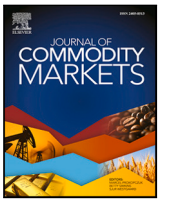
Journal of Commodity Markets 30 (2023) 100314

Contents lists available at ScienceDirect

Journal of Commodity Markets

journal homepage: www.elsevier.com/locate/jcomm



Regular article

Microstructure and high-frequency price discovery in the soybean complex

Xinquan Zhou [a,*], Guillaume Bagnarosa [b,c], Alexandre Gohin [c], Joost M.E. Pennings [d,e,g], Philippe Debie [d,f]

[a] Dublin City University, Ireland
[b] Rennes School of Business, France
[c] INRAe, France
[d] Wageningen University, Netherlands
[e] Maastricht University, Netherlands
[f] Wageningen Economic Research, Netherlands
[g] University of Illinois Urbana-Champaign, United States

ARTICLE INFO

Keywords:
Soybean
Futures market microstructure
Liquidity
Price discovery
High-frequency

ABSTRACT

We develop a theoretical framework and propose a relevant empirical analysis of the soybean-complex prices' cointegration relationships in a high-frequency setting. We allow for heterogeneous expectations among traders on the multi-asset price dynamics and characterize the resulting market behaviour. We demonstrate that the asset prices' autoregressive matrix rank and the speed of reversion towards the long-term equilibrium are related to the market realized and potential liquidity, unlike the cointegrating vector. Our empirical application to the soybean complex, where we control for volatility, supports our theoretical results when the price idleness of the different assets is properly accounted for. Our analysis further suggests that the presence of cointegration among assets is related to the time of day and the contract maturities traded at a given time.

1. Introduction

Financial markets offer the opportunity for a wide variety of economic agents to express their economic expectations. The resulting price-discovery process reflects the agents' respective levels of information and investment capacities. With the advent of electronic financial markets and automated trading, the development of index investing accelerated over the last two decades for different financial markets, including commodity markets. This new financial environment is often referred to as the financialization of commodity markets and raises questions about the influence of index investing on the real economy and the commodities' price-discovery process (Brogaard et al., 2018; Brown et al., 2020; Bond and García, 2021; Goldstein and Yang, 2022). Our paper contributes to this literature by developing a multivariate micro-economic equilibrium model for cointegrated assets with heterogeneous agent expectations. Furthermore, relying on our theoretical model, we study the potential effects of index investing on the soybean complex in a high-frequency setting. The latter is particularly interesting for our equilibrium study for two reasons: First, the derivatives exchanges offer futures contracts both on soybean and on its processed products, soyoil and soymeal, but their liquidity and market participants are different. Beyond the traditional hedgers and speculators, the well-known GSCI index,

* Correspondence to: Dublin City University Business School, Ballymun Road, Glasnevin, Dublin 9, Ireland.
E-mail addresses: xinquan.zhou5@mail.dcu.ie, xinquan.zhou@hotmail.com (X. Zhou).

https://doi.org/10.1016/j.jcomm.2023.100314
Received 17 May 2022; Received in revised form 29 November 2022; Accepted 18 January 2023
Available online 27 January 2023

for instance, only invests in soybean futures contracts[1] but does not include related soybean meal and soybean oil contracts.[2] Such behaviour could lead to short-term asynchronicity among the cointegrated markets and market inconsistencies, such as non-synchronous financial bubbles or even the oft-decried financialization of commodities markets (Basak and Pavlova, 2016; Shang et al., 2018). The second interesting aspect of the soybean complex is the cointegration among its three components. This has been demonstrated at daily frequency in several studies (Barrett and Kolb, 1995; Simon, 1999; Mitchell, 2010; Marowka et al., 2020). In this context, our paper is structured around two questions: First, what are the identification conditions under which the activities at high frequency by different agents ultimately coordinate futures prices and make their cointegration materialize over a longer time horizon, such as a trading day? Second, What are the driving forces behind cointegration if it exists?

We answer these questions by applying a novel volume-adapted price-cointegration framework to the soybean complex as we examine how the short-term market microstructure influences the price-discovery processes of multiple related assets. We indeed contribute to the price-discovery-process literature by extending the cointegration multivariate long-term price equilibrium model with a short-run microeconomic equilibrium framework which allows different groups of agents to invest independently – or not invest at all – in individual cointegrated markets. Furthermore, we contribute to the literature studying the price–volume relationship as our theoretical microstructure model establishes the link between information heterogeneity, the assets' traded volumes, and the strength of the cointegration relationships at high-frequency levels for multiple related assets. While existing works in this field are using both univariate or multivariate model settings (Epps and Epps (1976), Tauchen and Pitts (1983), He and Velu (2014), Duchin and Levy (2010), Darolles et al. (2017), Atmaz and Basak (2018)) when investigating the relationship between price and volume, none of them has considered heterogeneous beliefs in a multivariate cointegrated equilibrium setting. Studying the price and volume relationship through the lens of such a market-price equilibrium permits to clearly specify the impact of partially informed traders' decisions and the central role of manufacturers in the commodity price-discovery process and the related market efficiency. Subsequently, using both the high-frequency traded prices and aggregated quantity data (i.e., daily traded volumes and limit order books' daily average liquidity measurements) for the soybean complex, our empirical study confirms the theoretical model we proposed by revealing a relationship between price cointegration and the traded or available volumes in each individual asset's LOB.

Observing this cointegration relationship in a high-frequency setting turns out to be challenging, given that microstructure noise[3] as well as lagged information among agents and markets disturb the latent joint price-discovery process (Janzen and Adjemian, 2017; Couleau et al., 2019). To deal with microstructure noise and non-synchronicity among markets, many statistical models have been considered in high-frequency price-dynamic modelling. The state–space representation of the vector error correction model (VECM) described in Seong et al. (2013) considers the Expectation Maximization algorithm proposed by Dempster et al. (1977) to cope with mixed-frequency or asynchronous data in cointegrated time-series models. More recently, Buccheri et al. (2021b) demonstrated that this filtering methodology adequately deals with microstructure noise and the information lag that exists among markets at the high-frequency level. Employing a slightly different approach, our filtering model deals with the problems of price idleness.[4] Our empirical study demonstrates that the soybean complex is significantly cointegrated and close to the underlying physical relationships. On the contrary, a noise-sensitive approach, such as Johansen's inference method (Johansen, 1995), yields inconsistent and unstable cointegrating vectors upon convergence in comparison with physical relationships. Furthermore, our high-frequency analysis confirms the central role played by realized and potential market liquidity[5] in the asset prices' multivariate dynamics, in particular for cointegrated assets. This confirms the findings by Arzandeh and Frank (2019), which emphasize the interest in considering LOB information in the price-discovery process. To validate our cointegration framework, we demonstrate that it is the crush-spread-associated adjustment space rather than the cointegrating space that turns out to be closely related to market microstructure.

This paper is organized as follows: The second section is devoted to the theoretical framework, which highlights the potential drivers of price cointegration. Consistent with the synthesis by Behrendt and Schmidt (2021), we find a non-linear relationship between volumes and prices. The third section describes the statistical methodologies retained to test the stationarity of our price data and to identify price cointegration at an intraday level, taking into account price idleness as defined by Bandi et al. (2020). The fourth section provides a description of the soybean complex and the retained data. In the fifth section, our results are commented and analysed, where we find that our filtering method outperforms traditional methodologies that ignore issues associated with multivariate high-frequency data (i.e., microstructure noise and non-synchronicity). We empirically test the relationship between volume and the strength of the cointegration, add more potential explanatory variables, and conduct several robustness checks. Our empirical results show that volume is significantly related to crush-spread cointegration in certain markets, particularly soymeal. We demonstrate how this relationship may affect the hedging efficacy of different rollover approaches. A conclusion completes the analysis.

[1] See also the S&P GSCI Index methodology document available from: https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-gsci.pdf.

[2] As of June 2021, the Bloomberg commodity index, formerly known as the Dow Jones-UBS Commodity Index, is invested in these three assets with long-only positions of about 5% in soybean, 2.3% in soybean meal, and 2.9% in soybean oil. Thus, it does not observe the CME crush spread, nor the proportions commonly accepted in the crushing industry. The Thomson Reuters Commodity Research Bureau index only includes soybeans futures contracts.

[3] The microstructure frictions can be associated with the bid–ask bounces, the discreteness of the price grid but also the technique used to construct the high-frequency price dataset (Hansen and Lunde, 2006).

[4] Staleness is defined according to Bandi et al. (2020) as a lack of price adjustments yielding zero returns with a traded volume associated whereas idleness corresponds to staleness without any trading activity.

[5] Potential liquidity is reflected by the depth of the LOB and defined by Kyle (1985) as the size of an order-flow innovation currently required by the market participants to change the price of a given amount.

2. Theoretical framework

Since the seminal papers of Epps and Epps (1976) and Tauchen and Pitts (1983), a large and still living literature strives to model the relationship between time series of financial prices and traded volumes.[6] Assuming certain market frictions or market-microstructure characteristics, these models boil down to first representing how information flows change the price expectations of market participants. Then equilibrium rules lead to fluctuations in volumes and asset prices. By and large, the existing academic contributions mainly focus on univariate dynamics modelling (O'Hara, 2015). In this article, we develop a multi-asset theoretical model that takes into account long-term equilibrium relationships. By taking into account heterogeneity in market participants' expectations, our theoretical framework sheds light on the short to mid-term dynamics of cointegrated time series conditional on the market participants' typologies.

To a significant extent, our model, like the aforementioned literature, stresses the role of information and tends to refute the rational-expectations assumption. For instance, Fishe et al. (2014) show theoretically that the rational-expectations equilibrium implies zero correlation between price and position changes, which is usually contradicted by available data. To reproduce well-known empirical features of financial asset prices, the literature has relied on the "difference-of-opinion" hypothesis, whereby economic agents agree to disagree on, for instance, public information. In the early paper by Tauchen and Pitts (1983), the economic agents disagree in a linear manner on the expected price of one commodity; by contrast, Epps and Epps (1976) formulate a non-linear disagreement function around the expected price. Later, He and Velu (2014) extend the linear approach of Tauchen and Pitts (1983) to a multi-asset settings approach by assuming that certain market announcements impacting the common latent factors can jointly affect the traded volume or the price of several assets, proportional to their respective latent-factor loadings. However, in their model, He and Velu (2014) do not consider the effect of belief heterogeneity among participants while, as shown by Duchin and Levy (2010) through simulations, disagreement on expected prices or on the expected price-variance–covariance matrix has a significant impact on asset prices and traded volumes. In the same vein, recent papers introduce market frictions (Darolles et al., 2017), heterogeneous discount factors (Beddock and Jouini, 2020), or a continuum of economic agents (Atmaz and Basak, 2018). While all of these extensions provide richer relationships between price, volume, and volatility, none have considered heterogeneous beliefs in a multivariate cointegrated setting.

