### MODELING MICROSTRUCTURE NOISE USING HAWKES PROCESSES

Emmanuel Bacry<sup>1</sup>, Sylvain Delattre<sup>2</sup>, Marc Hoffmann<sup>3</sup>, Jean-François Muzy<sup>4,1</sup>.

<sup>1</sup> CMAP, Ecole Polytechnique, 91128 Palaiseau Cedex, France
 <sup>2</sup> Université Paris Diderot and CNRS-UMR 7599, France
 <sup>3</sup> ENSAE-CREST and CNRS-UMR 8050, 3, avenue Pierre Larousse, 92245 Malakoff Cedex, France.
 <sup>4</sup> CNRS-UMR 6134, Université de Corse, 20250 Corté, France

### **ABSTRACT**

Hawkes processes are used for modeling tick-by-tick variations of a single or of a pair of asset prices. For each asset, two counting processes (with stochastic intensities) are associated respectively with the positive and negative jumps of the price. We show that, by coupling these two intensities, one can reproduce high-frequency mean reversion structure that is characteristic of the microstructure noise. Moreover, in the case of two assets, by coupling the stochastic intensities corresponding to the positive (resp. negative) jumps of each asset, we are able to reproduce the Epps effect, i.e., the decorrelation of the increments at microscopic scales. At large scale our model becomes diffusive and converge towards a standard Brownian motion. Analytical closed-form formulae for the mean signature plot, the diffusive correlation matrix and the cross-asset correlation function at any time-scale are given. Empirical results are shown on futures Euro-Bund and Euro-Bobl high frequency data.

*Index Terms*— Microstructure noise, Hawkes processes, Bartlett spectrum, Signature plot, Epps effect.

## 1. INTRODUCTION

During the past decade, the explosion of the amount of available data associated with electronic markets has permitted important progress in the description of price fluctuations at the microstructure level. There is a fast growing literature devoted to the modeling of intra-daily (tick-by-tick) asset prices behavior (e.g., [5, 9]).

A key issue naturally emerging is the problem of volatility or covariance estimation (which is not addressed in the aforementioned literature which mainly focuss on trades arrivals dynamics). The discrete nature of time trade arrivals and of price variations (prices are point processes living on a tick grid) and the presence of the so-called noise microstructure (strong mean reversion at small scales) makes this question highly non trivial. It is generally answered through the study of two well documented stylized facts: the signature plot behavior and the Epps Effect [6].

**Signature plot.** Let X(t) be the price of some asset at time t (defined indifferently as the last traded price or the midprice between best bid and best offer in the order book). The

signature plot corresponds to the so-called realized volatility over a time period [0, T] at a scale  $\tau > 0$ , i.e.,

$$\widehat{C}(\tau) = \frac{1}{T} \sum_{n=0}^{T/\tau} |\delta_{\tau} X[n]|^2, \tag{1}$$

where  $\delta_{\tau}X[n] = X((n+1)\tau) - X(n\tau)$  The microstructure noise effect manifests through an increase of the observed daily variance when  $\tau$  decreases (see Fig. 1(b)). This leads to a simple paradox : on the one hand, the smaller  $\tau$ , the larger the dataset that can be used to estimate the volatility, on the other hand, the realized volatility (1) is not stable as  $\tau$  decreases. Many previous works addressed this paradox, the most popular being the additive model introduced by [7] (see also [1, 11]), which expresses the price as the sum of a Brownian motion (the latent price) and of a (microstructure) noise. The estimation problem becomes a denoising problem. However, though it certainly is a good model for describing microstructure noise effects at the scale of a few minutes, it cannot faithfully reproduce the data on a microscopic scale of a few seconds (e.g., price does not change on a discrete grid,  $C(\tau)$  is artificially forced to explode when  $\tau \to 0$ ).

