grimwood, Jens Jackwerth, Wojtek Krzanowski, Joao Pedro Nunes, Juan Carlos Mejia Perez, Chris Strickland and Robert Tompkins for helpful discussions. We are grateful to Stephen Taylor (the referee) for his constructive comments and to Menachem Brenner (the editor) for his suggestions. Part of this paper was written while the first author benefited from the Human Capital & Mobility Programme of the European Commission. Financial support from the Corporate Members of FORC is also gratefully acknowledged. Any remaining errors are our responsibility alone.

ities vary across different strikes (smiles or skews), and different times-to-expiration (term structure) for options at the same point in time. In addition, implied volatilities also vary, in a stochastic way, across different points in time for a given option. These results suggest that implied volatilities could be viewed as a two-dimensional surface which evolves over time.

A number of option pricing models have been proposed which give rise to smiles or skews, and to a term structure of implied volatilities, roughly similar to what is observed empirically. These models provide for stochastic volatility (Hull and White, 1987; Johnson and Shanno, 1987; Scott, 1987; Wiggins, 1987), or jump processes (Bates, 1988; Merton, 1976; or both (Bates, 1996; Scott, 1997). However, none of these models fits observed implied volatility patterns well (see Das and Sundaram, 1998) making it difficult to use them in practice to price and hedge exotic options. These problems have led to the recent literature on "smile consistent" no-arbitrage stochastic volatility models.

This "evolutionary" approach is similar to the Heath, Jarrow, Morton (HJM, 1992) methodology for stochastic interest rates and was originally inspired by it (Dupire, 1992, 1993). The models take today's option prices (or equivalently the implied volatility surface) as given and they let them evolve stochastically in such a way as to preclude arbitrage. This allows for correct pricing of standard options and is relevant to the pricing of exotic options. For example, Derman and Kani (1998) start from today's option prices and postulate a stochastic process for the forward volatility. They then find the no-arbitrage condition that its drift must satisfy. Ledoit and Santa-Clara (1998) propose a model where the implied rather than the local volatilities are used to obtain a simple no-arbitrage condition that the drift of the implied volatility process must satisfy.

Practical implementation of this class of models requires us to understand the dynamics of the implied volatility surface. In this paper, we explore three questions related to this: (1) how many factors are needed to explain the dynamics of the volatility surface? (2) what do these factors look like? and (3) how are these factors correlated with the innovation in the underlying asset's process?[1]

We are able to identify two factors which explain about 60% of the variance. The first factor is interpreted as an essentially parallel shift and the second as a Z-shaped twist. Our results are remarkably consistent across years. Moreover, we confirm their robustness by investigating the dynamics of individual smiles, as well. To implement a "smile-consistent" no-arbitrage stochastic volatility model for the pricing and hedging of futures options we need three factors. One is required for the underlying asset and the other two for the implied volatility. A related paper by Kamal and Derman (1997) has also investigated the same issue. They have used a different data set and a somewhat different methodology; their results are markedly different from ours. We comment further on the differences between the two studies in the concluding section of the paper.

The paper is organized as follows. First, we describe the data set used and how we have screened it. In the second section we describe the Principal Components Analysis technique, including why it was chosen, and how it will be applied to analyze the data. Next, we analyze the dynamics of smiles and of implied volatility surfaces, respectively. Further, we calculate the correlations between the changes of the principal components and the underlying asset price. Finally, we discuss the implications of our results.

1. The Data Set

1.1. Source Data

We use daily data on futures options on the Standard and Poor index (S&P 500) from the Chicago Mercantile Exchange (CME) for the years 1992–95. The primary data for this study are the transaction report "Stats Database," compiled daily by CME. This database contains the following daily information for each option traded: the date, the style (call or put), the options and futures expiration month, the exercise price, the number of contracts traded, the opening, closing, low and high future's and option's price, the opening, closing, low and high bid-ask future's and option's price, and the settlement price. We extract from this, for the purposes of our study, the closing options and futures prices.

In addition, we use London Euro-Currency interest rates (middle-rates) on the US dollar, obtained from Datastream, to proxy for the riskfree rate. Daily interest rates for 7-days, one-month, three months, six months and one year were used, while those for other maturities were obtained by linear interpolation.

Since the S&P500 options futures are American-style options, implied volatilities were calculated using the algorithm of Barone-Adesi and Whaley (1987).

1.2. Screening the Data

In the first stage, the raw data are screened for data errors. Observations which violate the arbitrage boundary conditions are excluded. We eliminate data where the option price is less than, or equal to, its intrinsic value. We also exclude options having a price of less than 10 cents, and we eliminate short term options with less than 10 days to expiry because they are very sensitive to small errors in the option price.

In a second stage, we exclude data likely to introduce errors into our volatility estimates. We construct our smiles by using out-of-the-money (OTM) puts for low strikes, and OTM calls for the high ones. The data on in-the-money (ITM) calls and puts are not used because they have high deltas; their prices and their implied volatilities are therefore very sensitive to the non-synchroneity problem (see Harvey and Whaley, 1991). Provided the put-call parity relationship holds (there is a simple arbitrage if it does not), puts and calls must have identical implied volatilities (in the absence of measurement error problems). Our procedure reduces the effects of errors due to non-synchronous data, without introducing any sources of bias.

We also decided to exclude implied volatilities calculated from options having a vega less than eight on the grounds of the probable magnitude of their measurement error.[2] For a given measurement error ΔC in the option price (e.g. arising from the bid-ask spread) the resulting error in the implied volatility $\Delta\sigma_{imp}$, is (approximately) $\frac{\Delta C}{vega}$. There is a trade-off between accuracy and number of observations that we exclude. The choice of eight for the vega cutoff was made following a preliminary examination of the data. As a result we are able to eliminate most of the noise and to retain about 40% of the observations for calls and 70% of the observations for puts.

2. Principal Components Analysis and Implied Volatilities

In this section, we outline the Principal Components Analysis (PCA) methodology used in this paper. PCA is used to explain the systematic behavior of observed variables, by means of a smaller set of unobserved latent random variables. Its purpose is to transform p correlated variables to an orthogonal set which reproduces the original variance-covariance structure.

In this paper, we apply PCA to decompose the variance-covariance structure of first differences of implied volatilities. To achieve this, we measure the daily differences of implied volatilities across different levels of moneyness and different ranges of days to expiry (expiry buckets). For example, one of our variables provides a time series of the first differences of implied volatilities which correspond to a moneyness level of (-1.5%) and expiry range 10–30. Typically, for each expiry bucket we have 7–10 levels of moneyness in each year, with 100–225 observations for each level.

