NCCC-134
APPLIED COMMODITY PRICE ANALYSIS, FORECASTING AND MARKET RISK MANAGEMENT

Microstructure and High-Frequency Price Discovery in the Soybean Complex

by

X. Zhou, G. Bagnarosa, A. Gohin, J. Pennings, and P. Debie

Suggested citation format:
Zhou, X., G. Bagnarosa, A. Gohin, J. Pennings, and P. Debie. 2022. "Microstructure and High-Frequency Price Discovery in the Soybean Complex." Proceedings of the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. [http://www.farmdoc.illinois.edu/nccc134].

Microstructure and High-Frequency Price Discovery in the Soybean Complex

Zhou X. (DCU), Bagnarosa G.(RSB and INRAE), Gohin A.(INRAE), Pennings J. (WUR), and Debie P. (WUR) *

Paper presented at the NCCC-134 Conference on Applied Commodity Analysis, Forecasting, and Market Risk Management, 2022.

* Corresponding Author. Dublin Business School, Dublin City University, Glasnevin, Dublin 9, Ireland. Email address: xinquan.zhou5@mail.dcu.ie

Microstructure and High-Frequency Price Discovery in the Soybean Complex

Abstract

We develop a theoretical framework and propose a relevant empirical analysis of the soybean complex prices cointegration relationship in a high-frequency setting. We allow for heterogeneous expectations among traders on the multi-asset price dynamics and characterize the resulting market behavior. 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.

Keywords: Soybean; Futures market microstructure; Liquidity; Price discovery; High-Frequency

1 Introduction

Financial markets offer the opportunity for a wide variety of economic agents to express their economic expectations. The resulting price-discovery process in these markets reflects the agents' respective levels of information and investment capacities. In a sense, Warren Buffett, CEO of Berkshire Hathaway, reaches the same conclusion, quipping that 'when you combine ignorance and leverage, you get some pretty interesting results'. In the long run, we would expect less informed traders to leave the market due to poorer investment performance. Conversely, we would expect greater diversity among economic agents on a short to mid-term horizon.

Using a novel price cointegration framework, we examine how the short-term market microstructure influences the price-discovery processes of multiple related assets, with an application to the soybean complex. We extend this 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. Our theoretical microstructure model establishes the link between information heterogeneity, the assets' traded volumes, and the strength of the cointegration relationship at high-frequency level. Subsequently, using the high-frequency Limit Order Book (LOB) data of the soybean complex, our empirical study confirms this theoretical model by revealing a relationship between price cointegration and the traded or available volumes in each individual asset's LOB.

At the empirical level, 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 & 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 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 daily

realised or potential volumes[1].

In the financial microeconomics literature, the types of investors are often differentiated according to their respective levels of information and risk aversion. This literature has mainly focused on the agricultural and energy commodities markets, where a wide spectrum of investor profiles, and thus information levels, generally interact. More precisely, partially informed or uninformed traders rub shoulders with arbitragers and manufacturers in these markets every day. While less informed traders are generally assumed to be genuinely uninformed (when studied in a single market), these investors will be considered partially uninformed in our model. By this we mean that they focus on a single asset without considering assets linked by cointegration relationships. We could typically associate to this investor profile 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. The well-known GSCI index, for instance, only invests in soybeans futures contracts[2] but does not include soybean meal and soybean oil contracts[3]. As a result, this index is 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. At best, such behaviour leads to short-term asynchronicity among the cointegrated markets; at worst it could lead to market inconsistencies, such as non-synchronous financial bubbles or even the oft-decried financialization of commodities markets (Basak & Pavlova 2016, Shang et al. 2018).

The second group of investors considered here are the arbitragers. They play an

[1]Potential volumes are reflected by the depth of the LOB and defined by Albert S. 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]See also the S&P GSCI Index methodology document available from:
https://www.spglobal.com/spdji/en/documents/methodologies/methodology-sp-gsci.pdf

[3]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.

important role in the derivatives industry: 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, consequently 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 usually remains within the arbitrager's risk capacity (Hong & Yogo 2012, Acharya et al. 2013). Only the last group of investors is likely to consider joint equilibrium relationships among multiple underlying assets in the physical markets: manufacturers. Unlike the other groups of investors, the manufacturers, through their commercial activities, 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 & Hayes 2022).

This activity at high frequency should ultimately coordinate futures prices and make their cointegration materialize over a longer time horizon, such as a trading day. However, observing this cointegration relationship in a high-frequency setting turns out to be challenging, given that microstructure noise[4] as well as lagged information among agents and markets disturb the latent joint price-discovery process (Janzen & Adjemian 2017, Couleau et al. 2019). This makes price cointegration inference difficult for the econometrician. 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 Maximisation algorithm proposed by Dempster et al. (1977) to cope with mixed-frequency or asynchronous data in cointegrated time-series models. More recently, Buccheri, Corsi & Peluso (2021) demonstrated that this filtering methodology adequately deals with microstructure noise and the information lag that exists among markets at the high-frequency level[5]. Employing a slightly

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

[5]It is worth noting that as far as the price discovery process is concerned, other methods have been considered in the economic literature to deal with prices staleness. For instance Janzen & Adjemian (2017) use a combination of the respective assets information and component shares as proposed by

different approach, our filtering model deals with the problems of price staleness and idleness[6]. Our empirical study demonstrates that the soybean complex is significantly cointegrated and close to the underlying physical relationships. To 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 in the assets prices multivariate dynamics, in particular for cointegrated assets. This confirms Arzandeh & Frank (2019) findings which emphasize the interest of considering LOB information in price discovery process. To validate our cointegration framework, we indeed demonstrate that, contrary to the cointegrating space, the crush-spread associated adjustment space turns out to be closely related to market microstructure.

This paper is organised as follows: The second section is devoted to the theoretical framework, which highlights the potential drivers of price cointegration. This theoretical framework is rooted in the large Mixture of Distribution Hypothetis (MDH) and Difference of Opinion Literature. Starting from Epps & Epps (1976), our main contribution is to develop a multi-market analysis with many potential participants. Consistent with the synthesis of Behrendt & 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. Furthermore, we contribute to the methodological literature by treating differently price staleness and idleness as defined by Bandi et al. (2020). In addition, we propose an adapted panel test to verify the intra-daily cointegration among asset prices in a high-frequency setting. The fourth section provides a description of the soybean complex and the retained data. Finally, our results are commented and analysed in the fifth section, where we contrast our main cointegration results with those obtained through traditional methodologies that ignore issues associated with multivariate high-frequency data set analysis (e.g. microstructure noise and non-synchronicity). We empirically test the relationship between volume and the strength of the cointegration, add more potential

Yang et al. (2003).

[6]Staleness is defined according to Bandi et al. (2020) as the frequency of zero returns with a traded volume associated whereas idleness corresponds to staleness but when trading activity is absent.

explanatory variables, and conduct several robustness checks. Our empirical results show that volume in certain markets, especially soymeal, is significantly related to crush-spread cointegration. We demonstrate how this relationship may affect the hedging efficacy of different rollover approaches and provide a conclusion in the final section.

2 Theoretical Framework

Since the seminal papers of Epps & Epps (1976) and Tauchen & Pitts (1983), a large and still living literature strives to model the relationship between time series of financial prices and traded volumes[7]. 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' typology.

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 through 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 & Pitts (1983), the economic agents disagree in a linear manner on the expected price of one commodity; by contrast, Epps & Epps (1976) formulate a non-linear disagreement function around the expected price. Later, He & Velu (2014) extend the linear approach of Tauchen & 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 factors

[7]Please refer to Behrendt & Schmidt (2021) for a literature review.

loadings. However, in their model He & Velu (2014) do not consider the effect of belief heterogeneity among participants while, as shown by Duchin & Levy (2010) through simulations, disagreement on expected prices or on the expected price variance-covariance matrix has a significant impact on assets price and traded volume. In the same vein, recent papers introduce market frictions (Darolles et al. 2017), heterogeneous discount factors (Beddock & Jouini 2020), or a continuum of economic agents (Atmaz & 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.

Our theoretical framework builds on the Epps & Epps (1976) framework. We consider $i = 1, \ldots, n$ commodity markets and $j = 1, \ldots, m$ 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 the agents' j 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 (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 :

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

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 & 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 & 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, e.g. 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

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

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 & Epps (1976) and appears as a convenient price normalisation. A complementary economic interpretation for storable commodity markets is that economic actors, while forming new price expectations after receiving new information, take into account the current market situation. For instance, when current prices are relatively low compared to historical prices, economic actors can 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 function of the current price's deviation from a long-term cointegration relationship. For instance, if certain actors, e.g. 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.

