Nonparametric State-Price Density Estimation using High Frequency Data

Jeroen Dalderop*
Faculty of Economics
University of Cambridge

January 20, 2016

Abstract

This paper studies the use of high frequency data to estimate the state-price density (SPD) implicit in option prices. Their large sample size allows estimation of the conditional SPD at any time point of interest, which can be directly used for model-free pricing and hedging and for computing market-implied conditional risk measures. We develop asymptotic theory for a time-varying kernel estimator when the trading times are modelled by point processes whose intensity goes to infinity. The pricing errors and strike prices are mixing, locally stationary time series, which can be weakly dependent with the trading times. Unlike realized volatility estimation, the market microstructural noise in recorded option prices is averaged out and there is no need to subsample the data. We apply the estimator to S&P 500 E-mini European call and put option mid quotes using an iterated plug-in bandwidth, and document the intraday dynamics of the SPD and derived quantities.

Keywords: Option Pricing, Kernel Regression, High Frequency Data, Local Stationarity
JEL Codes: C14, G13

1 Introduction

Option prices contain detailed information about the expectations and preferences of market participants. This information can be summarized in the state-price density (SPD), also known as the risk neutral density, of the underlying asset. This contains the value of Arrow-Debreu securities that pay out when the asset price at maturity falls within the infinitesimal interval

*Email address: jwpd2@cam.ac.uk. I thank my supervisor Oliver Linton for his guidance and support, as well as Seokyoung Hong, Ryoko Ito, Alexei Onatski, Eric Renault, Mike Tehranchi, Steve Thiele, and participants in the Econometrics Workshop at the University of Cambridge and the SoFiE Summer School 2015 for helpful comments. Financial support by CERF is gratefully acknowledged.

$(x, x + \mathrm{d}x)$ for varying levels of the outcome x. High values of the SPD in a particular part of the distribution correspond to a high marginal rate of substitution for wealth in these states.

Estimators of the SPD are useful in at least four contexts. First, they allow traders to compute the no-arbitrage price of new or illiquid derivatives simply by integration. They can also be used to filter out the noise in recorded option prices before converting to implied volatilities (Chen and Xu, 2014). Second, risk managers can compute market-implied values of risk measures such as Value-at-Risk and Expected Shortfall directly from the SPD for the maturity of interest (Aït-Sahalia and Lo, 2000). Third, event studies analyze changes in the SPD to assess the impact of economic events on investor beliefs and preferences (Beber and Brandt, 2006; Birru and Figlewski, 2012). Finally, together with an estimator of the objective density from historical data it allows computing empirical pricing kernels (Rosenberg and Engle, 2000). The finding that the implied Arrow-Pratt risk aversion coefficient does not monotonically decrease with wealth has been referred to as the 'pricing kernel puzzle' (Jackwerth, 2000).

The SPD can be nonparametrically estimated from cross sections of European option prices with varying strike prices. Such estimators are robust to misspecification of the asset return dynamics and associated risk premia (Broadie et al., 2007). Starting from Aït-Sahalia and Lo (1998), several studies estimate the pricing function, or equivalently the implied volatility surface, by nonparametric regression of observed option prices on the stock price, strike price, and time-to-maturity. The SPD then follows by taking the second derivative of the call pricing function with respect to the strike price using the result by Breeden and Litzenberger (1978).

Whilst parametric models can easily accommodate the dynamics of the pricing function in terms of some (possibly latent) state variables, the curse of dimensionality limits the number of state variables such as stochastic volatility and jump intensity that can be included in the nonparametric approach.[1] Time variation has therefore mainly been allowed for by splitting the sample of option prices into short time periods, such as a few months (Aït-Sahalia and Lo, 1998) or trading days (Härdle and Hlavka, 2009; Chen and Xu, 2014), during which the SPD for a given time-to-maturity is assumed constant. These time windows need to be sufficiently large for nonparametric estimators to perform well. With, say, daily data, this would require covering multiple maturity cycles, so that estimates of the SPD are best interpreted as an unconditional average over the subsample period, instead of being conditioned on information available at a specific time point.

The increasing availability of high frequency data allows estimating SPDs accurately without having to pool option prices over long time periods. Infill asymptotics can be used to show that estimators based on observations within a fixed interval are consistent when the sample size goes to infinity. This paper uses high frequency data to estimate the conditional SPD without having

[1] Song and Xiu (2014) estimate the SPD as a function of the VIX as a proxy for spot volatility. Semiparametric models with dynamic factors provide an alternative (Fengler et al., 2007).

to specify the conditioning information. As a result our estimator is able to measure changes in the option pricing function due to updates in investor beliefs or risk aversion in response to market events. The conditional density estimators can be directly used for model-free pricing and hedging of option portfolios and for extracting market-implied conditional quantiles.

Under the assumption that the pricing function varies smoothly over time, we use a kernel estimator that fits the regression curve around any time point of interest. This does not require smoothing with respect to time to maturity, which would use large bandwidths to fill the gaps between the maturities and lead to substantial biases. Smoothing in time requires a sampling window to be chosen for each time point of interest. We investigate a data driven choice of the sampling windows that automatically adapts to the speed of the information flow in the market. This makes it possible to optimize the appropriate bias-variance trade-off which would be neglected when using a fixed sampling period such as a day or a week.

For the theoretical treatment of the time-varying kernel regressor we generalize the asymptotic normality result in Vogt (2012) on locally stationary time series towards random sampling times. These are allowed to be endogenously determined with the covariate, motivated by recent work on the informational content of endogenous trading times (e.g. Li et al. (2009), Renault and Werker (2011)). In particular the sampling times follow a point process with a conditional intensity function that depends on its own history as well as that of the covariates. This framework can incorporate the main empirical features of trading times, such as clustering, interaction with covariates, and time-of-the-day effects that lead to non-stationary trading volume. Under appropriate mixing conditions on the point process and the covariate, when the mean intensity goes to infinity, the time-varying kernel estimator has the same asymptotic variance as when the observations are independent and the sampling times follow a nonstationary Poisson process. We illustrate these conditions by means of both self-exciting and covariate-driven Hawkes processes.

In the empirical application we estimate the conditional state-price density implicit in S&P 500 E-mini European call and put option quotes. We discuss dimension reductions such that the option pricing function can be represented as an unknown function of time and moneyness of the option only. We document the interday and intraday dynamics of the state-price density and derived quantities such as implied volatilities and higher moments, and its quantiles. The bandwidths is chosen by a an iterated plug-in bandwidth procedure with the cubic implied volatility surface in the first step. This leads to effective sampling windows which are typically about half an hour, yet exhibit substantial variation during the day.

The remainder of this paper is organized as follows. Section 2 provides the theoretical treatment of our time-varying kernel regressor under random sampling times. Section 3 discusses its application for estimating the state-price density, and provides empirical results. Section 4 concludes.

2 Time-varying kernel regression with random sampling times

In nonparametric time series models with stationary regressors the regression relationship between two random variables is unknown, yet constant over time. Consistency of commonly used kernel estimators can then obtained by letting the time span go to infinity under weak dependence (see e.g. Robinson (1983), Bosq (1996)). However, in various applications the joint distribution of the random variables may not be stationary, for example due to structural change that is not explicitly modelled with observed variables. In these cases the conditional expectation $E(Y_t|X_t = x) =: m(t, x)$ will be a function of both the time t and the state x. An increasing time span no longer guarantees consistency, as this requires a growing number of observations locally around the time point of interest.

In data environments with a large number of observations within small time windows, such as with high frequency data in finance, this local information on the dependency between the time series is in fact available. This motivates the use of large sample theory based on infill asymptotics. In particular, we study the asymptotic properties of a kernel smoother which smoothes in both the state and the time dimension, when the number of observations within a fixed time interval goes to infinity. As financial transaction and quote level data are known to be irregularly spaced and exhibit clustering, we allow for random sampling times modelled by point processes. The number of observations then grows to infinity stochastically, at a rate determined by the mean intensity.

There are many financial applications that use smoothing in the time domain in the infill setting, particularly to estimate spot volatility (e.g. Barndorff-Nielsen et al. (2008), Kristensen (2010)), spot covariation (e.g. Barndorff-Nielsen and Shephard (2004), Zhang (2011)), or time-varying betas for diffusions (Mykland et al., 2006). Typically the data are treated as discrete time observations from a continuous semimartingale, such as an Itô process. However, for data constructed in this way simultaneously smoothing in the state dimension is generally not possible. Continuity of the sample paths implies that the range of observed value of the semimartingale becomes arbitrary small when the time band vanishes. In the limit only the value at one particular time point is observed, from which no consistent estimator can be created. For example, the dependency between high frequency returns is typically measured by the quadratic covariation, which is a linear measure of dependence. A nonlinear regression model would require rescaling the intraday returns by the sample size. Yet the regression relation may not be scale-invariant so that standard asymptotic theory does not apply.

Instead, we treat our covariate as a piecewise constant continuous-time process, or a marked point process, whose number of jumps (marks) goes to infinity within any given time interval. This applies to the high frequency option pricing setting, where the covariates are the strike prices (or 'moneyness' ratios) of option trades which are scattered over its range in short time periods,

as will be discussed in section 3.[2] This ensures there is sufficient variation in the covariate around each point in time. To formalize this, mixing conditions that control the dependency of the time series need to be adapted to the infill asymptotic setting. Combined with a local stationarity condition on the covariate, this suffices to achieve consistent estimators of the regression function at each time point of interest.

Our setting is similar to that of Vogt (2012), who shows the consistency of the time-varying kernel regressor under local stationarity. Vogt (2012) considers long span asymptotics, but derives consistency in rescaled time. Therefore his results carry over to our setting by rescaling the time dimension to the unit interval. We generalize them to random sampling times that possibly depend on the covariates. The importance of random sampling times on estimators using high frequency data has been emphasized by Aït-Sahalia and Mykland (2003), and Duffie and Glynn (2004), among others. As endogenous sampling times contribute to the overall variation of our estimator we do not condition on them and instead derive the unconditional distribution. Li et al. (2009) have shown the importance of this difference for realized volatility estimators. The interaction with the covariate is modelled via the conditional intensity function which depends on both its own history as well as that of the covariates. There is rapidly growing interest to point process models for financial duration data, see Russell (1999) or Bauwens and Hautsch (2009) for an overview. Related estimation problems come from the literature on marked point processes, see for example Ellis (1991) and Pawlas (2009) for results on density estimation for stationary regressors observed at random time points.

2.1 Model

Let $\{(Y_{i,n}, X_{i,n}) : i \in \mathcal{Z}, n \in \mathcal{N}\}$ be a bivariate stochastic process, observed synchronously at the random times $\{t_{i,n} : i \in \mathcal{Z}, n \in \mathcal{N}\}$.[3] The sampling times $\{t_{i,n}\}$ are modelled as a point process with counting measure $N_n(a,b) = \#\{i : a \leq t_{i,n} \leq b\}$.[4] Define also the counting process $N_n(t) := N_n(0,t)$ of the number of events up to time $t \in \mathbb{R}$. Throughout we use double-index notation, where the second index n creates a sequence of point processes with increasing mean arrival rate. The information set for each n is described by the natural filtration $\mathcal{F}^N_{t,n}$ and $\mathcal{F}^x_{t,n}$ of the counting process $\{N_n(t)\}$ and the covariates $\{X_{i,n}\}$ observed before time t, respectively, and their joint filtration $\mathcal{F}_{t,n} = \mathcal{F}^N_{t,n} \bigcup \mathcal{F}^x_{t,n}$.

[2]The strike prices follow a discrete time series, as there is a finite set of values traded each day. However, the moneyness ratio divides the strike prices by the futures price, leading to a continuum of possible values.

[3]Asynchronous observations are often studied for spot covariance estimation using high frequency return data, see e.g. Bibinger et al. (2015); Linton et al. (2015). The synchronization error between option prices and asset prices is likely to be small as the latter are observed at much higher frequency, see section 3.3.

[4]Formally, $\{(t_{i,n}, Y_{i,n}, X_{i,n})\}$ is a univariate marked point process, with bivariate vector of marks $(Y_{i,n}, X_{i,n})$. In the case of a asynchronous sampling times, this is a bivariate marked point process, with univariate marks.

Consider the time-varying nonparametric regression model

$$Y_{i,n} = m(t_{i,n}, X_{i,n}) + \epsilon_{i,n}, \tag{1}$$

with $E(\epsilon_{i,n}|t_{i,n}, X_{i,n}) = 0$ and $\text{Var}(\epsilon_{i,n}|t_{i,n} = t, X_{i,n} = x) = \sigma^2(t,x)$. We are interested in estimating the conditional expectation function $m(t,x)$ of $Y_{i,n}$ given $t_{i,n} = t$ and $X_{i,n} = x$ on a fixed time interval $(0,T)$. The total number of observations in this time window is $N_n(0,T)$. Hereafter without loss of generality we set $T = 1$.

2.2 Estimator

A natural estimator for $m(\cdot)$ at the design point (t,x) is the Nadaraya-Watson (or locally constant) kernel estimator

$$\hat{m}_h(t,x;0) = \frac{\sum_{i=N_n(0)+1}^{N_n(1)} K_{h_t}(t - t_{i,n}) K_{h_x}(x - X_{i,n}) Y_{i,n}}{\sum_{i=N_n(0)+1}^{N_n(1)} K_{h_t}(t - t_{i,n}) K_{h_x}(x - X_{i,n})}, \tag{2}$$

with $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$ for some kernel function K. This estimator takes a weighted average over data points close to (t,x), where the weights are controlled by the bandwidths h_t and h_x for the time and state dimension, respectively. The kernel smoothing method can be generalized to fitting a local polynomial around (t,x), which yields automatic estimates of the partial derivatives. Fan and Gijbels (1996) recommend using a $p+1$ polynomial when the interest in the p-the derivative. While we implement the cubic polynomial in the empirical application, for ease of exposition this section only discusses the locally constant estimator.

For kernels with bounded support, say, $[-1,1]$, the sum in (2) has only $N_n(t - h_t, t + h_t)$ non-zero terms for any $t \in (h_t, 1 - h_t)$. In this case the sum needs only to be taken from $i = N_n(t - h_t) + 1$ to $i = N_n(t + h_t)$. We study the limiting distribution of this estimator when n goes to infinity, that is, when the expected number of observations within any nonzero interval becomes infinitely large.