In general, denote time by $t = 1, \ldots, T$ and let p be the number of variables. Such a variable is a $(T \times 1)$ vector $\mathbf{x}$. The purpose of the PCA is to construct Principal Components (PCs hereafter) as linear combinations of the vectors $\mathbf{x}$, orthogonal to each other, which reproduce the original variance-covariance structure. The first PC is constructed to explain as much of the variance of the original p variables, as possible (maximization problem). The second PC is constructed to explain as much of the remaining variance as possible, and so on. The coefficients with which these linear combinations are formed are called the *loadings*. In matrix notation

$$\mathbf{Z} = \mathbf{XA} \tag{1}$$

where $\mathbf{X}$ is a $(T \times p)$ matrix, $\mathbf{Z}$ is a $(T \times p)$ matrix of PCs, and $\mathbf{A}$ is a $(p \times p)$ matrix of loadings. The first order condition of this maximization problem results to

$$(\mathbf{X}'\mathbf{X} - l\mathbf{I})\mathbf{A} = \mathbf{0} \tag{2}$$

where l_i are the Lagrange multipliers. From equation (2) it is evident that the PCA is simply the calculation of the eigenvalues l_i and the eigenvectors of the variance-covariance matrix $\mathbf{S} = \mathbf{X}'\mathbf{X}$. Furthermore, the variance of the ith PC is given by the ith eigenvalue, and the sum of the variances of the PCs equals the sum of the variances of the $\mathbf{X}$ variables.

When both variables and components are standardized to unit length, the elements of $\mathbf{A}'$ are correlations between the variables and PCs and they are called *correlation loadings* (see Basilevsky, 1994) for more details). If we retain $r < p$ PCs then

$$\mathbf{X} = \mathbf{Z}_{(r)}\mathbf{A}'_{(r)} + \varepsilon_{(r)} \tag{3}$$

where $\varepsilon_{(r)}$ is a $(T \times p)$ matrix of residuals and the other matrices are defined as before having r rather than p columns. The percentage of variance of $\mathbf{x}$ which is explained by the retained PCs (*communality* of $\mathbf{x}$) is calculated from the correlation loadings. After retaining $r < p$ components, we look at equation (3) to examine the size of the communalities, and the meaning of the retained components.

PCA is a natural and parsimonious technique to identify the number and the interpretation of stochastic shocks that move the implied volatilities. It enables us to simplify the complex dynamics of the volatility surface, by identifying its most important components, without imposing any prior structure. This contrasts, for example, with the alternative regression analysis approach of estimating a specific function of time and moneyness (see Taylor and Xu, 1994). PCA is preferred to factor analysis, for our analysis, because of its variance maximization property.[3]

We investigate the dynamics of implied volatilities by performing PCA on the first differences of implied volatilities. This is because the empirical evidence shows that implied volatilities follow a mean reverting process with the first order autoregressive coefficient being very close to one (see for example Hsieh, 1993). Therefore, implieds are stationary with a near unit root. Hence, they need to be differenced once because PCA is misleading when applied to non-stationary variables (Frachot, Jansi and Lacoste, 1992).

The variables **x** of changes in implied volatilities to which we apply the PCA are indexed with the moneyness level $\frac{K-F_t}{F_t} * 100$ (moneyness metric). We choose this metric because there are theoretical reasons (see Heynen, 1994; Taylor and Xu, 1994) for smiles being a function of moneyness.

Next, we group the data into different buckets for distinct ranges of days to expiry. We need to control for the time to expiry because we expect the implied volatilities of the shorter dated options to vary more than those of the longer dated ones. This is a feature of various models (such as Stein, 1989) and earlier empirical studies e.g. Hsieh (1995) and Xu and Taylor (1994).[4] We fix six such intervals: 30–10, 60–30, 90–60, 150–90, 240–150, and 360–240 days to expiry.[5] The dynamics of individual smiles are analyzed (in the next section) by applying PCA separately to the expiry buckets. Then, in Section 4 we analyze the dynamics of the whole implied volatility surface by pooling the buckets together and applying the PCA on the whole data set.

To summarize, in our framework, each variable **x** is a time series collection of differences in implied volatilities for a given moneyness level, and within a certain expiry range. The first PC is the linear combination of these variables which contributes the most to explaining the overall variance of the changes in the implied volatility surface. The second PC is the linear combination (of the same variables) orthogonal to the previous one, which explains as much as possible of the remaining variance, and so on.

3. PCA on the Smiles

In this section we investigate the dynamics of individual smiles by performing a separate PCA on each expiry bucket (smile analysis). Within each PCA the variables have different moneyness levels, but the same expiry range.

3.1. Construction of the Moneyness Metric

To construct our variables, we first choose a grid of moneyness points for them. We measure implied volatilities at a number of different moneyness levels chosen as fixed percentage difference between the strike and the futures price F_t. This involves interpolation across the implied volatilities for these fixed variables, since F_t is always changing.

In choosing this grid we need to be careful not to make it too fine. If we allowed two different grid points to fall between adjacent strikes, they would both be interpolated from the same two data points. This would produce spurious dependence which would distort our results. We have therefore chosen moneyness levels which are slightly coarser than the coarsest spacing of the strikes.

The application of listwise deletion creates a trade-off between the number of variables on which we perform the PCA and the total number of observations. For each year, we choose the variables within a given expiry range so that we get (a) a sufficient number of moneyness variables (ideally not less than 7) i.e. a wide range of the smiles, and (b) a sufficient number of observations (no less than 100).

3.2. Number of Retained Principal Components and a First Interpretation

We now decide on the number of components to be retained and we look at their interpretation. Earlier researchers have used a variety of rules of thumb to determine the number of components to be retained. For example, they keep the components corresponding to eigenvalues larger than the mean of all the eigenvalues (mean eigenvalue rule of thumb), or they keep the components which explain 90% of the total variance.[6] As Basilevsky notes "such practice is statistically arbitrary, and seems to be prompted more by intuitive concepts of practicality and "parsimony," than by probabilistic requirements of sample-population inference."

We determine the number of components to be retained by looking at a range of criteria in an even handed way. First, we apply Velicer's (1976) non-parametric criterion.[7] Next, working with components retained under this criterion, we look at the communalities. Finally, we look at the interpretation of the PCs.

Velicer proposes a non-parametric method for selecting nontrivial PCs, i.e. components which have not arisen as a result of random sampling, measurement error, or individual variation. His method is based on the partial correlations of the residuals of the PCs model, after $r < p$ components have been extracted. The criterion can be described as follows: The variance–covariance matrix of the residuals $\varepsilon_{(r)}$ in equation (3) is given by

$$\varepsilon'_{(r)}\varepsilon_{(r)} = \mathbf{X}'\mathbf{X} - \mathbf{A}_{(r)}\mathbf{A}'_{(r)} \tag{4}$$

Let $\mathbf{D} = diag(\varepsilon'_{(r)}\varepsilon_{(r)})$. Then, $\mathbf{R}^* = \mathbf{D}^{-\frac{1}{2}}\varepsilon'_{(r)}\varepsilon_{(r)}\mathbf{D}^{-\frac{1}{2}}$ is the matrix of partial correlations of the residuals. If r^*_{ij} represents the ith row, jth column element of R^*, then the Velicer statistic is given by

$$f_r = \sum\sum_{i \neq j} \frac{r^{2*}_{ij}}{p(p-1)} = \sum\sum_{i=j} \frac{r^{2*}_{ij} - p}{p(p-1)} \tag{5}$$

and lies in the interval 0 to 1. The behavior of f_r is that it is decreasing until a number r^* and then it increases again. Velicer suggests that $r = r^*$ should be the number of components to be retained.

