This version: 9/2021.

1 Outline of the proposed method

We give an outline of the proposed method to help understand the R code.

The method begins with changes of variables:

$$x = \frac{\log(S_T/S) - r\tau}{\sigma\sqrt{\tau}} \tag{1}$$

and

$$z = \frac{\log(K/S) - r\tau}{\sigma\sqrt{\tau}}. \tag{2}$$

where S is the spot price of the underlying asset at time t, r and τ are the riskfree rate and time to maturity, and σ is the Black-Scholes implied volatility for the at-the-money call option. Because the implied volatility is always reported as a summary statistic, these two transformations are straightforward to compute. Let the density after the change of variables be $f(x)$. Then, the call and put options satisfy

$$C(z) = \int_{-\infty}^{\infty} S\left(e^{\sqrt{\tau}\sigma x} - e^{\sqrt{\tau}\sigma z}\right)^{+} f(x)dx$$

and

$$P(z) = \int_{-\infty}^{\infty} S\left(e^{\sqrt{\tau}\sigma z} - e^{\sqrt{\tau}\sigma x}\right)^{+} f(x)dx.$$

We propose to approximate $f(\cdot)$ using a Gauss-Hermite expansion. After estimating $f(\cdot)$, we recover $f^*(\cdot)$, i.e., the SPD for the price level, using

$$f^*(S_T) = \frac{1}{\sigma\sqrt{\tau}S_T} f\left(\frac{\log(S_T/S) - r\tau}{\sigma\sqrt{\tau}}\right).$$

The SPD of the return $R_T = \log(S_T/S)$ is also straightforward to obtain, given by

$$\frac{1}{\sigma\sqrt{\tau}} f\left(\frac{R_T - r\tau}{\sigma\sqrt{\tau}}\right).$$

The transformation in (1) achieves the following two goals at once. First, the division by $\sqrt{\tau}\sigma$ serves as a variance-stabilizing transformation to ensure that the dispersion of $f(x)$ is insensitive to changes in market conditions and the time to maturity. In particular, if the dispersion of $f^*(x)$ increases abruptly due to a large negative shock, the Black-Scholes volatility will rise instantly, and the dispersion of $f(x)$ will remain relatively unchanged. Because σ is computed by inverting the Black-Scholes formula, it is not treated as an unknown parameter in the estimation. Consequently, the estimation problem remains linear in parameters after the transformation. Second, the logarithmic transformation $\log(S_T/S)$ shifts the density's support from the positive axis to the real

line. If $f^*(x)$ is close to the log-normal density, $f(x)$ will be close to the normal density. This is important for obtaining an effective approximation.

Note that $f(x)$ has three features. First, its support has no fixed boundary. There is no simple rescaling of the data that can reduce the support of $f(x)$ to, say, $[0,1]$. Second, this density is closely related to the normal density. When the stochastic process for the underlying asset is a geometric Brownian motion with drift r and volatility σ, this density is exactly $N(0,1)$. Third, the density can have thicker tails than the normal distribution. These three features suggest that Hermite functions are suitable basis functions for approximating $f(x)$. Recall that the Hermite functions $\{h_j\}$ are the complete orthonormal system in $L^2(-\infty,\infty)$ given by

$$h_j(x) = \frac{H_j(x)}{(2^j j! \pi^{1/2})^{1/2}} e^{-x^2/2} \quad (j = 0, 1, 2, ...),$$

where

$$\int_{-\infty}^{\infty} h_j^2(x) dx = 1$$

for all $j = 0, 1, 2, ...,$

$$\int_{-\infty}^{\infty} h_i(x) h_j(x) dx = 0$$

for all $i \neq j$, and $\{H_j\}$ are the standard physicist's Hermite polynomials given by

$$H_j(x) = (-1)^j \, e^{x^2} \frac{d^j}{dx^j} e^{-x^2} \quad (j = 0, 1, 2, ...).$$