Define the piecewise constant stochastic process

$$X_{\{u\},n} := X_{N_n(u),n}, \qquad\qquad u \in (0,1)$$

and similarly for $Y\{u\}, n$ and $\epsilon_{\{u\},n}$. Note that $X_{\{u\},n}$ is a piecewise deterministic, cádlág function on the real line. It can be seen as the value of the covariate after the last jump before or at time u. With this notation, for a kernel with support $[-1,1]$, the estimator (2) can be written as a

stochastic integral with respect to the counting measure $N_n(t)$

$$\hat{m}_h(t,x;0) = \frac{\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)K_{h_x}(x-X_{\{u\},n})N_n(\mathrm{d}u)Y_{\{u\},n}}{\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)K_{h_x}(x-X_{\{u\},n})N_n(\mathrm{d}u)}. \tag{3}$$

2.3 Random sampling times

Assume that the counting processes are such that for every n and $t \in \mathbb{R}$

$$\begin{aligned}
\mathbb{P}(N_n(t,t+\Delta) = 1 | \mathcal{F}_t^N, \mathcal{F}_t^x) &= \lambda_n(t)\Delta + o_P(\Delta), \\
\mathbb{P}(N_n(t,t+\Delta) > 1 | \mathcal{F}_t^N, \mathcal{F}_t^x) &= o_P(\Delta),
\end{aligned}$$

for $\Delta \to 0$, which defines the conditional intensity function $\lambda_n(t)$. Note that $\lambda_n(t)$ depends on both past event times as well as past observed covariates.The second property is known as orderliness and ensures there are no simultaneous events, see Daley and Vere-Jones (2003) for details. It implies that

$$\mathrm{Var}(\mathrm{d}N_n(t)) = E\{(\mathrm{d}N_n(t))^2\} = E(\mathrm{d}N_n(t)) = E(\lambda_n(t))\mathrm{d}t.$$

Furthermore assume that $N_n(t)$ admits a covariance density, defined for $t \neq s$ by

$$\mu_n(t,s) = \frac{\mathrm{Cov}(N_n(t,t+\mathrm{d}t), N_n(s,s+\mathrm{d}s))}{\mathrm{d}t\mathrm{d}s},$$

while $\mu_n(t,t)$ is defined such that the function is continuous.

The following conditions require the average intensity and covariance density at each time point to be proportional to n and n^2, respectively[5]:

(C1) There exists a twice differentiable function $\nu(t)$ on $(0,1)$, with $\inf_{t\in(0,1)} \nu(t) > 0$ and $\sup_{t\in(0,1)} \nu(t) < C_1 < \infty$, such that for every n

$$E(\lambda_n(t)) = n\nu(t). \tag{4}$$

(C2) For each time $t \in (0,1)$, there exists a continuous function $\gamma_t : \mathbb{R} \mapsto \mathbb{R}$, with $\sup_{\tau\in\mathcal{R}} \gamma_t(\tau) < C_2 < \infty$ and

$$\eta(t) := \int_{-\infty}^{\infty} \gamma_t(\tau)\mathrm{d}\tau < \infty,$$

such that

$$\frac{\mu_n(t, t+\frac{\tau}{n})}{n^2} = \gamma_t(\tau).$$

This set-up allows for general nonstationarity of the mean intensity and covariance density of the

[5]More generally, we could require the intensity function to converge uniformly to its limit at some rate $a_n \to \infty$. All results would go through by appropriately rescaling the bandwidths.

point process. It reduces to the stationary case when $\nu(t) = \nu$ and $\gamma_t(\tau) = \gamma(\tau)$.

Example. A natural specification of the sequence of counting processes $\{N_n(t)\}$ is

$$N_n(t) = N\left(n\int_0^t \pi(s)\mathrm{d}s\right), \tag{5}$$

for $t \in (0,1)$, where $N(\cdot)$ is a stationary point process with intensity ρ and covariance density $\mu(\cdot)$. In this specification the index n 'speeds up the clock', whereas the function $\pi : (0,1) \mapsto \mathbb{R}_+$ captures a deterministic pattern in the mean intensity, satisfying without loss of generality $\int_0^1 \pi(s)\mathrm{d}s = 1$. It holds that

$$\begin{aligned} E(\lambda_n(t)) &= n\rho\pi(t) := n\nu(t), \\ \frac{\mu_n(t, t+\frac{\tau}{n})}{n^2} &= \pi^2(t)\mu(\tau\pi(t)) := \gamma_t(\tau). \end{aligned}$$

□

The unconditional local intensity $\nu(t)$ can be estimated as

$$\hat{\nu}_h(t) = \frac{1}{n}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)N_n(du), \tag{6}$$

for a kernel K with $\mu_1(K) = \int xK(x)dx = 1$, $\mu_2(K) = \int x^2K(x)dx < \infty$, and $R(K) = \int K^2(x)dx < \infty$. When $n \to \infty$, $h_t \to 0$, and $nh_t \to \infty$, this estimator has bias

$$E\left(\hat{\nu}_h(t) - \nu(t)\right) = \frac{1}{2}\mu_2(K)h_t^2\nu''(t) + o(h_t^2), \tag{7}$$

and variance (see Appendix)

$$\mathrm{Var}\left(\hat{\nu}_h(t)\right) = \frac{R(K)}{nh_t}\{\nu(t) + \eta(t)\} + o\left(\frac{1}{nh_t}\right). \tag{8}$$

Therefore the asymptotic MSE optimal bandwidth of $\hat{\nu}_h(t)$ is given by

$$h^*_{AMSE}(t) = \left(\frac{\{\nu(t) + \eta(t)\}R(K)}{n\nu''(t)^2\mu_2^2(K)}\right)^{\frac{1}{5}}. \tag{9}$$

This shows the optimal bandwidth increases with the long-run variance of the point process, and decreases with the local curvature of the base-line intensity.

Example. The linear self-exciting process by Hawkes (1971) has conditional intensity function

$$\lambda(t) = \phi + \int_{-\infty}^t g(t-u)N(\mathrm{d}u), \tag{10}$$

for some parameter ν and non-negative 'infectivity' function $g(\cdot)$. A common choice is the exponential model $g(u) = \alpha e^{-\beta u}$, in which case the process is stationary for $\frac{\alpha}{\beta} < 1$, with mean intensity $\rho = \frac{\phi\beta}{\beta-\alpha}$ and covariance density $\mu(t,s) = \frac{\alpha\rho(2\beta-\alpha)}{2(\beta-\alpha)} e^{-(\beta-\alpha)|t-s|}$ for any (t,s). In the speeding-up setting of (5) with $N(\cdot)$ a Hawkes process it holds that

$$\nu(t) = \frac{\pi(t)\phi\beta}{\beta-\alpha},$$
$$\eta(t) = \int_{-\infty}^{\infty} \pi^2(t)\mu(\tau\pi(t))\mathrm{d}\tau = \frac{\pi(t)\alpha\phi(2\beta-\alpha)}{\beta-\alpha}.$$

Hence when the bandwidth for the smoother is chosen based on the Hawkes model, (9) reduces to

$$h^*_{AMSE}(t) = \left(\frac{\pi(t)(1-\alpha/\beta)(1+\alpha(2-\alpha/\beta))R(K)}{n\phi\pi''(t)^2\mu_2^2(K)} \right)^{\frac{1}{5}}.$$

□

2.4 Asymptotic normality

For the asymptotic analysis we need to impose some mixing conditions that control the temporal dependence of the covariates and errors as well as the point process. Therefore we adapt the usual mixing conditions to the infill setting. For point processes it is common to assume that event counts within two non-overlapping time intervals become independent once there is sufficient time between them. When the expected number of observations grows with n within each time interval, the distance between two time intervals needs to be rescaled with n to maintain asymptotic independence. This is reflected in the mixing condition below. Define

$$\mathcal{F}^b_{a,n} = \sigma\{N_n(A), A \subseteq (a,b), X_{\{t\},n} : a \le t \le b\}.$$

Definition. (i) A marked point processes $(N(t), X_{\{t\}})$ with natural filtration $\mathcal{F}_t$ is α-mixing if $\alpha(s) \to 0$ when $s \to \infty$, where

$$\alpha(s) = \sup_{\substack{A\in\mathcal{F}^u_{-\infty}, B\in\mathcal{F}^\infty_v \\ |v-u|\ge s}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| \tag{11}$$

(ii) A sequence of marked point processes $(N_n(t), X_{\{t\},n})$ with family of natural filtrations $\{\mathcal{F}_{t,n}\}$ is α-mixing *in rescaled time* if $\alpha(s) \to 0$ when $s \to \infty$, where

$$\alpha(s) = \sup_{n\ge 1} \sup_{\substack{A\in\mathcal{F}^u_{-\infty,n}, B\in\mathcal{F}^\infty_{v,n} \\ |v-u|\ge \frac{s}{n}}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| \tag{12}$$

Part (i) in this definition is a standard mixing conditions for point processes to be found

for example in Cranwell and Weiss (1978), or for marked point processes in Pawlas (2009). The mixing condition in rescaled time in part (ii) differs from this one in that for each n, the supremum is taken over intervals whose gap is at least $\frac{s}{n}$. This requires the dependency to vanish proportional to the expected number of observations, or equivalently the integrated intensity, in between intervals, rather than their mere distance in time. In the context of high frequency trading, this requires the effect of market events in the past on the current activity to decrease proportionally to the number of trades, which reflects the speed at which new information is handled, in between them.

Example. (cont.) In the speeding-up setting of (5), it follows from the definition that α-mixing (in the usual sense) of the baseline process $N(t)$ ensures that the sequence of point processes $\{N_n(t)\}$ is α-mixing in rescaled time. □

Hence we impose the following condition on the point process:

(C3) $\{(N_n(t), X_{\{t\},n}, \epsilon_{\{t\},n})\}$ is α-mixing in rescaled time, with the mixing coefficients satisfying

$$\int_0^\infty w^\lambda \alpha(w)^{\frac{\delta}{2+\delta}} \mathrm{d}w < \infty, \qquad \text{for some } \delta > 0 \text{ and } \lambda > \frac{\delta}{2+\delta}. \tag{13}$$

Since we are interested in estimating a regression function locally in time, we do not need to impose global stationarity of the covariates. Instead, we only require the following local stationarity condition, which is an adaptation of the one in Vogt (2012) to allow for random sampling times:

(C4) The process $\{X_{i,n}\}$ observed at times $\{t_{i,n}\}$ is locally stationary (see e.g. Vogt (2012)), i.e. for each time point $t \in [0,1]$ there an exists a stationary process $\{X^*_{i,n}(t)\}$, such that

$$|X_{i,n} - X^*_{i,n}(t)| \leq \left(\left|t_{i,n} - t\right| + \frac{1}{n}\right) U_{i,n}(t) \quad \text{a.s.},$$

where $U_{i,n}(t)$ is a positive-valued process satisfying $E(U_{i,n}(t)^\rho \big| t_{i,n}) < C$ a.s. for some $\rho > 0$ and $C < \infty$ independent of t, i, and n.

Furthermore, we require the following regularity conditions:

(C5) The distribution function of the stationary process $X^*_{i,n}(t)$ conditional on being observed at time t, $\mathbb{P}(X^*_{\{t\},n}(t) \leq x | N_n(\mathrm{d}t) = 1)$, has continuously differentiable density $f(t,x)$ independent of n, and $0 < f(t,x) \leq C$ for all $(t,x) \in (0,1) \times S$ with S a compact subset of $\mathbb{R}$. Furthermore the density $f_{X_{i,n}}$, the conditional density $f_{\epsilon_{i,n}|X_{i,n}}$ and the joint conditional density $f_{\epsilon_{i,n},\epsilon_{j,n}|X_{i,n},X_{j,n}}$ are bounded for every i, j and n.

(C6) The kernel K is a density, symmetric around zero and with zero value outside $[-1,1]$. Define $R(K) = \int K^2(x)dx < \infty$ and $\mu_2(K) = \int x^2 K(x)dx < \infty$. Furthermore K is bounded and

Lipschitz continuous, i.e. there exists an $L < \infty$ such that $|K(u) - K(v)| \leq L|u - v|$ for every $u, v \in \mathbb{R}$.

(C7) The regression function $m(t, x)$ is twice continuously partially differentiable and Lipschitz continuous w.r.t. t and x. $\sigma(t, x) > 0$ is continuously differentiable in both arguments.

(C8) $E\left|\epsilon_{i,n}\right|^{2+\delta} < \infty$ for every i, n

(C9) The bandwidths satisfy $h_t \to 0$, $h_x \to 0$, $nh_t h_x \to \infty$, $nh_t^5 h_x = O(1)$, $nh_t h_x^5 = O(1)$, $h_t(nh_x)^{1-2r} \to 0$ with $r = \min(\rho, 1)$, and $(nh_t)^{1-\epsilon} h_x^{\frac{\lambda(2+\delta)+2+2\delta}{\lambda(2+\delta)+2}} = O(1)$ for some $\epsilon > 0$.

Theorem 1. Let (C1)-(C9) hold. Then for $(t, x) \in (h_t, 1 - h_t) \times S$

$$\sqrt{nh_t h_x}(\hat{m}_h(t, x; 0) - m(t, x) - B(t, x)) \xrightarrow{d} \mathrm{N}(0, V(t, x)), \tag{14}$$

with asymptotic bias

$$\begin{aligned} B(t, x) = {} & h_t^2 \mu_2(K) \left(\frac{\partial}{\partial t} m(t, x) \frac{\partial}{\partial t} \log(f(t, x)\nu(t)) + \frac{1}{2} \frac{\partial^2}{\partial t^2} m(t, x) \right) \\ & + h_x^2 \mu_2(K) \left(\frac{\partial}{\partial x} m(t, x) \frac{\partial}{\partial x} \log f(t, x) + \frac{1}{2} \frac{\partial^2}{\partial x^2} m(t, x) \right) + o(h_t^2 + h_x^2), \end{aligned} \tag{15}$$

and asymptotic variance

$$V(t, x) = \frac{R^2(K)\sigma^2(t, x)}{nh_t h_x \nu(t) f(t, x)} + o\left(\frac{1}{nh_t h_x} \right). \tag{16}$$

Proof. See the Appendix. □

This result shows that the asymptotic distribution of our estimator is the same as that of independent observations. The long run variance of the point processes does not contribute to the asymptotic variance here due to the local weighting of the covariate which effective reshuffles the data under the mixing condition. Compared to fixed, equally spaced sampling times, random sampling times lead to the additional term $\nu(t)$ in the asymptotic bias and variance expressions. A higher mean intensity leads to more observations and hence to a lower variance. This factor may depend on the distribution of the covariates when these contribute to the conditional intensity, as illustrated in the following example.

Example. Suppose that the conditional intensity is specified by the marked Hawkes process

$$\frac{\lambda_n(t)}{n} = \phi\pi(t) + \int_{-\infty}^{t} h\left(X_{\{u\},n}\right) g(n(t - u)) \mathrm{d}N_n(u),$$

where $h(\cdot)$ is a nonnegative, Lipschitz continuous function that captures the impact of the covariate, and $g(\cdot)$ is the infectivity function as before, now with its argument scaled by n. Then

it can be shown (see Appendix) that

$$\frac{E(\lambda_n(t))}{n} \to \frac{\phi\pi(t)}{1 - \int_S h(x)f(t,x)\mathrm{d}x \int_0^\infty g(\tau)\mathrm{d}\tau} =: \nu(t), \tag{17}$$

provided $\int_0^\infty g(\tau)\mathrm{d}\tau < \infty$, $\int_0^\infty \tau g(\tau)\mathrm{d}\tau < \infty$, and the stationary condition

$$\int_S h(x)f(t,x)\mathrm{d}x \int_0^\infty g(\tau)\mathrm{d}\tau < 1. \tag{18}$$

Hence the limiting local intensity can be expressed in terms of the density $f(t,x)$ of the stationary approximation process. The likelihood functions of this and similar models are available (Daley and Vere-Jones, 2003, §7.3) and can be used to estimate the parameters in $\nu(t)$. □

3 Estimating the State-Price Density from European options

The state-price density (SPD) is a central object in two main branches of asset pricing theory. In no arbitrage-based models, the SPD is the density of the Equivalent Martingale Measure (EMM), typically denoted $\mathbb{Q}$, under which all discounted price processes are martingales. The existence of the EMM follows from the absence of arbitrage, while its uniqueness follows from market completeness. In consumption-based equilibrium models, the SPD is the product of the density of the objective measure $\mathbb{P}$, and a stochastic discount factor or pricing kernel. The latter is proportional to the marginal rate of substitution over states of a representative agent.