Table 1 shows the number of components retained under Velicer's criterion and under the mean eigenvalue rule of thumb. It also shows the percentage of the variance explained by

Table 1. Principal components in the smile analysis: $r^* =$ number of components retained under Velicer's criterion (minimum $f_0, \dots f_3$), $\bar{l} =$ number of components retained under rule of thumb with percentage variance explained by components 1–3.

	Year	f_0	f_1	f_2	f_3	r^*	$\bar{l}$	1st PC	2nd PC	3rd PC
30–10	1992	0.4818	0.4792	0.4801	0.4807	1	1	73.30	11.00	5.50
	1993	0.5521	0.5491	0.5500	0.5508	1	1	77.80	8.10	6.00
	1994	0.6959	0.6936	0.6946	0.6949	1	1	85.70	6.40	2.30
	1995	0.4764	0.4735	0.4745	0.4752	1	1	73.00	11.10	5.90
60–30	1992	0.1816	0.1812	0.1810	0.1811	2	2	39.60	30.30	8.10
	1993	0.2065	0.2058	0.2058	0.2059	2	2	44.60	28.70	7.70
	1994	0.2761	0.2747	0.2751	0.2753	1	2	53.60	22.90	5.30
	1995	0.2189	0.2183	0.2183	0.2185	2	2	47.60	25.50	8.70
90–60	1992	0.2291	0.2276	0.2274	0.2278	2	2	48.80	27.30	9.00
	1993	0.2164	0.2133	0.2132	0.2137	2	2	46.80	30.60	9.70
	1994	0.2168	0.2153	0.2152	0.2154	2	2	44.20	29.60	7.40
	1995	0.1395	0.1383	0.1381	0.1384	2	2	39.70	25.50	9.50
150–90	1992	0.2326	0.2313	0.2316	0.2317	1	2	51.50	16.80	7.50
	1993	0.1989	0.1969	0.1972	0.1975	1	2	48.10	16.90	8.90
	1994	0.2578	0.2560	0.2563	0.2565	1	2	54.20	18.00	8.00
	1995	0.2008	0.1985	0.1988	0.1991	1	2	48.80	18.10	8.50
240–150	1992	0.4314	0.4302	0.4303	0.4307	1	1	69.20	11.90	7.60
	1993	0.4334	0.4322	0.4324	0.4327	1	2	68.70	10.40	6.50
	1994	0.3595	0.3585	0.3588	0.3590	1	2	63.40	16.90	7.40
	1995	0.3224	0.3211	0.3213	0.3215	1	2	60.70	12.80	7.30
360–240	1992	0.5668	0.5653	0.5657	0.5663	1	1	77.90	14.10	4.40
	1993	0.4933	0.4917	0.4922	0.4928	1	1	75.10	16.90	4.10
	1994	0.6852	0.6826	0.6839	0.6844	1	1	85.10	7.80	2.50
	1995	0.5932	0.5874	0.5894	0.5911	1	1	80.90	6.80	5.80

the first three components. These results come from estimating the PCs separately for six distinct expiry ranges and in four distinct one year periods. We can see that under Velicer's criterion we should retain either one or two PCs. This is more conservative than the mean eigenvalue rule of thumb which, in some expiries, retains one more PC than Velicer's. Note that $f_0 > f_1$, showing that at least one component can be extracted.

The first component explains between 40% and 86% of the variance, depending on which year and which expiry range we consider. The second component explains between 7% and 31%. The two components combined explain between 65% and 92% of the variance. The highest proportions are explained for the shortest and longest expiry date options.

Next, we look at the interpretation of the first three PCs.[8] If any PC appears to be mostly noise, then we will prefer to reject it. Figure 1 shows the correlation loadings of the first

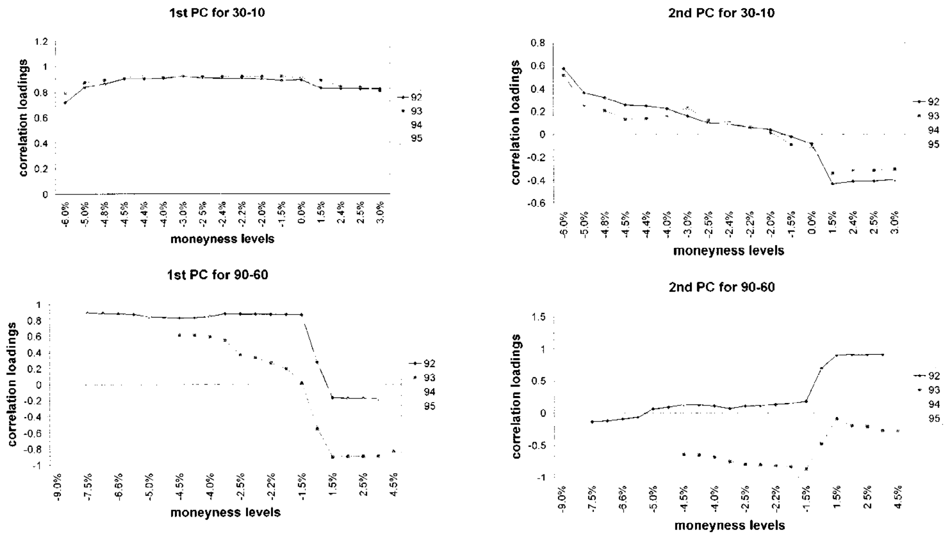

Figure 1. First and second PCs for expiry buckets 30–10 and 90–60.

and second PC for the 30–10 and 90–60 buckets, over the different years. For the sake of clarity, we have interpolated across the missing variables in these graphs. The figure shows that in every year, the first PC for the expiry range 30–10 is like a parallel shift with a slight attenuation at the edges. In the 90–60 range it has both positive and negative correlation loadings (it is like a Z-shape, i.e. slope with attenuation at the edges). The second PC, has a Z-shape in the expiry range 30–10 in every year. In the range 90–60 it has positive correlation loadings in 1993, and a mixture of both positive and negative loadings in the other years.

Finally, graphs for the third PC suggested strongly that it was just noise, confirming the interpretation from Velicer's procedure. Therefore, under the variety of criteria that we applied, we judge that we can only identify two shocks driving the implied volatility smiles of the S&P 500 Futures Options.

It is reassuring that the components are so consistent across the four years. For the 30–10 day expiry range they correspond nicely to a near parallel shift and to a change in slope. However, for the expiry range 90–60 the "raw" PCs, although consistent through time, do not have such a neat interpretation. In order to obtain a better interpretation we will utilize a rotation method, explained in the next section. We therefore defer the discussion of the results for further expiry dates until after we have explained this technique.