In the financial microeconomics literature, the types of investors are often differentiated according to their respective levels of information and risk aversion. While less informed traders are generally assumed to be genuinely uninformed (when studied in a single market), we propose instead to consider these investors partially informed. By this, we mean that they focus on a single asset without considering the other assets linked by cointegration relationships. We could typically associate this investor profile with the commodity index traders or the index-tracking ETFs who go long on a specific futures contract because of its appealing liquidity, ignoring the less liquid cointegrated futures contracts. As a result, these indices are trading in and focusing on a given market based on private or publicly available information without necessarily considering the structural relationships of the physical assets. The second group of partially informed investors concerns the cash-and-carry arbitragers. Enticed by a theoretical arbitrage-risk-free gain, they reduce the basis volatility and force the prices of the underlying asset and its derivatives to converge at maturity, guaranteeing hedging efficiency by mitigating any non-convergence risk (although storage frictions may lead to structural non-convergence, Garcia et al., 2015). Nevertheless, this strategy of cash-and-carry (or reverse cash-and-carry) arbitrage is generally deployed on a single-asset basis and remains within the arbitrager's risk capacity (Hong and Yogo, 2012; Acharya et al., 2013). Only the last group of investors, manufacturers, are likely to consider joint equilibrium relationships among multiple underlying assets in the physical markets and could thus be considered fully informed. Through their commercial activities, manufacturers have the capacity to build synchronous positions in the cointegrated physical markets and eventually hedge their margin exposure through opposite trades in the associated futures contracts (Li and Hayes, 2022).

Our theoretical framework builds on the (Epps and Epps, 1976) framework. We consider $i = 1, \dots, n$ commodity markets and $j = 1, \dots, m$ potentially risk-averse economic agents. Like many previous papers, we simplify the analysis by assuming CARA preferences, a zero risk-free rate, and a finite horizon. Rather than expressing the inverse demand, we start with agent j's demand for assets. With $Q_{j,t-1}$ representing the $(n \times 1)$ vector of demand of assets by agent j at time $t-1$, P_{t-1} the vector of asset prices at time $t-1$, S_j the expected price-covariance matrix by agent j (assumed constant over time), ξ_j their risk aversion[7] (constant as well) and $X_{j,t-1}$ their expected final-prices column vector for the n assets at time $t-1$, we obtain :

$$Q_{j,t-1} = (\xi_j S_j)^{-1}(X_{j,t-1} - P_{t-1}) \tag{1}$$

This can be rewritten as:

$$Q_{j,t-1} = \lambda_j(X_{j,t-1} - P_{t-1}) \tag{2}$$

If an agent never participates in market i, this is reflected by a corresponding null row i in the λ_j matrix.

At the equilibrium, we assume that $\sum_j Q_{j,t-1} = 0$. Then the economic agents receive new information and process it into a new price expectation, so $X_{j,t-1}$ becomes $X_{j,t}$. Agent j's demand changes from period $t-1$ to period t and likely creates a market disequilibrium, which can be restored through appropriate price changes. From period $t-1$ to period t, we thus have :

$$Q_{j,t} - Q_{j,t-1} = \lambda_j(X_{j,t} - X_{j,t-1} - (P_t - P_{t-1})) \tag{3}$$

[6] Please refer to Behrendt and Schmidt (2021) for a literature review.

[7] It is worth highlighting that our model allows for considering a flexible and realistic framework including heterogeneity in investor risk aversion. Although this is not equivalent to the CRRA preferences hypothesis, it still allows for indirectly coping with investors' initial wealth disparity.

or

$$V_j = \lambda_j(\delta_j - \Delta P) \tag{4}$$

with $\Delta P = P_t - P_{t-1}$, δ_j denoting the change in price expectations, and V_j the volume traded by agent j.

Then we follow (Epps and Epps, 1976) in specifying the change in price expectations as[8]:

$$\delta_j | P_{t-1} = \hat{\delta} + \alpha_j(P_{t-1}) ABS(\hat{\delta})^{(1/\gamma)} \tag{5}$$

with γ being a positive constant and α_j a $(n \times n)$ matrix of strictly positive IID random variables that potentially depends on current prices in nonlinear ways, while $\hat{\delta}$ corresponds to the average change of price expectations across economic agents. We thus impose that $\sum_j \alpha_j(P_{t-1}) = 0_{n \times n}$, such that $\sum_j \delta_j / m = \hat{\delta}$. This multi-asset framework allows for more general specification than (Epps and Epps, 1976), who assume that $\alpha(P_{t-1})$ is simply the inverse function.

Let us interpret this crucial specification by computing the extent of disagreement between one agent and the market participants before the price change:

$$\delta_j - \hat{\delta} = \alpha_j(P_{t-1}) ABS(\hat{\delta})^{(1/\gamma)} \tag{6}$$

The extent of disagreement increases with the absolute value of the average change of price expectations. The economic logic of this specification is the following: when all economic actors expect small (positive or negative) price changes, i.e. due to new public information, their disagreement is likely to be small. On the other hand, if some economic actors receive private information and formulate new price expectations that are very different from their previous expectations while other economic actors did not access this information, the new agents' price expectations will be much more widely dispersed around a new average change of price expectations.

The specification of the stochastic matrix α_j recognizes that there may be variation in the logic described above. For instance, if a substantial new piece of public information is received and similarly interpreted by all economic agents, the average change of price expectations can be high and disagreement low. Conversely, if many economic agents receive the same significant private information, interpret it differently, and consequently formulate new price expectations in opposite directions, the average of the new price expectations can be equal to that of the previous price expectations, despite a higher dispersion.

The presence of the inverse of current prices in the extent of disagreement is not economically interpreted by Epps and Epps (1976) and appears as a convenient price normalization. A complementary economic interpretation for storable commodity markets is that economic actors take into account the current market situation while forming new price expectations after receiving new information. For instance, when current prices are relatively low compared to historical prices, this may lead economic actors to believe that commodity stocks are plentiful, and thus spot or physical (as well as futures) prices cannot change significantly due to the mitigating effect of stocks (Williams et al., 1991). Accordingly, some economic actors should make limited efforts to gather and process information to form new price expectations. In such an environment, even though significant private information received by some market participants cannot lead to significant (physical and futures) price changes, it will lead to more dispersed price expectations around the average value. Conversely, when current prices are relatively high compared to historical prices, many, if not all, economic actors concerned by a potential price bubble will gather and process public information. In this instance, price expectations should be characterized by a lower dispersion once a new piece of information has been released, as all the economic agents will be looking for it.

This interpretation of the inclusion of current prices in the disagreement specification extends to our multi-asset case. Indeed, our general formulation $\alpha_j(P_{t-1})$ allows for rich specifications, where the current prices of some assets may impact the changes in price expectations of other assets. We could also imagine the changes in price expectations being a function of the current price's deviation from a long-term cointegration relationship. For instance, if certain actors, i.e. manufacturers, find that the current price levels are significantly spreading out from the long-term physical relationships, they will expect the prices to progressively revert towards their long-run equilibrium.

With a demonstration provided in Supplementary Material A, we derive the relationship between asset-price changes and traded volumes as follows:

$$\Delta P = \Omega(V_1)^{\gamma} Diag\left(Sgn(\hat{\delta})\right) + \left(\sum_j \lambda_j\right)^{-1} \sum_j \lambda_j \alpha_j(P_{t-1}) \Omega(V_1) \tag{7}$$

where:

$$\Omega(V_1) = \left[\alpha_1(P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1} \sum_j \lambda_j \alpha_j(P_{t-1})\right]^{-1} \lambda_1^{-1} V_1$$

Eq. (7) generalizes the price-change Equation (21) obtained by Epps and Epps (1976), which does not include the second term on the right hand side. This expression makes clear that we have a non-linear relationship between the price changes and the volumes traded by one agent participating in all markets.

[8] To avoid cumbersome notations, we will refrain from the conditional formulation in the remainder of the paper.

Proposition 1. *Let us assume a two-commodity setting, where three investors are characterized by different levels of risk aversion and/or different variance–covariance matrices are forecast, while retaining* $\gamma = 1$,[9] *as well as the particular specifications (20) for the traders' forecast matrices* $\alpha_{i=1,\dots,3}$. *Then, the assets' joint price dynamics are characterized by a vector-error-correction-model (VECM) relationship if and only if the matrix* $\boldsymbol{\Pi}^{\star}(V_1)$, *which is a function of the volume traded by agent 1, is low-rank in the following expression (for a demonstration, see Supplementary Material B):*

$$\begin{aligned}\Delta P &= \alpha_1 (P_{t-1})^{-1} \mathbf{A} V_1 + \mathbf{B} V_1 \\ &= \boldsymbol{\Pi}^{\star}(V_1) P_{t-1} + \mathbf{B} V_1\end{aligned} \tag{8}$$

where:

$$\begin{aligned}\boldsymbol{\Pi}^{\star} &= \boldsymbol{\Phi}\mathbf{A}V_1\beta' \\ \boldsymbol{\Phi} &= \begin{pmatrix}\phi_1 & 0 \\ 0 & \phi_2\end{pmatrix} \\ \mathbf{A} &= Diag\left(Sgn(\hat{\delta})\right)\left[I_2 - \lambda^{\star}\right]^{-1}\lambda_1^{-1} \\ \mathbf{B} &= \lambda^{\star}\left[I_2 - \lambda^{\star}\right]^{-1}\lambda_1^{-1} \\ \lambda^{\star} &= \left(\sum_j \lambda_j\right)^{-1}\begin{pmatrix}\lambda_1^{11} - \lambda_2^{11} & 0 \\ 0 & \lambda_1^{22} - \lambda_3^{22}\end{pmatrix}\end{aligned}$$

where I_2 *is an identity matrix of dimension* 2×2 *and* β *and* $\kappa(V_1) = \boldsymbol{\Phi}\mathbf{A}V_1$ *two low-rank matrices of dimension* (2×1).[10] *This means that, if we assume the cointegrating vector* β *to be stable over time, both elements of the vector* κ, *denoted* κ_1 *and* κ_2, *respectively, and associated with the speed of reversion towards the long-term cointegration relationship are a function of the volumes traded by agent 1 in both markets. By assuming that only one trader is trading in both markets, we also assume that this agent has no arbitrage limit and can thus match the number of contracts that the two other agents would like to sell or buy on each individual market. This explains why the theoretical relationship does not involve the other agents' positions or trades per asset but only their risk aversion and variance–covariance expectations. Thus, only the volumes associated with the agent participating in both markets are considered capable of affecting both prices and revealing a multivariate cointegration relationship in a high-frequency setting.*

Another important point to make for our empirical study is that, in studying the link between κ and traded volumes, we are conditioning our analysis to the assumption that the system is cointegrated. However, the matrix $\boldsymbol{\Pi}$ could itself be a function of traded volumes without being low rank, which would mean that the auto-regressive matrix of the asset prices would be a function of volumes but not necessarily the cointegration process itself. To demonstrate that the traded volumes play a determining role in the cointegration process itself, we thus need to verify that not only κ, but also the rank of matrix $\boldsymbol{\Pi}$ is related to volumes. Put differently, the rank of matrix $\boldsymbol{\Pi}$ – which determines whether or not a cointegration relationship exists – and the low-rank matrix κ must both be a function of the volumes traded of each asset to justify the conclusion that the cointegration process is intrinsically linked to the volumes traded in the financial markets.

