

Accounting for the Epps Effect: Realized Covariation, Cointegration and Common Factors

Jeremy Large*

JEREMY.LARGE@ECONOMICS.OX.AC.UK

All Souls College, University of Oxford, Oxford, OX1 4AL, U.K.

24 July 2007

Abstract

High-frequency realized variance approaches offer great promise for estimating asset prices' covariation, but encounter difficulties connected to the Epps effect. This paper models the Epps effect in a stochastic volatility setting. It adds dependent noise to a factor representation of prices. The noise both offsets covariation and describes plausible lags in information transmission. Non-synchronous trading, another recognized source of the effect, is not required. A resulting estimator of correlations and betas performs well on LSE mid-quote data, lending empirical credence to the approach.

*I am grateful to Neil Shephard for his support and advice. I thank warmly participants at the Stanford Institute of Theoretical Economics segment on High-Frequency Data and the Impact of Economic News, June 2007, for helpful comments, as well as those present at a workshop organized by Nour Meddahi in Imperial College, London, February 2007.

1 Introduction

The covariance of financial asset returns is of central importance in the theory of asset prices, and is a recurring theme throughout finance. Finding good empirical *ex post* estimates of covariance is a key step to understand it better. For this purpose, there is an opportunity to draw on recent advances in the study of *ex post* realized variances, see for example Barndorff-Nielsen and Shephard (2002), Andersen, Bollerslev, and Meddahi (2004) and Andersen, Bollerslev, Diebold, and Labys (2003). Indeed, a program of research was set out in Barndorff-Nielsen and Shephard (2004), to extend these advances from the univariate to the multivariate case, where they should yield good estimators of covariation, and thereby of covariance.

However this encounters a puzzling problem, studied comprehensively in Epps (1979) for US automobile stocks, whereby empirical correlations virtually disappear at high frequencies of the order of a minute, while being far from zero at moderate intraday frequencies. Epps' findings have been replicated extensively in financial markets.

The Epps (1979) effect has been widely associated with non-synchronous trading, when fresh observations of transactions prices do not arise simultaneously across markets, but are separated by e.g. a few seconds – see Scholes and Williams (1977) and Martens (2003). If non-synchronous trading is the source of the Epps effect, there is a challenging consequence for realized covariation estimation: namely, that robust high-frequency covariation estimators must take explicit account of the pure-jump nature of price processes. Indeed, Hayashi and Yoshida (2005) develops an ‘all-overlapping-returns’ estimator of covariation to do this (as named by F. Corsi), and Lunde and Voev (2006) and Zhang (2006) assess it when there is ‘contamination’ or measurement error.

However, evidence from Renó (2003) indicates that on equity and currency markets non-synchronous trading is not alone sufficient to explain Epps effects.¹ This paper goes further, providing examples where it is not necessary either. Instead, it provides an orthogonal and complementary explanation for the Epps effect. Starting from the

¹Renó (2003) implements Fourier analysis-based methods in Malliavin and Mancino (2002), as do Mancino and Renó (2005) and Precup and Iori (2005).

familiar common factor framework of multiple asset returns, it adds a dependent error term. This error produces the Epps effect by introducing lagged cross-correlation at the expense of contemporaneous correlation, and it also represents delays in information transmission between markets.

As this encompasses the Epps effect within a ‘semi-martingale plus noise’ framework, it is encouraging that the ongoing development of multivariate versions of estimators such as those in Aït-Sahalia, Mykland and Zhang (2005), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2006) and Zhang, Mykland, and Aït-Sahalia (2005) will be effective.

To confirm the empirical relevance of the approach, the paper develops it to the point of deriving its own estimators of correlations and betas (which are defined later), that should be robust to the Epps effect. Implementations of the correlation estimator on a group of stocks traded at the London Stock Exchange are favorable in a model which is well-specified even without taking account of non-synchronous trading, comparing well against a non-parametric alternative. To relate the approach to another important literature, resulting extensions of Hasbrouck (1995) information shares are developed, for the case of multiple assets with a single latent common pricing factor.

Write observed asset log-prices as a vector Y of length n . The quadratic variation process of Y , which is often denoted $[Y]_t$, is defined as follows:

$$[Y]_t = p\text{-}\lim \sum_{i=0}^{m-1} (Y_{s_{i+1}} - Y_{s_i})(Y_{s_{i+1}} - Y_{s_i})', \quad (1)$$

where $0 = s_0 < s_1 < s_2 < \dots < s_m = t$ determines a grid. In the probability limit, $m \rightarrow \infty$ and the grid’s maximum increment tends to zero. The off-diagonal elements of $[Y]_t$ are the realized covariations, while realized correlations and betas are simple ratios of elements in the matrix $[Y]_t$ to be defined later. They are zero when $[Y]_t$ is diagonal. In this framework, Epps’ puzzling finding is that $[Y]_t$ is (almost) diagonal despite considerable comovement in elements of Y .

Since, at least off-diagonal, $[Y]_t$ is therefore finite, we will be concerned with a model where the observed price Y is a semi-martingale. Concentrating only on periods free of the sporadic large jumps due to e.g. public announcements,² it follows that this semi-

²QV-related tests for these in Barndorff-Nielsen and Shephard (2006a) are implemented in Andersen,

martingale is continuous.

Therefore an SDE can be used to decompose the price vector Y into idiosyncratic and cross-sectionally dependent parts: let

$$dY_t = dL_t + \sigma_t dW_t, \quad (2)$$

where σ is a diagonal volatility process and W is a multivariate standard Brownian motion. As σ is diagonal, the term $\sigma_t dW_t$ captures only idiosyncratic innovations to the prices in Y . Assume that $L \perp\!\!\!\perp (\sigma, W)$. So all dependence in Y is through L .

The Epps effect can then be obtained by introducing a cointegrating relation, the theory of which was set out in important work by for example Engle and Granger (1987) and Johansen (1988), and applied in continuous time in Phillips (1991), Corradi (1997) and Comte (1998). Let L evolve according to

$$dL_t = \alpha\beta' L_t dt + \omega_t dZ_t, \quad (3)$$

where ω is another volatility process; α and β are full-rank $n \times (n - r)$ matrices with $r \in \mathbb{N}$; and Z is another multivariate standard Brownian motion. Later, (3) will be generalized further and technical assumptions will be made, but otherwise this largely completes the specification of the model.

Because of the cointegrating relationship in (3) the returns in Y can be closely correlated at moderate frequencies, such as hourly or less frequently still. Indeed, we will see using a relative of the Granger-Johansen representation, that Y has a familiar common factor representation (plus a noise term). However, substituting (3) into (2) and applying Itô algebra,

$$E[dY_t dY_t'] = (\sigma_t \sigma_t' + \omega_t \omega_t') dt. \quad (4)$$

Barndorff-Nielsen and Shephard (2002) and Meddahi (2002) show that $[Y]_t$ is the integral of (4). Therefore in any specification where not only σ , but also ω , is diagonal, the quadratic variation process $[Y]_t$ is diagonal as well.