3.3. The Rotation Method

We would like the first PC to be as close to a parallel shift, as possible (as in the 30–10 expiry). We use a rotation technique to achieve this. We form new rotated loadings from

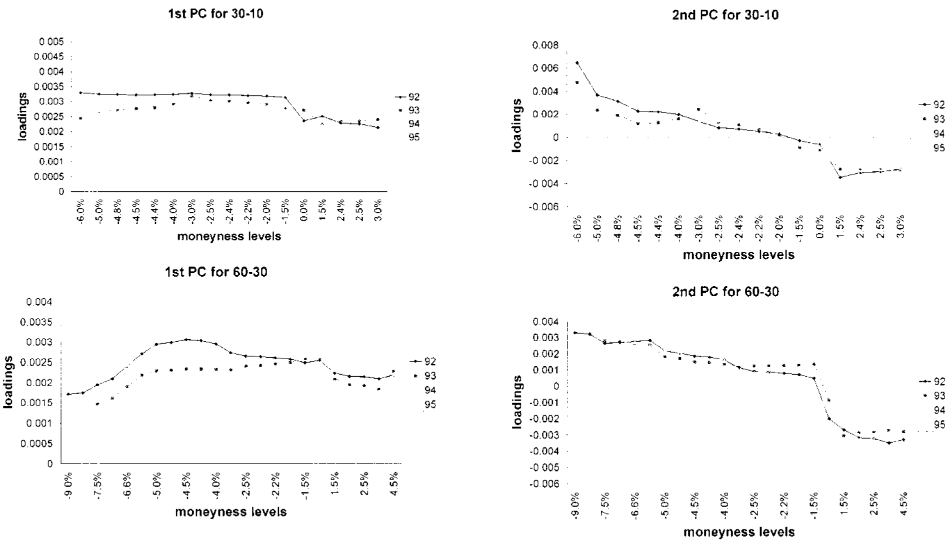

Figure 2. First and second PCs for expiry buckets 30–10 and 60–30.

the original ones **A**, by multiplying them by the rotation matrix

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

where θ is the angle of rotation. The rotated components still explain the same total amount of variance, as the unrotated ones, but the total variance may have been re-distributed between the two (see Basilevsky, (1994)) for the properties of orthogonal rotations). For the purposes of our study, we choose **T**, so that the loadings on the first rotated PC are as flat, as possible.[9] We accomplish this by using a regression to find the orthogonal rotation which minimizes the least squares distance between the loadings of the first PC and a vector of constants.

The results for the first and second rotated PC, for the different expiries, appear in Figures 2, 3 and 4. These graphs are scaled so that they show the effect of one standard deviation daily change on the volatility smile.

The first rotated PC provides positive loadings for all expiries and years with the single exception of 240–150 days in 1992. In the range 30–10, the parallel shift with the attenuation at the edges has been maintained. In the range 60–30 there are positive loadings in every year, with a much smoother shape than in the unrotated ones. Again, the shape indicates movement which is close to parallel. In the ranges 90–60, 150–90 and 360–240 there are positive loadings in every year. In the range 240–150 we see much greater attenuation for extreme moneyness, giving rise to a more triangular shape. This shape is also evident for one or two years in the adjacent expiry ranges. The figures also show that the magnitude

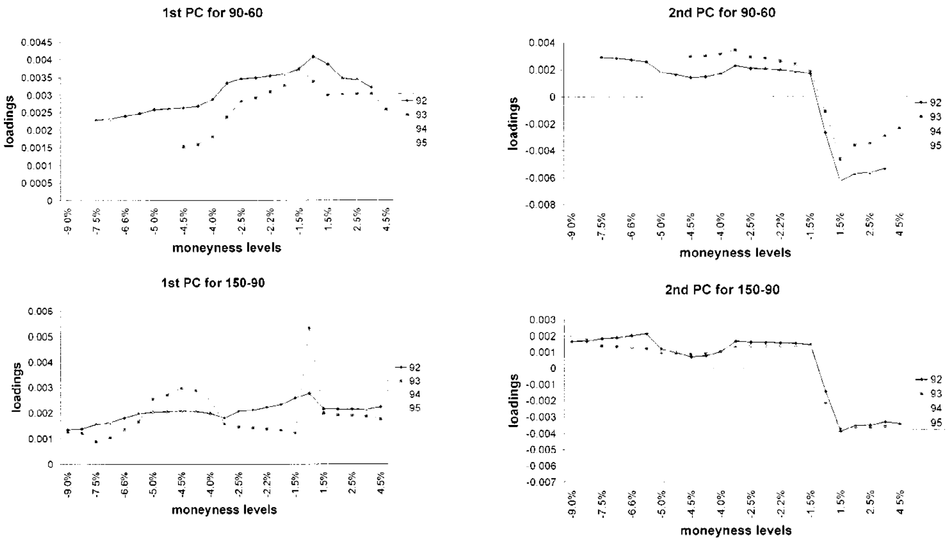

Figure 3. First and second PCs for expiry buckets 90–60 and 150–90.

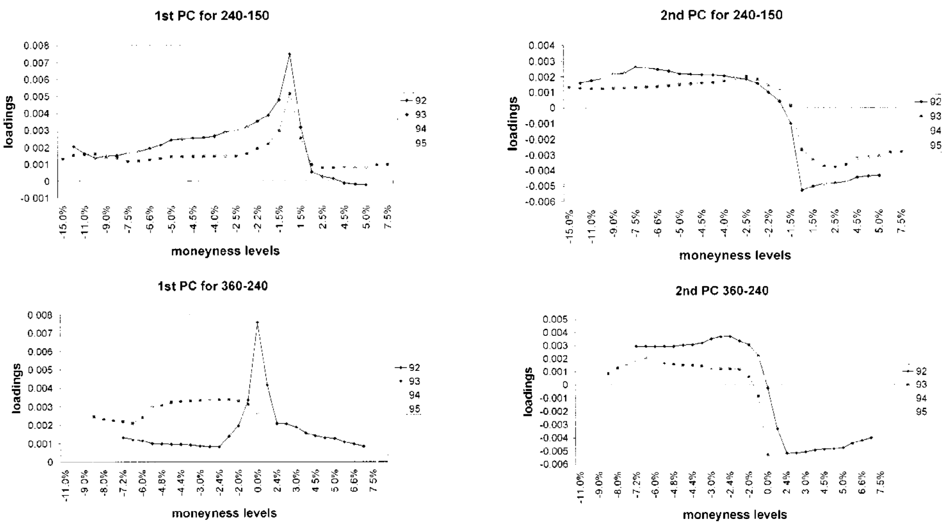

Figure 4. First and second PCs for expiry buckets 240–150 and 360–240.

of the first PC attenuates, as expected, with expiry. This is especially clear for the 30–10 and 60–30 expiries, but rather less so for the others, presumably because of less liquidity in these expiries.

The second rotated PC consistently produces a Z-shaped profile of loadings. This component does not show the clear attenuation with expiry that we find for the first component. Although its magnitude is greatest for the shortest expiry range, it changes rather little over the longer expiries.

In Table 2, we show what percentage of the variance is explained by those first and second rotated PCs (as well as the percentage for the original first component). We find that the parallel shock of the first component dominates in the shorter expiries of 30–10 and 60–30, while the Z-shaped shock explains more of the variance in the longer ones of 150–90, 240–150 and 360–240.