Whereas in our empirical study of the soybean crush spread we will assume that three partially informed agents trade the individual markets and the manufacturer intervenes simultaneously in the three related commodities (bean, oil and meal) meaning that n=3 and m=4. For the sake of notational clarity, in the following theoretical demonstration we will consider a simpler case with two commodities (n=2) and three agents (m = 3). The dimension extension of this model to the case of the soybean crush is straightforward. Agent 1 is operating on both markets, while agent 2 only intervenes in the first market and agent 3 only in the second, exactly as partially informed market operators, such as index traders, might behave. In this simplest scenario, we assume that the initial market situation is different from the long-term price equilibrium, which, in the case of cointegrated time series, can be represented by the product of cointegrating vector β and the time $t-1$ market prices P_{t-1}. Furthermore, we assume that only agent 1 receives and processes new information. Due to partial information, the other two agents will not change their price expectations. Accordingly, the disagreement among agents' price expectations as expressed in eq. (5) increases and then depends on the long-term price relationship, which is only known to agent 1. The following specification captures this market configuration:

$$\alpha_1(P_{t-1}) = \begin{bmatrix} \frac{1}{\phi_1 \beta' P_{t-1}} & 0 \\ 0 & \frac{1}{\phi_2 \beta' P_{t-1}} \end{bmatrix}, \quad \alpha_2(P_{t-1}) = \begin{bmatrix} -\frac{1}{\phi_1 \beta' P_{t-1}} & 0 \\ 0 & 0 \end{bmatrix},$$

$$\text{and} \quad \alpha_3(P_{t-1}) = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{1}{\phi_2 \beta' P_{t-1}} \end{bmatrix} \tag{7}$$

whereby ϕ_1 and ϕ_2 are positive random variables, both with expected values of one. We check that $\sum_{i=1}^{3} \alpha_i = 0$. With only two agents in each market, there is only one disagreement per market (which amounts to $\frac{1}{\phi_1 \beta' P_{t-1}}$ in the first market and $\frac{1}{\phi_2 \beta' P_{t-1}}$ in the second market). It should be clear that the term $ABS(\hat{\delta})$ in eq. (5) allows our first agent to be long or short, and to have lower or higher price expectations.

Having interpreted the changes in price expectations according to Epps & Epps (1976), we should now solve our model using the generalized $\alpha(P_{t-1})$ formulation, bearing in mind that long-term price cointegration could be included in this formulation. Summing over

all j equations (4), using (5) and the market equilibrium condition, we find :

$$\sum_j V_j = 0 \tag{8}$$

$$= \sum_j \left[\lambda_j \hat{\delta} + \lambda_j \alpha_j (P_{t-1}) ABS(\hat{\delta})^{1/\gamma} - \lambda_j \Delta P\right] \tag{9}$$

so

$$\Delta P = \hat{\delta} + \left(\sum_j \lambda_j\right)^{-1} \sum_j \lambda_j \alpha_j (P_{t-1}) ABS(\hat{\delta})^{1/\gamma} \tag{10}$$

This equation (10) is the generalized version of the Equation (20) in Epps & Epps (1976), where they assume $m = 2$, $\alpha_1 = -\alpha_2$ and $\lambda_1 = \lambda_2$. In this particular case, the second term to the right hand side is zero, and the price change is simply given by the average of the new price expectations. In our more general setting, the price change also depends on the m matrices λ, which include the risk aversion (or investment capacity) as well as the individual expected variance-covariance matrix for each economic agent intervening in the markets.

In the two-agent economy considered by Epps & Epps (1976), the volume of the market equals the volume of both agents because, if one agent is short, the other one has to be long. Proceeding to our more general economy, let us assume that one economic agent (e.g. agent 1, a risk-averse manufacturer) is participating in all markets, making λ_1 invertible. Then, from equation (4), we obtain:

$$\lambda_1^{-1} V_1 = \hat{\delta} + \alpha_1 (P_{t-1}) ABS(\hat{\delta})^{1/\gamma} - \Delta P \tag{11}$$

and, using equation (10), it can be simplified as:

$$\lambda_1^{-1} V_1 = \left[\alpha_1 (P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1} \sum_j \lambda_j \alpha_j (P_{t-1})\right] ABS(\hat{\delta})^{1/\gamma} \tag{12}$$

We thus obtain a relationship between the average change of price expectations and the volumes traded by agent 1:

$$\hat{\delta} = \left[\alpha_1 (P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1} \sum_j \lambda_j \alpha_j (P_{t-1})\right]^{-\gamma} \left(\lambda_1^{-1} V_1\right)^{\gamma} Diag\left(Sgn(\hat{\delta})\right) \tag{13}$$

We finally combine this last equation with equation (10) to write the relationship between asset prices changes and traded volumes as follows:

$$\Delta P = \Omega(V_1)^{\gamma} Diag\left(Sgn(\hat{\delta})\right) + \left(\sum_j \lambda_j\right)^{-1} \sum_j \lambda_j \alpha_j (P_{t-1}) \Omega(V_1) \tag{14}$$

where:

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

Equation (14) generalises the price-change equation (21) obtained by Epps & 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.

Proposition 1:

Let's assume a two-commodity setting, where three investors are characterised by different levels of risk aversion and/or different variance-covariance matrices are forecast, while retaining $\gamma = 1$, as well as the particular specifications 7 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 $\mathbf{\Pi}^\star(V_1)$, which is a function of the volume traded by agent 1, is low-rank in the following expression (for a demonstration, see Supplementary Material A):

$$\begin{aligned} \Delta P &= \alpha_1(P_{t-1})^{-1}\mathbf{A}V_1 + \mathbf{B}V_1 \\ &= \mathbf{\Pi}^\star(V_1)P_{t-1} + \mathbf{B}V_1 \end{aligned} \tag{15}$$

where:

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

where I_2 is an identity matrix of dimension 2×2 and β and $\kappa(V_1) = \mathbf{\Phi}\mathbf{A}V_1$ two low-rank matrices of dimension (2×1)[9]. This means that, if we assume the cointegrating vector

[9] 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.

β to be stable over time, both elements of the vector κ, denoted respectively κ_1 and κ_2, 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 $\mathbf{\Pi}$ could itself be function of traded volumes without being low-rank, which would mean that the auto-regressive matrix of the asset prices would be 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 $\mathbf{\Pi}$ is related to volumes. Put differently, the rank of matrix $\mathbf{\Pi}$ - which determines whether or not a cointegration relationship exists - and the low-rank matrix κ must both be 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 equation (15). 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 as well. Equation (15) thus simplifies to:

$$\Delta P = \alpha_1 (P_{t-1})^{-1} Diag\left(Sgn(\hat{\delta})\right) \lambda_1^{-1} V_1 \tag{16}$$

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. 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, once demonstrated the stability over time of the cointegrating vector β. 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 very 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 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 & 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). This paper, on the other hand, focuses on cointegration relationships stemming from the physical characteristics of each asset, such as the relationship between a given commodity and its by-products, where no genuine arbitrage gain is to be expected. The supply and demand disequilibrium associated with the commodity itself or its by-products could indeed consistently or temporarily change the associated spread levels. This rich strain of literature does, 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), Buccheri, Corsi & Peluso (2021) and Buccheri, Bormetti, Corsi & Lillo (2021)[10]. Initially proposed

[10]For the sake of completeness, a potential missing-data modification for their algorithm is mentioned in the technical appendix of Buccheri, Corsi & Peluso (2021).

by Shumway & 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), where observation noise is added to cope with microstructure noise, as described by Buccheri et al. (2019), Buccheri, Corsi & Peluso (2021) and Buccheri, Bormetti, Corsi & Lillo (2021). 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, allowing us to tackle in a different manner the inference biases stemming from the price idleness described in Bandi et al. (2020).

3.1 VECM State-Space Representation

Let us assume h cointegration relationships among n non-stationary financial assets prices; we will denote P_t the n dimension row vector of the asset prices at time t; then equation (15) can be written as the following vector error correction model (VECM):

$$\Delta P_t = c_0 + \Pi P_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta P_{t-j} + e_t \tag{17}$$

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 equation (15), denoted as c_0 here, can be removed and included as an intercept term in the cointegration relationships, as demonstrated in Lütkepohl (2005)[11]. This VECM formulation can thus be equivalently written as a VAR(p) model, such that (Lütkepohl 2005):

$$P_t = \sum_{j=1}^{p} \Phi_j P_{t-j} + e_t \tag{18}$$

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

Following Buccheri et al. (2019), Buccheri, Corsi & Peluso (2021) and Buccheri, Bormetti, Corsi & Lillo (2021), we assume that this discretized multivariate dynamics is

[11] As demonstrated in Lütkepohl (2005), if we keep the constant in equation (17), we should then constrain it for the model estimation, such that $c_0 = -\kappa\beta'\mu_0$, where μ_0 is the adjusted constant. In this way, we avoid generating a linear trend in the mean of P_t.

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 (18):

$$x_t = Fx_{t-1} + Ge_t \tag{19}$$

where $x_t = (P'_t, P'_{t-1}, \ldots, P'_{t-p+1})'$ 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:

$$y_t = H_t x_t + w_t \tag{20}$$

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 $q \times n$ observation design matrix, which converts the unobserved $n \times 1$ vector x_t into the $q \times 1$ imperfectly observed series y_t. This observation equation is different from the one proposed by Buccheri et al. (2019), Buccheri, Corsi & Peluso (2021) and Buccheri, Bormetti, Corsi & Lillo (2021), where the matrix $H_t = I_n$. In our state-space model, we thus distinguish the situation where one of the assets has simultaneously traded with the others or not. With $H_t = I_n$, the measurement error associated with an idle price is first assumed to be a 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 could not be considered price discreteness but conveying information closely related to the traded volumes (Bandi et al. 2020). Our empirical study buttresses this conclusion, showing that cointegration results differ significantly

depending on whether we assume that the matrix $H_t = I_n$ 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[12].