Since European options only identify the SPD, identification of the objective probability measure or the stochastic discount factor requires imposing further structure on at least one of them. This could be done by parametrizing the dynamics of the underlying asset or the utility function of a representative investor. The resulting theoretical option prices yield moment conditions which can be estimated by GMM techniques or simulation methods. However the classical log-normal Black-Scholes model as well as more flexible parametric models have been reported to miss important features of the data (Bakshi et al., 1997). As a result, estimated risk premia are notably sensitive to the precise model specification (Broadie et al., 2007).

Alternatively, the call pricing function can be estimated via nonparametric regression of observed call prices on their trade characteristics stock price, strike price, and time-to-maturity (Aït-Sahalia and Lo, 1998). The SPD is then obtained from the second derivative of call pricing function with respect to the strike price, using the result by Breeden and Litzenberger (1978). The availability of intraday option prices greatly enhances the applicability of data-intensive nonparametric methods for estimating the SPD. Proposed methods include kernel smoothers (Aït-Sahalia and Lo, 1998; Aït-Sahalia and Duarte, 2003), smoothing splines (Yatchew and Härdle, 2006), Hermite polynomial expansions (Xiu, 2014), model-guided nonparametric correction (Fan

and Mancini, 2009), and methods explicitly incorporating the no-arbitrage shape constraints (Birke and Pilz, 2009; Fengler and Hin, 2015).

An important difference in the literature is the horizon over which option prices are pooled together. Aït-Sahalia and Lo (1998) assume that call prices over a period of about a year are generated by the same call pricing function. Their estimate of the SPD can therefore be interpreted as the average SPD over the sample period, rather than a point estimate for a given point in time. Alternatively, Aït-Sahalia and Duarte (2003) show how a cross-section of option prices at a given point in time can suffice to get feasible estimates. They use shape constraints such as monotonicity and convexity to reduce the variance of their estimator. However, since the number of traded strikes is limited by the exchange, their estimator is vulnerable to small sample issues.

Yet whereas the SPD may not be constant in the long run, it likely does exhibit some persistency in the short run. The availability of high frequency options data allows increasing the sample size dramatically without introducing large biases due to structural change. The SPD is unlikely to vary heavily during small intradaily time windows. Indeed, by now there are a number of studies that use intradaily options data to estimate the call pricing function, e.g. Härdle and Hlavka (2009) and Chen and Xu (2014). However, the call pricing function may still vary throughout the day due to time-of-the-day effects, or indeed changes in the SPD due to new information. Empirical evidence of such intraday patterns comes from autocorrelation in the errors of a time-homogeneous model (Härdle and Hlavka, 2009).

This motivates the application of a time-varying regression model for the call pricing function. This makes it possible to estimate the SPD at any time point of interest, and to study the dynamics of the SPD over a given sampling period. Rather than having to set an arbitrary pooling window, such a model is naturally capable to balance the trade-off between increasing the sample size to reduce variance while controlling the bias due to time-variation in the SPD.

3.1 Identification

A European call option is a contingent claim which gives the owner the right but not the obligation to buy the stock at maturity time T for a given strike price K. Hence its payoff at maturity equals $C_T = (S_T - K)^+$. Absence of arbitrage and complete markets implies there exists a unique measure $\mathbb{Q}$ such that its price at time t can be written as (see e.g. Karatzas and Shreve (1998) for details)

$$C_t = e^{-r(T-t)} E^{\mathbb{Q}} \left((S_T - K)^+ | \mathcal{F}_t \right), \tag{19}$$

where r is the riskfree rate of the bond B. An analogue expression holds for the price of European put option with payoff $P_T = (K - S_T)^+$. Provided the measure $\mathbb{Q}$ admits a density $f_t^{\mathbb{Q}}$ with respect

to the Lebesque measure for information set $\mathcal{F}_t$, (19) can be expressed as

$$C_t = e^{-r(T-t)} \int_K^\infty (S_T - K) f_t^{\mathbb{Q}}(S_T) dS_T. \tag{20}$$

Differentiating with respect to K yields

$$\frac{\partial C_t}{\partial K} = -e^{-r(T-t)} \int_K^\infty f_t^{\mathbb{Q}}(S_T) dS_T. \tag{21}$$

By differentiating once more, Breeden and Litzenberger (1978) derive that

$$\frac{\partial^2 C_t}{\partial K^2} = \mathrm{e}^{-r(T-t)} f_{t,S_T}^{\mathbb{Q}}(K). \tag{22}$$

This relation identifies the SPD as the curvature of the call pricing function with respect to the strike price.

The fact that $f^{\mathbb{Q}}$ is a density also puts some restrictions on the call pricing function. In particular, from (21) it follows that

$$-e^{-r(T-t)} \leq \frac{\partial C_t}{\partial K} \leq 0, \tag{23}$$

while (22) implies that

$$\frac{\partial^2 C_t}{\partial K^2} \geq 0. \tag{24}$$

Hence the call pricing function must be decreasing and convex. Any local violation implies negative state-prices, and hence arbitrage opportunities. Further bounds can be obtained for the level of the call pricing function, see Aït-Sahalia and Duarte (2003) for details.

3.2 Estimation method

From (20) it can be seen that the no arbitrage price of a call option is determined by the SPD $f_t^{\mathbb{Q}}(S_T)$, and by the observable variables strike price K, and time-to-maturity $T-t$. The Breeden and Litzenberger (1978) result (22) gives the relation (up to integration constants) between the SPD and the price of a call option C as a function of its strike price K. Hence an estimator of the SPD follows directly from an estimator of the call pricing function.

However, both the call pricing function and the SPD are random, as they are a conditional expectation (resp. density) given information in the filtration $\mathcal{F}_t$. Therefore it is convenient to make a Markov assumption, which states that all relevant information from the past is summarized by the present level of some state variables. These can either be observed, such as the current stock price S_t, or unobserved, such as the spot volatility (Song and Xiu, 2012). As we wish to avoid specifying which state variables are relevant, we propose to use the current level of the

stock price as the only random state variable, but in addition include time itself in the regressor set. This means that the randomness in the call prices can only be generated by the underlying stock price and other observable trade characteristics, yet we are able to capture smooth changes in the functional form of the dependency of the call price on these variables. This leads to the nonparametric regression model

$$C_t = C(t, \tau, S_t, K_t, r_{t,T}) + \epsilon_t, \tag{25}$$

where ϵ_t is an error term that satisfies $E(\epsilon_t|\mathcal{F}_t) = 0$.[6]

3.2.1 Dimension reduction

Since nonparametric regression is vulnerable to the curse of dimensionality, we make some dimension reductions that are common in the literature. Firstly, since we deal with options written on the price of a futures contract F_T, rather than price of a stock S_T, the interest rate $r_{t,T}$ only affects the time discount factor and not the SPD of the futures price. [7][8] Therefore we work with the inversely discounted prices $C_t^d = e^{r_{t,T}(T-t)}C_t$, so that the interest rate $r_{t,T}$ drops out of the set of regressors:

$$C_t^d = C(t, \tau, F_t, K_t) + \epsilon_t. \tag{26}$$

Second, smoothing in the time-to-maturity dimension τ requires combining information on European option contracts with different maturities. Since the gaps between the maturities that are traded may take several weeks or even months, a long sampling period is needed to get repeated observations of the same value of τ. When working with high frequency data from only one or a few days, this gap cannot be closed and hence it is not possible to cover the whole term structure. Instead, high frequency studies typically produce separate estimators for a finite set of maturity dates (e.g. Härdle and Hlavka, 2009). Therefore we restrict attention to options with the same maturity date T, so that we do not smooth explicitly in time-to-maturity dimension τ. Note that when only one maturity date T is in the sample, then t and τ are not separately identified. In this case only the combined effect of time variation due to either time discounting or new information arising is identified. However, for short intradaily windows the time-to-maturity effect is likely to be small, and can be controlled for by an appropriate rescaling of the resulting densities.

[6] Even more generally, the option price may depend on the dividend yield $\delta_{t,T}$. For options that are written on the futures prices, these are incorporated and hence do not affect option prices separately. For options written on stock prices, they may do so. Still it can be reasonably assumed that the dividend yield enters the option price only via the futures price $F_{t,T} = S_t e^{(r_{t,T}-\delta_{t,T})\tau}$, see Aït-Sahalia and Lo (1998) for details.

[7] If the futures contract has the same maturity as the option, then the prices of a European option written on the stock and on the futures contract are the same (Aït-Sahalia and Lo, 1998). If the futures contract expires after the maturity of the option, this need no longer hold. Yet this only matters insofar the resulting estimate is now the SPD of the futures price, which only differs from that of the stock price by a known discount factor.

[8] Alternatively, for short, intradaily time windows the interest rate is unlikely to change heavily and can therefore safely be ignored.

Finally, we normalize the strike prices K_t by the futures price level F_t at the time of trade, which yields the 'moneyness' ratio $M = K/F$. The rationale for this is provided by Merton (1973, Thm. 9), who shows that if the distribution of the *return* of the underlying asset S is independent from the *level* of S, then the call pricing function is homogeneous of order one in F and K, i.e.

$$C_t(F, K) = FC_t(1, K/F). \tag{27}$$

The class of models that satisfies this assumption includes the widely used geometric Brownian motion and many stochastic volatility models (Joshi, 2001). Yet in general, this condition may not be easy to verify empirically as it requires measuring changes in the return distribution. Under the homogeneity assumption, the regression function in (26) for the normalized option price $\tilde{C}_t = C_t^d/F_t$ becomes

$$\tilde{C}_t = m(t, M_t) + \epsilon_t, \tag{28}$$

with $E(\epsilon_t|M_t) = 0$ and $\text{Var}(\epsilon_t|M_t) = \sigma^2(t, M_t)$. This reduces the regression model to a function of only two variables, time t and moneyness M_t.[9]

Moreover, homogeneity of F and K implies that the second derivative of the call pricing function becomes

$$\frac{\partial^2 C_t^d}{\partial K^2} = \frac{\partial^2 F_t m(t, K/F_t)}{\partial K^2} = \frac{1}{F_t}\frac{\partial^2 m(t, M)}{\partial M^2}. \tag{29}$$

Remembering the Breeden and Litzenberger (1978) result (22), and using the substitution

$$f^{\mathbb{Q}}_{F_T/F_t}(K/F_t) = f^{\mathbb{Q}}_{F_T}(K)F_t, \tag{30}$$

we derive that

$$f^{\mathbb{Q}}_{F_T/F_t}(x) = \frac{\partial^2 \tilde{m}(t, M)}{\partial M^2}\Big|_{M=x}. \tag{31}$$

This provides a direct relation between the SPD of the gross return F_T/F_t and our regression function $m(\cdot)$.

3.2.2 Time-varying kernel regression

The bivariate regression model (28) can be estimated using (2), which takes a weighted average over the observed call prices with weights depend on the distance in terms of time and moneyness. To apply the asymptotic theory derived in Section 2, assume that we have a sample of option prices C_{t_i}, strike prices K_i, and futures price level F_i, observed at normalized times $0 <= t_1 < ... < t_{N_n(1)} <= 1$ which are generated by a point process model with mean intensity proportional to n. Hence the observed sample size is $N_n(1)$, after normalizing $N_n(0) = 0$.

[9]The homogeneity assumption also has been used recently by Chen and Xu (2014), who use intradaily data to estimate the call pricing function nonparametrically as a function of moneyness and time-to-maturity only.

The locally constant estimator (2) can be generalized to local polynomial smoothing of order p, to remove leading bias terms while maintaining the same order of bias at the boundaries (Fan and Gijbels, 1996). In this case the coefficients of the approximating polynomial around the design point (t, x) are chosen to minimize the weighted least squares criterion

$$\sum_{i=1}^{N} \left(\tilde{C}_{t_i} - \sum_{k=0}^{p} \sum_{\{a,b:a+b=k\}} \beta_{a,b}(t,x)(t_i - t)^a (M_{t_i} - x)^b \right)^2 K_{h_t}(t_i - t) K_{h_M}(M_{t_i} - x), \tag{32}$$

with the weight of observation i given by $K_{h_t}(t_i - t)K_{h_M}(M_{t_i} - x)$. The estimated call price is then given by $\hat{\hat{m}}(t, x; p) = \hat{\beta}_{0,0}$, which is the constant coefficient of the fitted p-th order polynomial. The estimated first and second derivative are given by $\hat{\beta}_{0,1}$ and $2\hat{\beta}_{0,2}$, respectively. This requires that p is of order at least two. Fan and Gijbels (1996) recommend to choose to fit a polynomial of one degree higher than the derivative of interest, to eliminate a leading bias term. Since our object of interest, the state price density, is the second derivative we set $p = 3$ and fit a local cubic polynomial. The coefficient of the quadratic moneyness term then gives a direct estimate the SPD from higher coefficients in the polynomial. Using the locally constant estimator would require taking numerical differences twice, the properties of which are studied in Aït-Sahalia and Duarte (2003). The asymptotic normality result in Theorem 1 for locally constant regression can be adapted to the local polynomial setting, in which case the bias and variance order terms depend on higher order partial derivatives. This merely requires some further smoothness and differentiability conditions on the regression function m.

We now comment on the validity of the conditions of Theorem 1 for high frequency option prices, in particular on the local stationarity and the mixing conditions for the moneyness covariate. Recall that moneyness is the ratio of two components, the traded strike price and the futures prices at the time of trade. The range of strike prices that is listed on the exchange is typically set on a daily basis in reaction to movements in the underlying.[10] Hence moneyness is recentered around one, which guarantees stationarity from day to day. However, it may still be nonstationary within a certain day. To see this note for a given strike price K' that is listed on a certain day, we have $\Delta \log M_{t_i} = \Delta \log \frac{F_{t_i}}{K'} = \log \frac{F_{t_i}}{F_{t_{i-1}}}$. So if intraday log futures returns are stationary then $\log M_t$ has a unit root and hence M_t is nonstationary. Yet, also within a day, we might expect traded strike prices to vary along with the underlying. When this happens in such a way that the relative amount of trading at different levels of moneyness is (roughly) constant during the day, M_t is (locally) stationary. Regarding the mixing condition, the effective recentering of the moneyness level around one similarly rules out any long-term memory in the futures price level. Hence as long as the strike prices being traded do not posses some long-memory property,

[10] Our dataset is from the E-mini S&P 500, which lists strike prices at regular intervals within +-50% of the closing price of the futures on the previous day. See the contract specifications on the CME Group website for details.

the mixing condition for the moneyness covariate seems sensible. It would for example hold if the strike prices follow a finite-order Markov chain, regardless of the trading intensity. The next section describes some empirical properties of M_t to illustrate how these conditions apply to our setting.