Bollerslev, and Diebold (2006). See also Aït-Sahalia (2002).

Thus realized correlations entirely disappear in the high-frequency limit, an extreme case of the Epps (1979) effect. One insight for this is that the continuous process, L_t , may be cointegrated without cross-sectional covariation since its pairwise comovements are, at high frequencies, $o(dt^2)$ per duration of length dt . More moderate Epps effects can be captured by allowing ω to be non-diagonal.

The paper proceeds as follows: in a generalized setting, Section 2 motivates and explains how the proposed model controls for the Epps effect in a plausible way. Section 3 then discusses how to assess the model's fit with reality. Semi-parametric estimators of correlations and betas are proposed, which are robust to the Epps effect according to the model. These estimators are then implemented on pairs of London Stock Exchange (LSE) equities in Section 4 and their quality assessed. Section 5 formulates an application of the current model to Hasbrouck (1995) information shares. Section 6 concludes.

2 Theoretical motivations

An important motivation for this calculus is its accommodation of the Epps effect, as was outlined in the Introduction. In addition, it has two main theoretical motivations. First, in it Y has a permanent-transitory decomposition such that the permanent part, denoted Y^* , has a familiar common factor representation. Second, the error-correcting behavior of L_t in (3) has a natural interpretation in terms of delays in information transmission across markets. In this Section these theoretical motivations are described, before the question of empirical relevance is turned to in Section 3.

Evidently, generalizations of the model could have the same (and more) theoretical properties: for example, the parameters α and β might vary over time as stochastic processes, and a drift term μ_t might appear. To concentrate on the essential issues, however, I refrain from such generalizations.

2.1 Common factor representation of Y

Let Y be observed over the interval $[0, T]$, perhaps a trading day. Choose units of time conveniently, e.g. minutes, so that $T \in \mathbb{N}$ and may be large even over a single day. I assume from now on that σ and ω are adapted, stationary and uniformly bounded processes. Suppose $Y_0 = L_0$, and is a random initial value of (for simplicity) mean zero.

Let $\tau : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be a fixed time-change so that for all t , $\tau(t) \leq t$. Instead of (3), let

$$dL_t = \alpha\beta' L_{\tau(t)} dt + \omega_t dZ_t. \quad (5)$$

This generalization is without analytical cost, and will help later in estimations. For now however the reader is encouraged to concentrate, without loss of insight, on the case $\tau(t) = t$, which recovers (3).

Write α_{\perp} and β_{\perp} for orthogonal complements of α and β respectively. So $\alpha_{\perp}'\alpha = 0$ etc.

Proposition 2.1 *The price Y has the **common factor representation with error**:*

$$Y_t = \tilde{\beta}F_t + \int_0^t \sigma_u dW_u + \epsilon_t, \quad (6)$$

where $F = \alpha_{\perp}'L$ and $\tilde{\beta} = \beta_{\perp}(\alpha_{\perp}'\beta_{\perp})^{-1}$, and where

$$\epsilon = \alpha(\beta'\alpha)^{-1}(\beta'L). \quad (7)$$

Provided that $\beta'L$ is stationary, ϵ is a stationary error.

Proof. Recall that $\tilde{\beta}\alpha_{\perp} + \alpha(\beta'\alpha)^{-1}\beta' = \mathbf{I}$. Therefore, $L_t = \tilde{\beta}\alpha_{\perp}'L_t + \epsilon_t$. Now use (2). ■

The common factor representation may be interpreted as follows. Note that from (5),

$$dF_t := d(\alpha_{\perp}'L_t) = \alpha_{\perp}'\omega_t dZ_t, \quad (8)$$

so that F is a local martingale.

Therefore F may be interpreted as a vector of r local martingale pricing factors with stochastic volatility, that are common across the assets in Y , whereas $\tilde{\beta}$ gives the factor loadings. This is added to a local martingale vector of cumulated idiosyncratic, shocks to prices, $\int_0^t \sigma_u dW_u$. As $L \perp (\sigma, W)$, there is no covariation between this and F : in other words, the idiosyncratic shocks are uncorrelated with shocks to the common factors.

2.2 Error term and lagged information transmission

Interesting cases will be ones where $\beta' L$ is stationary. Then, a mean-zero stationary error term, ϵ_t , appears in the common factor representation of Proposition 2.1. When it deviates from zero, this indicates that $\beta' L_t \neq 0$. But, the condition $\beta' L_t = 0$ is the long-run relationship of the error-correcting process, L_t , as determined in (3). So, as ϵ_t reverts towards zero, the long-run relationship is restored, with ϵ_t 's dynamic properties determine the rate and way that this comes about.

Therefore, a non-zero error can be interpreted as indicating that common information has not been proportionately impounded into all prices: although they have been shocked, they have not yet reverted to their long-run relationship as defined by $\beta' L_t = 0$ or, equivalently, by $\epsilon_t = 0$. An example of this will be worked through in the empirical implementation of Section 4.

As the first two components in the common factor representation are local martingales, while ϵ is stationary, the representation also provides the aforementioned permanent-transitory decomposition of the log-price vector, given by

$$Y_t = Y_t^* + \epsilon_t, \quad (9)$$

which defines Y^* . If ω_t is always diagonal, then the error, ϵ , is dependent on Y^* in such a way as to 'cancel-out' the covariation between elements of Y^* . This follows from (4). It means that realized covariations are zero, consistently with an extreme Epps effect.

To have this consequence, ϵ must of course be closely related to F . However, note that unless β and α_{\perp} are collinear, ϵ is not progressively measurable with respect to the common factors, F , and the idiosyncratic effects.

2.3 Condition for stationary error

Typically, simple parameter restrictions ensure that $\beta' L$, and so ϵ , is stationary. For example:

Lemma 2.2 *Suppose $\tau(t) = t$ and that $-\beta' \alpha$ is positive definite. Then Y_0 can be given a distribution such that $\beta' L_t$ is a stationary process.*

Proof. Note that from (5),

$$d(\beta' L)_t = (\beta' \alpha)(\beta' L)_t + \beta' \omega_t dZ_t. \quad (10)$$

This specifies $\beta' L$ as an Ornstein-Uhlenbeck process, which is known to have an initial condition making it stationary provided that $-\beta' \alpha$ is positive definite. ■

3 Empirical relevance and comparison to data

If this model is realistic, then it should be possible to apply it to data in order to estimate covariations, correlations and betas that are free of the Epps effect. Estimation would involve specification checks that would further reinforce or challenge the model's fit with reality.