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 Maximisation 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:

$$\begin{aligned}\log \mathcal{L}(\theta; x_{1:T}, y_{1:T}) = &-\frac{1}{2}\log|\Lambda| - \frac{1}{2}(x_0 - \delta)'\Lambda^{-1}(x_0 - \delta) \\ &- \frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{t=1}^{T}(Ax_t - \Gamma B x_{t-1})'\Sigma^{-1}(Ax_t - \Gamma B x_{t-1}) \\ &- \frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T}(y_t - H_t x_t)'R^{-1}(y_t - H_t x_t) \end{aligned} \tag{21}$$

where:

$$A = \begin{bmatrix} I_n & -I_n & O_{n\times(np-2n)} \end{bmatrix};$$

$$\Gamma = \begin{bmatrix} \kappa & \Gamma_1 & \Gamma_2 & \dots & \Gamma_{p-1} \end{bmatrix};$$

and

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

[12]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 the Supplementary Material B, 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.

with a normalized $\beta = [I_h \ \beta_0']$, β_0 being the $(n-h) \times h$ matrix to be estimated[13], and $x_0 \sim N(\delta, \Lambda)$. The EM algorithm then consists in a two-step recursive procedure[14]:

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 $y_{1:T}$:

$$Q(\theta|\theta^l) = E_l\{log\mathcal{L}(\theta; x_{1:T}, y_{1:T}|y_{1:T})\} \tag{22}$$

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 Maximisation step: we maximize this conditional expectation of the complete-data log likelihood using the analytical gradient[15] 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 & Stoffer 1982) that we use to determine at each iteration when the algorithm should be stopped.

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:

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

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

[14] The Supplementary Material C provides a detailed description of the algorithm

[15] A detailed derivation of the gradient is provided in the Supplementary Material C

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:

$$\begin{aligned}\lambda_{LR}(r_0, r_1) &= -2 \log \left[\frac{sup_{\theta^{\star}_{r_0}} \mathcal{L}}{sup_{\theta^{\star}_{r_1}} \mathcal{L}} \right] \\ &= -2 \left[log\mathcal{L}(r_1) - log\mathcal{L}(r_0) \right] \end{aligned} \tag{23}$$

With a preliminary panel stationarity test, we ensure all the asset prices follows unit root processes, hence $rank(\Pi) < n$. Then we considered in our empirical study the likelihood ratio test statistic (23), with $r_0 = 0$ and $r_1 = 1$, to assess the p-value for a rank of Π equal to 0[16]. 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 & Poitras 1993, Simon 1999, Mitchell 2010, Liu & Sono 2016, Marowka et al. 2020, Li & Hayes 2022). This spread is often studied for its presence of cointegrated multivariate time series[17] 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 con-

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

[17] Based on long-term time series, Simon (1999), Mitchell (2010), and Liu & 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.

sidered, such as interest rates (Bradley & Lumpkin 1992, Dewachter & 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 equation (17), 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 database investigated covers the total trading activity of 2015, amounting to 243 trading days. The database is stored as a sequence of messages, whereby each message has a millisecond-resolution timestamp and contains 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.

Using these messages, the order book can be reconstructed, and a time series can be generated choosing any 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[18]. Moreover, we distinguish two periods within a trading day: the electronic trading session from 7PM to 7.45AM (session 1) and the market trading session from 8.30AM to 1.20PM (session 2), during which most of the trades take place. 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).

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

[18]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 available at the first second of each minute, those available at the 30th second are used. For this last data sample (30s Monthly Rollover), the Monthly Rollover has been retained as rolling technique.

deviates from the mid price. Equation (26) shows how to calculate the XLM:

$$XLM_{B,t}(V) = 10,000\frac{P_{B,t}(V) - MP_t}{MP_t} \tag{24}$$

$$XLM_{S,t}(V) = 10,000\frac{MP_t - P_{S,t}(V)}{MP_t} \tag{25}$$

$$XLM_t(V) = XLM_{B,t}(V) + XLM_{S,t}(V) \tag{26}$$

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.

At any point in time, there are always multiple contracts available for trading with a separate open interest. 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 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 & 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 contract's largest open-interest criteria (Carchano & Pardo 2009). This method ensures the rollover of all time series at the same time, while soybean is selected for being the largest contract in the soybean complex. The second rolling technique is an independent open-interest rollover (Independent Rollover), where each contract's open-interest crossover triggers the associated position roll. The final rolling technique is the monthly rollover (Monthly Rollover), which has been applied in most of the existing literature (Frank & Garcia 2011, Trujillo-Barrera & Garcia 2012, Gorton et al. 2013, Etienne et al. 2014, 2015, Dorfman & Karali 2015, Han et al. 2016, Fernandez-Perez et al. 2016, Fan et al. 2020). With this rolling technique, the current position 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 optimal rolling techniques designing.

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 apply it to the 243 trading days in 2015 for which we observe 1-minute data samples. According to table 1, based on calendar order[19], 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.

[19]Other ordering variables for the panels construction have been considered, for instance relative to daily traded volumes, the stationary hypothesis 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 critical values in one tail test are 1.282 for the 10% significance level, 1.645 for the 5% significance level, 2.326 for the 1% significance level. * p <0.10, ** p <0.05, *** p <0.01.

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 disturbs significantly the estimation of the dependence structure among time series (Lo & 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 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 lead effect of a specific asset on the other lagged assets, we applied an asset-based matching algorithm. This method consists in matching the last trading price of either the soybean, the soybean oil or the soybean meal (denoted respectively ZSmatch, ZLmatch and ZMmatch in the following) 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,

Table 2: Cointegration Test for Session 2 - 1 minute 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.

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 & Rush 1991), we increased the size of the sample 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 in comparison with the EM-algorithm approach. For sake of robustness we verified that these results are not affected by the various rolling techniques and the data-frequency choices[20]. Furthermore, we discovered the presence of a diurnal effect with regard to cointegration. A strong cointegration is indeed observed during session 2 trading hours which fades away during session 1[21]. Another interesting result is the stronger intraday cointegration we observe in average on USDA announcement days, although the number of observations available is limited.

[20] A description of the tests results is provided in the supplementary material E

[21] The results are provided in the supplementary material E.1

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 $\mathbf{\Pi} = \boldsymbol{\kappa}'\boldsymbol{\beta}^*$ in equation (17) 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 exist. To verify that the presence or absence of cointegrated time series is function of the traded volume, we analyze 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 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 logit stepwise regression with 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 & Seguin 1993), we propose a set of nine control variables stemming from high-frequency literature. This includes assessments of the average intraday realized variance, 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 tradeable 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.

The GLM stepwise regression results displayed in table 3 show that the daily volumes are significant in explaining the rank of matrix $\mathbf{\Pi}$ and thus the cointegration process in markets. It is worth highlighting that the bipower variation measure is statistically insignificant in the GLM stepwise regression. This result supports the notion that the information content regarding price volatility and traded volumes should not be mistakenly confounded when analyzing multivariate high-frequency price dynamics.

Moreover, the daily traded volumes for soymeal and soyoil tend to significantly and positively impact the rank of the $\mathbf{\Pi}$ matrix, meaning that higher traded volumes lead to a non-zero rank matrix and thus cointegration relationship. With regard to market efficiency and cointegration, the volume or liquidity of soybean are not necessarily the most important variables to monitor for exchanges, but rather the volumes of the byproducts[22].

Furthermore, the interpretation of these results, based on the limit-of-arbitrage theory, would be that the industrial agents who enforce time-series cointegration in our model have limited capacity for arbitrage and thus no capacity to intervene on a daily basis. The significance and the sign of the XLM_ZS index regression coefficient shed light on the relationship between asset price cointegration and the market's potential liquidity: when potential trading activity in the soybean market is noticeably promising (i.e. a low XLM_ZS index), cointegration tends also to be weaker or even absent. In other words, if arbitragers expect a high volume to be traded on the bean, relative to the byproduct, they prefer to stay out of the market, rather than bear the risk of an intraday widening of the soybean crush spread.

[22]As a robustness check, we also investigate in the Supplementary Material E.2 if the sampling frequency could have any impact on our results. As we proceeded to the GLM stepwise regression with 2-minutes data set, the conclusions turned out to be the same, that is, daily volumes significantly influence the cointegration process.

Table 3: Stepwise GLM Regression of Matrix Π Rank

Regressor	Coef[a]	P-Value	Marg.Eff.[b]
RV_ZS			
BV_ZS			
RV_ZL			
BV_ZL			
RV_ZM			
BV_ZM			
ZS_vol			
ZL_vol	4.5E-05	0.0224**	0.0001
ZM_vol	8.5E-05	0.0009***	0.0001
XLM_ZS	0.1397	0.0358**	0.1379
XLM_ZL			
XLM_ZM			
R2	0.27		
Adj-R2	0.26		

[a] Using daily sets of one-minute data from session 2 and the ZS open interest rollover technique, this table records AIC-based stepwise GLM regressions of the daily cointegrated/non-cointegrated binary variable 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. With * p <0.10, ** p <0.05, *** p <0.01.