3.4. Serial Correlation and the Use of the Differences

At this point we should examine the appropriateness of having chosen to work with first differences in implied volatilities. Performing PCA on the changes of implied volatilities was motivated by the idea that implied volatilities should be close to a random walk. The empirical literature on the serial correlation of the first differences of implied volatilities is slightly mixed. On the one hand, Brenner and Galai (1987) find that the daily average at-the-money implied volatility is autocorrelated, but the first differences of it are not, as we would expect. On the other hand, negative serial correlation in the first differences of implied volatilities is likely to arise through non-synchronous data (as Harvey and Whaley (1991) suggest) and from “bid-ask bounce” (see Roll, 1984).

Even though we have been careful to minimize the sources of measurement error in our data, serial correlation effects could hamper the efficiency of the PCA. We investigated this by applying the following alternative method of filtering out the serial correlation:

Let $\hat{\sigma}_t$ be the observed implied volatility which is measured with noise, and σ_t be the true implied volatility. Then, $\hat{\sigma}_t = \sigma_t + u_t$, where u_t is the measurement error. Moreover, assume that $\sigma_t = \phi\sigma_{t-1} + \varepsilon_t$, where ε_t are the white noise disturbances. The above two equations are equivalent to an ARMA(1,1) process (see Harvey, 1993), i.e. $\hat{\sigma}_t = \phi\hat{\sigma}_{t-1} + \eta_t - \vartheta\eta_{t-1}$, where η_t are white noise disturbances. We estimated this model and then we applied PCA to η_t, for several subsets of our data. After correcting for the serial correlation in this way, we obtained results from the PCA which were very similar to those already reported. Hence, we can conclude that PCA is reasonably robust to the fairly minor serial correlation present in our data set. We have preferred in the main body of the study to work with the original first differences of implied volatilities.

4. PCA on the Implied Volatility Surface

4.1. Redefining the Expiry Buckets

In order to look into the dynamics of implied volatilities as a surface for a given year (surface analysis), we first choose the expiry buckets, and then the moneyness levels for each bucket.

Table 2. Percentage of variance explained by the unrotated first PC and by the rotated PCs in the smile analysis.

Range	Year	Unrot. 1st PC	1st PC	2nd PC	Cumulative
30–10	1992	73.30%	73.30%	11.10%	84.40%
	1993	77.80%	77.80%	8.10%	85.80%
	1994	85.70%	85.70%	6.40%	92.10%
	1995	73.00%	73.00%	11.10%	84.10%
	average	77.45%	77.50%	9.20%	86.60%
60–30	1992	39.60%	36.00%	33.90%	69.90%
	1993	44.60%	44.40%	28.90%	73.30%
	1994	53.60%	53.20%	23.30%	76.50%
	1995	47.60%	44.70%	28.30%	73.10%
	average	46.35%	44.60%	28.60%	73.20%
90–60	1992	48.80%	40.00%	36.10%	76.10%
	1993	46.80%	33.50%	43.90%	77.40%
	1994	44.20%	37.90%	35.90%	73.90%
	1995	39.70%	31.80%	33.50%	65.20%
	average	44.88%	35.80%	37.40%	73.20%
150–90	1992	51.50%	19.90%	48.40%	68.30%
	1993	48.10%	19.60%	45.40%	65.00%
	1994	54.20%	34.00%	38.30%	72.30%
	1995	48.80%	25.50%	41.30%	66.90%
	average	50.65%	24.80%	43.40%	68.10%
240–150	1992	69.20%	22.20%	59.00%	81.20%
	1993	68.70%	15.80%	63.20%	79.00%
	1994	63.40%	17.20%	63.10%	80.30%
	1995	60.70%	25.10%	48.40%	73.50%
	average	65.50%	20.10%	58.40%	78.50%
360–240	1992	77.90%	14.90%	77.10%	92.00%
	1993	75.10%	50.70%	41.30%	92.00%
	1994	85.10%	24.50%	68.50%	92.90%
	1995	80.90%	16.80%	70.90%	87.70%
	average	79.75%	26.70%	64.50%	91.20%

Finally, all the variables are brought together to perform the PCA. However, the original expiry buckets gave rise to a problem with this pooled analysis. Their spacing was too fine, so that we had insufficient dates with data points in every expiry range. Therefore, we redefine the new expiry buckets to 90–10, 180–90 and 270–180 days to expiry. These intervals give us a satisfactory number of observations (not less than 100), and they permit us to measure smiles across a wide range (not less than 20 variables), once they are pooled together.

Table 3. Principal components in the surface analysis: $r^* =$ the number of components retained under Velicer's criterion (minimum of $f_0, \ldots f_3$), $\bar{l} =$ number of components retained under rule of thumb, with percentage of variance explained by components 1–3.

Year	f_0	f_1	f_2	f_3	r^*	$\bar{l}$	1st PC	2nd PC	3rd PC
1992	0.1847	0.1838	0.1836	0.1838	2	4	38.6	22.7	10.7
1993	0.1722	0.1714	0.1711	0.1713	2	4	34.5	26.7	10.2
1994	0.1894	0.1884	0.1884	0.1885	2	6	40.6	19.2	10.2
1995	0.1726	0.1713	0.1712	0.1714	2	4	39.2	18.3	9.4

4.2. Number of Retained Principal Components and a First Interpretation

We investigate the dynamics of the volatility surface just as we did for the individual smiles. In other words, we apply three criteria to decide how many components to retain, and then we use a rotation in order to obtain a clean interpretation. In Table 3, we show the results from applying Velicer's criterion and the mean eigenvalue rule of thumb. We also show the percentage of the variance explained by each one of the first three PCs. We can see that in every year, the test keeps two PCs, while the reduction in the dimensionality of the variables is legitimate, since $f_0 > f_1$. The mean eigenvalue rule of thumb retains too many PCs. This shows why we should not rely on such ad-hoc rules (see Jackson, 1991).

In Figure 5, we show the interpretation of the first two "pooled" PCs.[10] We can see that they do not have a consistent shape across years and expiries. For example, the first PC has a Z-shape in the range 90–10 for the years 1992 and 1993. However, in the same range it has a shift interpretation for the years 1994 and 1995. The second PC has a shift interpretation in the range 90–10 for the year 1992. On the other hand, in the same range, it has a Z-shape for the year 1995. The graphs for the third PC showed that it is noise.

Hence, all the criteria that we applied, suggest that two PCs are adequate in explaining the implied volatility surface dynamics. The two PCs explain on average, across the years, 60% of the total variance.

4.3. Interpretation of the Rotated PCs

As with the smile analysis, we will apply a rotation to our original PCs. In Figure 6 we show the shape of the first and second rotated PCs. The first PC moves the implied volatility surface consistently across ranges and across years and its shape "approaches" the shift interpretation. The figure also shows that the effect of the shift component attenuates with expiry, as expected. The second PC, in general, has a Z-shape across years, even though it has a shift interpretation in the range 90–10 in the years 1994 and 1995. There is not much sign of attenuation with expiry. More important, the number of shocks and their interpretation is the same for both the smiles and the surfaces.