To this end, it is natural, as in the univariate case, to focus on the variation in the permanent martingale price process, Y^* , defined in (9). Being a local martingale, its returns are free of lagged dependence, so that the lead-lag effect has been eliminated. As the sum of the idiosyncratic components with the loaded common factors, it can be viewed as an unobserved underlying or efficient price containing the fundamental news that passes into observed realized prices with varying delays. Now,

$$dY_t^* = \tilde{\beta} \alpha' \omega_t dZ_t + \sigma_t dW_t, \quad (11)$$

so the quadratic variation process of Y^* is

$$[Y^*]_t = \int_0^t \tilde{\beta} \alpha' \omega_u \omega_u' \alpha \tilde{\beta}' + \sigma_u \sigma_u' du. \quad (12)$$

Write $[Y^*]_{t,i,j}$ for the (i, j) th element of $[Y^*]_t$. Barndorff-Nielsen and Shephard (2004) defines the realized correlation between the i 'th and j 'th assets over $[0, T]$ as

$$\widetilde{Cor}_{i,j} = \frac{[Y^*]_{T,i,j}}{\sqrt{[Y^*]_{T,i,i}[Y^*]_{T,j,j}}}; \quad (13)$$

while the realized regression coefficient, or beta, of the i 'th asset on the j 'th asset is

$$\tilde{\beta}_{i,j} = \frac{[Y^*]_{T,i,j}}{[Y^*]_{T,j,j}}. \quad (14)$$

This section develops a special case of the model, within which estimators of correlations and betas can be derived. The next section, Section 4, implements the correlation estimator for pairs of similar stocks traded on the LSE and does specification testing on the model.

3.1 A discrete-time specification

Under the maintained assumption that the volatilities σ_t and ω_t are stationary and bounded over the course of a trading day, $[0, T]$, the underlying moments of $\widetilde{Cor}_{i,j}$ and $\widetilde{\beta}_{i,j}$ exist, and are given by

$$Cor_{i,j} = \frac{E[Y^*]_{T,i,j}}{\sqrt{E[Y^*]_{T,i,i}E[Y^*]_{T,j,j}}}; \text{ and } \beta_{i,j} = \frac{E[Y^*]_{T,i,j}}{E[Y^*]_{T,j,j}}. \quad (15)$$

This part specializes the model to the point where these underlying moments may be estimated. Essentially this involves a discretization. For this purpose, assume that even though Y is a continuous process over the entire interval $[0, T]$, it is sampled at discrete times $t = 0, 1, 2, \dots, T$.

Furthermore, suppose that $\tau(t) = [t]$ (where $[t]$ is the integer part of t). This gives a high-frequency but discrete-time representation of the data generating process, with the advantage of having known statistical properties. To see this, first define

$$\eta_t^Y = \Delta Y_t - \Delta L_t, \quad (16)$$

where $\Delta Y_t = Y_t - Y_{t-1}$. So η_t^Y is an n -variate martingale difference sequence. Setting $\tau(t) = [t]$ implies that at integer times, t , L_t has the error-correction form:

$$\Delta L_t = \alpha\beta' L_{t-1} + \eta_t^L, \quad (17)$$

where η_t^L is another martingale difference sequence. This follows on aggregating (5) between sampling times.

This specification has residuals, η_t^Y and η_t^L , with conditional heteroskedasticity of unknown law, and whose unconditional variances exist (the usual restrictions on fourth-order cumulants follow as σ and ω are bounded). Note that as $L \perp\!\!\!\perp (\sigma, W)$, we have that

$\eta^Y \perp\!\!\!\perp \eta^L$: that is, the two time-series η^Y and η^L are independent. Let $E[\eta_t^Y \eta_t^{Y'}] = \Sigma$, and let $E[\eta_t^L \eta_t^{L'}] = \Omega$. So,

$$E \left[\int_0^T \omega_u \omega_u' du \right] = T\Omega, \quad \text{and} \quad E \left[\int_0^T \sigma_u \sigma_u' du \right] = T\Sigma, \quad (18)$$

giving from (12),

Proposition 3.1 *In this discrete-time framework,*

$$E[Y^*]_T = T \left(\tilde{\beta} \alpha_{\perp}' \Omega \alpha_{\perp} \tilde{\beta}' + \Sigma \right). \quad (19)$$

Thus, if the parameters of the model can be consistently estimated, then $Cor_{i,j}$ and $\beta_{i,j}$ can be also: this is achieved by taking appropriate ratios of the elements in the matrix $\left(\widehat{\beta} \widehat{\alpha}_{\perp}' \widehat{\Omega} \widehat{\alpha}_{\perp} \widehat{\beta}' + \widehat{\Sigma} \right)$ (with the factor of T cancelling out in both cases), see (15).

Finally, I provide the appropriate parameter condition such that the error ϵ in the representation of Proposition 2.1 is stationary when observed at integer times. Let $\Psi = (\mathbf{I} + \beta' \alpha)$.

Lemma 3.2 *Suppose $\tau(t) = \lfloor t \rfloor$ and that Ψ , i.e. $(\mathbf{I} + \beta' \alpha)$, has roots inside the unit circle. Then Y_0 can be given a distribution such that $\{\epsilon_t : t \in \mathbb{N}\}$ is a stationary process.*

Proof. Note that from (17), when $\tau(t) = \lfloor t \rfloor$, we have that at integer times $t \in \mathbb{N}$,

$$\beta' L_t = \Psi(\beta' L)_{t-1} + \beta' \eta_t^L. \quad (20)$$

So $\{\beta' L_t : t \in \mathbb{N}\}$ is a heteroskedastic AR(1) process with stationary error, which has an initial condition making it stationary provided that Ψ has roots inside the unit circle. Finally, $\epsilon = \alpha(\beta' \alpha)^{-1}(\beta' L)$. ■

3.2 Consistency and asymptotic limit theory

There are two natural candidates for asymptotic limit theory in this context: the first is a standard large- T theory; while the second is an infill asymptotic theory as deployed in Aït-Sahalia, Mykland and Zhang (2005), Bandi and Russell (2006), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2006) and Zhang, Mykland, and Aït-Sahalia (2005), where sampling becomes arbitrarily more frequent over a fixed period of observation.

Large- T asymptotic theory has the benefit that it permits known discrete time-series results to be deployed quite readily. Such results about quasi-maximum likelihood estimation will in fact be very helpful. However, $[Y]_t$ cannot be estimated consistently without an alternative, infill asymptotic theory. Otherwise, for example, a momentary peak in volatility, which raises elapsed quadratic variation, may fall between observations and go unobserved in the large- T limit.³

Nevertheless, given the assumption of stationary volatility, the underlying moments, $Cor_{i,j}$ and $\beta_{i,j}$ in (15), are identifiable in a large- T framework, and this approach is taken here. Note that (16) and (17) define a multivariate linear process with unmodelled, stationary, conditional heteroskedasticity. Kuersteiner (2001) indicates that such processes can be consistently estimated by maximizing the Gaussian pseudo-likelihood (by PML). He demonstrates this in the univariate case, and suggests that the extension to a multivariate setting is straightforward.

This can be implemented using a state-space representation, with code from `ssfPack` for Ox introduced in Koopman, Shephard, and Doornik (1998).⁴ Maximization can be done using the MaxBFGS algorithm in Ox with numerical derivatives (see Doornik 2001).