[b] In this table, we display only the significant GLM coefficients using a logit regression with binomial distribution and we calculate the associated marginal effects as proposed by Greene (2003) and defined as:

$$\Lambda\left(\mathbf{x}'\beta\right) = \frac{e^{\mathbf{x}'\beta}}{1+e^{\mathbf{x}'\beta}}$$

$$\frac{d\Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)}{d\left(\mathbf{x}'\boldsymbol{\beta}\right)} = \frac{e^{\mathbf{x}'\beta}}{\left(1+e^{\mathbf{x}'\beta}\right)^2} = \Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\left[1-\Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\right]$$

$$\frac{\partial E[y \mid \mathbf{x}]}{\partial \mathbf{x}} = \Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\left[1-\Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right)\right]\boldsymbol{\beta}$$

where x is the regressor's vector and β is the regression coefficient vector.

As a complement to the linear GLM approach we propose a set of panel cointegration tests which allow to capture non-linear relationships. The panel cointegration test

proposed by Larsson et al. (2001)[23] has been used for our high-frequency study, relating this test to 8 panels on the different variables of interest. For instance, to study the non-linear impact of soybean traded volume on the cointegration test statistics, we ranked the 243 daily traded volumes for the contract ZS, created 8 sub-samples of 30 volume data points each (the last sample got 33 points), and applied the panel cointegration test to the associated cointegration results that we obtained from daily 1-minute asset prices. Table 4 shows the associated test statistics for each panel. This ranking of the data before analysing the panel allowed us to study the stability of the relationships among the different panels associated to the distribution of a given variable.

Table 4: ZS Open Interest Rollover Larson Panel Cointegration t-statistic

30 days panel[a]	1	2	3	4	5	6	7	8
Calendar order	33.01***	37.10***	16.66***	25.38***	54.22***	40.48***	-234.21	-178.86
ZS volume order	-267.69	-142.07	30.69***	47.89***	35.51***	42.10***	30.61***	41.44***
ZM volume order	-475.57	-5.29	36.36***	56.58***	42.95***	42.40***	55.07***	75.14***
ZL volume order	-470.16	9.55***	16.07***	33.30***	56.91***	60.37***	43.76***	77.75***
ZS BV order	-96.14	-64.66	-170.61	-9.45	38.64***	20.23***	40.63***	54.96***
ZM BV order	-282.20	-33.54	20.27***	-9.12	-5.14	23.25***	52.53***	53.71***
ZL BV order	34.74***	-93.31	-119.86	12.16***	-44.18	-67.94	65.08***	20.23***

[a] This table records the panel cointegration t-stat per 30 days panel using cointegration results based on the high-frequency filtering method and daily one-minute data from session 2. We first sort the daily t-stat by calendar order, volume order or bipower variation order of the three assets, from lowest value to highest value for 8 panels (30 days per panel and 33 days for the last panel). Secondly, we calculate the panel cointegration t-stat of each panel as proposed by Larsson et al. (2001), according to which the critical value in the one-tail test is 1.282 for the 10% significance level (denoted with *), 1.645 for the 5% significance level (denoted with **), and 2.326 for the 1% significance level (denoted with ***).

We carried out this panel cointegration analysis for different data frequencies and different panel sizes, using different rolling techniques and with respect to all the different variables studied previously. The results are always the same, showing a strong and

[23] As shown by Banerjee et al. (2004) multivariate panel cointegration tests can be substantially over-sized in presence of cross-unit cointegration. Nevertheless in our case the daily data considered within each panels are not necessarily consecutive dates and thus not affected by cross dependencies.

significant positive relationship between the rank of matrix Π and the traded volumes of soymeal and soyoil [24]. We clearly notice that only the lowest panel associated to the traded volume of soyoil is not panel cointegrated, whereas the following panels are showing a monotonic reinforcement of panel cointegration. This generally holds for the soymeal traded volume panels as well, whereas the monotonic increase is not observable for the other variables, especially the proxies of high-frequency volatility associated to the same asset. This table demonstrates 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 $\mathbf{\Pi}$ 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 $\mathbf{\Pi}$ and if 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. We indeed demonstrate that the traded volumes of specific assets directly influence the parameters of the matrix κ and, as such, modify the null space associated with the loading matrix. This, in turn, changes the assets' marginal dynamics. That being said, the economic interpretation one can develop directly from each element of κ

[24] Table 4 only displays the results for the traded volumes and the bipower variations, the other results being non significant are available on demand.

is questionable, given the identification challenges associated with the parameters κ and β. As a matter of fact, by multiplying both parameters' matrices by any other $h \times h$ square matrix, we obtain, by definition, the same conditional distribution but another interpretation of the individual parameters' values. To deal with this scaling problem and for a proper interpretation of the components of κ, we propose to study instead the linked relative contribution of each asset to the common factor and thus the assets' relative influence on the permanent component of the marginal assets dynamics. This so-called 'relative market-information share' 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], such that:

$$\kappa_{ij} = \frac{\kappa_{\perp,i}\Sigma_{ii}}{\kappa_{\perp,j}\Sigma_{jj}} \tag{27}$$

where $\kappa_{\perp,i}$ corresponds to the i-th element of the vector $\kappa_\perp$ orthogonal to the vector κ and Σ_{ii} denotes the innovations variance associated to the asset i.

This ratio also provided information about the influence of the spread components on each other. A high absolute ratio means that the numerator 'related assets' Granger causes the denominator 'related assets', provided the higher relative influence of the former on the common factor[26].

Following our theoretical model, if we assume that $\beta_\perp$ in the VAR representation associated matrix Ξ[27] is not related to the traded volumes, 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, the variance of asset prices being closely related to the volume traded (Epps & Epps 1976, among others.), we assume the scaling multiplier in (27), that is the Σ matrix components ratio, to equal one, and we focus our analysis on the squared value of the vector $\kappa_\perp$ components ratios[28], $\tilde{\kappa}_{ij} = (\kappa_{\perp,i})^2/(\kappa_{\perp,j})^2$. We then regressed it on the relative traded volumes and the same control variables that

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

[26] Alternative methods have been proposed in the literature (Janzen & Adjemian 2017, Hu et al. 2020) and could have been considered in this study.

[27] A formal definition of Ξ is provided in the Supplementary Material D

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

we previously considered: the high-frequency bipower variation measure, the realized variance measure, and the XLM index.

Table 5: Regression of $\tilde{\kappa}_{ij}$ ratios

Regressor	$\tilde{\kappa}_{1,2}$	$\tilde{\kappa}_{1,3}$	$\tilde{\kappa}_{2,3}$	$\tilde{\kappa}_{2,1}$	$\tilde{\kappa}_{3,1}$	$\tilde{\kappa}_{3,2}$
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

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 tables 5 validate our model assumption by displaying 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 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 within our model.

As we can see in table 6, the average and median values of the cointegrating vector components remain rather centered near the physical quantities relayed by the CME and displayed in this table. By comparison, if we consider the basic Johansen's cointegration test without dealing with the markets' non-synchronicity, one can clearly notice from table 6 that the average value is then significantly biased, 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 the cointegration is efficiently playing, it is mainly 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 level. 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 $\beta_{2,3}$).

Table 6: 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.

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 to 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.

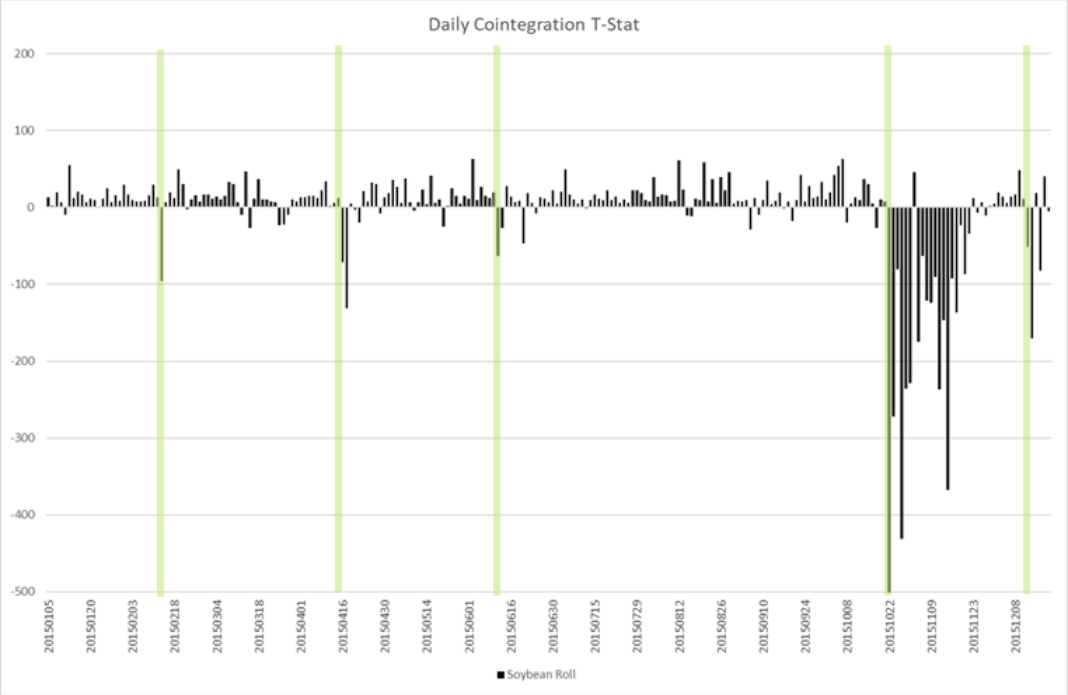

Figure 1: Using daily sets of one-minute data from session 2 and the ZS open interest rollover technique, this figure shows the daily Cointegration test statistics values. T-stats beyond 7.02 mean that the cointegration is statistically significant. The light green lines represent the rolling dates.