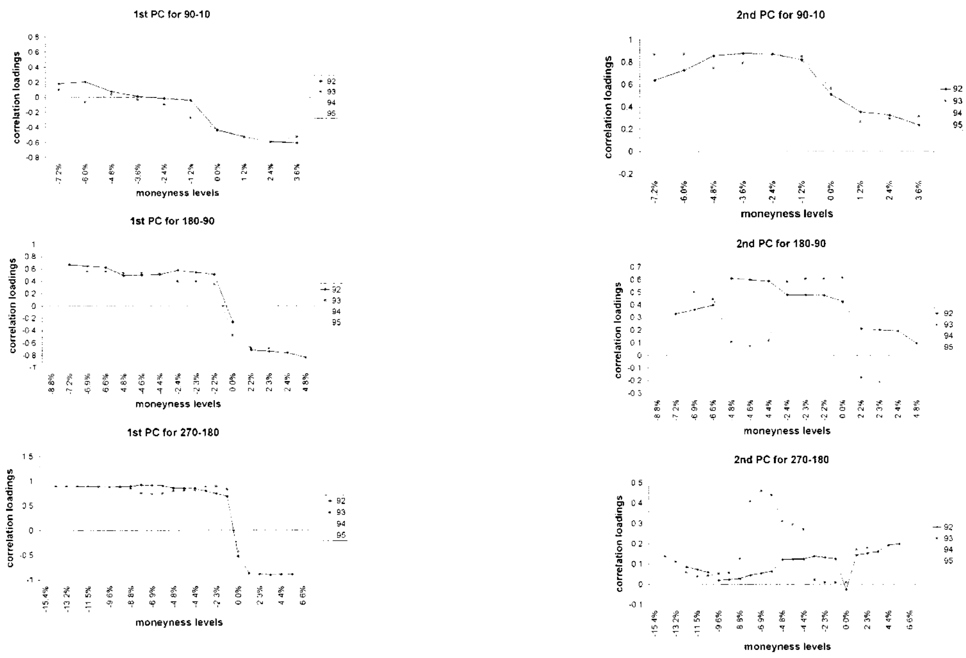

Figure 5. First and second rotated PCs for the implied volatility surface.

Table 4. Percentage of variance explained by the unrotated first PC and by the rotated PCs in the surface analysis.

Year	Unrot. 1st PC	1st PC	2nd PC	Cumulative
1992	38.6%	22.7%	38.6%	61.4%
1993	34.5%	26.9%	34.3%	61.2%
1994	40.6%	19.5%	40.4%	59.5%
1995	39.2%	18.4%	39.2%	57.5%
Average	38.2%	21.8%	38.1%	60.0%

Our analysis shows that the implementation of a "smile consistent" no-arbitrage stochastic volatility model suggests the need for three factors. One is required for the underlying asset and two more for the implied volatility. The loadings of the two PCs may be useful for volatility risk management, even though we do not know whether they will be stable (in fact, they exhibit some variability over the years). With this type of factor model, the definition of vega can be generalized to the sensitivities to each of the volatility shocks.

In Table 4 we show the percentage of the variance explained by the first and second rotated PCs (as well as the percentage for the original first component). We find that it is not the shift which has the dominant effect on the implied volatility surface, but the second PC.

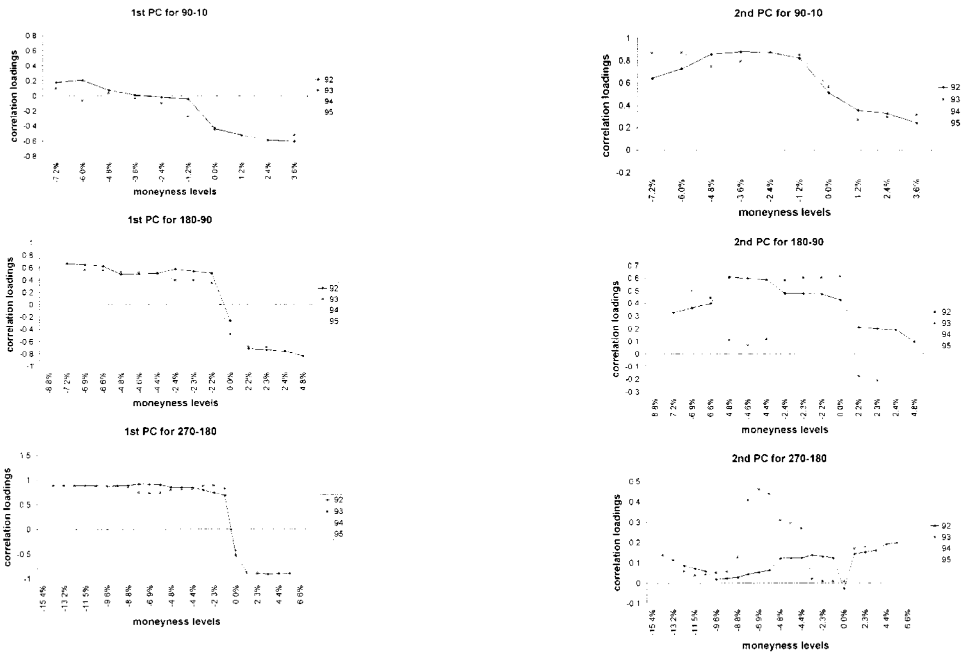

Figure 6. First and second rotated PCs for the implied volatility surface.

5. Correlations Between the Futures Price and the Principal Components

So far, we have found simple parameterizations of the dynamics of implied volatilities. In this section, we investigate the correlations between the percentage changes in the futures price and the changes of the principal components. Knowledge of these is necessary in order to complete the specification of the process for the evolution of implied volatilities.

Table 5 presents the correlations between the proportional changes of the futures price ($\frac{\Delta F_t}{F_t} = \frac{F_{t+1}-F_t}{F_t}$) and the changes of each one of the first two rotated principal components, for both the smile and the surface analysis. One asterisk is displayed when the coefficient is significant at 5% significance level, and two asterisks are displayed when the coefficient is significant at 10% significance level.

The first rotated PC from the smile analysis is negatively correlated with the underlying; the only exception to this is in the range 360–240 for the years 1993 and 1995. The correlations for the rotated second PC are always positive. Similarly, in the surface analysis, the correlations for the rotated first PC are negative and for the second PC are positive, apart from year 1993, where they are negative. We investigated the scatterplots for the anomalous correlations and found that although there were some outliers, removing them did not affect the sign of the correlation (though in some cases it became insignificantly different from zero).

Table 5. Correlations between proportional changes of the futures price with changes of the rotated PCs for the smile and the surface analysis.