**Epps effect.** In the case of two assets  $X_1(t)$ ,  $X_2(t)$  whose returns are strongly correlated at daily scales, the same kind of paradox raises when trying to estimate their correlation. A correlation coefficient estimator  $\widehat{\rho}(\tau)$  over a time period [0,T] can be naturally defined from high frequency price increments by

$$\widehat{\rho}(\tau) = \frac{\widehat{C}_{12}(\tau)}{\sqrt{\widehat{C}_{11}(\tau)\widehat{C}_{22}(\tau)}},\tag{2}$$

or

$$\widehat{C}_{12}(\tau) = \frac{1}{T} \sum_{n=0}^{T/\tau} \delta_{\tau} X_1[n] \delta_{\tau} X_2[n]$$
 (3)

where  $\widehat{C}_{11}(\tau)$  (resp.  $\widehat{C}_{22}(\tau)$ ) denotes the realized volatility (1) of  $X_1(t)$  (resp.  $X_2(t)$ ) at scale  $\tau$ . The Epps effect, first reported by Epps [6], corresponds to the fact that both  $\widehat{\rho}(\tau)$  and  $\widehat{C}_{12}$  decrease when  $\tau$  decreases and they almost vanishe at very high frequency (see Fig. 3(c)) Very few approaches address the Epps effect in the literature.

In this paper, we present a "fine-to-coarse" model that starts from the description of the changes of prices in con-

This research is part of the Chair Financial Risks of the Risk Foundation.

The financial data used in this paper have been provided by the company QuantHouse EUROPE/ASIA, http://www.quanthouse.com

tinuous time and that allows one, from the microscopic properties of the model to recover a large scale diffusion behavior and, at the same time, to reproduce both the signature plot behavior and the Epps effect. This model was previously introduced in our work [2]. In Section 2 we define the model in a bivariate context (i.e., two assets) and discuss briefly the diffusion behavior. Section 3 gives closed form expression for both the signature plot and the Epps effect in a particular case (see [2] for a more general case). Numerical simulations, in both univariate and bivariate cases, are presented in section 4 and comparisons to empirical data are provided in Section 5.

### 2. THE MODEL

We start from two point processes  $X_1(t)$  and  $X_2(t)$  which represent the prices of two assets. We define the two point processes  $N_1(t)$  and  $N_2(t)$  (resp.  $N_3(t)$  and  $N_4(t)$ ) which respectively correspond to the positive jumps and the negative jumps of  $X_1(t)$  (resp.  $X_2(t)$ ), i.e.,

$$X_1(t) = N_1(t) - N_2(t)$$
  
 $X_2(t) = N_3(t) - N_4(t)$ .

If  $N_1(t)$ ,  $N_2(t)$ ,  $N_3(t)$  and  $N_4(t)$  are four independent Poisson processes with intensity  $\mu$ , it is easy to show that  $\frac{X_i(tT)}{\sqrt{T}}$  diffuses at large scale  $(T \to +\infty)$  with diffusive variance  $2\mu$  (for both i=1,2). Consequently, according to (1), their mean signature plots are flat, i.e.  $\mathbb{E}[\widehat{C}_{11}(\tau)] = \mathbb{E}[\widehat{C}_{22}(\tau)] = 2\mu$ . In order to account for a non-flat signature plot, one needs to introduce some mean reversion in the small scales for each process independently. Let us point out that, along the same time, we also want to be able to control the correlation between the two assets returns. Both correlation structures (e.g., intra asset mean reversion or inter asset correlation) can be naturally introduced using Hawkes framework [8, 4].