Let θ be the vector of parameters, containing $\{\alpha, \beta, \Sigma, \Omega\}$. Let $s(\theta)$ be the score vector (evaluated on the data). Given consistency, the asymptotic covariance of θ is given by the usual limit theory when $T \rightarrow \infty$,

$$\sqrt{T} (\hat{\theta} - \theta) \xrightarrow{d} N(0, J^{-1} I J^{-1}), \quad (22)$$

where

$$I = \lim_{T \rightarrow \infty} \frac{1}{T} cov[s(\theta)], \quad (23)$$

³The problem is only exacerbated here, when the object of interest is $[Y^*]_t$, since Y^* is unobserved.

⁴Because of the moving average component to ΔL_t , the state variables are both ΔL_t and η_t^L . Write $\mathbf{0}_n$ for an $(n \times n)$ matrix of zeros, and \mathbf{I}_n for the $(n \times n)$ identity matrix. Write $\Phi = \mathbf{I}_n + \alpha' \beta$. Then the state space representation is

$$\begin{pmatrix} \Delta L_{t+1} \\ \eta_{t+1}^L \\ \Delta Y_t \end{pmatrix} = \begin{pmatrix} \Phi & -\mathbf{I}_n \\ \mathbf{0}_n & \mathbf{0}_n \\ \mathbf{I}_n & \mathbf{0}_n \end{pmatrix} \begin{pmatrix} \Delta L_t \\ \eta_t^L \end{pmatrix} + \begin{pmatrix} \eta_{t+1}^L \\ \eta_{t+1}^L \\ \eta_t^Y \end{pmatrix}. \quad (21)$$

and

$$J = \lim_{T \rightarrow \infty} \frac{1}{T} E \left[-\frac{\partial s}{\partial \theta} \right], \quad (24)$$

both evaluated at the truth. Working from (22), the Delta Method can be easily used to give limit theories for the estimates of $Cor_{i,j}$ and $\beta_{i,j}$, for the purposes of inference. In the upcoming implementation, J and I will be computed numerically, evaluating them at the MLEs and, in the case of I , using the well-known technique due to Newey and West (1987).

4 Empirical implementation

This section implements the model to study covariation between three relatively heavily-traded UK equities, AstraZeneca (hereafter AZ), GlaxoSmithKline (GSK) and Shell, which trade on the London Stock Exchange SETS limit order book. AZ and GSK are from the same industry, namely pharmaceuticals. This suggests that specific detailed industry information about AZ may have implications for GSK; and vice versa. The model provides a means to see if such information is transferred between prices with a delay. Shell, by contrast, is an oil and gas major.

The data runs from the start of October 2004 to the end of February 2005. Data was timed to the nearest second. Over this period, continuous trading ran from from 8am to 4:30pm each day, except for 24 December and 31 December, when markets closed at 12:30pm. These two days were excluded. AZ's best bid and ask pair changed on average every 22 seconds; while GSK's changed every 46 seconds; and Shell's changed every 35 seconds.

To minimize issues around nonsynchronicity, I followed Martens (2003) in using mid-quotes as proxies for underlying prices. Large (2005) emphasizes that quotes data is continuously observed by the econometrician, in contrast to the punctuated observations provided in transactions data. So, importantly for estimating covariation, prevailing mid-quotes can be observed simultaneously across markets. Mid-quotes have also been advocated in Barndorff-Nielsen and Shephard (2007) as superior to transaction prices for the purpose of minimizing bias in volatility estimators due to univariate market

microstructure noise induced, for example, by limit order book dynamics.

The three equities' mid-quote returns were then sampled every 90 seconds in logarithms. Overnight returns were excluded, as were mid-quote returns in the first 5 minutes of the trading day, when after-effects of the opening auction are known to induce unique market microstructure. This resulted in 336 observations per equity per day, on 102 trading days: a trivariate time-series of 34,272 observations. Recognizing that in the data there are jumps (see Barndorff-Nielsen and Shephard 2006a and 2006b), for each stock, those returns whose magnitudes exceeded six standard deviations were set to zero. They represented 0.10% of observations.

4.1 Fit to the whole sample

Before moving on to a day-by-day analysis, the model was fitted to the entire data set of 102 days. The dimensions of α and β , in particular the choice of r , should be determined by the data. If $r = 1$, then there are, flexibly, two common factors for the three assets; but there is a rather limited error term ϵ , of rank 1 since $\beta'L$ is univariate. Estimating the model with $r = 1$ was found to be mis-specified.

On the other hand, setting $r = 2$ met with better results. This allows for only one common factor, F , but permits richer dynamics for information transmission between markets. Thus, α and β are (3×2) . I now show that this specification fits the data well.

Recall that σ is diagonal, so that the idiosyncratic parts of assets' returns have zero covariation. I also assume from now on that these are mutually uncorrelated at the sampling interval of 90 seconds, so that Σ is diagonal. The main case ruled out by this is of a correlated crash or jump in prices, due to tail dependence.

Imposing the restriction that all elements of L are non-stationary, I normalize β so that

$$\beta = \begin{pmatrix} 1 & 1 \\ 1 & \dagger \\ \times & 0 \end{pmatrix}, \quad (25)$$

where $\dagger \neq 1$. Thus, the right-hand cointegrating relation regards AZ and GSK alone, while the other relates Shell to them, as an evenly-weighted pair.

Figure 1 reports the auto- and cross- correlations in the trivariate time-series of

returns, looking up to 12 lags, or 18 minutes into the past. Significant lagged effects exist. However, these are largely absent in the model's residuals. This indicates that even over a long period of five months, during which parameter instability might be suspected, the model is reasonably well-specified.

Despite the use of mid-quotes, unmodelled univariate microstructure noise remains an *a priori* concern here. Its moderate levels in the cases of AZ, GSK and Shell makes them particularly suited to this exercise. Indeed, Figure 1 indicates that the autocorrelation in AZ's, GSK's and Shell's mid-quote returns that might have been due to univariate noise is in fact accounted for by the current model.

Recall that (17) reads

$$\Delta L_t = \alpha \beta' L_{t-1} + \eta_t^L. \quad (26)$$

Once estimated, (17) takes the form

$$\Delta L_t = \begin{pmatrix} \Delta L_t^{AZ} \\ \Delta L_t^{GSK} \\ \Delta L_t^{Shell} \end{pmatrix} = \begin{pmatrix} -0.11 & -0.23 \\ -0.14 & 0.27 \\ 0.11 & 0.01 \end{pmatrix} \begin{pmatrix} 1 & 1 & -4.91 \\ 1 & -0.75 & 0 \end{pmatrix} L_{t-1} + \eta_t^L. \quad (27)$$

All parameters are significant at 1%. Put into the common factor representation of Proposition 2.1,

$$Y_t := \begin{pmatrix} Y_t^{AZ} \\ Y_t^{GSK} \\ Y_t^{Shell} \end{pmatrix} = \begin{pmatrix} 0.789 \\ 1.045 \\ 0.373 \end{pmatrix} F_t + \int_0^t \sigma_u dW_u + \epsilon_t. \quad (28)$$