3.3 Empirical results

3.3.1 Data Description

Our dataset consists of the transaction prices and bid and ask quotes of European options on a futures contract on the E-mini S&P 500 equity index, and of the underlying futures contract themselves. The bid and ask quotes are 'best' quotes, that is, only ask (bid) quotes entering the order book that are lower (higher) than the previously best quote are recorded. They are traded on the electronic GLOBEX platform and obtained from the CME Group. The sample period consists of all trading days between November 1 and November 29, 2013. The options are End-of-Month options, which deliver the futures contract at 3:00 PM Chicago Time on the last trading day of the month, cash-settled with the strike price. Note the futures contract is the nearest-expiring quarterly E-mini S&P 500 futures after maturity of the option. Hence the futures contract expires beyond the maturity time of the option. For example, for the November End-of-Month options the relevant futures contract is the December futures contract. However, since there exists a simple no-arbitrage relation between the futures prices and the equity index, see footnote 6, we can restrict attention to the SPD of the futures price at maturity of the option. This should be the same as the SPD of the equity index up to a known discount factor.

The options and futures on the E-mini S&P 500 are traded around the clock, except a daily maintenance interval from 4:15-5:00 PM. The time stamps are recorded in seconds. Since several transactions often take place within the same second, the data are ordered in terms of their trade sequence number. This numbering is reset at 5:00 PM every day, so that observations after 5:00 PM on a particular day are classified to the next trading day.

The dataset not only contains all transactions of option and futures contracts on each trading day, but also all updates on the highest bid price and the lowest ask price. Table 1 gives the number of transactions, bid updates, and ask updates, for futures and options on November 1, 2013.

Table 1: Number of observations for E-mini S&P 500 futures and option prices in November 2013.

	Futures	Calls	Puts
Transactions	6,922,393	11,329	15,713
Bid Prices	20,883,086	10,947,785	14,223,270
Ask Prices	20,883,071	11,472,708	15,395,095

As seen from the table, the number of updated bid and ask quotes of the futures and option prices is of the same order, with about twice as many observations for futures. However, whilst the number of futures transactions is in the order of millions, there are only a few ten thousands of option transactions. This indicates the huge potential of adding bid and ask prices to the sample. Other observations are that there are roughly as many bid updates as ask updates, and that put option are slightly more actively traded than call options.

Figure 1 shows a scatter plot of moneyness levels of option transactions during November 2013 versus their trading times. These are the realizations of the covariate used to construct our estimator. The unevenly spaced nature of the transaction times is clearly observed, in particular from the difference between day and night time (five vertical lines per week), and the weekends when the markets are closed. Note also that closer to maturity the traded moneyness levels are more centered around one.

Figure 2 shows time series plots of traded strike prices (upper panel), futures prices (mid panel), and their moneyness ratio (lower panel). The futures prices are synchronized with the option transactions by taking the average of the last recorded best bid and ask quotes. The upper panel shows an intradaily pattern in trading activity, which is high between the working hours 9:00AM-6:00PM and low during the night hours. In particular, during night hours almost only strike prices that are close to at-the-money are quoted, whereas during working hours almost the full range of listed strike prices is quoted upon. The scale of the mid panel shows that there were no large shocks in the futures price on this day, so that the time series of the moneyness ratio in the lower panel closely resembles the time series of the strike price. This suggests that intraday futures price movements are unlikely to induce severe nonstationarity in moneyness during intradaily time windows.

Figure 3 shows recorded call and put prices as a function of their moneyness level prices, for both transaction prices and best bid and ask quotes. Under homogeneity, the option pricing function should not be affected by scaling both the option price and the strike price by the futures price. The data clearly reflect that, by the absence of arbitrage, call prices are convexly decreasing with strike price and put prices are convexly increasing. By put-call parity (**??**), the put prices may be translated into call prices so that one convexly increasing function is displayed. The upper panel shows the best bid and ask quotes, from which it is clear that most activity is at strikes that are 'at-the-money', i.e. with moneyness levels around one. For some intervals of strike prices there are remarkably high levels of the best ask prices. This can be explained by

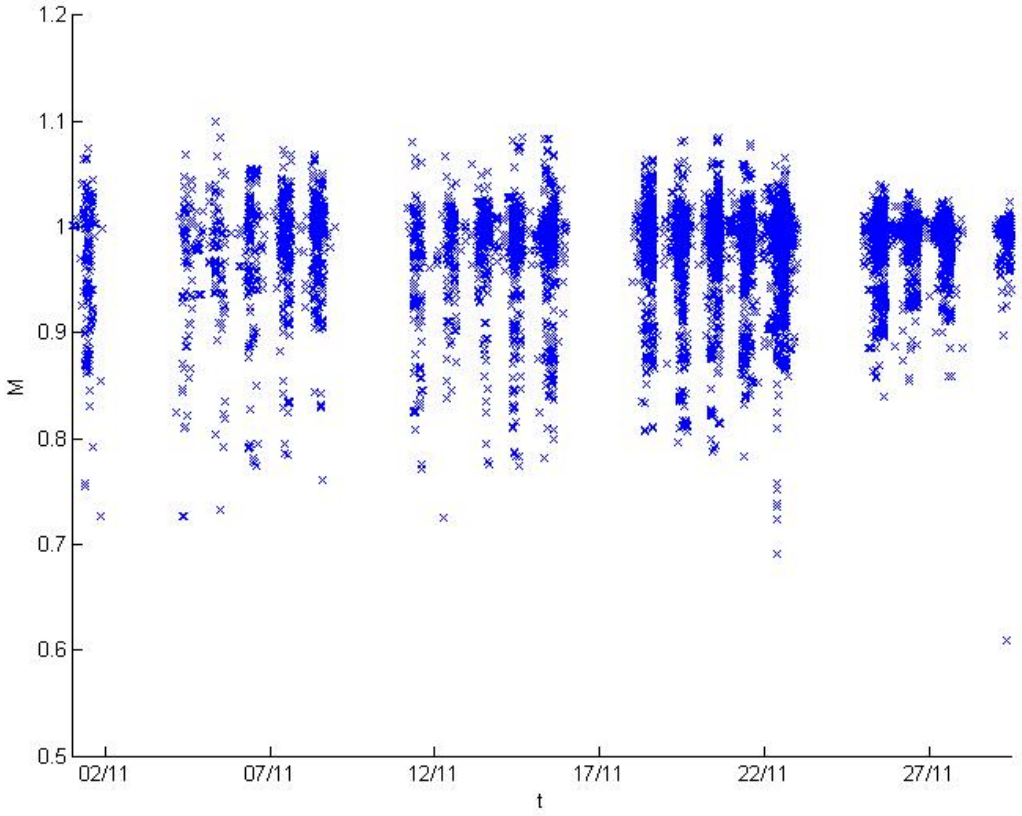


Figure 1: Time series plot of moneyness levels of options transactions during November 2013, for options expiring November 29, 2013.

a lack of liquidity during parts of the day. It warns us that bid and ask prices, or even their averages, can be poor proxies of the actual market price. The lower panel shows transaction prices, from which it is even more clear that trading activity is centred around at-the-money strike prices. The transaction prices, however, do not display similar outliers as the bid and ask quotes do.

Motivated by these observations we apply the following data filters. Only option prices quoted during 8.30AM-4PM Chicago time are included to prevent discontinuities in trading intensity which would violate the smoothness assumption of Theorem 1. Trading activity is highest during the working hours 8.30AM-6PM, yet including the last hour is problematic to the break between 4.15-5PM. Furthermore, as common in the literature, we filter out option quotes that violate the no arbitrage restrictions $\max(0, F_t - \mathrm{e}^{-r\tau}K) <= C_t(K)$ and $\max(0, \mathrm{e}^{-r\tau}K - F_t) <= P_t(K)$, as well as quotes for which the annualized implied volatility is outside the range $5 - 50\%$. Each remaining ask (bid) quote is then averaged with the latest bid (ask) quote with the same strike price, provided there are no more than 1000 quotes in between them to prevent including stale quotes. Matches with bid-ask spread larger than 50% of the resulting mid quote are excluded. To

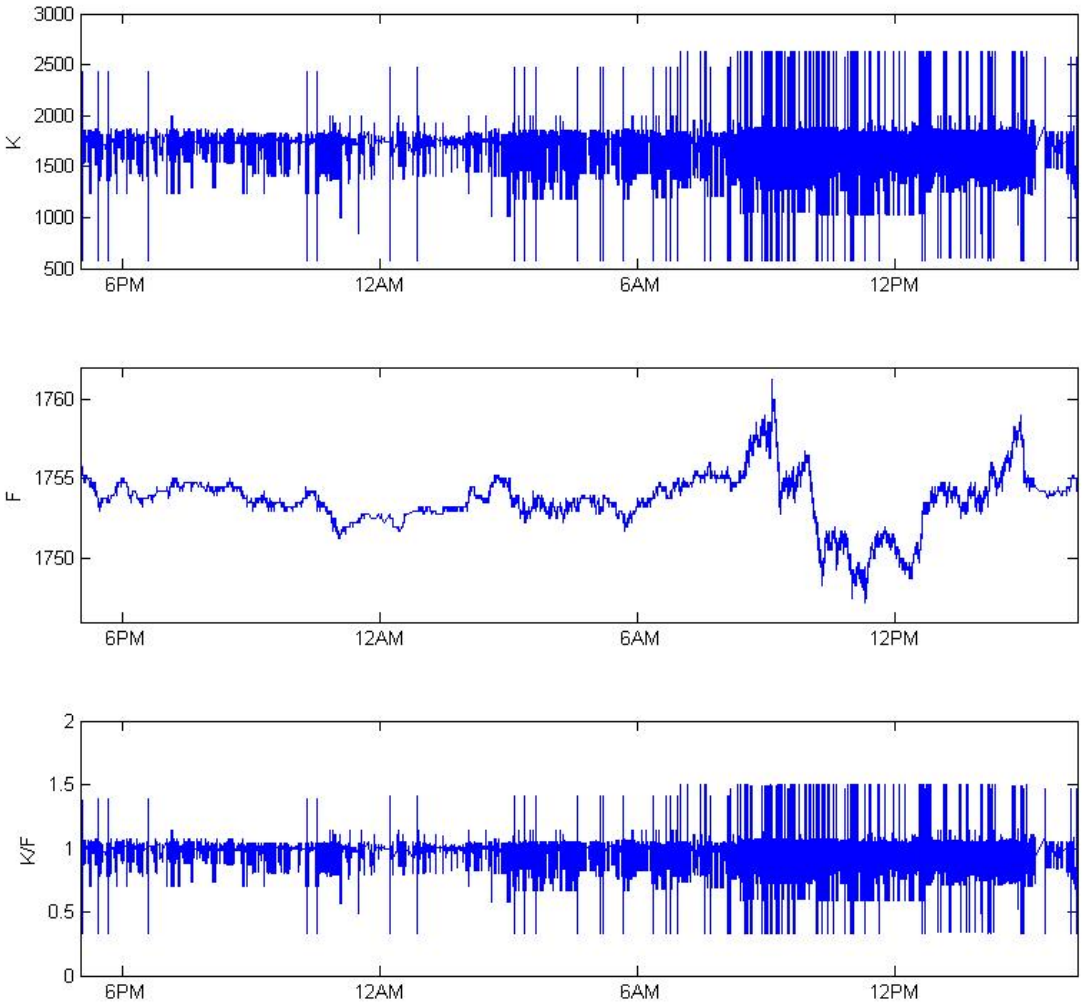


Figure 2: Time series plots of traded strike prices (upper panel), synchronized futures prices (mid panel), and their moneyness ratio (lower panel), for options traded on November 1, 2013.

normalize the option prices the mid quotes are then matched with the average of the synchronized futures prices of the bid and ask quotes. Typically these differ only by a few ticks.

3.3.2 Bandwidth choice

Implementing the kernel estimator requires a data driven choice of the bandwidths h_t and h_x. An important feature of high frequency data is that they are irregularly spaced and the arrival speed of new trades changes over time. While E-mini S&P 500 options are traded overnight, the trading activity during the night much lower than during the day. Also during the day there are some well known diurnal patterns. A globally chosen bandwidth would therefore lead to large

differences in the effective sample size used at different time points of interest. Also for the state dimension a global bandwidth, while stable and easy to compute using cross validation methods, is likely to perform poorly in sparse regions such as in the tails and in regions of high curvature. To balance this, we choose to use variable bandwidths $h(t,x)$. Therefore we adapt the plug-in bandwidth for the optimal asymptotic MSE bandwidth for local polynomial time series regression derived in Masry (1996) to our setting. This bandwidth varies with the point of interest (t,x), reflecting the different densities of the time points and regressors, the variation in curvature, as well as heteroskedasticity of the error variance. Masry (1996) assumes a diagonal bandwidth matrix for the multidimensional case, so that the amount of smoothing in each dimension is the same. This becomes a reasonable approximation once the regressors are rescaled to the unit interval. The optimal bandwidth for a p-order local polynomial around a design point x is given by

$$h_{n,\mathbf{j}}^{opt} = \left(\frac{\sigma^2(x)(d+2|\mathbf{j}|)(M^{-1}\Gamma M^{-1})_{i,i}}{2n\nu(t)f(t,x)(p+1-|\mathbf{j}|)\left[(M^{-1}B\mathbf{m}_{p+1})_i\right]^2} \right)^{\frac{1}{d+2(p+1)}}, \tag{33}$$

where M, Γ, and B only depend on the kernel, and $\mathbf{m}_{p+1}$ is a vector with partial derivatives of order $p+1$. The bandwidth depends on the partial derivative of interest indexed by $\mathbf{j}$, which for example for the second partial derivative with respect to the second regressor equals $\mathbf{j} = (0,2)$. Note Masry (1996) considers the case of a stationary time series observed at a regular interval, in which case the unconditional density of the regressors $f(x)$ features in the denominator. In our case this is replaced by $\nu(t)f(t,x)$, which accounts for the time-varying baseline intensity the observation times and for local stationarity of the regressors. The feasible plug-in bandwidth requires estimating the unknown partial derivatives $\mathbf{m}_{p+1}$, the error variance $\sigma^2(x)$, and the densities $\nu(t)$ and $f(t,x)$. Below we discuss the estimation of each of these.

- $\nu(t)$ - the baseline intensity of the trading times can be estimated in several ways. In the literature on autoregressive conditional duration (ACD) model, diurnal patterns in trading activity are often modelling using cubic splines (Engle and Russell, 1998; Russell, 1999). The literature on conditional intensity based models of self-exciting point process such as the Hawkes model has similarly used spline functions (Bowsher, 2007) or explicitly parametrized the baseline intensity Chen et al. (2013). As there is little guidance on how choose such a parametrization or to set the number of knots for the spline, we use the kernel estimator (6) with a plug-in bandwidth based on (9). Initial estimates of ν_t and ν_t'' are obtained using Silverman's rule-of-thumb density estimator. To estimate η_t we deseasonalize the durations with these density estimates and then compute the empirical autocovariance function as a function of the lag. The resulting bandwidth estimates are then smoothed over to prevent additional variability. Figures 4 show the resulting estimates of the baseline intensity.

