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

DRAFT - COMMENTS WELCOME

Jeremy Large*
JEREMY.LARGE@ECONOMICS.OX.AC.UK

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

25 June 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 a workshop organized by Nour Meddahi in Imperial College, London, on 22 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 \textit{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 \textit{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.

Like in the univariate case, market microstructure effects create difficulties when sampling high-frequency returns, as realized variance techniques prescribe. Unlike in the univariate case, two additional complicating factors arise: 1) nonsynchronous trading across markets, which causes fresh observations of transactions prices not to arise simultaneously across markets, but to be separated by e.g. a few seconds – see Scholes and Williams (1977), Martens (2003), Hayashi and Yoshida (2005), Lunde and Voev (2006) and Zhang (2006); and 2) a lead-lag effect – the focus of this paper. Each of these two factors has been found empirically to contribute individually to the puzzling effect first studied comprehensively in Epps (1979), whereby the empirical correlations in related assets’ returns are biased towards zero at high sampling frequencies.

This paper presents a parsimonious stochastic calculus which captures the lead-lag effect and the Epps effect, even without taking account of nonsynchronous trading. This model is a generalization of the familiar common factor framework of multiple asset returns, where the generalization essentially involves adding a dependent error. To assess the model’s empirical relevance, the paper develops it to the point of deriving 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, comparing well against a non-parametric alternative. So the calculus seems to be informative about the data. Tangentially, we
also have an opportunity to investigate an extension of Hasbrouck (1995) information shares to the case of multiple assets with a single latent common pricing factor.

The lead-lag effect is the phenomenon whereby information relevant to two or more asset prices is learned and acted upon sooner by traders on one market, than by traders at others. Prices on the first market therefore move in advance of others. This induces lagged cross-correlation into high frequency asset returns, at the expense of instantaneous cross-correlation: resulting, potentially, in zero realized covariations – and, consequently, in zero realized correlations and betas – see Reno (2003) and Sheppard (2005). This can help explain Epps (1979)’s finding among US automobile stocks, that empirical covariances virtually disappear at high frequencies of a few minutes, while being far from zero at moderate intraday frequencies. Subsequently, this finding has been replicated widely in equity markets.

More formally, let \( Y \) be a vector of \( n \) asset log-prices, which is a Brownian semi-martingale. The quadratic variation process of \( Y \), which is often denoted \([Y]_t\), is defined as follows:

\[
[Y]_t = p - \lim_{m \to \infty} \sum_{i=0}^{m-1} (Y_{s_{i+1}} - Y_{s_i})(Y_{s_{i+1}} - Y_{s_i})',
\]

where \( 0 = s_0 < s_1 < s_2 < \ldots < s_m = t \) determines a grid. In the probability limit, \( m \to \infty \) 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. Thus in this framework, Epps’ puzzling finding is that \([Y]_t\) is (almost) diagonal despite considerable comovement in elements of \( Y \).

To capture this, I use ideas developed in cointegration theory, see important work by for example Engle and Granger (1987) and Johansen (1988), and applied in continuous time in Phillips (1991) and Corradi (1997). Let

\[
dY_t = dL_t + \sigma_t dW_t,
\]

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

$$L \perp (\sigma, W).$$

(3)

All dependence in $Y$ is due to $L$, which evolves according to

$$dL_t = \alpha\beta' L_t dt + \omega_t dZ_t,$$

(4)

where $\omega$ is another volatility process; $\alpha$ and $\beta$ are full-rank $n \times (n - r)$ matrices with $r \in \mathbb{N}$; and $Z$ is another multivariate standard Brownian motion. Later, (4) will be generalized further and technical assumptions will be made, but otherwise this largely completes the specification of the model.

The model has the property that

$$E[dY_t|L_t] = \alpha\beta' L_t dt,$$

(5)

so that the elements of $Y_t$ move together whenever $Y_t$’s unobserved component, $L_t$, diverges from the relationship $\beta' L_t = 0$: giving correlated returns at moderate frequencies, for example hourly or less frequent still.