We first notice that the choice of the rolling technique, described in the data related section, 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. figure 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.

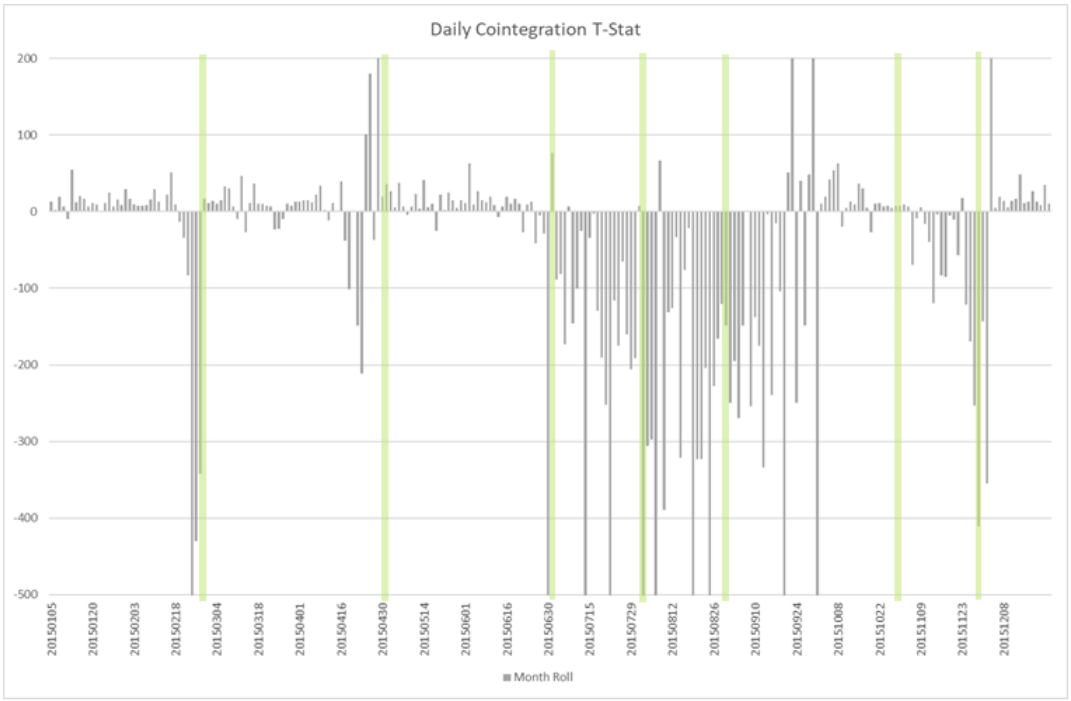

Figure 2: Using daily sets of one-minute data from session 2 and the end of the month rollover technique, this figure shows the daily Cointegration test statistics values. T-stats beyond 7.02 mean that the cointegration is statistically significant. The light green lines represent the rolling dates.

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[29].

Finally, we also notice 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 dates of rollover for each asset, which we leave for future studies.

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

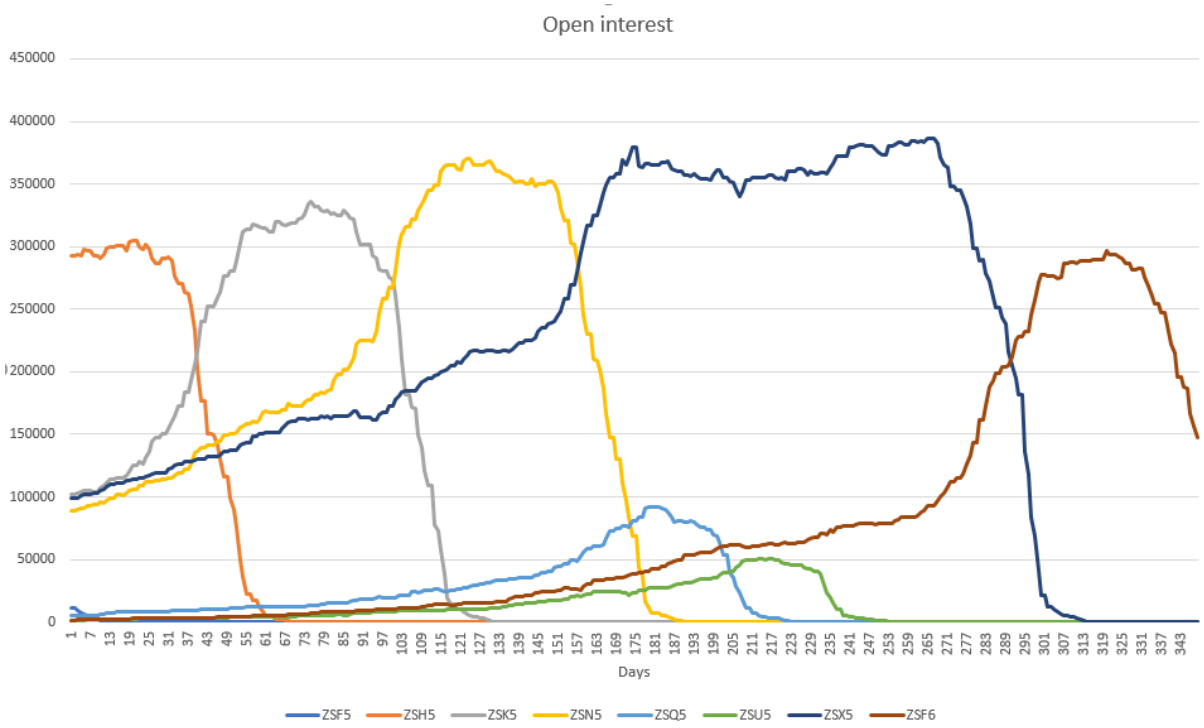

Figure 3: This figure displays the open interest associated to each contract maturity at a given time. The respective contracts correspond to the months of January (ZSF5), March (ZSH5), May (ZSK5), July (ZSK5), August (ZSQ5), September (ZSU5), November (ZSX5) for the year 2015, while the final contract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

6 Conclusion

We proposed 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 their capacity to enforce this equilibrium. We indeed demonstrate and observe that the traded volumes on commodity by-products, such as soybean meal and soybean oil, positively influence the rank of the auto-regressive matrix associated to the soybean complex dynamics and hence the presence or absence of intraday cointegration among related assets. To the contrary, important potential trading volumes in the main market, the soybean market in our case, or a lack of liquidity in the secondary markets, ie. 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 to volatility measures.

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 value. Another 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 contracts' 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.

Finally, 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.

References

Acharya, V. V., Lochstoer, L. A. & Ramadorai, T. (2013), 'Limits to arbitrage and hedging: Evidence from commodity markets', Journal of Financial Economics **109**, 441–465.

Albert S. Kyle (1985), 'Continuous Auctions and Insider Trading', Econometrica **53**(6), 1315–1335.

Arzandeh, M. & Frank, J. (2019), 'Price discovery in agricultural futures markets: Should we look beyond the best bid-ask spread?', American Journal of Agricultural Economics **101**(5), 1482–1498.

Atmaz, A. & Basak, S. (2018), 'Belief Dispersion in the Stock Market', The Journal of 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', Economic Modelling **26**(3), 549–559.

Baillie, R. T., Geoffrey Booth, G., Tse, Y. & Zabotina, T. (2002), 'Price discovery and common factor models', Journal of Financial Markets **5**(3), 309–321.

Bandi, F. M., Kolokolov, A., Pirino, D. & Renò, R. (2020), 'Zeros', Management Science **66**(8), 3466–3479.

Banerjee, A., Marcellino, M. & Osbat, C. (2004), 'Some cautions on the use of panel methods for integrated series of macroeconomic data', The Econometrics Journal **7**(2), 322–340.

Basak, S. & Pavlova, A. (2016), 'A model of financialization of commodities', The Journal of Finance **71**(4), 1511–1556.

Beddock, A. & Jouini, E. (2020), 'Live fast, die young: Equilibrium and survival in large economies', Economic Theory **71**, 961–996.

Behrendt, S. & Schmidt, A. (2021), 'Nonlinearity matters: The stock price – trading volume relation revisited', Economic Modelling **98**, 371–385.