- $f(t,x)$ - the density of the locally stationary regressor at time t is also estimated by a kernel

density estimator with a Silverman's rule-of-thumb. As time-variation of the moneyness-regressor within a day is small, we group observations from the same day together so that we do not need to smooth in the time dimension. The resulting density estimates are in Figure 5.

- $\mathbf{m}_{p+1}$ - for an initial estimate of the partial derivatives of the regression function we fit the ad hoc implied volatility model by Dumas et al. (1998) based on a cubic polynomial for the moneyness dimension:

$$\sigma^{IV}(t, m_t) = \beta_{0,t} + \beta_{1,t}(m_t - 1) + \beta_{2,t}(m_t - 1)^2 + \beta_{4,t}(m_t - 1)^3. \tag{34}$$

  We fit the parameters by ordinary least squares for nonoverlapping one-hour blocks in the sample. The partial derivatives are then simply computed from the numerical differentation of the Black-Scholes formula evaluated at different points on the estimated implied volatility surface. The errors of the fit can also be used to estimate the error variance function $\sigma^2(t, x)$.

We then estimate the regression function with the plug-in version of $h^{opt}_{n,\mathbf{j}}$. In practice the estimated bandwidth can be variable and hence might itself contribute to the variability of the regression estimate, in contradiction to the theory. This effect might be dampened by smoothing the resulting estimated bandwidth surface using a cross-validated bandwidth. Moreover, we can iterate the procedure by using the initial estimate of $\hat{m}(\cdot)$ and its derivatives to estimate $\mathbf{m}_{p+1}$ and $\sigma^2(t, x)$. This makes the result less sensitive to the initial parametric model (34). We find that changes are typically small after about three iterations.

3.3.3 Shape restrictions

An advantage of the local polynomial smoother is that the shape constraints (23) and (24) on the call pricing function can be directly implemented. Specifically, the monotonicity constraint (23) requires that $-1 \leq \beta_{0,1} \leq 0$ and the convexity constraints requires that $\beta_{0,2} \geq 0$. This leads to the restricted least squares problem of minimizing (32) subject to these constraints. We implement both constraints together with the positivity constraint for the call pricing function. As violations of these conditions point to arbitrage opportunities, the shape restrictions are unlikely to be violated for sufficiently large and clean datasets. However, the shape constraints may be useful in the tails where the state price density is close to zero and there are relatively few data points.

A drawback of the local polynomial smoother is that the number of local parameters rapidly increases in the multivariate setting. In the bivariate case, a local cubic polynomial requires ten parameters instead of three for the local linear case. If higher order terms have only a small impact these may lead to efficiency losses and numerical instabilities. Therefore we follow Yang

and Tschernig (1999) who recommend removing higher order cross terms that are not of direct interest. Specifically, we set all coefficients related to a second or third partial derivative of the time dimension equal to zero.

3.3.4 Estimation results

We now apply the local cubic estimator with the shape restrictions discussed above to our dataset. First we discuss the estimated components of the plug-in bandwidth.

Figure 4 shows the kernel-based estimate of the diurnal pattern in the arrival of option quotes between 8.30AM-3PM in November 2013. The E-mini S&P 500 option market pauses between 3.15-4PM and trading activity is drastically lower after that. We observe a recurrent yet time-varying intraday pattern with the highest trading activity occurring about one hour after opening and then gradually slowing down during the day. Figure 5 shows the kernel density estimate of the trading activity over different levels of moneyness. We see that trading activity peaks for moneyness levels around one, and the more so when getting closer to maturity. Furthermore trading activity for out-of-the-money put options and in-the-money call options (the left tail) is higher than their counterparts on the right tail, which shows that insurance against negative shocks drives a large part of trading activity. For most maturities there is very little trading for moneyness outside the interval 0.85-1.05, which corresponds to the range for which the exchange sets strike prices at small distances from each other. As local polynomial estimators are highly variable in regions with few observations, in the following we often restrict attention to moneyness levels within this interval.

Figure 6 shows the estimated implied volatility surface using the ad-hoc cubic least squares regression discussed in Dumas et al. (1998). The volatility smile or smirk can be seen from the large implied volatilities in the tails, particularly on the left. There is some steepening of this effect when close to maturity. Note also that the parameter for the cubic term is fairly instable which sometimes leads to unrealistically high or low implied volatilities for far out-of-the money options. This reflects the inability of a global parametric model to capture all features of the data.

Figure 7 shows the resulting iterated plug-in bandwidth surface based on the MSE optimal diagonal bandwidth (33) for estimating the second derivative with respect to moneyness. It uses the estimated densities from Figs. 4 and 5 as well as the iterated partial derivative estimates with a cubic polynomial implied volatility surface from Figure 6 in the first step. The bandwidths are larger the further we go into the tail as the data get more sparse and the state price density is relatively flat. The bandwidths are smallest for moneyness around one, reflecting the many data points and in particular the high curvature of the state price density in this region. Furthermore there is a kink for moneyness around 0.95 where the state price density is roughly linearly de-

creasing and hence curvature is low. The average computed bandwidth lies between 0.01-0.02 as a fraction per day which corresponds to about fifteen to thirty minutes for the time dimension. However there is considerable time variation in the bandwidth surface from day to day and even within the day, despite the intraday smoothing that we perform. Figure 8 compares the bandwidths, averaged over the moneyness dimension, with the estimated mean intensity at each time point. From this it is clear that the diurnal pattern is the main source of time variation of the plug-in bandwidths, apart from a decreasing trend when getting nearer to maturity. Ignoring this diurnal pattern is likely to lead to oversmoothing bias in active trading hours and high variance in low activity hours.

Figure 9 shows the resulting local cubic estimator of the call price function, the option delta, and the SPD with the bandwidth surface from Figure 7. The shape restrictions on the first and second derivative with respect to moneyness are imposed, and higher order derivatives with respect to time are set to zero for stability. The estimated call price function is a smooth surfaces which represent closely the payoff function $(F_T - K)^+$ of the European call options, but show a clear convex pattern which represents the time value of the options. The lower panel shows the estimated state price densities during the sample, the quantities that are of direct economic relevance as they indicate how valuable a unit payoff is to investors in different states of the economy (as proxied for by the S&P 500). These show a clear negative skewness and a fat left tail. The overall volatility or scale is as expected decreasing when nearing maturity, leading to highly spiked densities for the last days in the sample. To deal with the lack of observations outside the moneyness range 0.85-1.05, we paste the tails of the cubic IV surface model of Dumas et al. (1998) for visual illustration. The risk neutral probability mass that lies outside this range is typically smaller than 1%.

Figure 10 shows the dynamics of the implied volatility, skewness, and kurtosis of the state price density for the futures prices on November 29, 2013. Figure 11 shows the dynamics of the corresponding quantiles, where the mean return is normalized to the riskfree rate as dictated by arbitrage theory. Both show a decreasing scale of the distribution over time, yet in a non-monotonic way which is affected by intraday trading.

4 Conclusion

This paper propopes a method to estimate the the state-price density from high frequency options data. Emphasis has been on the time-varying nature of the SPD, and the choice of a time-smoothing bandwidth when trading activity varies among and within trading days. For this purpose asymptotic theory of a time-varying kernel regressor is derived that allows for random sampling times.

The empirical results show that the local cubic polynomial, subject to shape restrictions on the partial derivatives, is able to produce smooth implied density from option quote data. The plug-in bandwidth selection methods adapts to the strong diurnal pattern in trading activity. Changes in the risk neutral moment and quantiles can be directly extracted and used to assess the impact of intraday returns on the whole risk neutral distribution. Formal testing procedures can point out whether the resulting changes in the SPD are statistically significant. The economic significance can be tested by comparison of the pricing and hedging performance of the nonparametric model with parametric benchmarks.

Appendix

Proof of (8). Define for $(t,s) \in \mathbb{R}^2$ the continuous covariance measure

$$\mu_n^{(c)}(t,s) = \delta(t-s)n\nu(t) + n^2\gamma_t(n(s-t)), \tag{35}$$

with $\delta(\cdot)$ the Dirac delta function. Note $\mu_n^{(c)}(t,s)$ is symmetric since $\gamma_s(n(t-s)) = \gamma_s(n(t-s))$. Then

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{n}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)N_n(\mathrm{d}u)\right) &= \frac{1}{n^2}\int_{t-h_t}^{t+h_t}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)K_{h_t}(t-v)\mu_n^{(c)}(u,v)\mathrm{d}u\mathrm{d}v \\
&= \frac{1}{n}\int_{t-h_t}^{t+h_t} K_{h_t}^2(t-v)\nu(t)\mathrm{d}v \\
&\quad + \int_{t-h_t}^{t+h_t}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)K_{h_t}(t-v)\gamma_u(n(v-u))\mathrm{d}u\mathrm{d}v \\
&= \frac{\nu(t)}{nh_t}R(K) + o\left(\frac{1}{nh_t}\right) \\
&\quad + \int_{-1}^{1}\int_{-nh_t}^{nh_t} K(z)K\left(z+y/n\right)\gamma_{t+h_tz}(y)\mathrm{d}y\mathrm{d}z \\
&= \frac{R(K)}{nh_t}\left(\nu(t) + \int_{-\infty}^{\infty}\gamma_t(\tau)\mathrm{d}\tau\right) + o\left(\frac{1}{nh_t}\right),
\end{aligned}$$

using dominated convergence in the last step. □

Proof of Theorem 1. Firstly, we decompose

$$\hat{m}_h(t,x;0) - m(t,x) = \frac{\hat{g}^V(t,x) + \hat{g}^B(t,x)}{\hat{f}(t,x)}, \tag{36}$$

with

$$\hat{f}(t,x) = \int_{t-h_t}^{t+h_t} K_{h_t}(t-u)K_{h_x}(x - X_{\{u\},n})N_n(\mathrm{d}u)$$
$$\hat{g}^V(t,x) = \int_{t-h_t}^{t+h_t} K_{h_t}(t-u)K_{h_x}(x - X_{\{u\},n})\epsilon_{\{u\},n}N_n(\mathrm{d}u)$$
$$\hat{g}^B(t,x) = \int_{t-h_t}^{t+h_t} K_{h_t}(t-u)K_{h_x}(x - X_{\{u\},n})\left(m(u, X_{\{u\},n}) - m(t,x)\right)N_n(\mathrm{d}u)$$

The proof consists of the following three parts:

(i) $$\hat{f}(t,x) \xrightarrow{p} \nu(t)f(t,x)$$

(ii) $$\hat{g}^B(t,x) \xrightarrow{p} \nu(t)f(t,x)B(t,x)$$

(iii) $$\sqrt{nh_th_x}\hat{g}^V(t,x) \xrightarrow{d} \nu(t)f(t,x) \times \mathrm{N}(0, V(t,x)).$$

Throughout we use the convention $\mathrm{d}N_n(t) = N_n(t - \mathrm{d}t, t)$. C denotes some constant which may take different values in different places.For each of the parts we use the following asymptotic expansion. Similar to (7) and (8), it holds for general r that

$$\frac{1}{n}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)(t-u)^r N_n(du) = \begin{cases} h_t^r \mu_r(K)\nu(t) + o_P(h_t^r) & \text{if } r \text{ even} \\ h_t^{r+1}\mu_{r+1}(K)\nu_n'(t) + o_P(h_t^{r+1}) & \text{if } r \text{ odd.} \end{cases}$$

A Taylor expansion implies that for any continuously differentiable function $g : [0,1] \to \mathbb{R}, u \mapsto g(u)$,

$$\frac{1}{n}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)g(u)N_n(\mathrm{d}u) = g(t)\nu(t) + O_P\left(\frac{1}{\sqrt{nh_t}}\right) + o_P(h_t). \tag{37}$$

(i) The density estimator has expectation

$$\begin{aligned} E(\hat{f}(t,x)) &= \frac{1}{n}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)E\left(K_{h_x}(x - X_{\{u\},n})N_n(\mathrm{d}u)\right) \\ &= \frac{1}{n}\int_{t-h_t}^{t+h_t} K_{h_t}(t-u)E\left(K_{h_x}(x - X_{\{u\},n})\middle|\mathrm{d}N_n(u) = 1\right)\bar{\lambda}_n(u)\mathrm{d}(u). \end{aligned}$$

Each $X_{\{u\},n}$ can be approximated by its stationary counterpart $X^*_{\{u\},n}(u)$ at the time of observation $t_{i,n} = u$. For brevity from now on write $X^*_{\{u\},n}(u) = X^*_{\{u\},n}$, that is, unless mentioned otherwise we assume the process is approximated around its own observation time. Following Vogt (2012, Thm. 4.2), let $\bar{K}$ be a Lipschitz continuous function on $[-q,q]$, $q > 1$ with $\bar{K}(x) = 1$

for $x \in [-1, 1]$, and decompose

$$\begin{aligned} K_{h_x}(x - X_{\{u\},n}) =& \bar{K}_{h_x}(x - X_{\{u\},n})\{K_{h_x}(x - X_{\{u\},n}) - K_{h_x}(x - X^*_{\{u\},n})\} \\ &+ \{\bar{K}_{h_x}(x - X_{\{u\},n}) - \bar{K}_{h_x}(x - X^*_{\{u\},n})\}K_{h_x}(x - X^*_{\{u\},n}) \\ &+ K_{h_x}(x - X^*_{\{u\},n}). \end{aligned}$$

For the first term, note that

$$\begin{aligned} \left|K\left(\frac{x - X_{\{u\},n}}{h_x}\right) - K\left(\frac{x - X^*_{\{u\},n}}{h_x}\right)\right| &\leq C\left|K\left(\frac{x - X_{\{u\},n}}{h_x}\right) - K\left(\frac{x - X^*_{\{u\},n}}{h_x}\right)\right|^r \\ &\leq CL\left|\frac{X_{\{u\},n} - X^*_{\{u\},n}}{h_x}\right|^r \\ &\leq \left|\frac{U_{\{u\},n}(u)}{nh_x}\right|^r, \end{aligned}$$

by boundedness and Lipschitz continuity of K, and local stationarity of $\{X_{\{u\},n}\}$. Hence from the existence of conditional moments of $U_{\{u\},n}$ it follows that the first term is $O_P\left(\frac{1}{n^r h_x^r}\right)$. Similarly it follows that the second term is also $O_P\left(\frac{1}{n^r h_x^r}\right)$, so with a standard approximation for the third term it follows that

$$E(K_{h_x}(x - X_{\{u\},n})\big|N_n(\mathrm{d}u) = 1) = f(u, x) + o(h_x) + O\left(\frac{1}{n^r h_x^r}\right). \tag{38}$$

Using (37) with $g(u) = f(u, x)$ it follows that

$$E(\hat{f}_h(t, x)) = \nu(t)f(t, x) + o(h_t + h_x) + O\left(\frac{1}{n^r h_x^r}\right).$$