On the other hand, substituting (4) into (2) and applying Itô algebra,

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

(6)

Barndorff-Nielsen and Shephard (2002) and Meddahi (2002) show that $[Y]_t$ is the integral of (6). Therefore in any specification where not only $\sigma$, but also $\omega$, is diagonal, the quadratic variation process $[Y]_t$ is diagonal as well. 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, $o(dt^2)$. More moderate Epps effects can be captured by allowing $\omega$ to be non-diagonal.

The Introduction concludes with a brief survey of related literature. Bandi and Russell (2005) discusses the optimal sampling frequency of data when estimating covariation using a framework where microstructure effects are not explicit. A set of papers has emphasized the effects of nonsynchronous trading on realized covariation. Early treatments of this topic are in Scholes and Williams (1977) and Martens (2003).
latter mentions an estimator of realized covariation (‘all-overlapping-returns’, as suggested by Fulvio Corsi) that is robust to nonsynchronous trading, which is analyzed and developed to a far greater extent in Hayashi and Yoshida (2005). This estimator is assessed in the presence of both nonsynchronous trading and univariate ‘contamination’, a measurement error, in Lunde and Voev (2006). Zhang (2006) also studies this setting. In a parallel stream of this literature, Fourier analysis-based methods in Malliavin and Mancino (2002) are implemented in Mancino and Reno (2005), in Precup and Iori (2005) and in Reno (2003). Reno (2003) argues that nonsynchronous trading alone is inadequate to describe the data: the lead-lag effect is also needed. Sheppard (2005) proposes a theory of ‘scrambling’ to describe this.

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 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 (4) has a natural interpretation in terms of delays in information transmission across markets. In this Section these theoretical motivations are explored, before the question of empirical relevance is returned to in Section 3.

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 $\sigma$ and $\omega$ 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 (4), let

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

(7)

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 (4).

Write $\alpha_\perp$ and $\beta_\perp$ for orthogonal complements of $\alpha$ and $\beta$ respectively. So $\alpha_\perp' \alpha_\perp = 0$ etc.

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

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

(8)

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

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

(9)

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

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

The common factor representation may be interpreted as follows. Note that from (7),

$$dF_t := d(\alpha'_L L_t) = \alpha'_L \omega_t dZ_t,$$

(10)

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 containing cumulated idiosyncratic, shocks to prices, $\int_0^t \sigma_u dW_u$. From (3), 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 stationary error term, \( \epsilon_t \), appears in the common factor representation of Proposition 2.1. This describes lagged interactions between markets. When it deviates from zero, this may indicate that information has not been fully impounded into prices: they have been shocked, but have not yet reverted to their long-run relationship as defined by \( \beta' L_t = 0 \) or, equivalently, by \( \epsilon_t = 0 \).

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

\[
Y_t = Y_t^* + \epsilon_t,
\]

which defines \( Y^* \). If \( \omega_t \) is always diagonal, then the error, \( \epsilon \), is dependent on \( Y^* \) in such a way as to ‘cancel-out’ the covariation between elements of \( Y^* \). This follows from (6). It means that realized covariations are zero, consistently with an extreme Epps effect.

To have this consequence, \( \epsilon \) must naturally be closely related to \( F \). However, note that unless \( \beta \) and \( \alpha \) are collinear, \( \epsilon \) is not progressively measurable with respect to the common factors, \( F \), and the idiosyncratic effects.

2.3 Stationary error

Typically, simple parameter restrictions ensure that \( \beta' L \), and so \( \epsilon \), 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 (7),

\[
d(\beta' L)_t = (\beta' \alpha)(\beta' L)_t + \beta' \omega_t dZ_t.
\]

This specifies \( \beta' L \) as an Ohrnstein-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 (11). 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 underlying or efficient price containing the fundamental news that passes into realized prices with varying delays. Now,

$$dY^*_t = \tilde{\beta} \alpha'_t \omega_t dZ_t + \sigma_t dW_t,$$

so the quadratic variation process of $Y^*$ is

$$[Y^*]_t = \int_0^t \tilde{\beta} \alpha'_u \omega'_u \alpha_j' \tilde{\beta} + \sigma_u \sigma'_u du.$$  (14)