Range		1992	1993	1994	1995
Smile Analysis					
30–10	$\Delta PC1$	−0.40**	−0.52**	−0.65**	−0.42**
	$\Delta PC2$	0.12	0.01	0.05	−0.06
60–30	$\Delta PC1$	−0.35**	−0.38**	−0.58**	−0.41**
	$\Delta PC2$	0.24**	0.22**	0.25**	0.11
90–60	$\Delta PC1$	−0.36**	−0.49**	−0.49**	−0.37**
	$\Delta PC2$	0.25*	0.17	0.33**	0.37**
150–90	$\Delta PC1$	−0.28**	−0.31**	−0.19*	−0.32**
	$\Delta PC2$	0.33**	0.26**	0.37**	0.16
240–150	$\Delta PC1$	0.08	0.04	−0.28**	0.08
	$\Delta PC2$	0.28**	0.36**	0.37*	0.38**
360–240	$\Delta PC1$	−0.45**	0.33**	0.14	0.27*
	$\Delta PC2$	0.31**	0.35**	0.47**	0.52**
Surface Analysis					
90–10	$\Delta PC1$	−0.29**	−0.36**	−0.55**	−0.29**
	$\Delta PC2$	0.34**	−0.33**	0.16	0.26**
180–90	$\Delta PC1$	−0.29**	−0.28**	0.04	−0.29**
	$\Delta PC2$	0.35**	−0.43**	0.27**	0.26*
270–180	$\Delta PC1$	−0.28**	−0.28**	0.06	−0.31**
	$\Delta PC2$	0.35**	−0.44**	0.27**	0.26**

The negative correlation with the first PC indicates that implied volatilities rise on average when the market falls. These results are consistent with the leverage effect (see Christie, 1982), and the related regularity that implied volatilities are negatively correlated with the market index returns (Franks and Schwartz, 1991; Rubinstein, 1994; Schmalensee and Trippi, 1978). The size of the correlation coefficient changes over years and over expiries. This is not surprising in the light of the way correlations behave in other markets. However, the instability of the correlations poses problems for the implementation of models such as Ledoit and Santa-Clara's (1998).

6. Implications of the Research

In this paper, we have investigated the dynamics of implied volatilities of the S&P 500 Futures Options, by applying Principal Components Analysis (PCA). This provides important insights into the behavior of traded option prices and it is a prerequisite for the future development of models to realistically portray the dynamics of implied volatility surfaces.

We have investigated both the dynamics of the whole implied volatility surface, and the dynamics of individual smiles which provided us with a check of the robustness of our results.

After considering three criteria (Velicer's criterion, communalities, interpretation), we found that two components could be reliably extracted. We next applied a novel form of "Procrustes" rotation to the components, so as to obtain a clear interpretation for them. The first component can be interpreted as a parallel shift, and the second as a Z-shaped twist. The magnitude of the first shocks attenuates with expiry, while it is not obvious that this is the case for the second shock. The number of shocks and their interpretation is the same for both the smile and the surface analysis. On average, the two factors explained about 78% of the smiles variation and 60% of the surface variation. These results were in most cases very consistent for the four separate years 1992, 1993, 1994 and 1995.

Our results have further implications. First, running PCA analysis both on individual volatility smiles and on the whole volatility surface gave consistent results on both the number of factors (two) and their structure. Second, the results again confirm the findings of other authors, e.g. Buraschi and Jackwerth (1998) and Dumas, Fleming and Whaley (1998), that deterministic volatility models are not capable of representing the dynamics of the volatility surface. Third, our results are in contrast with Kamal and Derman's (1997). They analyze the dynamics of implied volatilities of over the counter (OTC) S&P 500 and Nikkei 225 Index options. They find that three PCs explain about 95% of the variance of the volatility surface. Their interpretation is a level of volatilities for the first PC, a term structure of volatilities for the second PC and a skew for the third. Their results suggest that a four factor model for pricing and hedging options under a stochastic volatility "smile consistent no-arbitrage pricing" type model is an appropriate one. Our work, identifies a model with one less factor, but cautions that the factor structure has higher dimensionality since it explains only 60% of the surface variation.[11] Despite considerable care and effort to eliminate the main sources of measurement error in the traded market data and to minimize its effects, the market futures option data seem to exhibit a noisier volatility structure than that estimated from the quotations for OTC index options provided by Goldman and Sachs traders. Portfolios of options modelled as riskless under a three or four factor models, may in fact exhibit substantial market risk.

Finally, just as Fung and Hsieh (1991) have shown that the typical smile structure depends on the particular underlying asset, it seems certain that the dynamics of the volatility surface must also be specific to the choice of asset. The proper test of all such models lies in out-of-sample hedging analysis. This is well beyond the scope of the current paper, but deserves to become a topic for future research.

Notes

1. This also parallels work on implementing the HJM term structure model where Litterman and Scheinkman (1988) and others, have applied Principal Components Analysis to analyze innovations in the yield curve.
2. Vega is defined as $Vega = \frac{\partial C}{\partial \sigma_{imp}}$, where C is the call option value and σ_{imp} its implied volatility.
3. Vector autoregressions (Hamilton, 1994) can also be used in order to investigate the dynamics of implied volatilities.

4. There are two other possible methods for accommodating the effect of the time to expiry on the variance of changes in implied volatilities. We could interpolate to get series for fixed expiries, or we could use a parametric model which uses the expiry date as an explanatory variable. We prefer grouping the data into buckets because the expiries in our data are too thinly spaced and because we do not want to impose any prior structure.
5. The expiry buckets were chosen, so as to cope with two constraints: (a) getting a sufficient amount of data for each range in order to perform the PCA, and (b) treating the missing observations that occur due to the screening criteria that we have applied. The literature on treating missing observations is vast, but as Anderson et al. (1983) note "The only real cure for missing data is to not have any." Therefore, we are not going to replace our missing values, because we do not know how this will bias our results. Instead, we are going to apply *listwise deletion* i.e. we will delete the whole day for which at least the observation for one variable is missing. Listwise deletion is applied after taking the first differences, so as differences in implieds are 1-day differences.
6. Litterman and Scheinkman (1988) apply PCA to the yield curve and they retain three components because these are explaining about 98% of the total variance. However, as Jackson (1991, page 44) notes "this procedure is not recommended. There is nothing sacred about any fixed proportion." For a description and discussion of the several rules of thumb, see Jackson (1991).
7. Most of the tests used for determining the number of PCs to be retained, are parametric based on the assumption of multivariate normality (For a review of these tests, see Basilevsky, 1994). However, application of the Bera–Jarque test showed that the null-hypothesis of univariate, and hence of multivariate normality, was rejected. Therefore, we decided to use a non-parametric method.
8. We also looked at the communalities explained by one and two PCs. One PC seems to need the assistance of the second, in order to explain a sufficient amount of the variability of each one of our moneyness variables. However, including the third PC did not increase the explained communalities significantly. Hence, the communalities criterion suggests that we should retain two PCs.
9. A rotation which tries to reveal a "targeted" interpretation is called a "Procrustes" rotation (see Jackson, 1991). Our choice of targeting on a flat "shift" loading seems to be novel.
10. Retaining two PCs gives rise to sufficiently high communalities for most of the variables. The exceptions are for the at-the-money variables in the ranges 180–90 and 270–180, in year 1994. In these ranges, the explained by the two PCs communalities were low (only 4.71% and 1.23%, respectively). However, even if we add the third PC, the explained communalities would have risen only to 17.82% and 4.02%, respectively. Therefore, the communalities criterion suggests that we should keep two PCs in the surface analysis.
11. It is worth noting that our model has similarities to that used recently in a time series analysis of index returns by Gallant, Chien and Tauchen (1998). Using daily data on close-to-close price movements and the high/low spread, they find that a model with two stochastic volatility shocks plus the underlying asset component, fits the data very well, and in particular it mimics the long-memory feature of volatility.