Let  $\lambda_i(t)$  be the stochastic intensities of the point processes  $N_i(t)$ , i = 1, 2, 3, 4, i.e.,

$$\lambda_i(t) = \lim_{\Delta \to 0} \Delta^{-1} \mathbb{E} \left[ N_i(t + \Delta) - N_i(t) \mid \mathcal{F}_t \right]$$
 (4)

where  $\mathcal{F}_t$  is the filtration generated by  $\{N_i(s)\}_{i=1..4,s< t}$ . The joint law of the processes  $N_i(t)$  are characterized by

$$\lambda_i(t) = \mu_i + \sum_{j=1}^4 \int_{-\infty}^t \varphi_{ij}(t-s) \ dN_j(s) \ i = 1, ..., 4, \quad (5)$$

where  $\mu_i$  are exogenous intensities and  $\varphi_{ij}(t)$  causal positive kernels which account for the mutual and cross excitations of positive/negative parts of the couple of assets. In the following, we will choose  $\varphi_{ij}(t) = \alpha_{ij}e^{-\beta_{ij}t}1_{\mathbb{R}^+}(t)$   $(\alpha_{ij},\beta_{ij}>0)$ . Eq. (5) fully define the processes  $N_i$  that can be shown to be stationary and stable under the conditions  $||\varphi_{ij}||_1 = \frac{\alpha_{ij}}{\beta_{ij}} < 1$ . In order to account for mean reversion and cross coupling between the two assets, we do not consider all possible mutual and cross excitations. More precisely, we choose the matrix  $\Phi = \{\varphi_{ij}\}_{1 \le i,j \le 4}$  of the following form

$$\mathbf{\Phi} = \begin{pmatrix} 0 & \varphi_{12} & \varphi_{13} & 0 \\ \varphi_{12} & 0 & 0 & \varphi_{13} \\ \varphi_{31} & 0 & 0 & \varphi_{34} \\ 0 & \varphi_{31} & \varphi_{34} & 0 \end{pmatrix}. \tag{6}$$

Thus  $\varphi_{12}$  (resp.  $\varphi_{34}$ ) is responsible for the mean reversion coupling of  $X_1(t)$  (resp.  $X_2(t)$ ) and  $\varphi_{13}$  and  $\varphi_{31}$  are responsible for the  $X_1(t), X_2(t)$  coupling.

Let us point out that, in a forthcoming paper, we prove (in the very general N-variate process case), that properly normalized the so-defined process diffuses at large scales. More precisely, if we choose the normalization

$$X_i^{(T)}(t) = \frac{1}{\sqrt{T}} X_i(tT),$$
 (7)

one can show that the process  $\{X_i(t)\}_{i=1...N,0 \le t \le 1}$  converges in law towards a multivariate Brownian motion when T goes to infinity.

### 3. MEAN SIGNATURE PLOT AND EPPS EFFECT

Let  $\{C_{kl}(\tau)\}_{1 \le k,l \le 2}$  be the covariance matrix with entries

$$C_{kl}(\tau) = \text{Cov}\left[\Delta_{\tau} X_k(t_0), \Delta_{\tau} X_l(t_0)\right]. \tag{8}$$

Thanks to the stationarity of the increments of  $X_k$ , this matrix does not depend on  $t_0$ . In particular, we recover from it the mean signature plots of both assets

$$\mathbf{E}[\hat{C}_{11}(\tau)] = C_{11}(\tau)/\tau \ , \ \mathbf{E}[\hat{C}_{22}(\tau)] = C_{22}(\tau)/\tau$$
 (9)

as well as the mean Epps effect:

$$\rho(\tau) = \frac{C_{12}(\tau)}{\sqrt{C_{11}(\tau)C_{22}(\tau)}} \text{ or } \mathbf{E}[\hat{C}_{12}(\tau)] = C_{12}(\tau)/\tau.$$
 (10)

After some algebra [2], one can actually obtain the exact expression for the mean signature plot and the mean Epps effect using the approach initiated in [3, 8].