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

$$\tilde{\text{Cor}}_{i,j} = \frac{[Y^*]_{T,i,j}}{\sqrt{[Y^*]_{T,i,i}[Y^*]_{T,j,j}}};$$  (15)

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

$$\tilde{\beta}_{i,j} = \frac{[Y^*]_{T,i,j}}{[Y^*]_{T,j,j}}.$$  (16)

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 $\sigma_t$ and $\omega_t$ are stationary and bounded over the course of a trading day, $[0, T]$, the underlying moments of $\tilde{\text{Cor}}_{i,j}$ and
\( \tilde{\beta}_{i,j} \) exist, and are given by
\[
\text{Cor}_{i,j} = \frac{E[Y^*|T,i,j]}{\sqrt{E[Y^*|T,i,i]E[Y^*|T,j,j]}}, \quad \text{and} \quad \beta_{i,j} = \frac{E[Y^*|T,i,j]}{E[Y^*|T,j,j]}.
\]

(17)

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 \( t = 0, 1, 2, ..., T \).

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
\[
\eta^Y_t = \Delta Y_t - \Delta L_t,
\]
where \( \Delta Y_t = Y_t - Y_{t-1} \). So \( \eta^Y_t \) is an \( n \)-variate martingale difference sequence. Setting \( \tau(t) = [t] \) implies that at integer times, \( t \), \( L_t \) has the error-correction form:
\[
\Delta L_t = \alpha \beta' L_{t-1} + \eta^L_t,
\]
where \( \eta^L_t \) is another martingale difference sequence. This follows on aggregating (7) between sampling times.

This specification has residuals, \( \eta^Y_t \) and \( \eta^L_t \), with conditional heteroskedasticity of unknown law, and whose unconditional variances exist (the usual restrictions on fourth-order cumulants follow as \( \sigma \) and \( \omega \) are bounded). Note that from (3), \( \eta^Y \perp \perp \eta^L \): that is, the two time-series \( \eta^Y \) and \( \eta^L \) are independent. Let \( E[\eta^Y_t \eta^Y_{t'}] = \Sigma \), and let \( E[\eta^L_t \eta^L_{t'}] = \Omega \). So,
\[
E \left[ \int_0^T \omega_u \omega'_u du \right] = T\Omega, \quad \text{and} \quad E \left[ \int_0^T \sigma_u \sigma'_u du \right] = T\Sigma,
\]
giving from (14),

**Proposition 3.1** In this discrete-time framework,
\[
E[Y^*|T] = T \left( \tilde{\beta}' \alpha' \Omega \alpha \tilde{\beta} + \Sigma \right).
\]

(21)
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 $(\hat{\beta}'\tilde{\Omega}\hat{\alpha}'\tilde{\beta} + \tilde{\Sigma})$ (with the factor of $T$ cancelling out in both cases), see (17).

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

**Lemma 3.2** Suppose $\tau(t) = \lfloor t \rfloor$ and that $\Psi$, i.e. $(I + \beta'\alpha)$, 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 (19), when $\tau(t) = \lfloor t \rfloor$, we have that at integer times $t \in \mathbb{N}$,

$$\beta' L_t = \Psi(\beta' L)_{t-1} + \beta' \eta^L_t.$$  \hspace{1cm} (22)

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 $\Psi$ 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 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.\(^1\)

\(^1\)The problem is only exacerbated here, when the object of interest is $[Y^*]_t$, since $Y^*$ is unobserved.
Nevertheless, given the assumption of stationary volatility, the underlying moments, \( Cor_{i,j} \) and \( \beta_{i,j} \) in (17), are identifiable in a large-\( T \) framework, and this approach is taken here. Note that (18) and (19) 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).\(^2\) Maximization can be done using the MaxBFGS algorithm in Ox with numerical derivatives (see Doornik 2001).

Let \( \theta \) 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 \( \theta \) is given by the usual limit theory when \( T \to \infty \),

\[
\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, J^{-1}IJ^{-1}),
\]

where

\[
I = \lim_{T \to \infty} \frac{1}{T} \text{cov}[s(\theta)],
\]

and

\[
J = \lim_{T \to \infty} \frac{1}{T} \text{E} \left[ \frac{\partial s}{\partial \theta} \right],
\]

both evaluated at the truth. Working from (24), 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).