Its variance is given by

$$\begin{aligned} \mathrm{Var}(\hat{f}_h(t, x)) &= \frac{1}{(nh_t h_x)^2}\mathrm{Var}\left(\int_{t-h_t}^{t+h_t} K\left(\frac{t - u}{h_t}\right) K\left(\frac{x - X_{\{u\},n}}{h_x}\right) \mathrm{d}N_n(u)\right) \\ &= \frac{1}{(nh_t h_x)^2}\int_{t-h_t}^{t+h_t} K^2\left(\frac{t - u}{h_t}\right) \mathrm{Var}\left(K\left(\frac{x - X_{\{u\},n}}{h_x}\right) \mathrm{d}N_n(u)\right) \\ &\quad + \frac{1}{(nh_t h_x)^2}\iint\limits_{\substack{t-h_t \leq u,v \leq t+h_t, \\ u \neq v}} K\left(\frac{t - u}{h_t}\right) K\left(\frac{t - v}{h_t}\right) \\ &\qquad \times \mathrm{Cov}\left(K\left(\frac{x - X_{\{u\},n}}{h_x}\right) \mathrm{d}N_n(u), K\left(\frac{x - X_{\{v\},n}}{h_x}\right) \mathrm{d}N_n(v)\right) \\ &\equiv (V) + (CV). \end{aligned}$$

The variance increments can be computed as, with the shorthand $K_{u,n}(x) = K((x - X_{\{u\},n})/h_x)$,

$$\begin{aligned}\mathrm{Var}\left(K_{u,n}(x)\mathrm{d}N_n(u)\right) &= E(K_{u,n}^2(x)(\mathrm{d}N_n(u))^2) - E(K_{u,n}(x)\mathrm{d}N_n(u))^2 \\ &= E(K_{u,n}^2(x)\big|\mathrm{d}N_n(u) = 1)n\nu(t)\mathrm{d}u - O(\mathrm{d}u)^2, \\ &= nh_x f(u,x)R(K)\nu(t)\mathrm{d}u + o(nh_x\mathrm{d}u),\end{aligned}$$

so that

$$(V) = \frac{R^2(K)f(t,x)\nu(t)}{nh_t h_x} + o\left(\frac{1}{nh_t h_x}\right).$$

For the covariance increments we use the law of total covariance to get, for $u < v$,

$$\begin{aligned}\mathrm{Cov}\left\{K_{u,n}(x)\mathrm{d}N_n(u), K_{v,n}(x)\mathrm{d}N_n(v)\right\} &= E\left\{\mathrm{Cov}\left(K_{u,n}(x), K_{v,n}(x)|N_n(\mathrm{d}u), N_n(\mathrm{d}v)\right)N_n(\mathrm{d}u)N_n(\mathrm{d}v)\right\} \\ &\quad + \mathrm{Cov}\left\{E\left(K_{u,n}(x)|N_n(\mathrm{d}u), N_n(\mathrm{d}v)\right)N_n(\mathrm{d}u),\right. \\ &\quad\quad \left. E\left(K_{v,n}(x)|N_n(\mathrm{d}u), N_n(\mathrm{d}v)\right)N_n(\mathrm{d}v)\right\} \\ &= (\star) + (\star\star).\end{aligned}$$

In terms of $h_n(u,v) = \mathbb{P}(N_n(\mathrm{d}v = 1|N_n(\mathrm{d}u = 1)$ and $h_u(\tau) := \frac{h_n(u,u+\tau/n)}{n}$ for $u < v$,

$$(\star) = \mathrm{Cov}\left\{K_{u,n}(x), K_{v,n}(x)|N_n(\mathrm{d}u) = N_n(\mathrm{d}v) = 1\right\}n^2\nu(u)h((u-v)/n)\mathrm{d}u\mathrm{d}v,$$

while the nonnegativity of the jumps implies that we can bound

$$\big|(\star\star)\big| \leq Ch_x^2\big|\mathrm{Cov}\left\{N_n(\mathrm{d}u), N_n(\mathrm{d}v)\right\}\big| = Ch_x^2\big|\mu_n(u,v)\big|.$$

Now note the equality of events $\{N_n(\mathrm{d}u) = N_n(\mathrm{d}v) = 1\} = \{t_{N_n(u),n} = u, t_{N_n(v),n} = v,\}$. Since by definition $t_{N_n(u),n} \leq u$ for every u, and the observation times becomes arbitrarily dense when $n \to \infty$, it follows that $t_{N_n(u),n} \xrightarrow{p} u$. Then conditioning on $t_{N_n(u),n}$ and $t_{N_n(v),v}$, implies that for every $u, v \in (0,1)$ and $x, y \in S$

$$\left|f_{X_{\{u\},n},X_{\{v\},n}|N_n(\mathrm{d}u)=N_n(\mathrm{d}v)=1}(x,y) - f_{X_{\{u\},n},X_{\{v\},n}}(x,y)\right| \to 0.$$

This implies that the joint distribution of $X_{\{u\},n}$ and $X_{\{v\},n}$, which are the last observed values at time u and v, resp., is asymptotically independent from the fact that there are jumps at u and v. In particular, in the limit their covariance is not affected by these jumps.

Now the mixing conditions and Davydov's lemma bound the covariance by

$$\begin{aligned}\left|\mathrm{Cov}\left(K_{u,n}(x), K_{v,n}(x)\right)\right| &\leq 8\alpha(n(v-u))^{\frac{\delta}{2+\delta}} E(K_{u,n}(x)^{2+\delta})^{\frac{1}{2+\delta}} E(K_{v,n}(x)^{2+\delta})^{\frac{1}{2+\delta}} \\ &\leq C\alpha(n(v-u))^{\frac{\delta}{2+\delta}} h_x^{\frac{2}{2+\delta}}. \end{aligned} \tag{39}$$

Hence

$$\begin{aligned} nh_t h_x\left|(CV)\right| &\leq \frac{nC}{h_t h_x} \int_{t-h_t}^{t+h_t} \int_{v}^{t+h_t} K\left(\frac{t-u}{h_t}\right) K\left(\frac{t-v}{h_t}\right) \mathrm{Cov}\left(K_{u,n}(x), K_{v,n}(x)\right) \nu(u) h((u-v)/n) \mathrm{d}u\mathrm{d}v \\ &\quad + \frac{Ch_x}{nh_t} \int_{t-h_t}^{t+h_t} \int_{v}^{t+h_t} K\left(\frac{t-u}{h_t}\right) K\left(\frac{t-v}{h_t}\right) \left|\mu_n(u,v)\right| \mathrm{d}u\mathrm{d}v \\ &= \frac{2Cnh_t}{h_x} \int_{-1}^{1} \int_{0}^{1-z} K\left(z+y\right) K\left(z\right) \mathrm{Cov}\left(K_{z,n}(x), K_{z+y,n}(x)\right) \nu(t+h_t z) h(ny) \mathrm{d}y\mathrm{d}z \\ &\quad + Cnh_t h_x \int_{-1}^{1} \int_{0}^{1-z} K\left(z+y\right) K\left(z\right) \left|\gamma_{t+h_t z}(ny)\right| \mathrm{d}y\mathrm{d}z \\ &= \frac{2C}{h_x} \int_{-1}^{1} \int_{0}^{nh_t(1-z)} K\left(z+\frac{w}{nh_t}\right) K\left(z\right) \mathrm{Cov}\left(K_{t+h_t z,n}(x), K_{t+h_t z+\frac{w}{n},n}(x)\right) \nu(t+h_t z) h(w) \mathrm{d}w\mathrm{d}z \\ &\quad + Ch_x \int_{-1}^{1} \int_{0}^{nh_t(1-z)} K\left(z+\frac{w}{nh_t}\right) K\left(z\right) \left|\gamma_{t+h_t z}(w)\right| \mathrm{d}w\mathrm{d}z \\ &\equiv (*) + (**), \end{aligned}$$

We now bound the separate terms. We split up (*) using a sequence m_n such that $m_n \to \infty$ and $m_n h_x \to 0$, and use covariance inequality (39) to bound the covariance between terms far apart. Combined with boundedness of the kernel this yields

$$\begin{aligned} (*) &\leq Cm_n h_x + Ch_x^{\frac{-\delta}{2+\delta}} \int_{-1}^{1} \int_{m_n}^{nh_t(1-z)} \alpha(w)^{\frac{\delta}{2+\delta}} \nu(t+h_t z) h(w) \mathrm{d}w\mathrm{d}z \\ &\leq Cm_n h_x + Ch_x^{\frac{-\delta}{2+\delta}} m_n^{-\gamma} \int_{-1}^{1} \int_{m_n}^{nh_t(1-z)} w^{\gamma} \alpha(w)^{\frac{\delta}{2+\delta}} \nu(t+h_t z) h(w) \mathrm{d}w\mathrm{d}z \\ &\to 0, \end{aligned}$$

for some $\gamma > 1$ by setting $m_n = h_x^{\frac{\delta}{\gamma(2+\delta)}}$, using dominated convergence and the rate condition on the mixing coefficients (13). For the second term we find

$$\frac{(**)}{h_x} \to C \int_{-1}^{1} K^2(z) \int_0^{\infty} \gamma_t(w) \mathrm{d}w \mathrm{d}z < \infty,$$

again by dominated convergence, so that $(**) = O(h_x)$. Combining terms, we conclude that $(CV) = o\left(\frac{1}{nh_th_x}\right)$, which means that the covariance terms are of smaller order than the variance terms. Hence $\mathrm{Var}(\hat{f}(t,x)) \to 0$, and since the bias vanishes as well this proves the first part.

(ii) For the leading bias term of the estimator is given by, write in terms of the stationary approximation process, similar to (38),

$$\begin{aligned}
&E\left(K_{h_x}(x - X_{\{u\},n})(m(u, X_{\{u\},n}) - m(t,x))\mathrm{d}N_n(u)/n\right) \\
&= E\left(K_{h_x}(x - X^*_{\{u\},n})(m(u, X^*_{\{u\},n}) - m(t,x))\big|\mathrm{d}N_n(u) = 1\right)\nu(u)\mathrm{d}u + O\left(\frac{1}{n^r h_x^r}\right).
\end{aligned}$$

Then using a Taylor approximation of $m(u, X^*_{\{u\},n})$ around $m(t,x)$ and (37) we find

$$\begin{aligned}
E(\hat{g}^B(t,x)) = {} & h_t^2 \mu_2(K)\left(\frac{\partial m(t,x)}{\partial t}\frac{\partial f(t,x)\nu(t)}{\partial t} + \frac{1}{2}\frac{\partial^2 m(t,x)}{\partial t^2} f(t,x)\nu(t)\right) \\
& + h_x^2 \mu_2(K)\left(\frac{\partial m(t,x)}{\partial x}\frac{\partial f(t,x)}{\partial x}\nu(t) + \frac{1}{2}\frac{\partial^2 m(t,x)}{\partial x^2} f(t,x)\nu(t)\right) + o(h_t^2 + h_x^2) + O\left(\frac{1}{n^r h_x^r}\right).
\end{aligned}$$

From the Lipschitz condition on $m(t,x)$ it follows that $\mathrm{Var}(\hat{g}^B(t,x)) = o\left(\frac{1}{nh_th_x}\right)$, which ensures that the variance of the bias term vanishes.

(iii) For the variance term, write

$$\begin{aligned}
nh_th_x\mathrm{Var}(\hat{g}^V(t,x)) &= \frac{1}{nh_th_x}\mathrm{Var}\left(\int_{t-h_t}^{t+h_t} K\left(\frac{t-u}{h_t}\right) K\left(\frac{x - X_{\{u\},n}}{h_x}\right)\epsilon_{\{u\},n}\mathrm{d}N_n(u)\right) \\
&= \frac{1}{nh_th_x} E\left(\int_{t-h_t}^{t+h_t} K^2\left(\frac{t-u}{h_t}\right) K^2\left(\frac{x - X_{\{u\},n}}{h_x}\right)\epsilon^2_{\{u\},n}\mathrm{d}N_n(u)\right) \\
&\quad + \frac{1}{nh_th_x} E\left(\iint_{\substack{t-h_t \le u,v \le t+h_t, \\ u \neq v}} K\left(\frac{t-u}{h_t}\right) K\left(\frac{t-v}{h_t}\right) K\left(\frac{x - X_{\{u\},n}}{h_x}\right)\right. \\
&\qquad \left. \times K\left(\frac{x - X_{\{v\},n}}{h_x}\right)\epsilon_{\{u\},n}\epsilon_{\{v\},n}\mathrm{d}N_n(u)\mathrm{d}N_n(v)\right) \\
&= R^2(K) f(t,x)\sigma^2(t,x)\nu(t) + o(1),
\end{aligned}$$

using arguments similar as for computing $\mathrm{Var}(\hat{f}_h(t,x))$, except in addition to bound the covariance term we use

$$E\left|K_{u,n}(x)\epsilon_{\{u\},n}\right|^{2+\delta} = E\left(E\left(\epsilon^{2+\delta}_{\{u\},n}(x)|X_{\{u\},n}\right)K^{2+\delta}_{u,n}(x)\right) < Ch_x.$$

The asymptotic normality follows from applying large-small block arguments to the integral form of $\hat{g}^V(t,x)$. In particular, write

$$\sqrt{nh_th_x}\hat{g}^V(t,x) = \frac{1}{\sqrt{nh_th_x}}\int_{-nh_t}^{nh_t} K\left(\frac{w}{nh_t}\right)K\left(\frac{x - X_{\{t+\frac{w}{n}\}}}{h_x}\right)\epsilon_{\{t+\frac{w}{n}\}}N_n(t+\mathrm{d}w/n),$$

and divide the integral in k_n large blocks $\zeta_{j,n}$ of length τ_l, and k_n small blocks $\eta_{j,n}$ of length τ_s, which integrate, respectively, from $-nh_t + (j-1)(\tau_l+\tau_s)$ to $-nh_t + (j-1)(\tau_l+\tau_s) + \tau_l$, and from $-nh_t + (j-1)(\tau_l+\tau_s) + \tau_l$ to $-nh_t + j(\tau_l+\tau_s)$. Then

$$\sqrt{nh_th_x}\hat{g}^V(t,x) = \frac{1}{\sqrt{nh_th_x}}\sum_{j=1}^{k_n}\left(\zeta_{j,n} + \eta_{j,n}\right),$$

where

$$k_n = \frac{2nh_t}{\tau_l + \tau_s}.$$

We set the block sizes such that $\tau_l \to \infty$, $\tau_s \to \infty$, $\frac{\tau_s}{\tau_l} \to 0$, $\frac{\tau_l}{\sqrt{nh_th_x}} \to 0$, and $k_n\alpha(s_n) \to 0$. Based on the choice in Fan and Yao (2002, Thm. 2.22), it can be shown that

$$\tau_l = \frac{\sqrt{nh_th_x}}{\log nh_t}, \qquad \tau_s = \left(\sqrt{nh_t/h_x}\log nh_t\right)^{\frac{\delta}{(\lambda+1)(2+\delta)}}$$

satisfies these conditions under the bandwidth conditions in the theorem and the mixing condition. The small blocks are asymptotically negligible since

$$\mathrm{Var}\left(\frac{1}{\sqrt{nh_th_x}}\sum_{j=1}^{k_n}\eta_{j,n}\right) < \frac{Ck_n\tau_s}{nh_t} \to 0.$$

The large blocks are asymptotically independent by the Volkonskii-Rozanov lemma

$$\left|E\left(\exp\left(it\frac{\sum_{j=1}^{k_n}\zeta_{j,n}}{\sqrt{nh_th_x}}\right)\right) - \prod_{j=1}^{k_n}E\left(\exp\left(it\frac{\zeta_{j,n}}{\sqrt{nh_th_x}}\right)\right)\right| \leq 16(k_n-1)\alpha(\tau_s) \to 0.$$