In this paper, for the sake of simplicity, we only present the closed form expressions in the "fully symetric" case, i.e., we assume  $\mu_1=\mu_2=\mu_3=\mu_4=\mu$ ,  $\varphi_{13}=\varphi_{31}$ ,  $\varphi_{12}=\varphi_{34}$  and  $\beta_{ij}=\beta$  for all i,j. In the following, we set  $\Gamma_{ij}=||\phi_{ij}||_1=\frac{\alpha_{ij}}{\beta_{ij}}$ 

**Proposition 3.1.** Assuming  $\Gamma_{ij} < 1$  (stability condition), one has  $\Lambda = 2\mathbb{E}[\lambda_i(t)] = \frac{\mu}{1 - \Gamma_{12} - \Gamma_{13}}$ , and

$$\begin{split} \frac{C_{11}(\tau)}{\tau} = & \Lambda + \frac{RC^{+}}{2G^{+}} + \frac{RC^{-}}{2G^{-}} - R\frac{C^{+}G^{-2} + C^{-}G^{+2}}{2G^{-2}G^{+2}\tau} \\ & + R\frac{C^{-}G^{+2}\mathrm{e}^{-\tau\,G^{-}} + C^{+}G^{-2}\mathrm{e}^{-\tau\,G^{+}}}{2G^{-2}G^{+2}\tau}, \\ \frac{C_{12}(\tau)}{\tau} = & \frac{-RC^{+}}{2G^{+}} + \frac{RC^{-}}{2G^{-}} + \frac{R\left(C^{+}\lambda_{2}{}^{2} - C^{-}G^{+2}\right)}{2G^{-2}G^{+2}\tau} \\ & + \frac{R\left(-C^{+}G^{-2}\mathrm{e}^{-\lambda_{1}\tau} + C^{-}G^{+2}\mathrm{e}^{-G^{-\tau}}\right)}{2G^{-2}G^{+2}\tau} \end{split}$$

where

$$R = \frac{\beta \mu}{\Gamma_{12} + \Gamma_{13} - 1} , G^{\pm} = \beta (1 + \Gamma_{12} \pm \Gamma_{13}) ,$$

$$C^{\pm} = \frac{(2 + \Gamma_{12} \pm \Gamma_{13})(\Gamma_{12} \pm \Gamma_{13})}{1 + \Gamma_{12} \pm \Gamma_{13}}$$

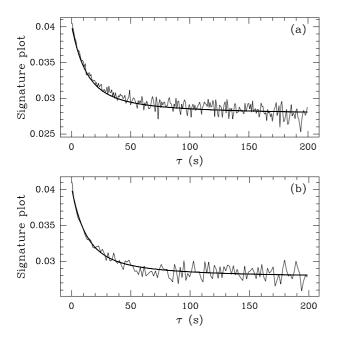


Fig. 1. (a) Estimated  $\hat{C}_{11}(\tau)$  and theoretical (Eq. (9)) analytical shape of the signature plot of a numerical simulation (42 hours long) using the parameters  $\mu=0.16$ ,  $\alpha=0.024$ ,  $\beta=0.11$ . (b) Daily signature plot  $\hat{C}_{11}(\tau)$  (computed on the dataset I described in Section 5) and theoretical fit (with Eq. (9)) using the estimator  $\theta_{reg}=(\mu=0.16,\alpha=0.023,\beta=0.11)$  (Section 4.1).

In the case, there is no coupling between the assets (i.e.,  $\phi_{31}=\phi_{13}=0$  and consequently  $G^\pm=G$  and  $C^\pm=C$ ), one can get the closed form for the mean signature plot from the last proposition :

$$\mathbb{E}[\widehat{C}_{11}(\tau)] = \Lambda + \frac{RC}{G} - \frac{RC}{G^2 \tau} (1 - e^{-\tau G}) \qquad (11)$$

From this last expression, one sees that there is a cross-over in the mean signature plot from the microstructural variance  $V_0 = \mathbb{E}[\widehat{C}(0)] = \Lambda$ , to the (smaller) diffusive variance  $V_\infty = \mathbb{E}[\widehat{C}(\infty)] = \Lambda/(1+\Gamma_{12})^2$ . In the same way, in the case the coupling term is non zero, one can easily get the asymptotic correlation coefficient