Finally, we notice that the traded volumes also impact the constant term in Eq. (8). Nevertheless, if we assume that all traders are characterized by the same level of risk aversion and forecast the same variance–covariance matrix, that is $\lambda_{j=1,\dots,3} = \lambda_1$, the matrices $\lambda^{\star}$ equal zero and, by definition, the matrix B equals zero as well. Eq. (8) thus simplifies to:

$$\Delta P = \alpha_1 (P_{t-1})^{-1} Diag\left(Sgn(\hat{\delta})\right)\lambda_1^{-1} V_1 \tag{9}$$

which also points to a VECM relationship, though without the constant term.

To empirically validate our model and the associated hypotheses, we propose to test and investigate the dynamics of the intra-daily cointegration among assets as a function of the daily traded volume and order-book depth of individual assets. Once the stability over time of the cointegrating vector β has been demonstrated, we will also investigate how, under the hypothesis of cointegrated time series, the dimension of the adjustment space spanned by the loading vector κ can be affected by traded volumes. Nevertheless, studying dynamics at such a high level of granularity has its inherent statistical challenges, including microstructure noise and asynchronicity of traded prices. To address these challenges, cointegration dynamics must be written in a state–space form.

3. Econometric models

Our theoretical model leads to a VECM model that links prices and volumes for cointegrated assets. The same VECM has already been considered in the high-frequency literature on the econometric representation of the price-discovery process between two closely related securities. Initially adapted by Hasbrouck (1995) to describe the joint dynamics of closely linked securities traded

[9] This hypothesis is not a necessary condition to express the cointegration relationship as a function of the traded volumes. Indeed, if we assume that $\gamma \neq 1$, we then obtain a nonlinear error-correction model (NEC) with a polynomial functional (Escribano and Mira, 2002; Escribano, 2004; Tjøstheim, 2020).

[10] In our two-commodity setting, we obtain a (2×1) vector. Nevertheless, if more assets are taken into account, two matrices, $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$, of dimension $(n \times h)$ are thus obtained, whereby n denotes the number of assets and h the number of cointegration relationships among the assets.

in different markets, the cointegration model has ever since been considered in high-frequency settings to capture the lead–lag relationship between related assets, such as underlying spot prices and related futures or options prices, or equities issued by the same company in different markets (Foucault et al., 2017; Hasbrouck, 2019; Brugler and Comerton-Forde, 2019). These studies generally used cointegration to represent very high-frequency joint dynamics resulting from financial arbitrage strategies, such as cash-and-carry or triangular arbitrage (Foucault et al., 2017). Conversely, this paper focuses on cointegration relationships stemming from the physical characteristics of each asset, such as the relationship between a given commodity and its byproducts, where no genuine arbitrage gain is to be expected. The supply and demand disequilibrium associated with the commodity itself or its byproducts could indeed consistently or temporarily change the associated spread levels. This rich strain of literature does not, however, shed light on high-frequency data features such as asynchronicity, microstructure noise, or price staleness and idleness, which we need to take into account in order to reduce the risk of model misspecification.

Our model can be cast in a state–space formulation of the VECM model, whereby the idle prices are considered as missing data, unlike in Buccheri et al. (2019, 2021b) and Buccheri et al. (2021a).[11] Initially proposed by Shumway and Stoffer (1982) and extended to the cointegrated processes by Seong et al. (2013), missing-data models consist in filling the database using a latent-process expected mean, conditional on given parameters. We use the same filtering technique proposed by Seong et al. (2013), whereby observation noise is added to cope with microstructure noise, as described by Buccheri et al. (2019, 2021b) and Buccheri et al. (2021a). Nevertheless, our model should not be confused with the model proposed in the latter two contributions, as the information associated with zero returns is treated differently in our model when the traded volume associated is null or positive.[12]

3.1. VECM state–space representation

Let us assume h cointegration relationships among n non-stationary financial asset prices; we will denote P_t the n dimension row vector of the asset prices at time t; then Eq. (8) can be written as the following vector-error-correction model (VECM):

$$\Delta P_t = c_0 + \Pi P_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta P_{t-j} + e_t \tag{10}$$

where the low-rank matrix $\Pi = \kappa\beta'$ can be decomposed into two rank h-matrices κ and β of dimensions $(n \times h)$ and $e_t \sim N(0, \Sigma)$. The constant term in Eq. (8), denoted as c_0 here, can be removed and included as an intercept term in the cointegration relationships, as demonstrated in Lütkepohl (2005).[13] This VECM formulation can thus be equivalently written as a VAR(p) model, such that Lütkepohl (2005):

$$P_t = \sum_{j=1}^{p} \Phi_j P_{t-j} + e_t \tag{11}$$

where $\Phi_1 = I_n + \Pi + \Gamma_1$, $\Phi_j = \Gamma_j - \Gamma_{j-1}$ for $j = 2, \dots, p-1$ and $\Phi_p = -\Gamma_{p-1}$.

Following Buccheri et al. (2019, 2021b) and Buccheri et al. (2021a), we assume that this discretized multivariate dynamics is latent in a high-frequency setting and thus inaccurately observed on account of the ubiquitous microstructure noise present in financial markets. A state–space representation is thus fully justified, with the transition equation following from expression (11):

$$x_t = F x_{t-1} + G e_t \tag{12}$$

where $x_t = (P_t', P_{t-1}', \dots, P_{t-p+1}')'$ and where we define F as the following $np \times np$ transition matrix:

$$F = \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_p \\ I_n & O_n & \cdots & O_n \\ O_n & I_n & \cdots & O_n \\ \vdots & \vdots & \cdots & \vdots \\ O_n & O_n & \cdots & O_n \end{bmatrix}$$

and the $np \times n$ matrix G as:

$$G = \begin{bmatrix} I_n \\ O_{(np-n)\times n} \end{bmatrix}$$

The following expression corresponds to the observation equation:

$$y_t = H_t x_t + w_t \tag{13}$$

where w_t is a zero mean, normally distributed uncorrelated $q \times 1$ noise vector with R as $q \times q$ covariance matrix. Moreover, H_t corresponds to a $q \times np$ observation design matrix, which converts the unobserved $np \times 1$ vector x_t into the $q \times 1$ imperfectly observed

[11] For the sake of completeness, a potential missing-data modification for their algorithm is mentioned in the technical appendix of Buccheri et al. (2021b).

[12] Using a continuous time semi-martingale model, Bandi et al. (2020) indeed differentiated between the impact of idleness and staleness upon parameter estimations.

[13] As demonstrated in Lütkepohl (2005), if we keep the constant in Eq. (10), we should then constrain it for the model estimation, such that $c_0 = -\kappa\beta'\mu_0$, where μ_0 is the adjusted constant. This avoids generating a linear trend in the mean of P_t.

series y_t. It is worth highlighting that the dimensions of the matrix H_t may change over time. Indeed, $q < n$ when all the assets are not trading simultaneously. This observation equation is different from the one proposed by Buccheri et al. (2019, 2021b) and Buccheri et al. (2021a), where the matrix $H_t = H = \left[I_n, O_{n\times(np-n)}\right]$. In our state–space model, we thus distinguish the situations where one of the assets has simultaneously traded with the others or not. With $H = \left[I_n, O_{n\times(np-n)}\right]$, the measurement error associated with an idle price is first assumed to be zero-mean and finite-variance white noise. Furthermore, this measurement error is mixed with the potential observation error when the cointegrated assets are simultaneously trading. This assumption can have significant impact on the cointegration model's estimation and the interpretation of results, as demonstrated in our empirical study. This is due to the fact that price idleness and staleness convey relevant information related to the data-generation process (Bandi et al., 2020). Our empirical study buttresses this conclusion, showing that cointegration results differ significantly depending on whether we assume that the matrix $H = \left[I_n, O_{n\times(np-n)}\right]$ or not. Provided that this state–space formulation is linear and Gaussian, we can apply the conventional Kalman filter and Kalman smoother, under the assumption that the parameters $\theta = \{\Phi_j, \Sigma, R\}$ are known.[14]

3.2. Model estimation

3.2.1. The EM algorithm

Whereas the rank and the parameters denoted θ are assumed to be known in the filtering and smoothing steps described in the previous section, Dempster et al. (1977) developed an Expectation Maximization algorithm, which consists in maximizing the complete data log likelihood and which assumes all data $x_{1:T}$ to be available, conditional to the data $y_{1:T}$ that we observed:

$$\begin{aligned}\log \mathcal{L}(\theta; x_{1:T}, y_{1:T}) = &-\frac{1}{2}\log|\Lambda| - \frac{1}{2}(x_0-\delta)'\Lambda^{-1}(x_0-\delta)\\ &- \frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{t=1}^{T}(Ax_t - \Gamma Bx_{t-1})'\Sigma^{-1}(Ax_t - \Gamma Bx_{t-1})\\ &- \frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T}(y_t - H_t x_t)'R^{-1}(y_t - H_t x_t)\end{aligned} \tag{14}$$

where:

$$A = \left[I_n \ -I_n \ O_{n\times(np-2n)}\right];$$

$$\Gamma = \left[\kappa \ \Gamma_1 \ \Gamma_2 \ \dots \ \Gamma_{p-1}\right];$$

and

$$B = \begin{bmatrix} \beta' & O_{h\times n} & O_{h\times n} & \dots \\ I_n & -I_n & O_n & \dots \\ O_n & I_n & -I_n & \dots \\ \vdots & \vdots & \vdots & \ddots \\ O_n & \dots & \dots & I_n \end{bmatrix}$$

with a normalized $\beta = [I_h \ \beta_0']$, β_0 being the $(n-h)\times h$ matrix to be estimated,[15] and $x_0 \sim N(\delta, \Lambda)$. The EM algorithm then consists in a two-step recursive procedure[16]:

(i) the Expectation step: a given set of parameters θ^l associated with the l-iteration is used to calculate the expected value of the complete-data log-likelihood, conditional on θ^l, represented by the operator E_l, and the observed data $y_{1:T}$:

$$Q(\theta|\theta^l) = E_l\{log\mathcal{L}(\theta; x_{1:T}, y_{1:T}|y_{1:T})\} \tag{15}$$

where the latent process-expectation and covariance-matrix estimators conditional on the observed data are provided by the combination of the Kalman filter and smoother.

(ii) The Maximization step: we maximize this conditional expectation of the complete-data log likelihood using the analytical gradient.[17] to obtain a new set of parameters θ^{l+1} that we use in the next iteration of the algorithm. We then go back to the E-step.

This iterative procedure has been shown to provide a non-decreasing likelihood towards the maximum incomplete-data log-likelihood innovations form (Dempster et al., 1977; Shumway and Stoffer, 1982) that we use to determine at each iteration when the algorithm should be stopped.