Bessembinder, H. & Seguin, P. J. (1993), 'Price volatility, trading volume, and market depth: Evidence from futures markets', Journal of Financial and Quantitative Analysis **28**(1), 21–39.

Bierens, H. J. & Martins, L. F. (2010), 'Time-varying cointegration', Econometric Theory **26**(5), 1453–1490.

Bradley, M. G. & Lumpkin, S. A. (1992), 'The treasury yield curve as a cointegrated system', The Journal of Financial and Quantitative Analysis **27**(3), 449–463.

Brugler, J. & Comerton-Forde, C. (2019), 'Comment on: Price discovery in high resolution', Journal of Financial Econometrics **19**(3), 1–8.

Buccheri, G., Bormetti, G., Corsi, F. & Lillo, F. (2019), 'Comment on: Price discovery in high resolution', Journal of Financial Econometrics **19**(3), 1–13.

Buccheri, G., Bormetti, G., Corsi, F. & Lillo, F. (2021), 'A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics', Journal of Business & Economic Statistics **39**(4), 920–936.

Buccheri, G., Corsi, F. & Peluso, S. (2021), 'High-Frequency Lead-Lag Effects and Cross-Asset Linkages: A Multi-Asset Lagged Adjustment Model', Journal of Business & Economic Statistics **39**(3), 605–621.

Büyükşahin, B. & Robe, M. A. (2014), 'Speculators, commodities and cross-market linkages', Journal of International Money and Finance **42**, 38–70.

Carchano, O. & Pardo, A. (2009), 'Rolling over stock index futures contracts', Journal of Futures Markets **29**(7), 684–694.

Chen, G.-M., Firth, M. & Meng Rui, O. (2002), 'Stock market linkages: Evidence from latin america', Journal of Banking & Finance **26**(6), 1113–1141.

Christensen, K., Hounyo, U. & Podolskij, M. (2018), 'Is the diurnal pattern sufficient to explain intraday variation in volatility? a nonparametric assessment', Journal of Econometrics **205**(2), 336–362.

Couleau, A., Serra, T. & Garcia, P. (2019), 'Microstructure noise and realized variance in the live cattle futures market', American Journal of Agricultural Economics **101**(2), 563–578.

Couleau, A., Serra, T. & Garcia, P. (2020), 'Are Corn Futures Prices Getting "Jumpy"?', American Journal of Agricultural Economics **102**(2), 569–588.

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

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', Journal of the Royal Statistical Society: Series B (Methodological) **39**(1), 1–22.

Dewachter, H. & Iania, L. (2011), 'An extended macro-finance model with financial factors', The Journal of Financial and Quantitative Analysis **46**(6), 1893–1916.

Dorfman, J. H. & Karali, B. (2015), 'A nonparametric search for information effects from usda reports', Journal of Agricultural and Resource Economics **40**(1), 124–143.

Duchin, R. & Levy, M. (2010), 'Disagreement, Portfolio Optimization, and Excess Volatility', Journal of Financial and Quantitative Analysis **45**(3), 623–640.

Engle, R. F. & Sokalska, M. E. (2012), 'Forecasting intraday volatility in the US equity market. Multiplicative component GARCH', Journal of Financial Econometrics **10**(1), 54–83.

Epps, T. W. (1979), 'Comovements in stock prices in the very short run', Journal of the American Statistical Association **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.

Etienne, X. L., Irwin, S. H. & Garcia, P. (2014), 'Bubbles in food commodity markets: Four decades of evidence', Journal of International Money and Finance **42**, 129–155.

Etienne, X. L., Irwin, S. H. & Garcia, P. (2015), '$25 spring wheat was a bubble, right?', Agricultural Finance Review **75**, 114–132.

Fan, J. H., Fernandez-Perez, A., Fuertes, A.-M. & Miffre, J. (2020), 'Speculative pressure', Journal of Futures Markets **40**(4), 575–597.

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

Fishe, R. P. H., Janzen, J. P. & Smith, A. (2014), 'Hedging and Speculative Trading in Agricultural Futures Markets', American Journal of Agricultural Economics **96**(2), 542–556.

Foucault, T., Kozhan, R. & Tham, W. W. (2017), 'Toxic arbitrage', The Review of Financial Studies **30**(4), 1053–1094.

Frank, J. & Garcia, P. (2011), 'Bid-ask spreads, volume, and volatility: Evidence from livestock markets', American Journal of Agricultural Economics **93**(1), 209–225.

Garcia, P., Irwin, S. H. & Smith, A. (2015), 'Futures market failure?', American Journal of Agricultural Economics **97**(1), 40–64.

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', European Financial Management **21**(1), 52–78.

Gonzalo, J. & Granger, C. (1995), 'Estimation of common long-memory components in cointegrated systems', Journal of Business & Economic Statistics **13**(1), 27–35.

Gorton, G. B., Hayashi, F. & Rouwenhorst, K. G. (2013), 'The Fundamentals of Commodity Futures Returns', Review of Finance **17**(1), 35–105.

Greene, W. H. (2003), Econometric Analysis, 5th edn, Prentice Hall.

Hadri, K. (2000), 'Testing for stationarity in heterogeneous panel data', The Econometrics Journal **3**(2), 148–161.

Hakkio, C. S. & Rush, M. (1991), 'Cointegration: How short is the long run?', Journal of International Money and Finance **10**(4), 571–581.

Han, Y., Hu, T. & Yang, J. (2016), 'Are there exploitable trends in commodity futures prices?', Journal of Banking & Finance **70**, 214–234.

Hansen, P. R. & Lunde, A. (2006), 'Realized variance and market microstructure noise', Journal of Business Economic Statistics **24**(2), 127–161.
URL: *http://www.jstor.org/stable/27638860*

Harris, L. (1986), 'A transaction data study of weekly and intradaily patterns in stock returns', Journal of Financial Economics **16**(1), 99–117.

Hasbrouck, J. (1995), 'One security, many markets: Determining the contributions to price discovery', The journal of Finance **50**(4), 1175–1199.

Hasbrouck, J. (2019), 'Rejoinder on: Price discovery in high resolution*', Journal of Financial Econometrics **19**(3), 465–471.

He, X. & Velu, R. (2014), 'Volume and Volatility in a Common-Factor Mixture of Distributions Model', Journal of Financial and Quantitative Analysis **49**(1), 33–49.

Hong, H. & Yogo, M. (2012), 'What does futures market interest tell us about the macroeconomy and asset prices?', Journal of Financial Economics **105**(3), 473–490.

Hu, Z., Mallory, M., Serra, T. & Garcia, P. (2020), 'Measuring price discovery between nearby and deferred contracts in storable and nonstorable commodity futures markets', Agricultural Economics **51**(6), 825–840.

Janzen, J. P. & Adjemian, M. K. (2017), 'Estimating the location of world wheat price discovery', American Journal of Agricultural Economics **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', Journal of Futures Markets **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', The Journal of Finance **75**(1), 377–417.

Koop, G., Leon-Gonzalez, R. & Strachan, R. W. (2011), 'Bayesian inference in a time varying cointegration model', Journal of Econometrics **165**(2), 210–220.

Larsson, R., Lyhagen, J. & Löthgren, M. (2001), 'Likelihood-based cointegration tests in heterogeneous panels', The Econometrics Journal **4**(1), 109–142.

Li, Z. & Hayes, D. J. (2022), 'The hedging pressure hypothesis and the risk premium in the soybean reverse crush spread', Journal of Futures Markets **42**(3), 428–445.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/fut.22285*

Liu, Q. W. & Sono, H. H. (2016), 'Empirical properties, information flow, and trading strategies of china's soybean crush spread', Journal of Futures Markets **36**(11), 1057–1075.

Lo, A. W. & MacKinlay, A. C. (1990), 'An econometric analysis of nonsynchronous trading', Journal of Econometrics **45**(1), 181–211.

Lütkepohl, H. (2005), New Introduction to Multiple Time Series Analysis, New York : Springer, Berlin.

Marowka, M., Peters, G. W., Kantas, N. & Bagnarosa, G. (2020), 'Factor-augmented Bayesian cointegration models: A case-study on the soybean crush spread', Journal of the Royal Statistical Society: Series C (Applied Statistics) **69**(2), 483–500.

Mitchell, J. (2010), 'Soybean Futures Crush Spread Arbitrage: Trading Strategies and Market Efficiency', Journal of Risk and Financial Management **3**(1), 63–96.

O'Hara, M. (2015), 'High frequency market microstructure', Journal of Financial Economics **116**(2), 257–270.

Rechner, D. & Poitras, G. (1993), 'Putting on the crush: Day trading the soybean complex spread', Journal of Futures Markets **13**(1), 61–75.

Seong, B., Ahn, S. K. & Zadrozny, P. A. (2013), 'Estimation of vector error correction models with mixed-frequency data', Journal of Time Series Analysis **34**(2), 194–205.

Shang, Q., Mallory, M. & Garcia, P. (2018), 'The components of the bid-ask spread: Evidence from the corn futures market', Agricultural Economics **49**(3), 381–393.