By expressing $f(x)$ in terms of $h_j(x)$, we obtain

$$f(x) \sim \sum_{j=0}^{\infty} \beta_j h_j(x),$$

where

$$\beta_j = \int_{-\infty}^{\infty} f(x) h_j(x) dx. \tag{3}$$

Let $\{y_i, z_i\}$ $(i = 1, ..., n)$ denote the sample of observed option prices and transformed strike prices at time t (see (2)). Options with zero open interests or zero transaction volumes can be excluded from the sample to avoid stale information. We assume that the data are ordered such that the first n_c observations are call options, and the remaining $n - n_c$ observations are put options. Then,

$$y_i = \begin{cases} \int_{-\infty}^{\infty} S \left(e^{\sqrt{\tau}\sigma x} - e^{\sqrt{\tau}\sigma z_i} \right)^+ f(x) dx + \varepsilon_i & \text{for } i = 1, ..., n_c, \\ \int_{-\infty}^{\infty} S \left(e^{\sqrt{\tau}\sigma z_i} - e^{\sqrt{\tau}\sigma x} \right)^+ f(x) dx + \varepsilon_i & \text{for } i = n_c + 1, ..., n, \end{cases}$$

where S denotes the spot price of the underlying asset at time t, and ε_i represents deviations from the theoretical prices due to various factors.

The proposed estimation procedure is based on a Gauss-Hermite series approximation to $f(x)$:

$$f(x) \approx \sum_{j=0}^{J} \beta_j h_j(x),$$

where β_j is defined in (3) and J is the truncation order.
STEP 1. For $j = 0, ..., J$, compute

$$x_{i,j} = \begin{cases} \int_{-\infty}^{\infty} S\left(e^{\sqrt{\tau}\sigma x} - e^{\sqrt{\tau}\sigma z_i}\right)^+ h_j(x)dx & \text{for } i = 1, ..., n_c, \\ \int_{-\infty}^{\infty} S\left(e^{\sqrt{\tau}\sigma z_i} - e^{\sqrt{\tau}\sigma x}\right)^+ h_j(x)dx & \text{for } i = n_c + 1, ..., n. \end{cases}$$

STEP 2. Let

$$x_i = (x_{i,0}, ..., x_{i,J})' \text{ and } \beta = (\beta_0, ..., \beta_J)'.$$

Solve the following optimization problem

$$\min_{\beta \in \mathcal{H}_J} \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \beta' Q_\alpha \beta,$$

where

$$\mathcal{H}_J = \left\{ \beta \in R^{J+1} \colon \inf_{x \in \mathbb{R}} \sum_{j=0}^{J} \beta_j h_j(x) \geq \eta \right\},$$

Q_α is a $(J+1)$-dimensional regularization matrix, and η is a small negative constant. Let $\hat{\beta}_0, ..., \hat{\beta}_J$ be the solutions. Compute the SPD estimate as

$$\hat{f}(x) = \sum_{j=0}^{J} \hat{\beta}_j h_j(x).$$

2 The R code

There are three R files in this folder.

1. `main.R`: the main code for estimation;

2. `kfold.R` and `SValpha.R`: supportive files for choosing the penalization parameter α.

There are detailed comments inside these files, explaining the operations involved. Below, we provide some additional details for main.R. This file is organized into eight short sections for ease of adaptation. Section 1: load packages and functions; Section 2: load data; Section 3: data processing; Section 4: set key control parameters; Section 5: set other control parameters; Section 6: select the penalization parameter; Section 7: carry out the estimation; Section 8: display the estimated SPD in a figure. When applying this code, the user need to modify Section 2, and

possibly Section 3. Other parts can be left unchanged. Among the remaining sections, Section 4 is the most important. We provide some additional information below. This section involving setting the values of four control parameters:

- `method`: 1 for the Tikhonov regularization and 2 the modified regularization; 1 is recommended.
- `degree`: order of the Hermite polynomial, J. The default is ceiling(2*(n/log(n))^0.2), with n the sample size.
- `negtol`: this is the η parameter defined above. It stabilizes the estimates by preventing the density from dipping below η. The default is -0.001. This constraint plays a crucial role in the estimation.
- `into1`: this parameter activates an equality constraint, forcing the estimate to integrate to 1 exactly. Set to 1 to make this constraint active. Otherwise, set to 0. This constraint has an effect in some cases when the sample size is small. The default value is 0.