So AZ and GSK are more heavily loaded onto the common factor than Shell, with loadings of 0.789 and 1.045 relative to a loading of 0.373 for Shell. This suggests that the common factor heavily weights the pharmaceutical sector. The quantity $\widehat{\Psi} = (\mathbf{I} + \hat{\beta}' \hat{\alpha})$ is the estimated autoregressive parameter in (20), and therefore also the estimated autoregressive parameter of ϵ . This is informative about the *rate* of information transmission between markets. It is almost diagonal:

$$\mathbf{I} + \hat{\beta}' \hat{\alpha} = \begin{pmatrix} 0.20 & -0.001 \\ -0.001 & 0.56 \end{pmatrix}. \quad (29)$$

The eigenvalue of 0.56 suggests that information flows between AZ and GSK with a half-life of about the sampling interval, of 90 seconds. However, the eigenvalue of 0.2

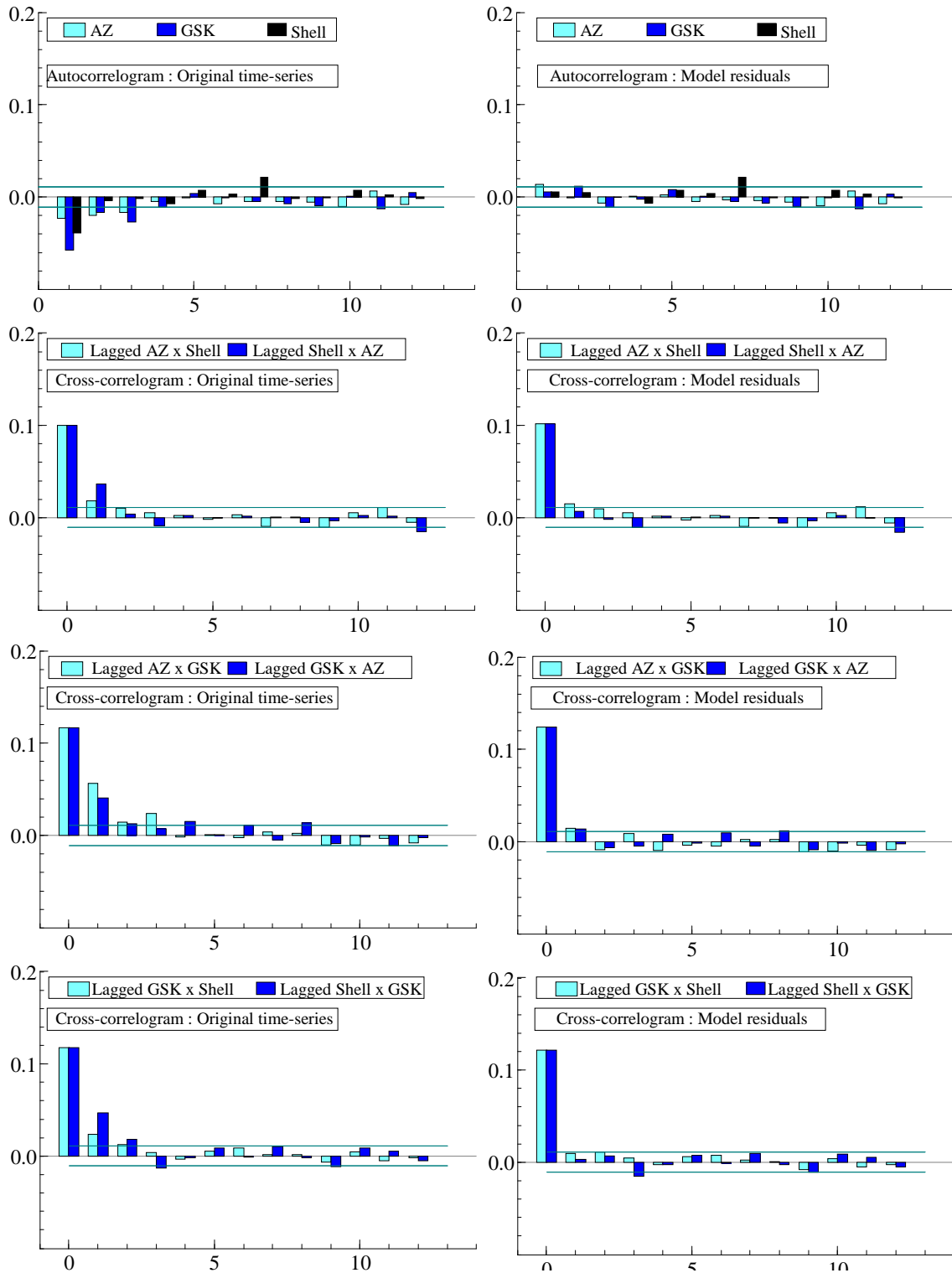


Figure 1: Left side: autocorrelograms and cross-correlograms of 90 second returns in AstraZeneca, GSK and Shell stock prices on the LSE SETS system. Right hand side: the same for the residuals from the model fitted to this data.

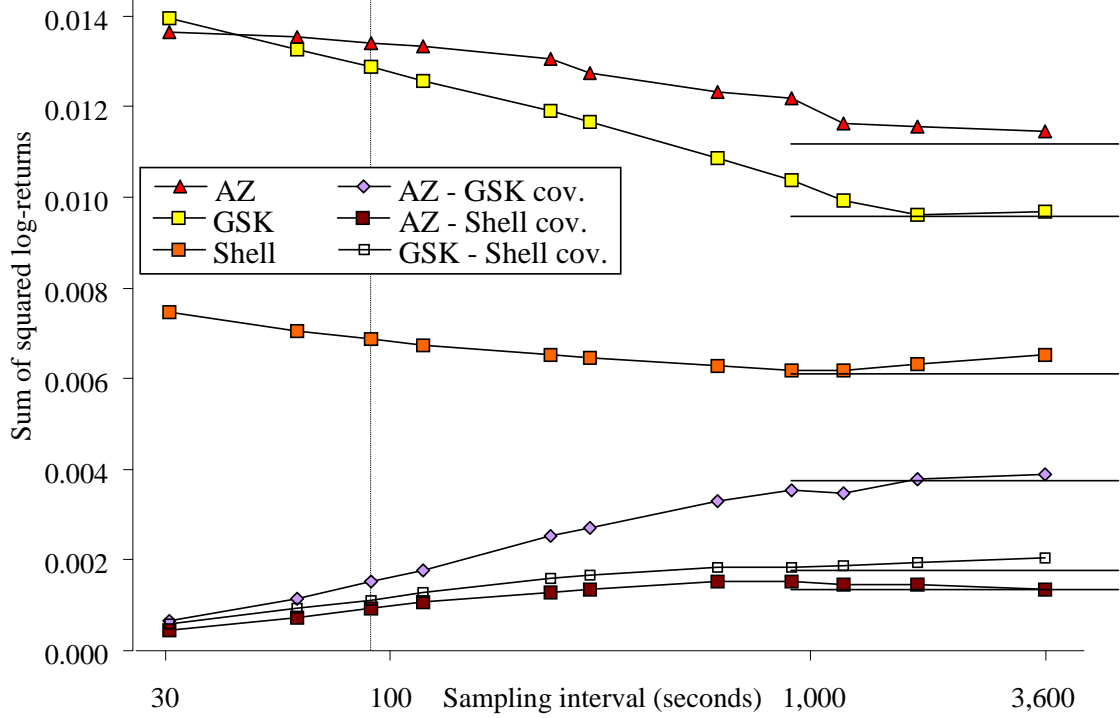


Figure 2: Empirical volatility signature plots and correlation signature plots of the time-series of AstraZeneca, GSK and Shell log mid-quotes. The heights of the horizontal bars on the right-hand side of the graph show the model estimates in (30).

indicates that information passes between these equities and Shell much more quickly. This suggests that information pertinent to Shell, as well as to AZ and GSK, is closely observed by market participants, and transfers easily between markets.