\(^2\)Because of the moving average component to \( \Delta L_t \), the state variables are both \( \Delta L_t \) and \( \eta_t^L \). Write \( 0_n \) for an \( (n \times n) \) matrix of zeros, and \( I_n \) for the \( (n \times n) \) identity matrix. Write \( \Phi = I_n + \alpha^t\beta \). Then the state space representation is

\[
\begin{pmatrix}
\Delta L_{t+1} \\
\eta_{t+1}^L \\
\Delta Y_t
\end{pmatrix} = \begin{pmatrix}
\Phi & -I_n \\
0_n & 0_n \\
I_n & 0_n
\end{pmatrix} \begin{pmatrix}
\Delta L_t \\
\eta_t^L \\
\eta_t^Y
\end{pmatrix} + \begin{pmatrix}
\eta_{t+1}^L \\
\eta_{t+1}^Y \\
\eta_t^Y
\end{pmatrix}.
\]
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 $\alpha$ and $\beta$, 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 $\epsilon$, 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, $\alpha$ and $\beta$ are $(3 \times 2)$. I now show that this specification fits the data well.

Recall that $\sigma$ 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 $\Sigma$ 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 $\beta$ so that

$$\beta = \begin{pmatrix} 1 & 1 \\ 1 & \dagger \\ \times & 0 \end{pmatrix},$$

(27)

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
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.
noise is in fact accounted for by the current model.

Recall that (19) reads

\[ \Delta L_t = \alpha' L_{t-1} + \eta_{t}^L. \]  

(28)

Once estimated, (19) takes the form

\[ \Delta L_t = \begin{pmatrix} \Delta L_{AZ}^t \\ \Delta L_{GSK}^t \\ \Delta L_{Shell}^t \end{pmatrix} = \begin{pmatrix} -0.11 & -0.23 \\ -0.14 & 0.27 \\ 0.11 & 0.01 \end{pmatrix} \begin{pmatrix} 1 & 1 & -4.91 \\ 1 & -0.75 & 0 \end{pmatrix} \begin{pmatrix} L_{t-1} \\ \eta_{t}^L \end{pmatrix}. \]  

(29)

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

\[ Y_t := \begin{pmatrix} Y_{AZ}^t \\ Y_{GSK}^t \\ Y_{Shell}^t \end{pmatrix} = \begin{pmatrix} 0.789 \\ 1.045 \\ 0.373 \end{pmatrix} F_t + \int_0^t \sigma_u dW_u + \epsilon_t. \]  

(30)

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 \( \hat{\Psi} = (I + \hat{\beta}' \hat{\alpha}) \) is the estimated autoregressive parameter in (22), and therefore also the estimated autoregressive parameter of \( \epsilon \). This is informative about the rate of information transmission between markets. It is almost diagonal:

\[ I + \hat{\beta}' \hat{\alpha} = \begin{pmatrix} 0.20 & -0.001 \\ -0.001 & 0.56 \end{pmatrix}. \]  

(31)

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

\[ 100 \times T \left( \hat{\beta}' \hat{\alpha} + \hat{\Sigma} \right) = \begin{pmatrix} 1.118 & 0.376 & 0.134 \\ 0.376 & 0.958 & 0.178 \\ 0.134 & 0.178 & 0.611 \end{pmatrix}. \]  

(32)
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 (29). These are close to the right-hand tails of the respective signature plots, lending credence to the model.

4.2 A digression: information shares

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 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

$$\Delta F_t = \alpha_\perp \eta_t^L,$$

so that

$$\text{var} [\Delta F_t] = \alpha_\perp' \Omega \alpha_\perp.$$  \hspace{1cm} (34)

Hasbrouck (1995) proposes that the question, which market contributes most to the innovation in $F$, be settled by decomposing $\text{var} [\Delta F_t]$ into a contribution from each market. Expanding (34), we have

$$\text{var} [\Delta F_t] = \begin{pmatrix} \frac{1}{2} \alpha_{AZ}^2 & \frac{1}{2} \alpha_{GSK}^2 & \frac{1}{2} \alpha_{Shell}^2 \end{pmatrix}' \begin{pmatrix} \text{var} \left[ \eta_t^{L,AZ} \right] \\ \text{var} \left[ \eta_t^{L,GSK} \right] \\ \text{var} \left[ \eta_t^{L,Shell} \right] \end{pmatrix} + \text{covariance terms}. \hspace{1cm} (35)$$