References

Anderson, A., A. Basilevsky, and A. Hum. (1983). "Missing Data: A Review of the Literature." In Rossi, P., J. Wright, and A. Anderson, *Handbook of Survey Research*, Academic Press.

Barone-Adesi, G., and R. E. Whaley. (1987). "Efficient Analytic Approximation of American Option Values," *Journal of Finance* 42, 301–320.

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods, Theory and Applications*. Wiley Series in Probability and Mathematical Statistics.

Bates, D. S. (1988). "Pricing Options under Jump-Diffusion Processes," Working Paper, Rodney L. White Center for Financial Research.

Bates, D. S. (1996). "Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options," *Review of Financial Studies* 9, 69–107.

Black, F., and M. Scholes. (1973). "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* 81, 637–654.

Brenner, M., and D. Galai. (1987). "On the Prediction of the Implied Standard Deviation," *Advances in Futures and Options Research* 2, 167–177.

Buraschi, A., and J. C. Jackwerth. (1998). "Explaining Option Prices: Deterministic vs. Stochastic Models," Working Paper, London Business School.

Christie, A. A. (1982). "The Stochastic Behavior of Common Stock Variances, Value, Leverage and Interest Rate Effects," *Journal of Financial Economics* 10, 407–432.

Das, S., and R. Sundaram. (1998). "Of Smiles and Smirks: A Term-Structure Perspective," *Journal of Financial and Quantitative Analysis*, forthcoming.

Derman, E., and I. Kani. (1998). "Stochastic Implied Trees: Arbitrage Pricing with Stochastic Term and Strike Structure of Volatility," *International Journal of Theoretical and Applied Finance* 1, 61–110.

Dumas, B., J. Fleming, and R. E. Whaley. (1998). "Implied Volatility functions: Empirical Tests," *Journal of Finance*, forthcoming.

Dupire, B. (1992). "Arbitrage Pricing with Stochastic Volatility," Working Paper, Societé Generale Division Options, Paris.

Dupire, B. (1993). "Model Art," *Risk* 6, 118–124.

Frachot, A., D. Jansi, and V. Lacoste. (1992). "Factor Analysis of the Term Structure: A Probabilistic Approach," NER, Bank of France.

Franks, J. R., and E. S. Schwartz. (1991). "The Stochastic Behavior of Market Variance Implied in the Prices of Index Options," *The Economic Journal* 101, 1460–1475.

Fung, K. H., and D. A. Hsieh. (1991). "Empirical Analysis of Implied Volatility: Stocks, Bonds and Currencies," Working Paper, Fuqua School of Business, Duke University.

Gallant, A. R., H. Chien, and G. Tauchen. (1998). "Calibrating Volatility Diffusions and Extracting Integrated Volatility," Working Paper, Duke University.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.

Harvey, A. (1993). *Time Series Models*. Second Edition, Harvester Wheatsheaf.

Harvey, C. R., and R. E. Whaley. (1991). "S&P 100 Index Option Volatility," *Journal of Finance* 46, 1551–1561.

Heath, D., R. A. Jarrow, and A. Morton. (1992). "Bond Pricing and the Term Structure of Interest Rates: A New Methodology For Contingent Claims Valuation," *Econometrica* 60, 77–105.

Heynen, R. (1994). "An Empirical Investigation of Observed Smile Patterns," *Review of Futures Markets* 13, 317–354.

Hsieh, D. A. (1993): "Estimating the Dynamics of Volatility," *Proceedings of the Conference on Financial Innovation: 20 Years of Black/Scholes and Merton*, Fuqua School of Business, Duke University, Durham, NC.

Hsieh, D. A. (1995). "Nonlinear Dynamics in Financial Markets: Evidence and Implications," *Financial Analyst Journal* 51, 55–62.

Hull, J., and A. White. (1987). "The Pricing of Options on Assets with Stochastic Volatilities," *Journal of Finance* 42, 281–300.

Jackson, E. (1991). *A User's Guide to Principal Components*. Wiley Series in Probability and Mathematical Statistics.

Johnson, H., and D. Shanno. (1987). "Option Pricing when the Variance is Changing," *Journal of Financial and Quantitative Analysis* 22, 143–151.

Kamal, M., and E. Derman. (1997). "The Patterns of Change in Implied Index Volatilities," Goldman Sachs, Quantitative Strategies Notes.

Ledoit, O., and P. Santa-Clara. (1998). "Relative Pricing of Options with Stochastic Volatility," Working Paper, University of California, Los Angeles.

Litterman, R., and J. Scheinkman. (1988). "Common Factors Affecting Bond Returns," Goldman Sachs, Financial Strategies Group.

Merton, R. C. (1976). "Option Pricing when Underlying Stock Returns are Discontinuous," *Journal of Financial Economics* 3, 125–144.

Roll, R. (1984). "A Simple Implicit Measure of the Effective Bid-Ask Spread," *Journal of Finance* 39, 1127–1139.

Rubinstein, M. (1985). "Non-Parametric Tests of Alternative Option Pricing Models," *Journal of Finance* 40, 455–480.

Schmalensee, R., and R. R. Trippi. (1978). "Common Stock Volatility Expectations Implied By Option Premia," *Journal of Finance* 33, 129–147.

Scott, L. O. (1987). "Option Pricing when the Variance Changes Randomly: Theory, Estimation and an Application," *Journal of Financial and Quantitative Analysis* 22, 419–438.

Scott, L. O. (1997). "Pricing Stock Options in a Jump-Diffusion Model with Stochastic Volatility and Interest Rates: Applications of Fourier Inversion Methods," *Mathematical Finance* 7, 413–426.

Stein, J. (1989). "Overreactions in the Options Market," *Journal of Finance* XLIV, 1011–1023.

Taylor, S. J., and X. Xu. (1994). "The Magnitude of Implied Volatility Smiles: Theory and Empirical Evidence for Exchange Rates," *Review of Futures Markets* 13, 355–380.

Velicer, W. (1976). "Determining the Number of Components from the Matrix of Partial Correlations," *Psychometrica* 41, 321–327.

Wiggins, J. B. (1987). "Option Values under Stochastic Volatility," *Journal of Financial Economics* 19, 351–372.

Xu, X., and S. J. Taylor. (1994). "The Term Structure of Volatility Implied by Foreign Exchange Options," *Journal of Financial and Quantitative Analysis* 29, 57–74.