[14] Bear in mind that the matrices Φ_j include the parameters of sub-matrices κ, β, and Γ_j. The lag used for the VECM model is determined with the Bayesian Information Criterion (BIC). Furthermore, in Supplementary Material C, a detailed description can be found of both the filter and the smoother used to estimate the conditional expectation, as well as of the conditional covariance matrix associated with the latent process.

[15] It is interesting to note that $Ax_t = \Delta P_t$, while $\Gamma Bx_{t-1} = \Pi P_{t-1} + \sum_{j=1}^{p-1}\Gamma_j \Delta P_{t-j}$.

[16] Supplementary Material D provides a detailed description of the algorithm.

[17] A detailed derivation of the gradient is provided in Supplementary Material D.

3.2.2. The rank estimation

While we have thus far assumed the rank of the $\Pi = \kappa\beta'$ matrix to be known, we perform the conditional likelihood ratio test to estimate it conditionally with respect to $\hat{\theta}$, the EM-estimated parameters, and the observed data $y_{1:T}$. For this likelihood ratio test, we postulate the following null hypothesis:

$$H_0 : rank(\Pi) = r_0 \text{ with } 0 \leq r_0 < n$$

where r_0 is the specific matrix rank to be tested. The alternative hypothesis is:

$$H_1 : r_0 < rank(\Pi) \leq r_1$$

Using the respective complete-data log likelihoods associated with $\theta^{\star}_{r_0}$, the EM-estimated optimal set of parameters assuming $rank(\Pi) = r_0$, and $\theta^{\star}_{r_1}$, which denotes the optimal parameters with $rank(\Pi) = r_1$, the LR statistic $\lambda_{LR}(r_0, r_1)$ is equal to:

$$\begin{aligned}\lambda_{LR}(r_0, r_1) &= -2\log\left[\frac{sup_{\theta^{\star}_{r_0}}\mathcal{L}}{sup_{\theta^{\star}_{r_1}}\mathcal{L}}\right] \\ &= -2\left[log\mathcal{L}(r_1) - log\mathcal{L}(r_0)\right]\end{aligned} \tag{16}$$

With a preliminary panel-stationarity test, we ensure all the asset prices follow unit-root processes, hence $rank(\Pi) < n$. Then we considered in our empirical study the likelihood-ratio-test statistic (16), with $r_0 = 0$ and $r_1 = 1$, to assess the p-value for a rank of Π equal to 0.[18] This rank-associated probability is then considered an ordinal number to detect whether or not asset prices are cointegrated on a daily basis and to establish the strength of this cointegration relationship throughout a given day. Regarding the non-standard asymptotic distribution of $\lambda_{LR}(r_i, r_{i+1})$ under the null hypothesis, we refer to the 99% critical value of 7.02 as provided by the table (15.1) in Johansen (1995).

4. Data: The soybean complex

For our empirical study, we use the soybean crush spread, a well-known commodity complex that has been extensively studied in the futures markets literature (Johnson et al., 1991; Rechner and Poitras, 1993; Simon, 1999; Mitchell, 2010; Liu and Sono, 2016; Marowka et al., 2020; Li and Hayes, 2022). This spread is often studied for its presence of cointegrated multivariate time series[19] and also because soybean futures are among the most traded commodity derivatives contracts in the world, with a double quotation on the US and Chinese derivatives markets. Other cointegrated financial assets could have been considered, such as interest rates (Bradley and Lumpkin, 1992; Dewachter and Iania, 2011) or equities (Chen et al., 2002; Awokuse et al., 2009).

For this study, we have used the data from the soybean complex (soybeans, soybean oil, and soybean meal) quoted at the CME (Chicago Mercantile Exchange). Matching the product codes at the CME Globex, we abbreviate soybean to ZS, soybean oil to ZL, and soybean meal to ZM; in Eq. (10), the prices vector z_t will observe the same order, such that z_{1t} represents the soybean price, z_{2t} stands for the soyoil price, and z_{3t} denotes the soymeal price, all at time t.

The high-frequency data used in this study covers the total trading activity of 2015, amounting to 243 trading days. The data is retrieved from the CME, which stores the data in a sequence of messages, each with a millisecond-resolution timestamp and representing an update of the security. Such an update can be an executed trade, a change in the limit order book, or the daily open-interest statistic. Note that these messages only arrive at updates; hence, the frequency of updates (messages) is based on, and reflects the activity in the market.

By iterating these messages sequentially, and updating the LOB accordingly, the LOB can be reconstructed at any point in time. Next, this time series of LOBs with irregular time intervals can be resampled into any arbitrarily chosen snapshot size. In this study, we opted for one-minute snapshots in order to limit the Epps effect (Epps, 1979), which describes sample correlation bias as moving towards zero as the data frequency in the analysis increases.[20] Moreover, we distinguish two periods within a trading day: the electronic trading session from 7 PM to 7.45 AM (session 1) and the market trading session from 8.30 AM to 1.20 PM (session 2). While the latter trading session is shorter, it contains the most trading activity.[21] For the robustness check, we use multiple methods to generate snapshots, which are described below. In addition, the XLM is calculated for each snapshot, so as to measure the liquidity of the market at any point in time (Gomber et al., 2015).

[18] For the selection of the model, we also tested in our empirical study the presence of more than one cointegration relationship using appropriate r_0 and r_1.

[19] Based on long-term time series, Simon (1999), Mitchell (2010), and Liu and Sono (2016) demonstrate in their empirical studies the existence of a stationary combination of soybean, soyoil, and soymeal futures prices. This cointegration relationship can be interpreted as a long-term market-price equilibrium for the so-called crush spread, combined with transitory seasonality and a consistent trend. More recently, Marowka et al. (2020) presented evidence that the crush-spread cointegrating vector and the associated cointegrating space display significant time instability on a yearly basis, which is detrimental to soybean processors who hedge their physical exposure on financial markets.

[20] In order to identify any side effects from the method used to generate the one-minute snapshot time series, a robustness check has been added in the supplementary materials: Instead of collecting the asset prices for each snapshot at the first second of each minute, asset prices at the 30th second are used. For this data sample (30s Monthly Rollover), the Monthly Rollover technique has been used.

[21] The CME closed its agricultural futures trading pits in July 2015 while leaving open the options trading pits. In our case, the "market trading session" corresponds to agricultural commodities' trading hours on the floor of the exchange until 2 July and on its electronic trading platform (Globex).

The XLM (Exchange Liquidity Measure) calculates the round-trip liquidity premium for buying and selling a chosen volume, i.e., how much the average transaction price deviates from the mid-price. Eq. (19) shows how to calculate the XLM:

$$XLM_{B,t}(V) = 10,000\frac{P_{B,t}(V) - MP_t}{MP_t} \tag{17}$$

$$XLM_{S,t}(V) = 10,000\frac{MP_t - P_{S,t}(V)}{MP_t} \tag{18}$$

$$XLM_t(V) = XLM_{B,t}(V) + XLM_{S,t}(V) \tag{19}$$

where, MP_t is the mid-price at time t, $P_{B,t}(V)$ is the average transaction price of a buy-initiated market order with dollar value V at time t, $P_{S,t}(V)$ is the average transaction price of a sell-initiated market order with dollar value V at time t, and $XLM_t(V)$ the round-trip liquidity premium at time t. In this research, the dollar value is set to be the median dollar value of the order book spanning the full year.

After processing (i.e., reconstructing and resampling into snapshots) each individual futures contacts, the multiple time series are merged into a single non-ending time series. This technique is called rollover. Merging of contracts is required since, at any point in time, there are multiple futures contracts available for trading with a separate open interest and maturity date. For example, soybean futures are available at the CME – in each specific year – as January, March, May, July, August, September, and November contracts, whereas soymeal and soyoil have additional October and December contracts but no November contracts. Different rollover techniques can be considered to combine all contracts, i.e., create a single time series per commodity that captures the most relevant information (Carchano and Pardo, 2009).

For the robustness check, three different rollover techniques will be compared in this paper. The first rolling technique is based on the soybean open interest (ZS Open interest). All three contracts are rolled to the next maturity based on the soybean open-interest crossover day (i.e., the first day on which the next futures contract has a higher open-interest value) (Carchano and Pardo, 2009). This method ensures that the rollover of all time series occurs at the same time, with the open interest of soybean being the determining factor. The second rolling technique is an independent open-interest rollover (Independent Rollover), where each contract's open-interest crossover triggers the associated roll position. The final rolling technique is the monthly rollover (Monthly Rollover), which has been applied in most of the existing literature (Frank and Garcia, 2011; Trujillo-Barrera and Garcia, 2012; Gorton et al., 2013; Etienne et al., 2014, 2015; Dorfman and Karali, 2015; Han et al., 2016; Fernandez-Perez et al., 2016; Fan et al., 2020). With this rolling technique, the current contract is rolled to the second nearby contract at the end of the month preceding contract expiration.

5. Econometric results

To validate the theoretical model proposed in this article and thus demonstrate the relationship between asset-price cointegration and individual traded volumes, we divided our results analysis into four subsets. We first verify the intraday non-stationarity of the marginal dynamics, as well as the cointegration among these dynamics. We take this opportunity to demonstrate how time of day may affect the joint stationarity of the soybean complex. Following our Proposition 1, we then validate the hypothesis that the rank of matrix Π is indeed a function of the volume traded on each market. In particular, we demonstrate that the presence or absence of cointegration among the soybean-complex components at the high-frequency level – and thus the intraday efficiency of the complex futures markets – is related to the volumes traded in each of these markets.

Furthermore, as stated in our proposition, the presence in the markets of traders with sufficient arbitrage capacity to enforce the cointegration relationship should manifest through κ, the speed of reversion towards the long-term trend, whereas the cointegrating vector should, on average, remain close to the physical weights following from the industrial soybean trituration. We thus verify in the following that the intraday cointegrating vector and loading matrix display such features. Finally, we demonstrate how our findings could influence the design of optimal rolling techniques.

5.1. Stationarity and cointegration of the high-frequency soybean complex

For the unit-root test, we considered the panel test introduced by Hadri (2000) and applied it to the 243 trading days in 2015 for which we observe 1-minute data samples. According to Table 1, based on calendar order,[22] we demonstrate that the three intraday-price time series are non-stationary for almost all panels, which justifies the performance of a high-frequency cointegration analysis.

As mentioned earlier, one of the main problems when studying the joint dynamics of high-frequency time series is the non-synchronicity of the markets, which hampers the estimation of the dependence structure among the time series significantly (Lo and MacKinlay, 1990). This impact on the estimation of the parameters manifests itself when comparing an EM-algorithm-based estimate with a basic Johansen approach at a high-frequency level. To carry out the Johansen test on non-synchronous high-frequency data, we had to apply ad hoc matching, which implies matching the price of a given asset that has just traded with the last-traded price of the other assets (denoted below as 'all price'), depending on the frequency considered. In addition, to investigate the potential

[22] Other ordering variables for panel construction have been considered, for instance relative to daily traded volumes. The stationary hypothesis was always rejected.