In the case where $\Omega$ is diagonal, there are no covariance terms, so that (35) is the sum of three contributions, one from each market. However in the current application to AZ, GSK and Shell, we find that $\hat{\Omega}$ 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 (35) is

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

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, \(\text{var}[\eta_t^{L,\text{Shell}}]\), but this is offset by a higher contribution to price discovery in \(\alpha_t\), which arises, broadly, because Shell’s price process is ‘nearer’ to being a local martingale.

### 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 \(\Omega\) 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 (18) and (19),

\[
E[\Delta Y_t \Delta Y_t'] = \Sigma + \Omega + \alpha \text{cov}[^t\beta L_{t-1}]\alpha'.
\]

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

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

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

\[
\frac{1}{T} \sum_{t=1}^{T} \Delta Y_t \Delta Y_t' = \hat{\Sigma} + \hat{\Omega} + \hat{\alpha} \left( \sum_{l=0}^{\infty} \hat{\Psi}^l (\hat{\beta}'\hat{\Omega}\hat{\beta}) \hat{\Psi}^l \right) \hat{\alpha}'.
\]
The matrix $\hat{\Omega}$ is thereby determined as an implicit function of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\Sigma}$, and the data. In a second step, always respecting this constraint, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\Sigma}$ were varied to maximize the pseudo-likelihood.

### 4.4 Day-by-day estimation

The previous parts have 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 $\alpha$ and $\beta$ fixed at the estimates in (29), fitting the model to day $\delta$ alone provides estimates of $\Omega$ and $\Sigma$ for that day, so that daily correlation and betas may be estimated: denote them respectively $\hat{\text{Cor}}_{i,j,\delta}$ and $\hat{\beta}_{i,j,\delta}$. To help assess the quality of these estimates, I next implement a forecasting comparison between $\hat{\text{Cor}}_{1,2,\delta}$ and a non-parametric alternative, $\hat{\text{Cor}}_{1,2,\delta}^{NP}$. So in this part, for brevity, I only look at the estimated correlation between 1) GSK returns and 2) AZ returns.

### 4.5 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 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
Figure 3: The daily time-series of $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ and $\hat{\text{Cor}}_{1,2,\delta}$ for AstraZeneca and GSK equity returns over 102 trading days between October 2004 and February 2005.

returns $\{r_{i,\delta}^{30 \text{ min}} : i = 1...316\}$. The subsampling estimator of $E[Y^\ast]_T$ is given by

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

(40)

so that $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ is the (1-2) correlation coefficient of (40). Figure 3 plots the time-series of $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ and $\hat{\text{Cor}}_{1,2,\delta}$. To the eye, it appears that there was a dip in correlation in the later part of the observed 5-month period.

4.6 Results of the forecasting comparison

I fitted an unrestricted in-sample VAR to the bivariate daily series of $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ and $\hat{\text{Cor}}_{1,2,\delta}$, 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 $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ is better forecast than $\hat{\text{Cor}}_{1,2,\delta}$.

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 depen-
dence of $\hat{\text{Cor}}_{1,2,\delta}$ on the lagged regressors was significant at 1%, whereas the dependence of $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ 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 $\hat{\text{Cor}}_{1,2,\delta}$ nor $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ emerged as a more significant lagged regressor. However, the dependence of $\hat{\text{Cor}}_{1,2,\delta}$ on the lagged regressors was significant at 10% (with a p-value of 0.059), whereas the dependence of $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ 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 $\hat{\text{Cor}}_{1,2,\delta}$ nor $\hat{\text{Cor}}_{1,2,\delta}^{NP}$ could be significantly explained by lagged regressors: suggesting that over the observed period there was little serial dependence in these equities’ daily correlations.

5 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 the 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.

To assess the empirical relevance of this calculus, 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