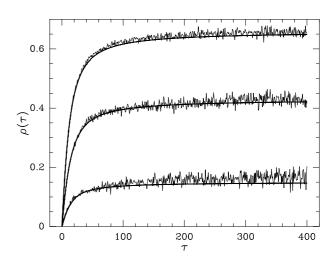
$$\rho(\tau) \to \frac{2\Gamma_{13}(1+\Gamma_{12})}{1+\Gamma_{13}^2+2\Gamma_{12}+\Gamma_{12}^2} \text{ as } \tau \to \infty.$$
(12)

# 4. NUMERICAL SIMULATIONS AND ESTIMATION

### 4.1. Univariate model

A uni-dimensionnal model can be obtained by canceling the coupling term  $(\phi_{31}=\phi_{13}=0)$  and by just looking at the process  $X(t)=X_1(t)$ . This model has only 3 parameters, namely:  $\theta=(\mu,\alpha,\beta)$  (where,  $\phi_{12}(t)=\alpha \mathrm{e}^{-\beta t}$ ). We simulated a realization of the process  $X_1(t)$  over T=42 hours with  $\theta=(\mu=0.016,\alpha=0.023,\beta=0.11)$  using a thinning method (see [10]). Let us note that  $\mu$ ,  $\alpha$  and  $\beta$  are all expressed in the the same unit, namely seconds  $^{-1}$ . These particular values were chosen to match the estimated parameters

on real data (see Fig. 1(b)). The estimation of the parameters can be processed in two very different ways depending on what is the focus of the model. On the one hand, if one is mainly interested in the ability of the model to reproduce the mean signature plot, the parameters can be estimated using a best fit of the realized signature plot (minimizing the mean square error between the theoretical and the realized mean signature plot) which leads to the estimation  $\theta_{reg}$ . On the other hand, if the goal of the model is not simply to reproduce the signature plot behaviour but to mimick the arrival times themselves, it is more natural to consider the Maximum Likelihood Estimator (MLE) instead  $\theta_{MLE}$ . This is possible since there is a closed formula for the likelihood [2]. Both estimations of the parameters match accurately the true parameter values ( $\theta_{reg} \simeq \theta_{MLE} = (0.16, 0.024, 0.11)$ ). Let us point out that, whereas  $\widehat{\theta}_{reg}$  can be performed on uniformly sampled data,  $\widehat{\theta}_{MLE}$  needs to have access to the point process itself. For that reason, when applied to real data, we expect it to be much more stable than  $\theta_{MLE}$ . Moreover, the regression estimator has the advantage to be computationally faster than the MLE.



**Fig. 2**. Epps effect for simulation samples. The estimated Epps effect (Eq. (2)) are compared to the theoretical analytical curves (Eq. (10)). From top to bottom the large scale cross correlation between the two asset returns increases from 0.15 to 0.65.

### 4.2. Bivariate model

The "fully symetric" bivariate model (as described before Proposition 1) has 4 parameters namely :  $\theta = (\mu, \alpha_{12}, \alpha_{13}, \beta)$ . A realization of this bivariate process  $(X_1, X_2)$  over T=20 hours has been simulated (using a thinning method [10]) for three sets of parameters such that, according to Eq. (12), the asymptotic correlations between the (large scale) increments of  $X_1$  and  $X_2$  are, respectively  $\rho \simeq 0.15$ ,  $\rho \simeq 0.40$  and  $\rho \simeq 0.65$ . In Fig. 2, the estimated (Eq. (2)) and theoretical (Eq. (10)) Epps effect are plotted for these three sets of parameters (the asymptotic values of the curves do match the theoretical asymptotic value of  $\rho$ ). Estimations were obtained using regression estimation minimizing both the MSE of each signature plots and of the Epps effect.