Shumway, R. H. & Stoffer, D. S. (1982), 'An Approach To Time Series Smoothing and Forecasting Using The EM Algorithm', Journal of Time Series Analysis **3**(4), 253–264.

Simon, D. P. (1999), 'The soybean crush spread: Empirical evidence and trading strategies', Journal of Futures Markets **19**(3), 271–289.

Tauchen, G. E. & Pitts, M. (1983), 'The Price Variability-Volume Relationship on Speculative Markets', Econometrica **51**(2), 485–505.

Trujillo-Barrera, M. M. & Garcia, P. (2012), 'Volatility spillovers in u.s. crude oil, ethanol, and corn futures markets', Journal of Agricultural and Resource Economics **37**(2), 247–262.

Williams, J. C., Wright, B. D. et al. (1991), Storage and commodity markets., Cambridge university press.

Yang, J., Zhang, J. & Leatham, D. (2003), 'Price and volatility transmission in international wheat futures', Annals of Economics and Finance **4**(1), 37–50.

A Demonstration Proposition 1

If we consider the equation (12):

$$\lambda_1^{-1}V_1 = \left[\alpha_1(P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1}\sum_j \lambda_j\alpha_j(P_{t-1})\right] ABS(\hat{\delta})^{1/\gamma} \tag{28}$$

we thus obtain the following expression for the sum of each group of investor-expected price changes:

$$\hat{\delta} = \left[\alpha_1(P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1}\sum_j \lambda_j\alpha_j(P_{t-1})\right]^{-\gamma} \left(\lambda_1^{-1}V_1\right)^{\gamma} Diag\left(Sgn(\hat{\delta})\right) \tag{29}$$

by combining it with the equation (10), we then obtain:

$$\begin{aligned}\Delta P = &\left[\alpha_1(P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1}\sum_j \lambda_j\alpha_j(P_{t-1})\right]^{-\gamma} \left(\lambda_1^{-1}V_1\right)^{\gamma} Diag\left(Sgn(\hat{\delta})\right) \\ &+ \left(\sum_j \lambda_j\right)^{-1}\sum_j \lambda_j\alpha_j(P_{t-1}) \left[\left(\alpha_1(P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1}\sum_j \lambda_j\alpha_j(P_{t-1})\right)^{-\gamma} \left(\lambda_1^{-1}V_1\right)^{\gamma}\right]^{1/\gamma} \\ &= \Omega(V_1)^{\gamma} Diag\left(Sgn(\hat{\delta})\right) + \left(\sum_j \lambda_j\right)^{-1}\sum_j \lambda_j\alpha_j(P_{t-1})\Omega(V_1)\end{aligned}$$

where:

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

If we assume that $\gamma = 1$ and assume that

$$\alpha_1 = \begin{bmatrix} \frac{1}{\phi_1\beta' P_{t-1}} & 0 \\ 0 & \frac{1}{\phi_2\beta' P_{t-1}} \end{bmatrix} \tag{30}$$

$$\alpha_2 = \begin{bmatrix} -\frac{1}{\phi_1\beta' P_{t-1}} & 0 \\ 0 & 0 \end{bmatrix} \tag{31}$$

and

$$\alpha_3 = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{1}{\phi_2\beta' P_{t-1}} \end{bmatrix} \tag{32}$$

we can then rewrite (14) as:

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

Provided that:

$$\begin{aligned}
\alpha_1(P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1} \sum_j \lambda_j \alpha_j(P_{t-1}) &= \begin{pmatrix} \frac{1}{\phi_1 \beta' P_{t-1}} & 0 \\ 0 & \frac{1}{\phi_2 \beta' P_{t-1}} \end{pmatrix} \\
&\quad - \left(\sum_j \lambda_j\right)^{-1} \left[\begin{pmatrix} \frac{\lambda_1^{11}}{\phi_1 \beta' P_{t-1}} & 0 \\ 0 & \frac{\lambda_1^{22}}{\phi_2 \beta' P_{t-1}} \end{pmatrix} + \begin{pmatrix} -\frac{\lambda_2^{11}}{\phi_1 \beta' P_{t-1}} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & -\frac{\lambda_3^{22}}{\phi_2 \beta' P_{t-1}} \end{pmatrix} \right] \\
&= \alpha_j(P_{t-1})(P_{t-1}) - \left(\sum_j \lambda_j\right)^{-1} \left[\begin{pmatrix} \lambda_1^{11} - \lambda_2^{11} & 0 \\ 0 & \lambda_1^{22} - \lambda_3^{22} \end{pmatrix} \alpha_1(P_{t-1}) \right] \\
&\equiv [I - \lambda^\star]\, \alpha_1(P_{t-1}) \qquad (33)
\end{aligned}$$

with:

$$\lambda^\star = \left(\sum_j \lambda_j\right)^{-1} \begin{pmatrix} \lambda_1^{11} - \lambda_2^{11} & 0 \\ 0 & \lambda_1^{22} - \lambda_3^{22} \end{pmatrix} \qquad (34)$$

We can subsequently rewrite the expression of $\Omega(V_1)$, such that:

$$\Omega(V_1) = \alpha_1(P_{t-1})^{-1} \left[I - \lambda^\star\right]^{-1} \lambda_1^{-1} V_1$$

and thus simplify the equation (14), such that:

$$\begin{aligned}
\Delta P &= \left[Diag\left(Sgn(\hat{\delta})\right) + \lambda^\star \alpha_1(P_{t-1}) \right] \Omega(V_1) \\
&= Diag\left(Sgn(\hat{\delta})\right) \alpha_1(P_{t-1})^{-1} \left[I - \lambda^\star\right]^{-1} \lambda_1^{-1} V_1 + \lambda^\star \left[I - \lambda^\star\right]^{-1} \lambda_1^{-1} V_1 \\
&= \alpha_1(P_{t-1})^{-1} \mathbf{A} V_1 + \mathbf{B} V_1 \qquad (35) \\
&= \begin{bmatrix} \phi_1 \beta' P_{t-1} & 0 \\ 0 & \phi_2 \beta' P_{t-1} \end{bmatrix} \mathbf{A} V_1 + \mathbf{B} V_1 \qquad (36) \\
&= \mathbf{\Pi}^\star(V_1) P_{t-1} + \mathbf{B} V_1 \qquad (37)
\end{aligned}$$

where:

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

B Kalman Filter and Kalman Smoother

To take into account both the microstructure noise in the observation equation (20) and the innovation covariance structure of the cointegrated process in the transition equation (19), we combine the Kalman filter and smoother proposed by Shumway & Stoffer (1982) and Seong et al. (2013), such that for $t = 1, \ldots, T$:

$$x_t^{t-1} = Fx_{t-1}^{t-1}, \quad P_t^{t-1} = FP_{t-1}^{t-1}F' + G\Sigma G'$$

$$x_t^t = x_t^{t-1} + K_t(y_t - H_t x_t^{t-1}), \quad P_t^t = P_t^{t-1} - K_t H_t P_t^{t-1},$$

where $x_t^s = E_l(x_t|y_{1:s})$, $P_t^s = cov_l(x_t|y_{1:s})$ and $P_{t,t-1}^s = cov_l(x_t, x_{t-1}|y_{1:s})$. E_l and cov_l are the conditional expectation and conditional covariance given θ^l the set parameters at iteration l. Moreover, based on Shumway & Stoffer (1982), where microstructure noise is taken into account, the Kalman gain is defined as follows:

$$K_t = P_t^{t-1}H_t'(H_t P_t^{t-1}H_t' + R_t)^{-1} \tag{38}$$

We set the initial values $x_0^0 \sim N(\delta^{(l)}, \Lambda)$. When we enter the Kalman smoother procedure, with microstructure noise taken into consideration, we denote the backward recursions as follows:

$$re_t = H_t'(H_t P_t^{t-1}H_t' + R_t)^{-1}(y_t - H_t x_t^{t-1}) + L_t' re_{t+1} \tag{39}$$

$$Re_t = H_t'(H_t P_t^{t-1}H_t' + R_t)^{-1}H_t + L_t' Re_{t+1}L_t \tag{40}$$

where we set initial values $r_{T+1} = 0$, $R_{T+1} = 0$ and $L_t = F(I_s - K_t H_t)$.

Following Seong et al. (2013), we write the smoothing equations as follows:

$$x_t^T = x_t^{t-1} + P_t^{t-1} r e_t, \quad P_t^T = P_t^{t-1} - P_t^{t-1} R e_t P_t^{t-1} \quad for \quad t = 1, \ldots, T, \tag{41}$$

$$P_{t+1,t}^T = (I_s - P_{t+1}^t R e_{t+1}) L_t P_t^{t-1} \quad for \quad t = 1, \ldots, T-1. \tag{42}$$

C EM algorithm and Gradient

C.1 EM Algorithm

In order to maximise the complete-data log-likelihood, we consider the EM algorithm, which consists in a two-step recursive method.