Table 1
ZS open interest rollover panel stationarity test.

30 days panel[a]	ZS	dif(ZS)	ZL	dif(ZL)	ZM	dif(ZM)
1	5006.1	−7.5***	4829.5	1.9***	4602.3	−4.5***
2	4488.9	1.0***	3573.5	0.8***	4681.2	1.9**
3	4693.7	3.3	3792.2	8.5	4813.0	−29.7***
4	4381.8	−11.6***	3604.6	−32.6***	85.1	−40.5***
5	5314.4	−1.6***	4547.3	−3.7***	4983.6	0.0***
6	5685.0	−5.3***	4730.4	−3.3***	5677.9	−11.1***
7	4297.1	−11.3***	406.0	−33.4***	305.7	−40.2***
8	5211.1	−4.4***	710.6	−40.3***	379.2	−40.3***

[a]This table records panel stationarity test statistics for 30-day samples of session 2 one-minute data considering the ZS open interest rollover technique. First, we sort the T-stat of the KPSS test according to the calendar order of the three assets, from the lowest value to the highest value for 8 panels (30 days per panel). Second, we calculate the panel stationarity test statistics for each panel (Hadri, 2000), where the critical values in the one-tail test are 1.282 for the 10% significance level, 1.645 for the 5% significance level, and 2.326 for the 1% significance level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2
Cointegration test for session 2 — 1 min data.

Johansen cointegration test[a]	Day	Month	Quarter
ZS open interest rollover_all price[b]	64	1	0
ZS open interest rollover_ZSmatch	69	6	1
ZS open interest rollover_ZLmatch	70	6	1
ZS open interest rollover_ZMmatch	68	5	1
Missing data filtered cointegration test[a]	**Day**	**Month**	**Quarter**
ZS open interest rollover[c]	180	11	3

[a]This table shows the number of cointegrated days, months and quarters for 2015. It compares the results of the Johansen approach and the missing-data filtering techniques described in Section 3.1.
[b]'All price' represents the trading-time approach, whereby the idle prices are not considered missing and are still assumed to be fair prices for the given assets until the next trade. 'ZSmatch', 'ZLmatch', and 'ZMmatch' represent the trading-time approach, where we match the trades for a particular asset with the idle prices of the two other assets.
[c]In this data set, idle prices are not considered informative and have been removed. As such, only the missing-data filtering technique can be applied.

lead effect of a specific asset on the other lagged assets, we applied an asset-based matching algorithm. This method matches the last trading price of either soybeans, soybean oil, or soybean meal (hereinafter denoted as ZSmatch, ZLmatch, and ZMmatch, respectively) with the most recent trading prices of the two other assets. We thus constrained our sample time stamp to a specific asset and presumed the lead–lag structure of the data.

Hardly any cointegration was detected using Johansen's approach, regardless of the matching method applied. After filtering for microstructure noise and time-series asynchronicity, however, the number of cointegrated days detected becomes significantly higher, as shown in Table 2. Since we know that the frequency of the data versus the period of data acquisition can impact the estimation of cointegration models (Hakkio and Rush, 1991), we increased the sample size to facilitate the detection of cointegration by Johansen's model. Whichever sample scheme we considered – daily, monthly, or quarterly – Johansen's approach without data filtering performed poorly compared to the EM-algorithm approach. For the sake of robustness, we verified that these results were not affected by the various rolling techniques and the data-frequency choices.[23] Furthermore, we discovered the presence of a diurnal effect with regard to cointegration. A high level of cointegration is indeed observed during session 2 trading hours, which fades away during session 1.[24] Another interesting result is the stronger intraday cointegration observed on average on USDA announcement days, although the number of observations available is limited.

5.2. Intraday cointegration and traded volumes

To determine whether the intraday price-cointegration process depends on the volume traded of each asset, we first have to verify that the rank of the product matrix $\boldsymbol{\Pi} = \boldsymbol{\kappa}'\boldsymbol{\beta}^*$ in Eq. (10) is a function of the daily traded volumes. The Granger representation theorem indeed states that an error-correction representation exists if low-rank matrices $\boldsymbol{\kappa}'$ and $\boldsymbol{\beta}^*$ both occur. To verify that the presence or absence of cointegrated time series is a function of the traded volume, we analyse the daily likelihood-ratio-test time series calculated based on the intraday asset-price vectors, in particular the likelihood ratio of the null-rank versus the rank-one hypothesis. This specific ratio indicates whether the matrix is statistically closer to a null-rank matrix or a rank-one matrix. If the

[23] A description of the test results is provided in Supplementary Material G.
[24] The results are provided in Supplementary Material G.1.

matrix rank is zero, all of the soybean complex components are integrated but no cointegration has been statistically detected. Conversely, if the matrix is rank one, at least one cointegration relationship has been detected.

The interest of the likelihood-ratio-based statistic we proposed to retain is that we know its asymptotic distribution and can thus determine a set of critical values for the test. Studying intraday data on a daily basis allows us to detect the presence of cointegration and then compare it with the daily traded volumes.

We apply a stepwise logit regression with a binomial distribution (denoted GLM) to model the cointegrated/non-cointegrated binary variable as a function of the daily traded volumes. To interpret the coefficients for each regressor, we report the associated marginal effects (Greene, 2003).

Since traded volumes can be closely related to a market's price volatility (Bessembinder and Seguin, 1993), we propose a set of nine control variables stemming from high-frequency literature. This includes assessments of the average intraday variance realized, bipower variation, and the XLM index for each of the three components of the soybean-crush spread. The XLM index allows us to distinguish between the influence of the daily traded volumes and the average depth of the book order, which could be defined as the average potential tradable volume for each individual market. While bipower variation and realized variance, as defined in Couleau et al. (2020), are two measures of integrated volatility, bipower variation offers the specificity of being a robust metric for identifying rare jumps as well as a model-free estimator of integrated variance. One of the alternatives to this estimator is the realized variance measure.

GLM stepwise regression can be found in Supplementary Material E.1, the result of which shows that daily volumes are significant in explaining the rank of matrix Π and thus the cointegration process in markets and that volatility measurement, such as the bipower variation, is statistically insignificant. We also demonstrate that the volumes of soybean byproducts are the most important variables for exchanges to monitor. As a complement to the linear GLM approach, in Supplementary Material E.2, we propose a set of panel cointegration tests provided by Larsson et al. (2001) that allows for the capturing of non-linear relationships. We again demonstrate that the traded volumes integrate information that volatility measures are not taking into account.

5.3. Long-run equilibrium dynamics and traded volumes

While, in the previous section, we demonstrated how the rank of matrix Π is positively related to the assets' traded volumes, the following subsections provide a detailed analysis of the joint and marginal dynamics components that cause this phenomenon. Conditional on the (low) rank of matrix Π and whether the time series are cointegrated, we can rewrite this matrix as the product of two $h \times n$ sub-matrices κ and β, with $h < n$. The former, i.e. the loading matrix, is interpreted as the adjustment of each asset's prices to the long-run equilibrium or error-correction term. The latter, i.e. the cointegrating vector, renders the integrated initial data stationary. Following our theoretical model, the expected value of the cointegrating vector should equal the trituration-associated weights as expected and enforced by the traders that intervene in the individual markets for all three crush-spread components. Conversely, the loading matrix should be a function of the volume traded for each asset, provided there is at least one cointegration relationship.

5.3.1. Adjustment space

In this section, we investigate how the traded volumes impact the cointegration process. To do so, we simultaneously study their impact on two related components of the cointegration process by calculating 'Relative market-information share.' This measurement was proposed by Hasbrouck (1995), while (Baillie et al., 2002) demonstrated that it is equivalent to the ratio of the respective components of the vector $\kappa_{\perp}$ weighted by the variance–covariance matrix of the innovations.[25]

Following our theoretical model, if we assume that $\beta_{\perp}$ in the VAR representation associated matrix Ξ[26] is not related to the traded volumes, then the relative market information share should thus, through κ and $\kappa_{\perp}$ components, be a function of the $ij_{\{i\neq j;i,j=1,2,3\}}$ relative traded volumes Vol_i/Vol_j. Furthermore, given that the variance of asset prices is closely related to the volume traded (Epps and Epps, 1976, among others.), we assume the scaling multiplier in (51), which is the Σ matrix components' ratio, to be equal to one, and we focus our analysis on the squared values of the ratios of the vector $\kappa_{\perp}$'s components,[27] $\tilde{\kappa}_{ij} = (\kappa_{\perp,i})^2/(\kappa_{\perp,j})^2$. We then regressed it on the relative traded volumes and the same control variables that we previously considered: the high-frequency bipower-variation measure, the realized-variance measure, and the XLM index.

We could thus conclude that, if the volumes traded in the byproducts' markets are sufficiently high relative to those in the bean market (meaning a simultaneous increase of Vol_ZL/Vol_ZS and decrease of Vol_ZL/Vol_ZM), the meal price will more significantly Granger cause the other market prices (higher $\tilde{\kappa}_{3,2}$ and $\tilde{\kappa}_{3,1}$). However, the more disconnected the volume of the bean market from that of the byproducts' markets (meaning a simultaneous decrease of Vol_ZL/Vol_ZS and increase of Vol_ZL/Vol_ZM), the less related the three markets (lower $\tilde{\kappa}_{3,2}$, $\tilde{\kappa}_{1,2}$ and $\tilde{\kappa}_{3,1}$).

The results displayed in Table 3 validate our model's assumption by showing significant linear relationships between three $\tilde{\kappa}_{ij}$ ratios and three traded volume ratios, whereas all other ratios or control variables prove to be insignificant. In addition, these linear relationships show that the contribution of soybean to the common factor relative to that of soyoil is positively related to the ratio of the traded volumes of soyoil and soybean. This means that the higher the volume of soyoil relative to soybean, the more soybean

[25] Please refer to Supplementary Material F for more details.

[26] A formal definition of Ξ is provided in Supplementary Material F.

[27] According to Baillie et al. (2002), the ratio $\tilde{\kappa}_{ij}$ is equivalent to the relative information share of market i versus market j, as defined in Hasbrouck (1995), provided that the scaling multiplier, i.e. the Σ matrix components ratio, is set to one and that there is no correlation between the error terms $\Sigma_{ij} = 0$.

Table 3
Regression of $\tilde{\kappa}_{ij}$ ratios.