The quantity $T \left(\widehat{\beta} \widehat{\alpha}_\perp' \widehat{\Omega} \widehat{\alpha}_\perp \widehat{\beta}' + \widehat{\Sigma} \right)$ is an estimate of the expected quadratic variation matrix, $E[Y]_T$. It is (when multiplied by 100)

$$100 \times T \left(\widehat{\beta} \widehat{\alpha}_\perp' \widehat{\Omega} \widehat{\alpha}_\perp \widehat{\beta}' + \widehat{\Sigma} \right) = \begin{pmatrix} 1.118 & 0.376 & 0.134 \\ 0.376 & 0.958 & 0.178 \\ 0.134 & 0.178 & 0.611 \end{pmatrix}. \quad (30)$$

To assess the plausibility of this estimate, Figure 2 presents empirical volatility signature plots (see Andersen, Bollerslev, Diebold, and Labys 2000) and on the same chart the corresponding covariation signature plot for this time-series. It shows the realized variances of AZ, GSK and Shell at various sampling frequencies, over the observed period. Throughout, returns exceeding 6 standard deviations were set to zero, and the data was subsampled five times, at five evenly-spaced lags. The vertical dotted line is at a sampling interval of 90 seconds, the frequency of the sampled data used in the fitted

model.

The right-hand tails of the signature plots give reasonably accurate estimates of the assets' quadratic variations and covariation. It is well known that realized variance is upwards-biased at high frequency, and this is indicated here by the upwards trends in the volatility signature plots as the sampling interval converges (leftwards) to zero. Similarly, the Epps effect manifests itself in the downwards trend in the covariation signature plots as the interval converges to zero. The horizontal lines record the model estimates given in (30). These are close to the right-hand tails of the respective signature plots, lending credence to the model.

4.2 Day-by-day estimation

The previous part presented a fairly well-specified semi-parametric model that spanned 102 trading days. To understand better its performance at the daily level, this section takes the estimated model to each day separately.

Index each day of the sample by $\{\delta : 1 \leq \delta \leq 102\}$. Holding α and β fixed at the estimates in (27), fitting the model to day δ alone provides estimates of Ω and Σ for that day, so that daily correlation and betas may be estimated: denote them respectively $\widehat{Cor}_{i,j,\delta}$ and $\widehat{\beta}_{i,j,\delta}$. To help assess the quality of these estimates, I next implement a forecasting comparison between $\widehat{Cor}_{1,2,\delta}$ and a non-parametric alternative, $\widehat{Cor}_{1,2,\delta}^{NP}$. So in this part, for brevity, I only look at the estimated correlation between 1) GSK returns and 2) AZ returns.

4.2.1 Subsampled correlation

Bandi and Russell (2005) discuss using simple realized volatility measures to estimate covariation. To get a nonparametric estimator of covariation, I follow this approach, but also subsample the data, which Zhang, Mykland, and Aït-Sahalia (2005) indicates will improve accuracy.

Bandi and Russell (2005) derives an optimal choice of sampling interval when in the presence of microstructure noise, on the basis of a mean square error criterion. By contrast, I use a very naive criterion to select the sampling interval: specifically, I note

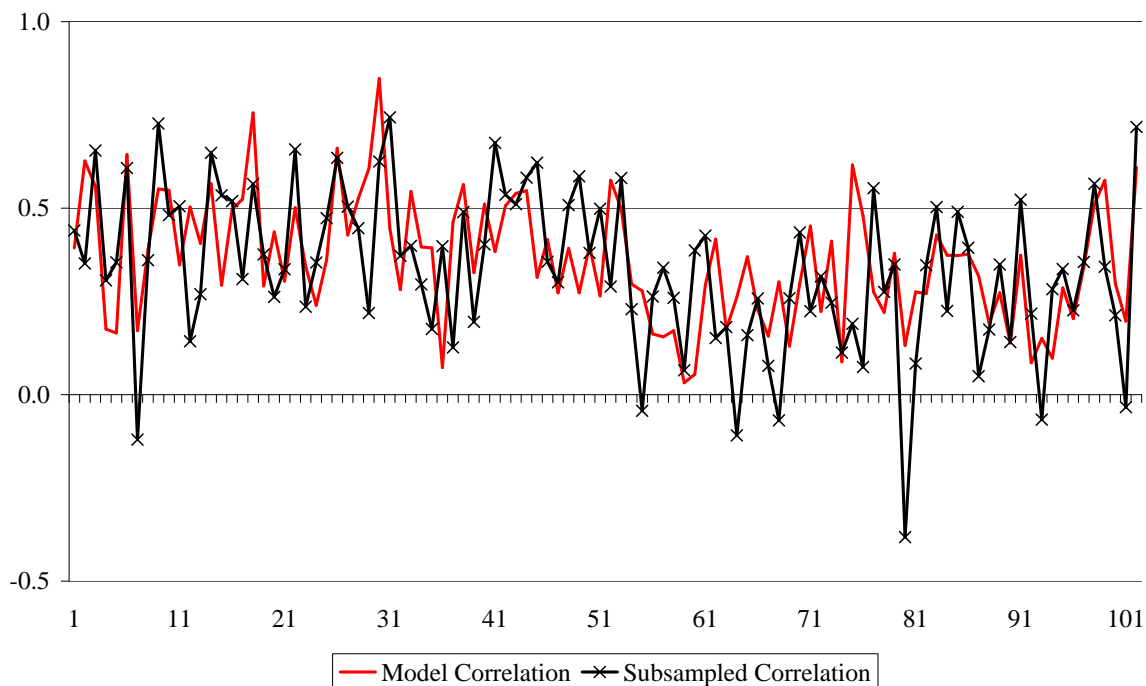


Figure 3: The daily time-series of $\widehat{Cor}_{1,2,\delta}^{NP}$ and $\widehat{Cor}_{1,2,\delta}$ for AstraZeneca and GSK equity returns over 102 trading days between October 2004 and February 2005.

that in Figure 2, squared returns over 30 minutes or more are not appreciably biased by the Epps effect, or by univariate microstructure noise.