It can be verified that

$$\frac{1}{nh_th_x}\sum_{j=1}^{k_n}\mathrm{Var}\left(\zeta_{j,n}\right) \to \mu_2^2(K)f(t,x)\sigma^2(t,x)\nu(t) > 0,$$

by decomposing

$$\mathrm{Var}\left(\sqrt{nh_th_x}\hat{g}^V(t,x)\right) = \frac{1}{nh_th_x}\sum_{j=1}^{k_n}\mathrm{Var}(\zeta_{j,n}+\eta_{j,n}) + \frac{1}{nh_th_x}\sum_{j=1}^{k_n}\sum_{k\neq j}\mathrm{Cov}(\zeta_{j,n}+\eta_{j,n},\zeta_{k,n}+\eta_{k,n}),$$

and showing that all other terms on the right hand side vanish, in particular using that for adjacent blocks

$$\frac{1}{nh_th_x}\mathrm{Cov}(\zeta_{j,n},\eta_{j,n}) < C\frac{\tau_s}{nh_t}.$$

The Lindeberg condition

$$\frac{1}{nh_th_x}\sum_{j=1}^{k_n}E\left(\zeta_{j,n}^2\mathbb{1}_{\{|\zeta_{j,n}|>\sqrt{nh_th_x}\epsilon\}}\right)\to 0,$$

is satisfied for any $\epsilon>0$, since $\tau_l = o(\sqrt{nh_th_x})$ causes $\{|\zeta_{j,n}|>\sqrt{nh_th_x}\epsilon\}$ to become an empty set for large n. The Lindeberg-Feller central limit theorem gives the required result. □

Proof of (17).

$$\begin{aligned}\frac{E(\lambda_n(t))}{n} &= \phi\pi(t) + \int_{-\infty}^{t} E\left(h(X_{\{u\},n})|N_n(\mathrm{d}u)=1\right)g(n(t-u))E(\lambda_n(u))\mathrm{d}u,\\ &= \phi\pi(t) + \int_0^{\infty} E\left\{h\left(X_{\left\{t-\frac{w}{n}\right\},n}|N_n(t+\mathrm{d}w/n)=1\right)\right\}g(w)E(\lambda_n(t-w/n))\mathrm{d}w,\end{aligned}$$

where we approximate with the stationary process around time t (all expectations are given $N_n(t+\mathrm{d}w/n)=1$)

$$\begin{aligned}E\left|h\left(X_{\left\{t-\frac{w}{n}\right\},n}\right) - h\left(X^*_{\left\{t-\frac{w}{n}\right\},n}(t)\right)\right| &\le LE\left|X_{\left\{t-\frac{w}{n}\right\},n} - X^*_{\left\{t-\frac{w}{n}\right\},n}(t)\right|\\ &\le CE\left|X_{\left\{t-\frac{w}{n}\right\},n} - X^*_{\left\{t-\frac{w}{n}\right\},n}(t)\right|^r\\ &\le C\left(\frac{1+w}{n}\right)^{\rho}E\left|U_{\left(t-\frac{w}{n}\right),n}(t)\right|^r\\ &\le C\left(\frac{1+w}{n}\right)^{r},\end{aligned}$$

where the second inequality uses the compactness of S, and we recall $r=\min(\rho,1)$. Hence

$$\nu(t) := \lim_{n\to\infty}\frac{E(\lambda_n(t)}{n} = \phi\pi(t) + \nu(t)\int_S h(x)f(t,x)\mathrm{d}x\int_0^{\infty}g(w)\mathrm{d}w,$$

using the stationarity of $X^*_{\{t\},n}$ and the integrability conditions on $g(\cdot)$. □

References

Aït-Sahalia, Y. and Duarte, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116(1-2):9–47.

Aït-Sahalia, Y. and Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance*, 53(2):499–547.

Aït-Sahalia, Y. and Lo, A. W. (2000). Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, 94(1-2):9–51.

Aït-Sahalia, Y. and Mykland, P. A. (2003). The effects of random and discrete sampling when estimating continuous–time diffusions. *Econometrica*, 71(2):483–549.

Bakshi, G., Cao, C., and Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of Finance*, 52(5):2003–2049.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.

Barndorff-Nielsen, O. E. and Shephard, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925.

Bauwens, L. and Hautsch, N. (2009). *Modelling financial high frequency data using point processes*. Springer.

Beber, A. and Brandt, M. W. (2006). The effect of macroeconomic news on beliefs and preferences: Evidence from the options market. *Journal of Monetary Economics*, 53(8):1997–2039.

Bibinger, M., Hautsch, N., Malec, P., and Reiss, M. (2015). Estimating the spot covariation of asset prices statistical theory and empirical evidence.

Birke, M. and Pilz, K. F. (2009). Nonparametric option pricing with no-arbitrage constraints. *Journal of Financial Econometrics*, 7(2):53–76.

Birru, J. and Figlewski, S. (2012). Anatomy of a meltdown: The risk neutral density for the s&p 500 in the fall of 2008. *Journal of Financial Markets*, 15(2):151–180.

Bosq, D. (1996). *Nonparametric statistics for stochastic processes*. Springer.

Bowman, A. W. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations: the kernel approach with S-Plus illustrations*. Oxford University Press.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912.

Breeden, D. T. and Litzenberger, R. H. (1978). Prices of state-contingent claims implicit in option prices. *The Journal of Business*, 51(4):621–51.

Broadie, M., Chernov, M., and Johannes, M. (2007). Model specification and risk premia: Evidence from futures options. *The Journal of Finance*, 62(3):1453–1490.

Chen, F., Hall, P., et al. (2013). Inference for a nonstationary self-exciting point process with an application in ultra-high frequency financial data modeling. *Journal of Applied Probability*, 50(4):1006–1024.

Chen, S. X. and Xu, Z. (2014). On implied volatility for options: Some reasons to smile and more to correct. *Journal of Econometrics*, 179(1):1–15.

Cranwell, R. and Weiss, N. (1978). A central limit theorem for mixing stationary point processes. *Stochastic Processes and their Applications*, 8(2):229–242.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes*, volume 1. Springer.

Duffie, D. and Glynn, P. (2004). Estimation of continuous-time markov processes sampled at random time intervals. *Econometrica*, 72(6):1773–1808.

Dumas, B., Fleming, J., and Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *The Journal of Finance*, 53(6):2059–2106.

Ellis, S. P. (1991). Density estimation for point processes. *Stochastic processes and their applications*, 39(2):345–358.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, pages 1127–1162.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fan, J. and Mancini, L. (2009). Option pricing with model-guided nonparametric methods. *Journal of the American Statistical Association*, 104(488).

Fan, J. and Yao, Q. (2002). *Nonlinear time series*, volume 2. Springer.

Fengler, M. R., Härdle, W. K., and Mammen, E. (2007). A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics*, 5(2):189–218.

Fengler, M. R. and Hin, L.-Y. (2015). Semi-nonparametric estimation of the call-option price surface under strike and time-to-expiry no-arbitrage constraints. *Journal of Econometrics*, 184(2):242–261.

Härdle, W. and Hlavka, Z. (2009). Dynamics of state price densities. *Journal of Econometrics*, 150(1):1–15.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Jackwerth, J. C. (2000). Recovering Risk Aversion from Option Prices and Realized Returns. *Review of Financial Studies*, 13(2).

Joshi, M. (2001). Log-type models, homogeneity of option prices and convexity.

Karatzas, I. and Shreve, S. E. (1998). *Methods of mathematical finance*, volume 39. Springer.

Kristensen, D. (2010). Nonparametric filtering of the realized spot volatility: a kernel-based approach. *Econometric Theory*, 26(01):60–93.

Li, Y., Mykland, P. A., Renault, E., Zhang, L., and Zheng, X. (2009). Realized volatility when sampling times are possibly endogenous.

Linton, O., Park, S., and Hong, S. (2015). Estimating the quadratic covariation matrix for an asynchronously observed continuous time signal masked by additive noise. *Journal of Econometrics*.

Masry, E. (1996). Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, 65(1):81–101.

Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics*, 4(1):141–183.

Mykland, P. A., Zhang, L., et al. (2006). Anova for diffusions and ito processes. *The Annals of Statistics*, 34(4):1931–1963.

Pawlas, Z. (2009). Empirical distributions in marked point processes. *Stochastic Processes and their Applications*, 119(12):4194–4209.

Renault, E. and Werker, B. J. (2011). Causality effects in return volatility measures with random times. *Journal of Econometrics*, 160(1):272–279.

Robinson, P. M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis*, 4(3):185–207.

Rosenberg, J. and Engle, R. F. (2000). Empirical pricing kernels. *Journal of Financial Economics*, (99-014).

Russell, J. R. (1999). Econometric modeling of multivariate irregularly-spaced high-frequency data. *Manuscript, GSB, University of Chicago*.

Song, Z. and Xiu, D. (2012). A tale of two option markets: State-price densities implied from s&p 500 and vix option prices. *Unpublished working paper, Federal Reserve Board and University of Chicago*.

Song, Z. and Xiu, D. (2014). A tale of two option markets: State-price densities implied from s&p 500 and vix option prices.

Vogt, M. (2012). Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40(5):2601–2633.

Xiu, D. (2014). Hermite polynomial based expansion of european option prices. *Journal of Econometrics*, 179(2):158–177.

Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 793–815.

Yatchew, A. and Härdle, W. (2006). Nonparametric state price density estimation using constrained least squares and the bootstrap. *Journal of Econometrics*, 133(2):579–599.

Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

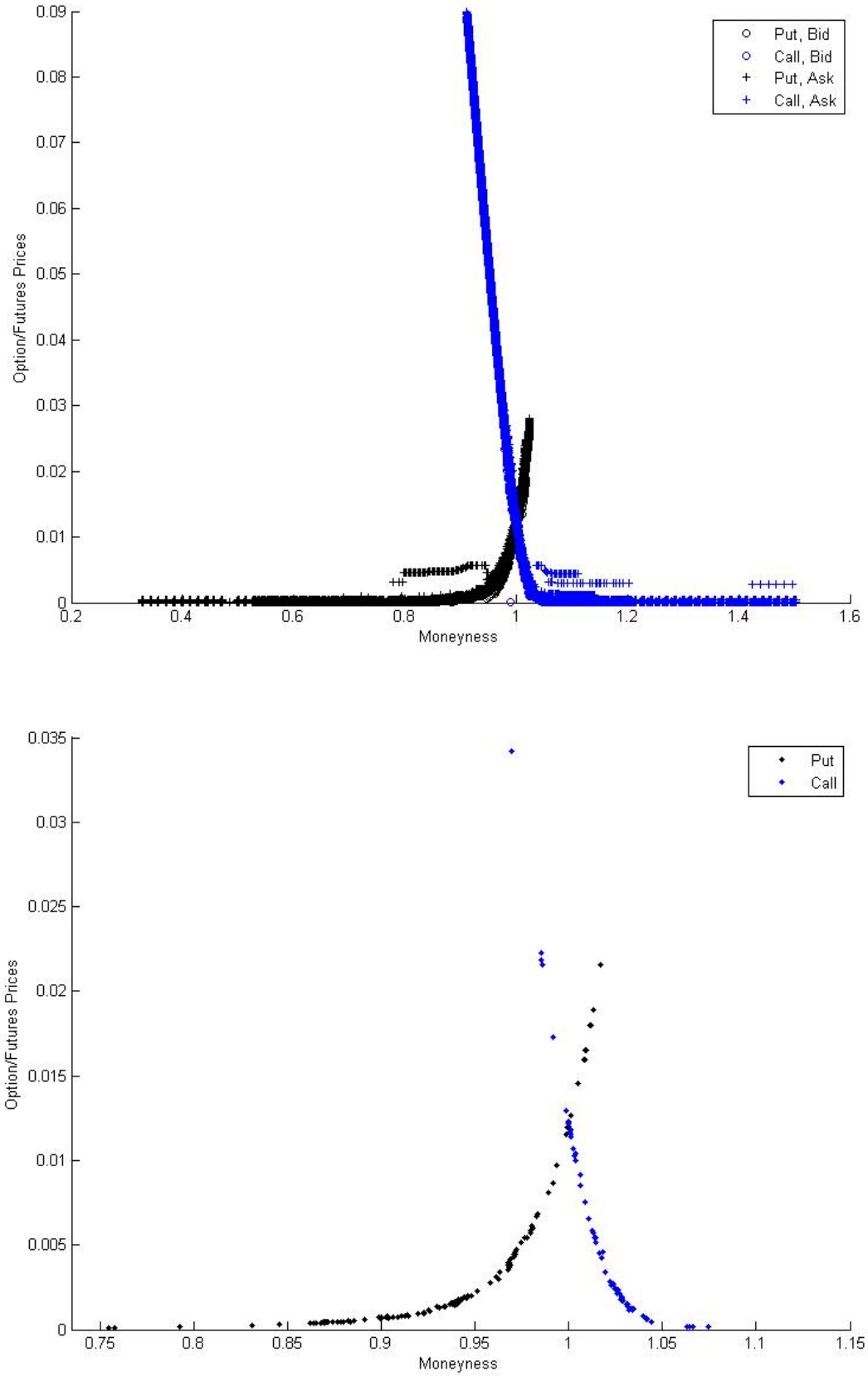


Figure 3: Call and put prices per unit of futures price as a function of moneyness, in terms of best bid and ask quotes (upper panel) and transaction prices (lower panel), for November 1, 2013.

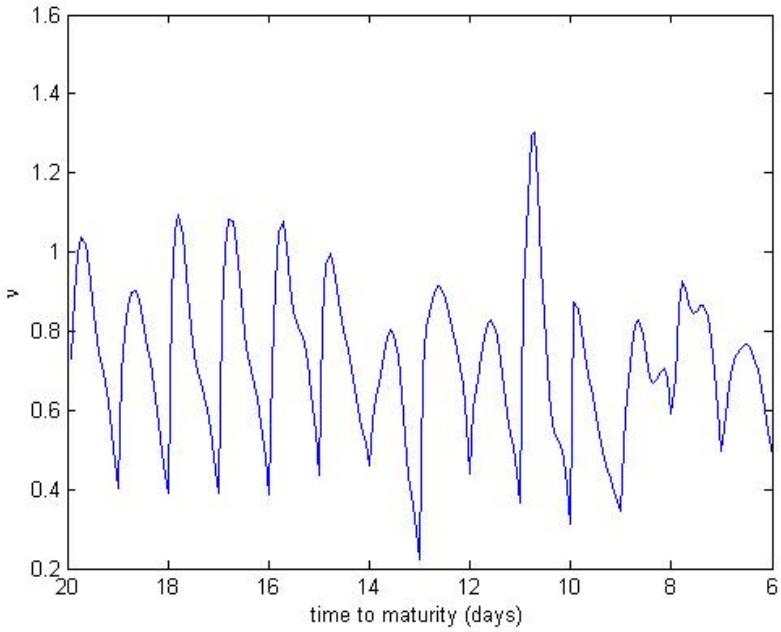


Figure 4: Kernel estimate of the diurnal pattern in the arrival of option quotes between 8.30AM-3PM from November 1 to November 20, 2013, using the plug-in version of the MSE optimal bandwidth with a rule-of-thumb bandwidth for the initial estimate. Quotes happening in the same second are treated as one observation with the number of quotes as its weight.