Regressor[a]	$\tilde{\kappa}_{1,2}$	$\tilde{\kappa}_{1,3}$	$\tilde{\kappa}_{2,3}$	$\tilde{\kappa}_{2,1}$	$\tilde{\kappa}_{3,1}$	$\tilde{\kappa}_{3,2}$
Vol_ZS	0.097	0.613	−0.579	−0.528	−0.453	0.673
Vol_ZL	0.331	1.366	−0.525	−0.498	−1.078	1.107
Vol_ZM	−0.586	−0.115	−0.442	−0.443	−0.365	0.186
Vol_ZS/Vol_ZL	0.806	−0.42	−0.249	−0.22	−0.706	1.312
Vol_ZS/Vol_ZM	0.643	0.366	−0.255	−0.213	−0.204	0.396
Vol_ZL/Vol_ZM	0.748	1.785	−0.197	−0.161	−2.71***	0.251
Vol_ZL/Vol_ZS	2.5***	0.908	−0.03	−0.047	−0.974	2.47**
Vol_ZM/Vol_ZS	−1.019	−0.691	0.19	0.1	−0.639	−0.836
Vol_ZM/Vol_ZL	−0.016	−1.114	−0.07	−0.093	−1.034	1.072
RV_ZS	−0.215	0.215	−0.309	−0.281	−0.884	−0.336
RV_ZL	1.008	1.249	−0.668	−0.626	−1.732	1.859
RV_ZM	−0.096	−0.063	−0.515	−0.479	−0.726	0.273
BV_ZS	−0.337	0.704	−0.346	−0.304	−0.756	−0.454
BV_ZL	0.765	1.663	−0.634	−0.592	−1.734	1.67
BV_ZM	−0.475	0.213	−0.518	−0.478	−0.32	−0.275
XLM_ZS	0.055	−0.046	−0.988	−0.949	−0.116	−0.412
XLM_ZL	0.766	0.157	−0.954	−0.872	−1.362	1.138
XLM_ZM	0.156	−0.038	−1.103	−1.07	−0.917	0.562

[a]Using daily sets of one-minute data from session 2 and the ZS open interest rollover technique, this table records the coefficients and p-values associated to the regression of $\tilde{\kappa}_{ij}$ on the daily realized variance (RV), bipower variation (BV), traded volumes (Vol), and XLM index (XLM); ZS stands for soybean, ZL for soybean oil, and ZM for soybean meal. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4
Missing data filter — daily cointegrating vector β descriptive statistics.

	Missing data filter[a]			Johansen model[a]		
	ZS	ZL	ZM	ZS	ZL	ZM
Median	1	−0.11	−0.19	1	−0.13	−0.17
Average	1	−0.09	−0.21	1	−0.29	−0.02
quartile 25%	1	−0.17	−0.24	1	−0.24	−0.25
quartile 75%	1	−0.06	−0.14	1	−0.05	−0.07
StDev	0	0.28	0.28	0	0.5	0.47
CME (physical)	1	−0.11	−0.22	1	−0.11	−0.22

[a]Using daily sets of one-minute data from session 2 and the ZS open-interest rollover technique, this table records descriptive statistics associated to the daily cointegrating vectors components for cointegrated days only (based on the missing data filtered and the Johansen cointegration test). ZS stands for soybean, ZL for soybean oil, and ZM for soybean meal.

will Granger cause soyoil (positive sign of Vol_ZL/Vol_ZS coefficient in the $\tilde{\kappa}_{1,2}$ regression). We find the same interpretation for the relative contribution of soymeal relative to soyoil. If the traded volumes of soyoil relative to soybean increase, we can expect soymeal to even more significantly Granger cause the soyoil dynamics (positive sign of Vol_ZL/Vol_ZS coefficient in the $\tilde{\kappa}_{3,2}$ regression). Nevertheless, the lower the traded volumes of soymeal relative to soyoil, the less soymeal prices will Granger cause soybean prices (negative sign of Vol_ZL/Vol_ZM coefficient in the $\tilde{\kappa}_{3,1}$ regression).

5.3.2. Cointegrating vector

In this subsection, we investigate whether the cointegrating vector remains stable over time and close to the physical weights resulting from the trituration of soybeans. Furthermore, we verify that the cointegrating vector does not depend on the assets' traded volumes, as assumed in our model.

As shown in Table 4, the average and median values of the cointegrating vector components remain rather centred near the physical quantities relayed by the CME and displayed in this table. By comparison, when considering the basic Johansen's cointegration test without dealing with the markets' non-synchronicity, one can clearly notice from Table 4 that the average value is significantly biased in this case, while the standard deviation is twice the value obtained with an appropriate state–space formulation and the filtering technique described earlier.

To validate the initial hypothesis of our theoretical model, we need to demonstrate that, when there is efficient cointegration, it mainly occurs through the adjustment space and not the cointegrating space, which is preserved from the disequilibrium in traded volumes. To this end, and as for the κ vector components, we investigate whether there is any statistically significant linear relationship between the ratios of the components of the cointegrating vector β and the traded volume ratios combined with the usual control variables. The results are available on demand, but no significant relationship has been found for any of the ratios at the 1% or 5% critical levels. At the 10% critical level, the β components ratios start to be slightly affected by the relative volumes (Vol_ZL/Vol_ZM and Vol_ZL/Vol_ZS), but this concerns two pairs of β vector components whose associated $\tilde{\kappa}_{ij}$ ratios were not related to traded volume (namely $\beta_{1,3}$ and $\beta_{2,3}$).

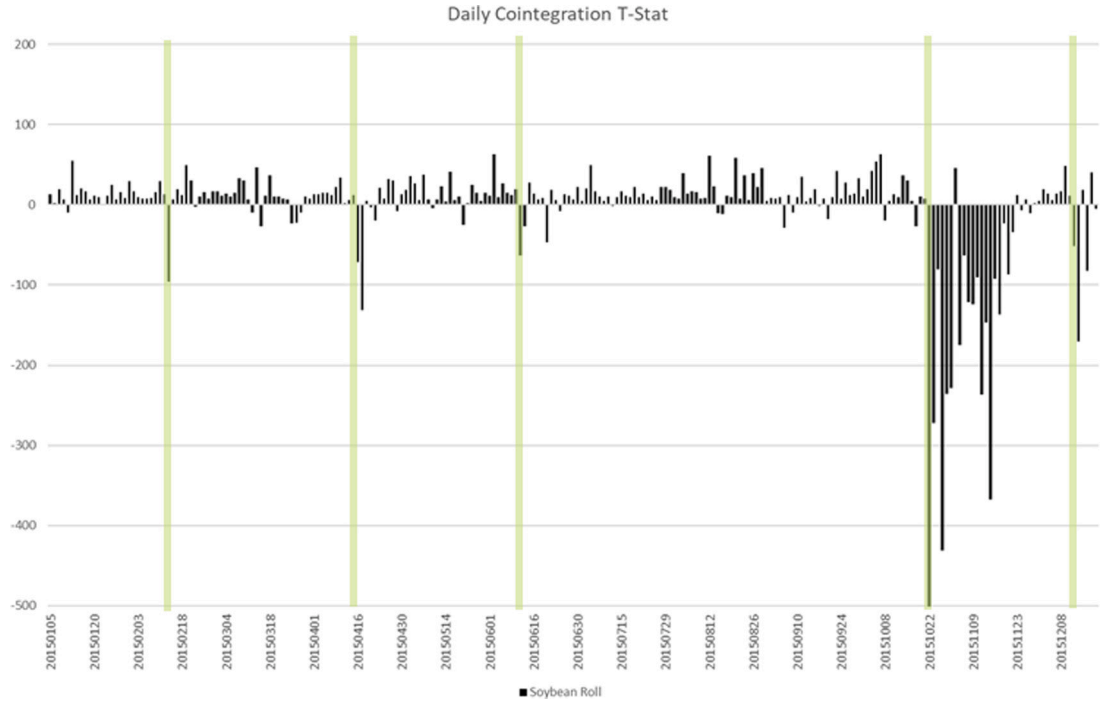


Fig. 1. Using daily sets of one-minute data from session 2 and the ZS open-interest rollover technique, this figure shows the values of the daily cointegration test statistics. T-stats beyond 7.02 mean that the cointegration is statistically significant. The light green lines represent the rolling dates.

5.4. Futures contracts rollover and cointegration relationships

Another observable consequence of the relationship between the strength of the intraday cointegration and the traded volumes associated with each market concerns the optimal rollover periods of futures contracts. As we can see in charts 1 and 2, the soybean-roll (ZS Open interest) and the month-end (Monthly Rollover) methods show particular differences before the month of October. Provided that intermediary but less-traded maturities are in place for the months of August (ZSQ5) and September (ZSU5), we should expect a conflict between these maturities and the highest open-interest contract for October (ZSX5), which trades at the same time.

We first notice that the choice of the rolling technique, described in Section 4, creates a very significant difference in cointegration strength among derivatives assets, as measured by the daily rank-test statistics. The month-end rollover approach suffers from the traded-volume weakness that characterizes the previously mentioned, less liquid intermediary maturities. The soybean-roll technique skips these contracts and directly trades the November contract (ZSX5) since it benefits from a higher open interest over the same period of time (cf. Fig. 3). This result underpins our theoretical model, which tells us that the traded volumes are key variables in understanding and modelling multi-asset joint dynamics.

Moreover, our model also states that some agents seek to enforce cointegration among assets and, to this end, build their expectations on the dynamics in the physical markets and on the fundamental or physical properties of the underlying commodities or assets. The futures with maturity in November (ZSX5) are indeed generally preferred by the crushing industry as they correspond to the new crop season in the northern hemisphere.[28]

Finally, we also noticed that the soybean-roll technique is not necessarily optimal for crush-spread hedging near the end of the year and might be improved by considering different rollover dates for each asset, which we leave for future studies.

6. Conclusion

Great efforts have been made in the academic literature to quantify the impact of commercial and non-commercial investors' behaviours on market prices by scrutinizing the Commitment of Traders (COT) report published on a weekly basis by the CFTC (for instance Fishe et al., 2014; Büyükşahin and Robe, 2014; Kang et al., 2020). However, these results are contingent on the CFTC's investors' classification and the reports' weekly frequency of publication. This paper chooses a different approach using high-frequency LOB data, whereby the typology of investors submitting market or limit orders is not pre-defined. Our approach

[28] Though maturity in May corresponds to the South American new crop season, it does not have the same effect on intermediary maturities (cf. Fig. 3).