As already mentioned, I sampled each trading day (excluding the first five minutes) so as to give 336 returns of 90 seconds each. This provides $(336 - 20) = 316$ distinct aggregated returns over 20 periods, i.e. over 30 minutes. Call these overlapping, bivariate returns $\{r_{i,\delta}^{30 \text{ min}} : i = 1 \dots 316\}$. The subsampling estimator of $E[Y^*]_T$ is given by

$$\frac{1}{20} \sum_{i=1}^{316} (r_{i,\delta}^{30 \text{ min}}) (r_{i,\delta}^{30 \text{ min}})', \quad (31)$$

so that $\widehat{Cor}_{1,2,\delta}^{NP}$ is the (1-2) correlation coefficient of (31). Figure 3 plots the time-series of $\widehat{Cor}_{1,2,\delta}^{NP}$ and $\widehat{Cor}_{1,2,\delta}$. To the eye, it appears that there was a dip in correlation in the later part of the observed 5-month period.

4.2.2 Results of the forecasting comparison

I fitted an unrestricted in-sample VAR to the bivariate daily series of $\widehat{Cor}_{1,2,\delta}^{NP}$ and $\widehat{Cor}_{1,2,\delta}$, which gives a useful assessment of forecasting effectiveness and forecastability, even in this short sample. To capture effects arising up to a week in the past, I included five lags. The results indicate that $\widehat{Cor}_{1,2,\delta}$ is better forecast than $\widehat{Cor}_{1,2,\delta}^{NP}$. While this may be due to serial dependence in $\widehat{Cor}_{1,2,\delta}^{NP}$'s measurement error, another explanation is that $\widehat{Cor}_{1,2,\delta}$ more successfully describes serial dependence in correlation over time.

Without exception, the VAR(5) passed an array of specification tests on its residuals at 5%. All lagged terms were individually insignificant at 10%. Nevertheless, the dependence of $\widehat{Cor}_{1,2,\delta}$ on the lagged regressors was significant at 1%, whereas the dependence of $\widehat{Cor}_{1,2,\delta}^{NP}$ on lagged regressors was insignificant at 10%.

I repeated the entire analysis for the equity pair, Shell and BP, which are oil and gas majors listed on the LSE. For these equities, in the VAR(5) neither $\widehat{Cor}_{1,2,\delta}$ nor $\widehat{Cor}_{1,2,\delta}^{NP}$ emerged as a more significant lagged regressor. However, the dependence of $\widehat{Cor}_{1,2,\delta}$ on the lagged regressors was significant at 10% (with a p-value of 0.059), whereas the dependence of $\widehat{Cor}_{1,2,\delta}^{NP}$ on lagged regressors was insignificant at 10%.

I likewise analyzed the pair HBOS and RBS, both UK retail banks listed on the LSE. In this case, neither $\widehat{Cor}_{1,2,\delta}$ nor $\widehat{Cor}_{1,2,\delta}^{NP}$ could be significantly explained by lagged regressors: suggesting that over the observed period there was little serial dependence in these equities' daily correlations.

4.3 A numerical technique using realized variance

Even with $n = 3$ assets, this model encounters numerical problems due to many parameters. In particular, there are $O(n^2)$ parameters, as the variance-covariance matrix Ω is not diagonal, and must be fully estimated. To reduce the size of the parameter set, a two-step procedure was adopted in estimation. This takes as its starting point the following consequence of the model for the realized variance estimator: from (16) and

(17),

$$E[\Delta Y_t \Delta Y_t'] = \Sigma + \Omega + \alpha \text{cov}[\beta' L_{t-1}] \alpha'. \quad (32)$$

The matrix $\text{cov}[\beta' L_{t-1}]$ is the unconditional covariance of an OU process of autoregressive parameter $\Psi = (\mathbf{I} + \beta' \alpha)$ and innovation variance $\beta' \Omega \beta$, which is known to be

$$\text{cov}[\beta' L_{t-1}] = \sum_{l=0}^{\infty} \Psi^l (\beta' \Omega \beta) \Psi^{l'}. \quad (33)$$

In a first step, therefore, the following moment constraint was imposed:

$$\frac{1}{T} \sum_{t=1}^T \Delta Y_t \Delta Y_t' = \widehat{\Sigma} + \widehat{\Omega} + \widehat{\alpha} \left(\sum_{l=0}^{\infty} \widehat{\Psi}^l (\widehat{\beta}' \widehat{\Omega} \widehat{\beta}) \widehat{\Psi}^{l'} \right) \widehat{\alpha}'. \quad (34)$$

The matrix $\widehat{\Omega}$ is thereby determined as an implicit function of $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\Sigma}$, and the data. In a second step, always respecting this constraint, $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\Sigma}$ were varied to maximize the pseudo-likelihood. These can be used as very accurate starting values for a final, unrestricted, maximization.

5 Information share of a common factor

In important work by Hasbrouck (1995) a methodology is outlined for determining the contributions to price discovery, called *information shares*, of several markets trading the same security. The realized price processes on these markets are viewed as cointegrated, with the underlying price of the security identified as being the single common trend. A similar approach is adopted in the case of the best bid and best ask quotes on a single market in Engle and Patton (2004) and Hansen and Lunde (2006).

The current model gives a means to extend the methodology in Hasbrouck (1995) to the case where multiple assets share a common factor, whose price is discovered through trade in those assets. In the current notation, the common factor is F . Being a local martingale, innovations to F represent permanent updates to its price. Such innovations are given by

$$\Delta F_t = \alpha_{\perp}' \eta_t^L, \quad (35)$$

so that

$$\text{var}[\Delta F_t] = \alpha_{\perp}' \Omega \alpha_{\perp}. \quad (36)$$

Hasbrouck (1995) proposes that the question, which market contributes most to the innovation in F , be settled by decomposing $var[\Delta F_t]$ into a contribution from each market. Expanding (36), we have

$$var[\Delta F_t] = \begin{pmatrix} \alpha_{AZ}^2 \\ \alpha_{GSK}^2 \\ \alpha_{Shell}^2 \end{pmatrix}' \begin{pmatrix} var \left[\eta_t^{L,AZ} \right] \\ var \left[\eta_t^{L,GSK} \right] \\ var \left[\eta_t^{L,Shell} \right] \end{pmatrix} + \text{covariance terms}. \quad (37)$$

In the case where Ω is diagonal, there are no covariance terms, so that (37) is the sum of three contributions, one from each market. However in the current application to AZ, GSK and Shell, we find that $\widehat{\Omega}$ is not diagonal, so that (at a sampling frequency of 90 seconds) a low level of contemporaneous correlation or price discovery is found across the three markets. In these circumstances, it is at best possible to place bounds on the information shares of the various markets.

The estimated form of (37) is

$$var[\Delta F_t] = 0.133 = \begin{pmatrix} 0.181 \\ 0.113 \\ 0.706 \end{pmatrix}' \begin{pmatrix} 0.128 \\ 0.207 \\ 0.034 \end{pmatrix} + \text{covariance terms}, \quad (38)$$

so that the contributions of each equity are roughly equal to one another, coming to 0.023. These place roughly equal lower bounds on the information shares of each of the three markets, of $0.023/0.133 = 17\%$.