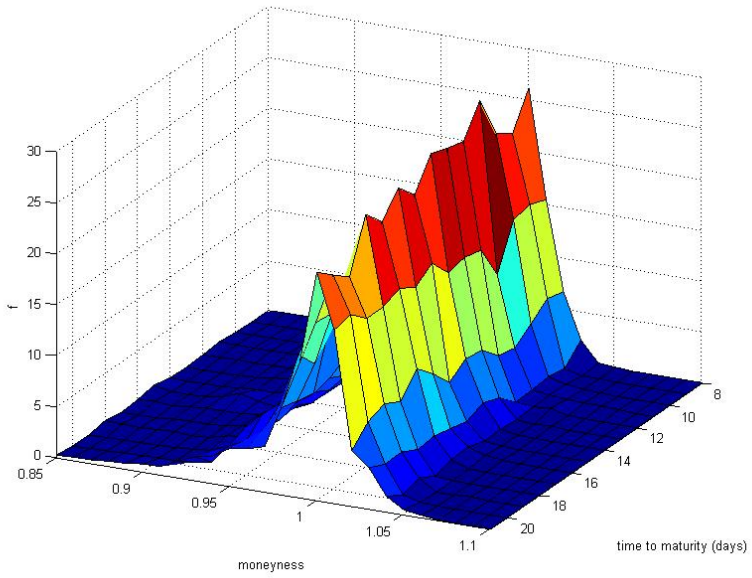


Figure 5: Kernel density estimate of the trading activity over different levels of moneyness during 8.30AM-3PM for each day between November 1 to November 20, 2013, using the rule-of-thumb bandwidth of (Bowman and Azzalini, 1997, p. 31) based on the median absolute deviation.

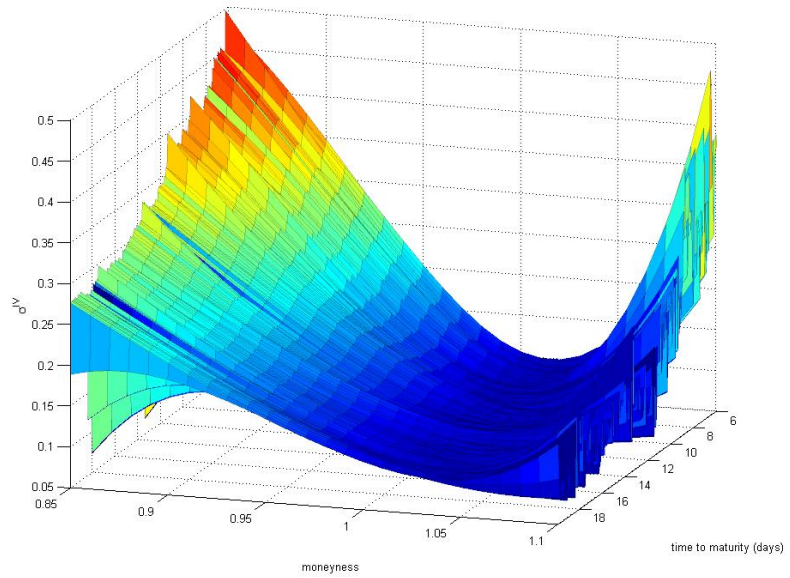


Figure 6: Estimate of the implied volatility surface using the ad-hoc cubic least squares regression for November 2013, using non-overlapping one-hour blocks to allow for time-varying parameters.

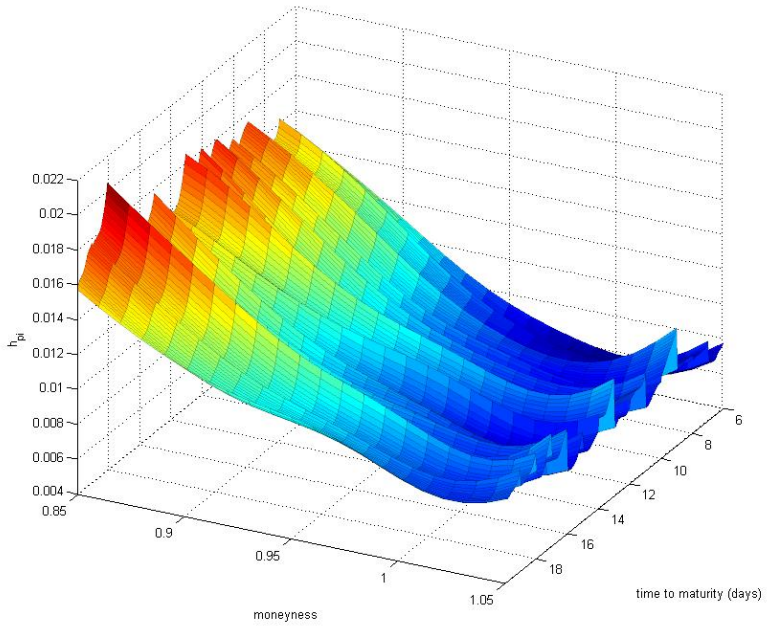


Figure 7: Plug-in bandwidth surface for November 2013 after three iterations, based on the MSE optimal diagonal bandwidth for estimating the second derivative with respect to moneyness. In each step the bandwidth surface has been smoothened to prevent additional variability for each day using a cross-validated bandwidth.

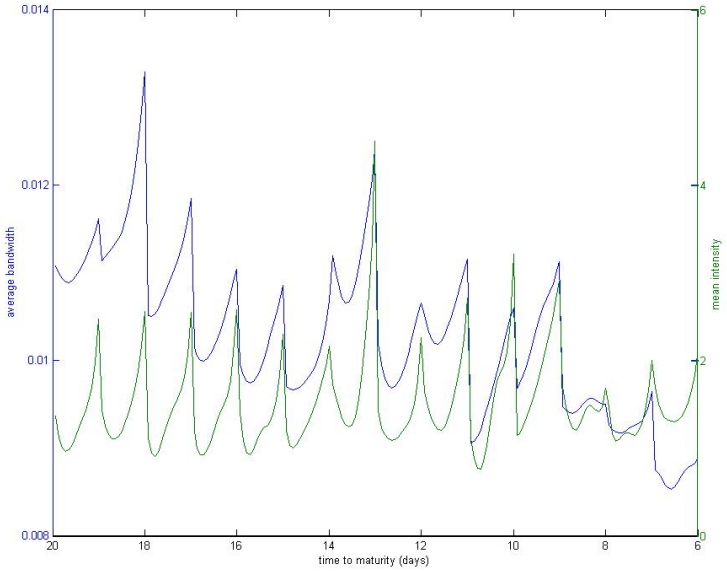
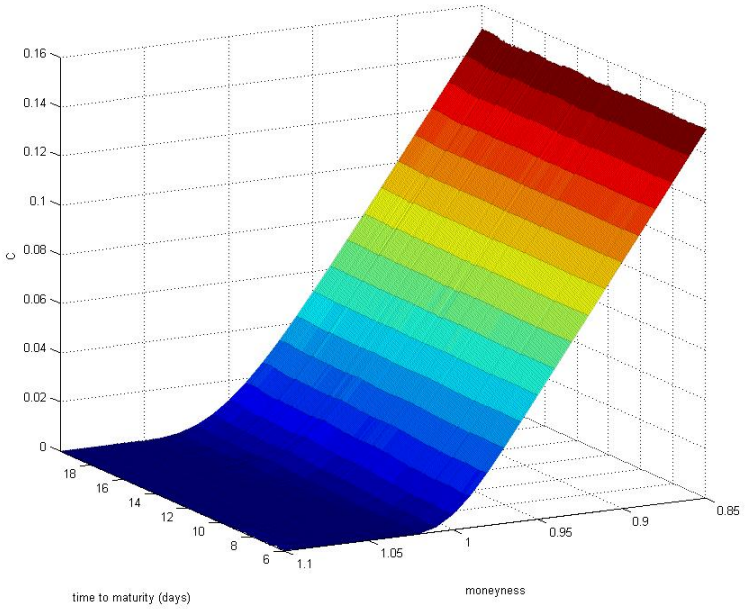
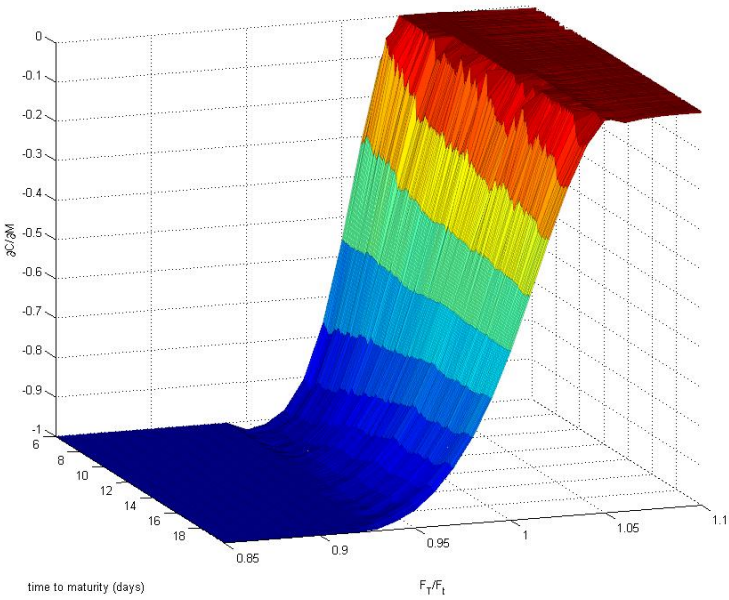


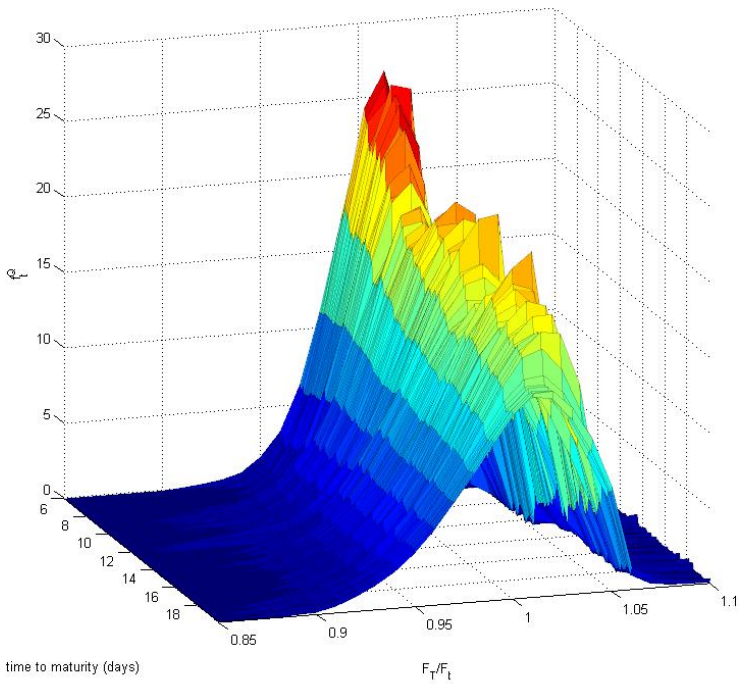
Figure 8: Comparison of the average plug-in bandwidth over different moneyness levels at any point in time with the mean intensity over time for November 2013.



(a) Call price function.



(b) Dynamics of the call price delta.



(c) Dynamics of the state price density.

Figure 9: Time-varying local polynomial smoother of call price function (upper panel), option delta (mid) and SPD (lower), using mid quotes of E-mini S&P 500 options expiring on November 29, 2013, with the bandwidth surface from Figure 7. Shape restrictions on the first and second derivative w.r.t. moneyness have been imposed, and higher order derivatives w.r.t. time have been set to zero for stability.

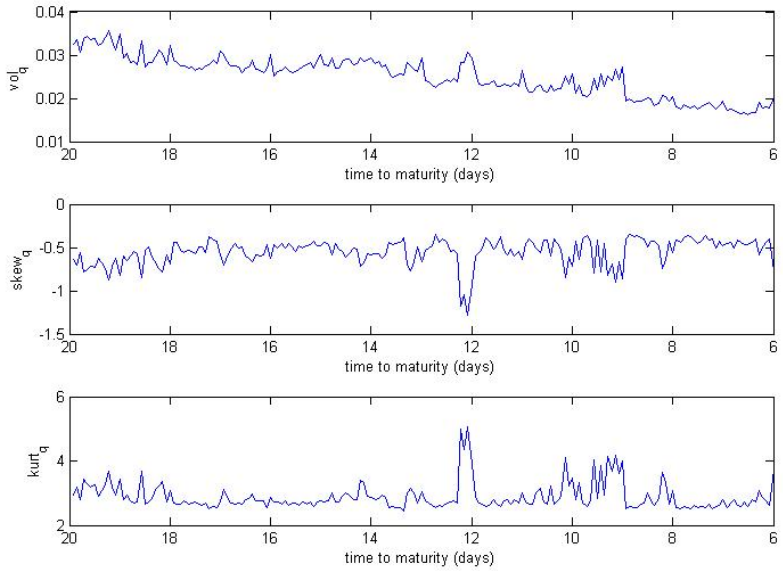


Figure 10: Dynamics of the implied volatility, skewness, and kurtosis of the state price density for the futures price at maturity November 29, 2013.

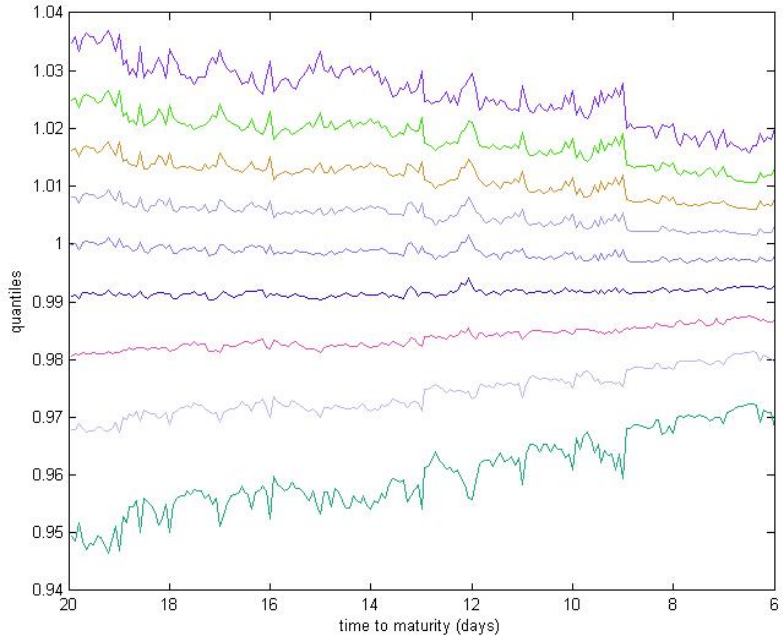


Figure 11: Dynamics of the implied 10-90% quantiles of the state price density for the futures price at maturity November 29, 2013, with mean normalized to the riskfree rate.