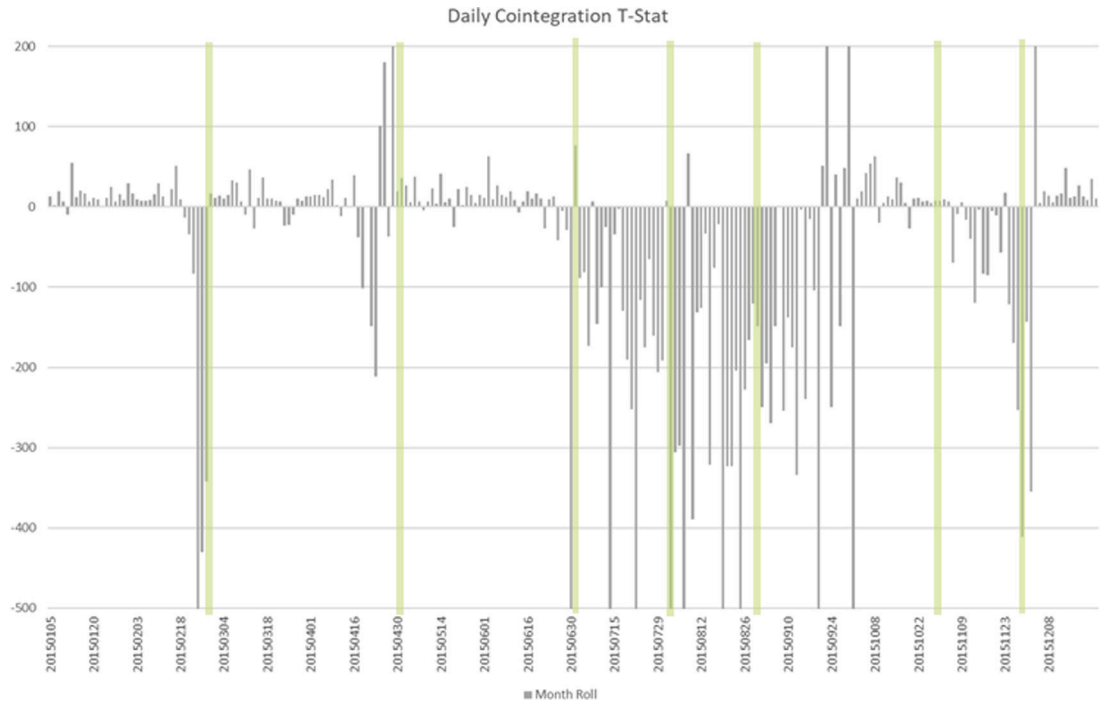


Fig. 2. Using daily sets of one-minute data from session 2 and the end-of-the-month rollover technique, this figure shows the values of the daily cointegration test statistics. T-stats beyond 7.02 mean that the cointegration is statistically significant. The light green lines represent the rolling dates.

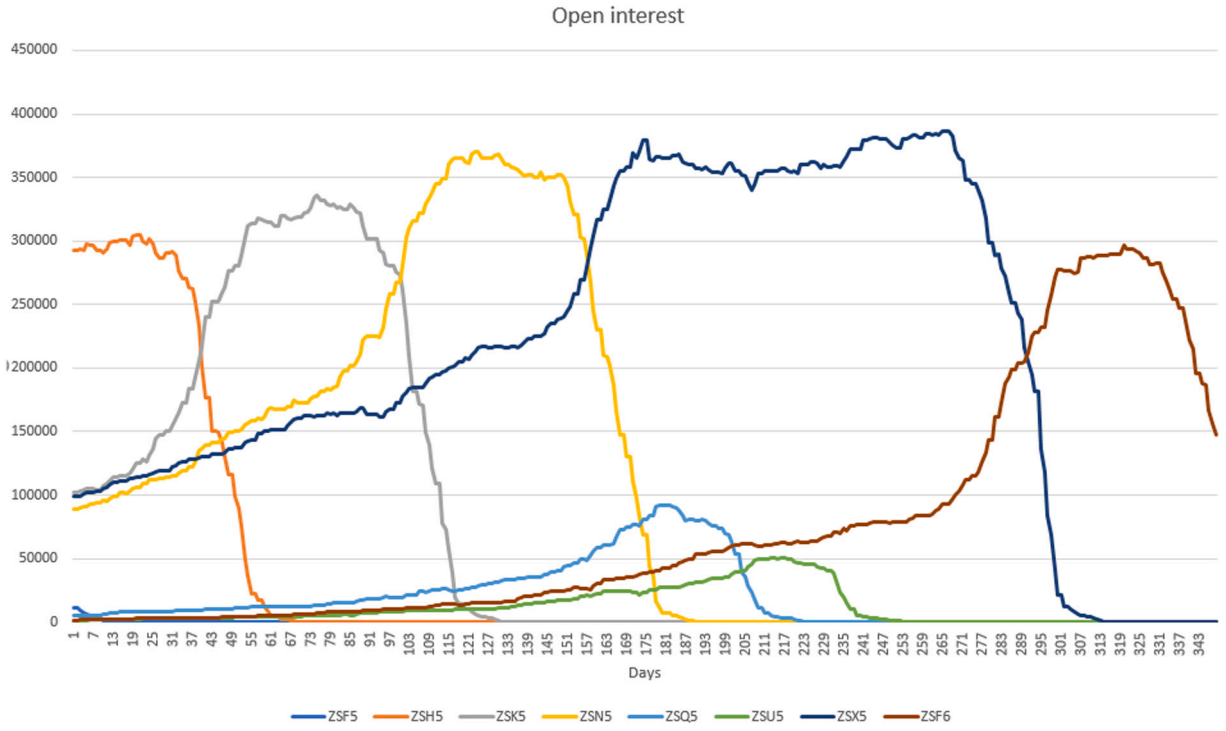


Fig. 3. This figure displays the open interest associated to each contract maturity at a given time. The respective contracts correspond to the months of January (ZSF5), March (ZSH5), May (ZSK5), July (ZSK5), August (ZSQ5), September (ZSU5), November (ZSX5) of the year 2015, while the final contract corresponds to January 2016 (ZSF6).

utilizes both traded prices and aggregated quantity data (i.e., daily traded volumes and limit order books' daily average liquidity measurements) to shed light on the relationship between price cointegration and the realized or potential daily volumes.

We specify the influence of hedgers and partially informed traders through a high-frequency price-cointegration framework in which market microstructure influences the price-discovery processes of several interrelated assets. Our market-equilibrium model demonstrates how partially informed traders, who only focus on some rather than all of these markets, may influence long-term structural relationships such as price cointegration. We show that an agent with a global view of the markets is necessary to restore the equilibrium, which raises the question of partially informed traders' capacity to enforce this equilibrium. We indeed demonstrate and observe that the traded volumes on commodity byproducts, such as soybean meal and soybean oil, positively influence the rank of the auto-regressive matrix associated with the soybean complex dynamics and hence the presence or absence of intraday cointegration among related assets. Conversely, important potential trading volumes in the main market, the soybean market in our case, or a lack of liquidity in the secondary markets, i.e., soybean meal and oil, act as a counterbalance and may discourage the fully informed traders from building correcting trades to enforce the cointegration relationship.

Furthermore, it has been empirically proven in this article that traded volumes, epitomizing disagreement on market expectations, mainly influence the speed of convergence towards the stationary cointegrated joint process, rather than the cointegrating vector itself. This finding underpins the relevance of a time-varying cointegrated relationship with respect to market liquidity, so as to model a dynamic market equilibrium among interrelated assets. Consequently, we can confirm that asset prices may deviate from the market equilibrium and that market liquidity conveys crucial information about the joint dynamics of asset values, which is complementary to the information associated with volatility measures.

From a methodological standpoint, we show that, at high-frequency granularity, filtering techniques are necessary to observe cointegration relationships and that Epps effects, microstructure noise, and idle prices significantly affect parameter estimation. Several robustness tests using different data sets and methodologies confirm our findings and support our theoretical model.

Interestingly, for the regulator and the exchanges, our paper shows that the traded volumes, the associated open interest, and the liquidity of a market should be considered within a multivariate setting. As a consequence, before issuing or authorizing any new derivative contract on a commodity, the relative sizes of interrelated markets should be taken into consideration as the latter may significantly influence the price-discovery process of the former and thus its interest for hedgers, especially when large index products include one asset and not the others. We also demonstrate that the role of traded volumes in the market's capacity to revert to equilibrium, and thus enforce the cointegration of asset prices, shrinks during electronic trading sessions. This diurnal phenomenon questions the importance of having 24-hour access to electronic markets, when it only contributes to adding noise and does not convey information about assets' fundamental values. As far as the hedgers are concerned, our last micro-economic study based on futures-contracts rollover further revealed that the presence of cointegration among assets is related to the contract maturities traded at a given time. Indeed, some intermediary contract maturities are not considered by informed traders throughout the year, thus leaving unused their capacity to counterbalance on an intraday basis the disturbing influence of partially informed traders. An interesting extension in future work could focus on considering CRRA preferences and the bid–ask spread.

CRediT authorship contribution statement

Xinquan Zhou: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software (econometrics), Writing – original draft. **Guillaume Bagnarosa:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software (econometrics), Supervision, Validation, Writing – original draft. **Alexandre Gohin:** Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Validation, Writing – original draft. **Joost M.E. Pennings:** Conceptualization, Formal analysis, Project administration, Supervision, Validation, Writing – review & editing. **Philippe Debie:** Data curation, Software (data cleaning), Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Joost Pennings reports financial support was provided by Province of Limburg in the Netherlands and Foundation Commodity Risk Management Expertise Center (CORMEC). Guillaume Bagnarosa reports financial support was provided by City of Rennes (Rennes Metropole) in France. These financial supports have been only used for market price database acquisition, as such none of the coauthors has any conflict of interest to declare.

Data availability

Data will be made available on request.

Acknowledgments

Supported by Rennes Metropole, Grant ID: 20C0653.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jcomm.2023.100314.

References

Acharya, V.V., Lochstoer, L.A., Ramadorai, T., 2013. Limits to arbitrage and hedging: Evidence from commodity markets. J. Financ. Econ. 109, 441–465.

Arzandeh, M., Frank, J., 2019. Price discovery in agricultural futures markets: Should we look beyond the best bid-ask spread? Am. J. Agric. Econ. 101 (5), 1482–1498.

Atmaz, A., Basak, S., 2018. Belief dispersion in the stock market. J. Finance 73 (3), 1225–1279.

Awokuse, T.O., Chopra, A., Bessler, D.A., 2009. Structural change and international stock market interdependence: Evidence from Asian emerging markets. Econ. Model. 26 (3), 549–559.

Baillie, R.T., Geoffrey Booth, G., Tse, Y., Zabotina, T., 2002. Price discovery and common factor models. J. Financial Mark. 5 (3), 309–321.

Bandi, F.M., Kolokolov, A., Pirino, D., Renò, R., 2020. Zeros. Manag. Sci. 66 (8), 3466–3479.

Barrett, W.B., Kolb, R.W., 1995. Analysis of spreads in agricultural futures. J. Futures Mark. 15 (1), 69–86.

Basak, S., Pavlova, A., 2016. A model of financialization of commodities. J. Finance 71 (4), 1511–1556.

Beddock, A., Jouini, E., 2020. Live fast, die Young: Equilibrium and survival in large economies. Econom. Theory 71, 961–996.

Behrendt, S., Schmidt, A., 2021. Nonlinearity matters: the stock price – trading volume relation revisited. Econ. Model. 98, 371–385.

Bessembinder, H., Seguin, P.J., 1993. Price volatility, trading volume, and market depth: Evidence from futures markets. J. Financ. Quant. Anal. 28 (1), 21–39.

Bond, P., García, D., 2021. The equilibrium consequences of indexing. Rev. Financ. Stud. 35 (7), 3175–3230.

Bradley, M.G., Lumpkin, S.A., 1992. The treasury yield curve as a cointegrated system. J. Financial Quant. Anal. 27 (3), 449–463.

Brogaard, J., Ringgenberg, M.C., Sovich, D., 2018. The economic impact of index investing. Rev. Financ. Stud. 32 (9), 3461–3499.