While information shares may be equal, their sources differ markedly. Shell has a much lower innovation variance, $var \left[\eta_t^{L,Shell} \right]$, but this is offset by a higher contribution to price discovery in α_{\cdot} , which arises, broadly, because Shell's price process is 'nearer' to being a local martingale.

6 Conclusion

This paper presents a model that captures the salient features of the intriguing Epps effect: namely, that empirical covariances among asset returns die away to zero at the highest frequencies, although they are significantly non-zero at the moderate intraday and interday frequencies that are most relevant to asset pricing theories. The model

represents prices as being loaded onto a common set of martingale pricing factors, which deliver the economically significant covariation. As usual, each price also has an idiosyncratic component, meaning that prices nevertheless diverge from one another arbitrarily over time.

Proposition 2.1 produces the Epps effect by adding to this common factor representation an error term which offsets covariation at high frequencies. As such, the error is the vehicle for lagged dependence among observed asset price returns. Consequently its dependence on underlying prices, and its serial dependence, are essential features. This contrasts to a literature on univariate realized variance estimation, where much is gained from a simpler assumption that microstructure noise is independent of the underlying martingale, and serially uncorrelated.

Nevertheless, where such estimators have been adapted so as to handle dependent noise, this paper provides a foundation for their use in the multivariate context: for it derives a ‘Brownian semi-martingale plus noise’ representation of a vector of prices observed at high-frequency.

To assess the empirical relevance of this representation, it is specialized to the point where it offers estimates of correlations and betas. The estimator of correlation is implemented at a daily frequency on pairs of London Stock Exchange equities, and is used to investigate Hasbrouck (1995) information shares in a multi-asset setting with a common factor. The estimator is found 1) to perform reasonably well in a forecasting setting, in comparison to a nonparametric alternative, and 2) to arise from a well-specified time-series model. Therefore the proposed model is in line with reality, to the extent that it fits a range of moments in the data, including ones that are most pertinent to the correlation in asset returns.

References

- Aït-Sahalia, Y. (2002). Telling from discrete data whether the underlying continuous-time model is a diffusion. *Journal of Finance* 57, 2075–2112.
- Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005). How often to sample a

- continuous-time process in the presence of market microstructure noise. *Review of Financial Studies* 18, 351–416.
- Andersen, T. G., T. Bollerslev, and F. Diebold (2006). Roughing it up: including jump components in the measurement, modeling and forecasting of return volatility. Forthcoming, *Review of Economics and Statistics*.
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2000). Great realizations. *Risk* 13, 105–108.
- Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2003). Modeling and forecasting realized volatility. *Econometrica* 71, 579–626.
- Andersen, T. G., T. Bollerslev, and N. Meddahi (2004). Analytic evaluation of volatility forecasts. *International Economic Review* 45, 1079–1110.
- Bandi, F. M. and J. R. Russell (2005). Realized covariation, realized beta and microstructure noise. Unpublished paper, Graduate School of Business, University of Chicago.
- Bandi, F. M. and J. R. Russell (2006). Separating market microstructure noise from volatility. *Journal of Financial Economics* 79, 655–692.
- Barndorff-Nielsen, O., P. Hansen, A. Lunde, and N. Shephard (2006). Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. Unpublished Paper: Nuffield College, Oxford.
- Barndorff-Nielsen, O. and N. Shephard (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society B* 64, 253–280.
- Barndorff-Nielsen, O. and N. Shephard (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica* 72, 885–925.
- Barndorff-Nielsen, O. and N. Shephard (2006a). Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics* 4, 1–30.

- Barndorff-Nielsen, O. and N. Shephard (2006b). Impact of jumps on returns and realised volatility: econometric analysis of time-deformed Lévy processes. *Journal of Econometrics* 131, 217–252.
- Barndorff-Nielsen, O. and N. Shephard (2007). Variation, jumps, market frictions and high frequency data in financial econometrics. Forthcoming in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress* (edited by Richard Blundell, Persson Torsten and Whitney K Newey).
- Comte, F. (1998). Discrete and continuous time cointegration - le cas multidimensionnel. *Journal of Econometrics* 88, 207–226.
- Corradi, V. (1997). Comovements between diffusion processes: characterization, estimation and testing. *Econometric Theory* 13, 646–666.
- Doornik, J. (2001). *An Object-Oriented Matrix Programming Language*. London: Timberlake Consultants Ltd.
- Engle, R. F. and C. W. J. Granger (1987). Cointegration and error correction: Representations, estimation and testing. *Econometrica* 55, 251–276.
- Engle, R. F. and A. J. Patton (2004). Impacts of trades in an error-correction model of quote prices. *Journal of Financial Markets* 7, 1–25.
- Epps, T. W. (1979). Comovements in stock prices in the very short run. *Journal of the American Statistical Association* 74, 291–296.
- Hansen, P. and A. Lunde (2006). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics* 24, 127–281. The 2005 Invited Address with Comments and Rejoinder.
- Hasbrouck, J. (1995). One security, many markets: determining the contributions to price discovery. *Journal of Finance* 50, 1175–1198.
- Hayashi, T. and N. Yoshida (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* 11, 359 – 379.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–254.

- Koopman, S. J., N. Shephard, and J. A. Doornik (1998). Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal* 1, 1–55.
- Kuersteiner, G. (2001). Optimal instrumental variables estimation for ARMA models. *Journal of Econometrics* 104, 359–405.
- Large, J. (2005). Estimating quadratic variation when quoted prices jump by a constant increment. Nuffield College Economics Group working paper W05.
- Lunde, A. and V. Voev (2006). Integrated covariance estimation using high-frequency data in the presence of noise. Unpublished paper: Department of Marketing and Statistics, Aarhus School of Business.
- Malliavin, P. and M. E. Mancino (2002). Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics* 6, 49–61.
- Mancino, M. and R. Renó (2005). Dynamic principal component analysis of multivariate volatilities via Fourier analysis. *Applied Mathematical Finance* 12, 187–199.
- Martens, M. (2003). Estimating unbiased and precise realized covariances. Unpublished paper.
- Meddahi, N. (2002). A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics* 17, 479–508.
- Newey, W. K. and K. D. West (1987). A simple, positive semi-definite heteroskedasticity and autocorrelation consistent variance covariance matrix. *Econometrica* 55, 703–708.
- Phillips, P. (1991). Error correction and long-run equilibrium in continuous time. *Econometrica* 59, 967–980.
- Precup, O. V. and G. Iori (2005). Cross-correlation measures in high-frequency domain. Unpublished paper: Department of Economics, City University.
- Renó, R. (2003). A closer look at the Epps effect. *International Journal of Theoretical and Applied Finance* 6, 87–102.

- Scholes, M. and J. Williams (1977). Estimating betas from nonsynchronous trading. *Journal of Financial Economics* 5, 309–327.
- Zhang, L. (2006). Estimating Covariation: Epps Effect, Microstructure Noise. Unpublished paper: Department of Finance, University of Illinois at Chicago.
- Zhang, L., P. Mykland, and Y. Aït-Sahalia (2005). A tale of two timescales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association* 100, 1394–1411.