### 5. COMPARISONS TO EMPIRICAL DATA

The data that have been used in this paper consist in tick-bytick last traded price time series. Dataset I (resp. II) corresponds to 21 days (resp. 41 days) between 11/01/2009 to 12/15/2009 (resp. 06/01/2009 to 08/01/2009). For each day, only the most liquid maturity has been used. The prices are either Eurex Euro-Bund futures contracts or Eurex Euro-Bobl futures contracts which correspond respectively to long-term or medium-term debt instrument (in Germany). The liquidity (and the volatility) is highly seasonal during the day and since our model does not account for such a seasonality, we shall restrict the data to the intraday period 9am to 11am. All computations have been made on last traded prices of buy orders only. Choosing sell orders would not have change the results, however, taking into account in the same time-series both buy and sell orders would lead to a highly bouncing artefact that shall not be able to be captured by our modelling approach.

Fig. 1(b) shows the daily signature plot of the Euro-Bund and its fit using the univariate model. It has been computed by averaging the value of  $\widehat{C}(\tau)$  (Eq. (1)) computed independantly on each day . Fig. 3(a) (resp. 3(b)) shows the daily signature plot of the Euro-Bund (resp. Euro-Bobl) and Fig. 3(c) shows the Epps effect  $\widehat{C}_{12}(\tau)$  (Eq. (3)). The fits (solid line) were obtained using the bivariate model  $\theta_{reg}$  estimations (in the non symetric case described in [2]). Given the simplicity of the model, one can consider that the model captures fairly well both variance and covariance features of assets from small to large time scales simultaneously.

# 6. REFERENCES

- [1] Y. Ait-Sahalia, P.A. Mykland, and L. Zhang. How often to sample a continuous-time process in the presence of market microstructure noise. *The Review of Financial Studies*, 18:351–416, 2005.
- [2] E.Bacry, S.Delattre, M.Hoffmann and J. F.Muzy. Modelling microstructure noise with mutually exciting point processes. *Submitted to Quantitative Finance*, 2010.
- [3] M. S. Bartlett. The spectral analysis of point processes. *Journal of the Royal Statistical Society. Series B.*, 25:264–296, 1963.
- [4] D.J. Daley and D. Vere-Jones. *An introduction to the theory of point processes*. Springer series in statistics, 2005.
- [5] R. F. Engle and J. R. Russell. Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66:1127–1162, 1998.
- [6] T. W. Epps. Comovements in stock prices in the very short run. Journal of the American Statiscal Association, 74:291– 298, 1979.
- [7] A. Gloter and J. Jacod. Diffusion with measurement errors. i. and ii. ESAIM: Prob. & Stat., 5:225-242 and 243-260, 2001.
- [8] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90, 1971.
- [9] P. Hewlett. Clustering of order arrivals, price impact and trade path optimisation. *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, 2006.
- [10] Y. Ogata. On lewis simulation method for point processes. *IEEE Information Theory*, 27:23–31, 1981.
- [11] L. Zhang, P. Mykland, and Y. Ait-Sahalia. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *J. Amer. Statist. Assoc.*, 472:1394–1411, 2005.

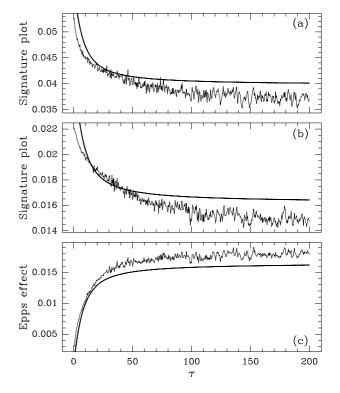


Fig. 3. Signature plots and Epps effect of the Euro-Bund and Euro-Bobl returns as functions of the time scale  $\tau$ . The computations have been made using the dataset II described in Section 5. (a) Daily signature plot  $\widehat{C}_{11}(\tau)$  associated with the Euro-Bund (b) Same as in (a) for the Euro-Bobl ( $\widehat{C}_{22}(\tau)$ ). (c) Epps effect  $\widehat{C}_{12}(\tau)$  between Euro-Bund and Euro-Bobl. In (a), (b) and (c) the solid lines represent a  $\theta_{reg}$  fit of the empirical curves according to the bivariate model.