Brown, D.C., Davies, S.W., Ringgenberg, M.C., 2020. ETF arbitrage, non-fundamental demand, and return predictability*. Rev. Finance 25 (4), 937–972.

Brugler, J., Comerton-Forde, C., 2019. Comment on: Price discovery in high resolution. J. Financ. Econom. 19 (3), 1–8.

Buccheri, G., Bormetti, G., Corsi, F., Lillo, F., 2019. Comment on: Price discovery in high resolution. J. Financ. Econom. 19 (3), 1–13.

Buccheri, G., Bormetti, G., Corsi, F., Lillo, F., 2021a. A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics. J. Bus. Econom. Statist. 39 (4), 920–936.

Buccheri, G., Corsi, F., Peluso, S., 2021b. High-frequency lead-lag effects and cross-asset linkages: a multi-asset lagged adjustment model. J. Bus. Econom. Statist. 39 (3), 605–621.

Büyükşahin, B., Robe, M.A., 2014. Speculators, commodities and cross-market linkages. J. Int. Money Finance 42, 38–70.

Carchano, O., Pardo, A., 2009. Rolling over stock index futures contracts. J. Futures Mark. 29 (7), 684–694.

Chen, G.-M., Firth, M., Meng Rui, O., 2002. Stock market linkages: Evidence from latin america. J. Bank. Financ. 26 (6), 1113–1141.

Couleau, A., Serra, T., Garcia, P., 2019. Microstructure noise and realized variance in the live cattle futures market. Am. J. Agric. Econ. 101 (2), 563–578.

Couleau, A., Serra, T., Garcia, P., 2020. Are corn futures prices getting "Jumpy"? Am. J. Agric. Econ. 102 (2), 569–588.

Darolles, S., Le Fol, G., Mero, G., 2017. Mixture of distribution hypothesis: Analyzing daily liquidity frictions and information flows. J. Econometrics 201 (2), 367–383.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1), 1–22.

Dewachter, H., Iania, L., 2011. An extended macro-finance model with financial factors. J. Financial Quant. Anal. 46 (6), 1893–1916.

Dorfman, J.H., Karali, B., 2015. A nonparametric search for information effects from USda reports. J. Agric. Res. Econ. 40 (1), 124–143.

Duchin, R., Levy, M., 2010. Disagreement, portfolio optimization, and excess volatility. J. Financ. Quant. Anal. 45 (3), 623–640.

Epps, T.W., 1979. Comovements in stock prices in the very short run. J. Amer. Statist. Assoc. 74 (366a), 291–298.

Epps, T.W., Epps, M.L., 1976. The stochastic dependence of security price changes and transaction volumes: Implications for the mixture-of-distributions hypothesis. Econometrica 44 (2), 305–321.

Escribano, A., 2004. Nonlinear error correction: the case of money demand in the united kingdom (1878–2000). Macroecon. Dyn. 8 (1), 76–116.

Escribano, A., Mira, S., 2002. Nonlinear error correction models. J. Time Series Anal. 23 (5), 509–522.

Etienne, X.L., Irwin, S.H., Garcia, P., 2014. Bubbles in food commodity markets: Four decades of evidence. J. Int. Money Finance 42, 129–155.

Etienne, X.L., Irwin, S.H., Garcia, P., 2015. $25 Spring wheat was a bubble, right? Agric. Finance Rev. 75, 114–132.

Fan, J.H., Fernandez-Perez, A., Fuertes, A.-M., Miffre, J., 2020. Speculative pressure. J. Futures Mark. 40 (4), 575–597.

Fernandez-Perez, A., Fuertes, A.-M., Miffre, J., 2016. Commodity markets, long-run predictability, and intertemporal pricing. Rev. Finance 21 (3), 1159–1188.

Fishe, R.P.H., Janzen, J.P., Smith, A., 2014. Hedging and speculative trading in agricultural futures markets. Am. J. Agric. Econ. 96 (2), 542–556.

Foucault, T., Kozhan, R., Tham, W.W., 2017. Toxic arbitrage. Rev. Financ. Stud. 30 (4), 1053–1094.

Frank, J., Garcia, P., 2011. Bid-ask spreads, volume, and volatility: Evidence from livestock markets. Am. J. Agric. Econ. 93 (1), 209–225.

Garcia, P., Irwin, S.H., Smith, A., 2015. Futures market failure? Am. J. Agric. Econ. 97 (1), 40–64.

Goldstein, I., Yang, L., 2022. Commodity financialization and information transmission. J. Finance 77 (5), 2613–2667.

Gomber, P., Schweickert, U., Theissen, E., 2015. Liquidity dynamics in an electronic open limit order book: An event study approach: liquidity dynamics in an electronic open limit order book. Eur. Financial Manag. 21 (1), 52–78.

Gorton, G.B., Hayashi, F., Rouwenhorst, K.G., 2013. The fundamentals of commodity futures returns. Rev. Finance 17 (1), 35–105.

Greene, W.H., 2003. Econometric Analysis, fifth ed. Prentice Hall.

Hadri, K., 2000. Testing for stationarity in heterogeneous panel data. Econom. J. 3 (2), 148–161.

Hakkio, C.S., Rush, M., 1991. Cointegration: How short is the long run? J. Int. Money Finance 10 (4), 571–581.

Han, Y., Hu, T., Yang, J., 2016. Are there exploitable trends in commodity futures prices? J. Bank. Financ. 70, 214–234.

Hansen, P.R., Lunde, A., 2006. Realized variance and market microstructure noise. J. Bus. Econom. Statist. 24 (2), 127–161, URL http://www.jstor.org/stable/27638860.

Hasbrouck, J., 1995. One security, many markets: determining the contributions to price discovery. J. Finance 50 (4), 1175–1199.

Hasbrouck, J., 2019. Rejoinder on: Price discovery in high resolution*. J. Financ. Econom. 19 (3), 465–471.

He, X., Velu, R., 2014. Volume and volatility in a common-factor mixture of distributions model. J. Financ. Quant. Anal. 49 (1), 33–49.

Hong, H., Yogo, M., 2012. What does futures market interest tell us about the macroeconomy and asset prices? J. Financ. Econ. 105 (3), 473–490.

Janzen, J.P., Adjemian, M.K., 2017. Estimating the location of world wheat price discovery. Am. J. Agric. Econ. 99 (5), 1188–1207.

Johansen, S., 1995. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press on Demand.

Johnson, R.L., Zulauf, C.R., Irwin, S.H., Gerlow, M.E., 1991. The Soybean complex spread: An examination of market efficiency from the viewpoint of a production process. J. Futures Mark. 11 (1), 25–37.

Kang, W., Rouwenhorst, K.G., Tang, K., 2020. A tale of two premiums: The role of hedgers and speculators in commodity futures markets. J. Finance 75 (1), 377–417.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53 (6), 1315–1315–1335.

Larsson, R., Lyhagen, J., Löthgren, M., 2001. Likelihood-based cointegration tests in heterogeneous panels. Econom. J. 4 (1), 109–142.

Li, Z., Hayes, D.J., 2022. The hedging pressure hypothesis and the risk premium in the Soybean reverse crush spread. J. Futures Mark. 42 (3), 428–445, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/fut.22285.

Liu, Q.W., Sono, H.H., 2016. Empirical properties, information flow, and trading strategies of China's Soybean crush spread. J. Futures Mark. 36 (11), 1057–1075.

Lo, A.W., MacKinlay, A.C., 1990. An econometric analysis of nonsynchronous trading. J. Econometrics 45 (1), 181–211.
Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer, New York, Berlin.
Marowka, M., Peters, G.W., Kantas, N., Bagnarosa, G., 2020. Factor-augmented Bayesian cointegration models: A case-study on the Soybean crush spread. J. R. Stat. Soc. Ser. C. Appl. Stat. 69 (2), 483–500.
Mitchell, J., 2010. Soybean futures crush spread arbitrage: trading strategies and market efficiency. J. Risk Financial Manag. 3 (1), 63–96.
O'Hara, M., 2015. High frequency market microstructure. J. Financ. Econ. 116 (2), 257–270.
Rechner, D., Poitras, G., 1993. Putting on the crush: Day trading the Soybean complex spread. J. Futures Mark. 13 (1), 61–75.
Seong, B., Ahn, S.K., Zadrozny, P.A., 2013. Estimation of vector error correction models with mixed-frequency data. J. Time Series Anal. 34 (2), 194–205.
Shang, Q., Mallory, M., Garcia, P., 2018. The components of the bid-ask spread: Evidence from the corn futures market. Agric. Econ. 49 (3), 381–393.
Shumway, R.H., Stoffer, D.S., 1982. An approach to time series smoothing and forecasting using the EM algorithm. J. Time Series Anal. 3 (4), 253–264.
Simon, D.P., 1999. The Soybean crush spread: Empirical evidence and trading strategies. J. Futures Mark. 19 (3), 271–289.
Tauchen, G.E., Pitts, M., 1983. The price variability-volume relationship on speculative markets. Econometrica 51 (2), 485–505.
Tjøstheim, D., 2020. Some notes on nonlinear cointegration: A partial review with some novel perspectives. Econometric Rev. 39 (7), 655–673.
Trujillo-Barrera, M.M., Garcia, P., 2012. Volatility spillovers in U.S. crude oil, ethanol, and corn futures markets. J. Agric. Res. Econ. 37 (2), 247–262.
Williams, J.C., Wright, B.D., et al., 1991. Storage and Commodity Markets. Cambridge University Press.

Further reading

Banerjee, A., Marcellino, M., Osbat, C., 2004. Some cautions on the use of panel methods for integrated series of macroeconomic data. Econom. J. 7 (2), 322–340.
Bierens, H.J., Martins, L.F., 2010. Time-varying cointegration. Econom. Theory 26 (5), 1453–1490.
Christensen, K., Hounyo, U., Podolskij, M., 2018. Is the diurnal pattern sufficient to explain intraday variation in volatility? A nonparametric assessment. J. Econometrics 205 (2), 336–362.
Engle, R.F., Sokalska, M.E., 2012. Forecasting intraday volatility in the US equity market. multiplicative component GARCH. J. Financ. Econom. 10 (1), 54–83.
Gonzalo, J., Granger, C., 1995. Estimation of common long-memory components in cointegrated systems. J. Bus. Econom. Statist. 13 (1), 27–35.
Harris, L., 1986. A transaction data study of weekly and intradaily patterns in stock returns. J. Financ. Econ. 16 (1), 99–117.
Hu, Z., Mallory, M., Serra, T., Garcia, P., 2020. Measuring price discovery between nearby and deferred contracts in storable and nonstorable commodity futures markets. Agric. Econ. 51 (6), 825–840.
Koop, G., Leon-Gonzalez, R., Strachan, R.W., 2011. Bayesian inference in a time varying cointegration model. J. Econometrics 165 (2), 210–220.
Vollmer, T., Herwartz, H., von Cramon-Taubadel, S., 2020. Measuring price discovery in the European wheat market using the partial cointegration approach. Eur. Rev. Agricu. Econ. 47 (3), 1173–1200.