Step 1: The Expectation step:

The complete data log-likelihood can be written in two different manners:

$$\begin{aligned} \log \mathcal{L}(\theta; x_{1:T}, y_{1:T}) = & -\frac{1}{2} \log |\Lambda| - \frac{1}{2} (x_0 - \delta)' \Lambda^{-1} (x_0 - \delta) \\ & - \frac{T}{2} \log |\Sigma| - \frac{1}{2} \sum_{t=1}^{T} (A x_t - \Gamma B x_{t-1})' \Sigma^{-1} (A x_t - \Gamma B x_{t-1}) \\ & - \frac{T}{2} \log |R| - \frac{1}{2} \sum_{t=1}^{T} (y_t - H_t x_t)' R^{-1} (y_t - H_t x_t) \end{aligned} \tag{43}$$

where:

$$A = \begin{bmatrix} I_n & -I_n & O_{n \times (np-2n)} \end{bmatrix};$$

$$\Gamma = \begin{bmatrix} \kappa, \Gamma_1 & \Gamma_2 & \ldots & \Gamma_{p-1} \end{bmatrix};$$

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

with a normalized $\beta = [I_h \ \beta_0']$, β_0 being the $(n-h) \times h$ matrix to be estimated and $x_0 \sim N(\delta, \Lambda)$. We can also decompose the second term to facilitate the gradient derivations with respect to the sub-matrix β_0:

$$\begin{aligned}\log \mathcal{L}(\theta; x_{1:T}, y_{1:T}) = &-\frac{1}{2}\log|\Lambda| - \frac{1}{2}(x_0-\delta)'\Lambda^{-1}(x_0-\delta) \\ &-\frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{t=1}^{T}\left[Ax_t - \Gamma^* B^* x_{t-1} - \left\{\kappa \otimes (Dx_{t-1})'\right\}\mathrm{vec}(\beta_0)\right]' \\ &\Sigma^{-1}\left[Ax_t - \Gamma^* B^* x_{t-1} - \left\{\kappa \otimes (Dx_{t-1})'\right\}\mathrm{vec}(\beta_0)\right] \\ &-\frac{T}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{T}(y_t - H_t x_t)'R^{-1}(y_t - H_t x_t) \end{aligned} \tag{44}$$

Where:

$$\Gamma^* = \left[\kappa V_1' \ \Gamma_1 \ \Gamma_2 \ \dots \ \Gamma_{p-1}\right];$$

$$D = \left[V_2' \ 0_{(n-h)\times(np-n)}\right];$$

$$B^* = \begin{bmatrix} I_n & -I_n & O_n & \dots \\ O_n & I_n & -I_n & \dots \\ \vdots & \vdots & \vdots & \ddots \\ O_n & \dots & \dots & I_n \end{bmatrix}$$

where we define two sub-matrices:

$$V_1' = [I_h \ 0_{h\times(n-h)}]$$

and:

$$V_2' = [0_{(n-h)\times h} \ I_{(n-h)}]$$

such that $V = [V_1 \ V_2] = I_n$.

We then consider the expectation of this complete-data log likelihood, conditional on the observed data set $y_{1:T}$ and the set of parameters $\theta^{(l)} = \{\kappa^{(l)}, \beta_0^{(l)}, \gamma^{(l)}, \Sigma^{(l)}, R^{(l)}\}$:

$$Q(\theta|\theta^l) = E_l\{log\mathcal{L}(\theta; x_{1:T}, y_{1:T}|y_{1:T})\} \tag{45}$$

Step 2: The Maximisation step:

we maximise the conditional expectation $Q(\theta|\theta^l)$ with respect to κ, β_0, γ, Σ and R. Using the analytical gradient described in C.2, we thus update the set of parameters θ from θ^l

to θ^{l+1} and then go back to the Expectation step.

Regarding the initial setting of the algorithm, we start with a set of parameters $\theta^{(0)}$ such that $\kappa^{(0)}$, $\beta_0^{(0)}$ and $\Gamma^{(0)}$ equal the Johansen's associated estimators while $\Sigma^{(0)}$ and $R^{(0)}$ initial values are set to $[0.0001^2, 0; 0, 0.0001^2]$. Furthermore, the initial values of the latent process used for the Kalman filter, $x_0^0 \sim N(\delta^{(0)}, \Lambda)$, are normally distributed with $\delta^{(0)} = y_{1:p}$ and $\Lambda = I_n$. Then, we iteratively update the parameters using the expectation and maximisation steps (the hyper parameter $\delta^{(0)}$ should also be updated and replaced by the estimator x_0^T, the conditional expected value given $\theta^{(l+1)}$ that we obtained through the Kalman filter and smoother algorithm described in the Supplementary Material B) until the improvement of the log-likelihood innovations form (Shumway & Stoffer 1982):

$$\begin{aligned}\log \mathcal{L}(\theta^{(l+1)}; y_{1:T}) = &-\frac{1}{2}\sum_{t=1}^{T} \log \left| H_t P_t^{t-1} H_t' + R^{(l+1)} \right| \\ &- \frac{1}{2}\sum_{t=1}^{T} \left(y_t - H_t x_t^{t-1}\right)' \left(H_t P_t^{t-1} H_t' + R^{(l+1)}\right)^{-1} \left(y_t - H_t x_t^{t-1}\right) \end{aligned} \tag{46}$$

is less than a predetermined constant. Where $x_t^s = E_{(l+1)}(x_t|y_{1:s})$ and $P_t^s = cov_{(l+1)}(x_t|y_{1:s})$ conditional on $\theta^{(l+1)}$.

C.2 EM Gradient Derivation

In order to update β_0, Γ, Σ and R, we consider the following demonstrations of the complete log-likelihood gradient with respect to the parameters θ. For the ease of notation, we consider that:

$$M_{jk} = \sum_{t=1}^{T} E_l(x_{t-j} x_{t-k}'|y_{1:T}) = \sum_{t=1}^{T} (P_{t-j,t-k}^T + x_{t-j}^T x_{t-k}'^T) \tag{47}$$

for $j, k = 0, 1$.

C.2.1 Derivation with Respect to β_0

Considering the complete-data log-likelihood formula (46) and if we assume that:

$$\begin{aligned} M &= \left[A x_t - \Gamma^* B^* x_{t-1} - \left\{\kappa \otimes (D x_{t-1})'\right\} \text{vec}(\beta_0)\right] \\ M^* &= \left[\left\{\kappa \otimes (D x_{t-1})'\right\} \text{vec}\,(\beta_0)\right] \end{aligned} \tag{48}$$

then, basing on the following properties:

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$$
$$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$$
$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$$

we can write:

$$M' = (Ax_t)' - (\Gamma^* B^* x_{t-1})' - \operatorname{vec}(\beta_0)' \{\kappa' \otimes (Dx_{t-1})\}$$
$$M^{*\prime} = \operatorname{vec}(\beta_0)' \{\kappa' \otimes (Dx_{t-1})\}$$

Furthermore, provided that:

$$\mathbf{A} = Tr(\mathbf{A}) \qquad \text{if } \mathbf{A} = a \text{ (scalar)}$$

we have:

$$\log \mathcal{L} = Tr(\log \mathcal{L})$$

Then, provided that:

$$Tr(\mathbf{A} + \mathbf{B}) = Tr(\mathbf{A}) + Tr(\mathbf{B})$$
$$\partial(\mathbf{A} + \mathbf{B}) = \partial(\mathbf{A}) + \partial(\mathbf{B})$$
$$\partial \mathbf{A} = 0 \qquad \text{(if } \mathbf{A} \text{ is a constant)}$$
$$(\Sigma^{-1})' = \Sigma^{-1} \qquad (\Sigma \text{ is a covariance matrix})$$
$$\frac{\partial}{\partial \mathbf{X}} Tr(\mathbf{AX}) = \frac{\partial}{\partial \mathbf{X}} Tr(\mathbf{XA}) = \mathbf{A}'$$
$$\frac{\partial}{\partial \mathbf{X}} Tr(\mathbf{AX}') = \frac{\partial}{\partial \mathbf{X}} Tr(\mathbf{X}'\mathbf{A}) = \mathbf{A}$$
$$\frac{\partial}{\partial \mathbf{X}} Tr(\mathbf{X}'\mathbf{AX}) = 2\mathbf{AX}$$

we thus obtain:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \operatorname{vec}(\beta_0)} &= -\frac{1}{2}\sum_{t=1}^{T}\frac{\partial}{\partial \operatorname{vec}(\beta_0)}\left[Tr\left(M'\Sigma^{-1}M\right)\right] \\
&= -\frac{1}{2}\sum_{t=1}^{T}\frac{\partial}{\partial \operatorname{vec}(\beta_0)}\left[Tr\left(-M^{*\prime}\Sigma^{-1}(Ax_t) - (Ax_t)'\Sigma^{-1}M^* \right.\right. \\
&\quad + M^{*\prime}\Sigma^{-1}(\Gamma^* B^* x_{t-1}) + (\Gamma^* B^* x_{t-1})'\Sigma^{-1}M^* \\
&\quad \left.\left. + M^{*\prime}\Sigma^{-1}M'\right)\right] \\
&= -\frac{1}{2}\sum_{t=1}^{T} -2\left(\{\kappa' \otimes (Dx_{t-1})\}\,\Sigma^{-1}\left(Ax_t - \Gamma^* B^* x_{t-1}\right)\right) \\
&\quad - \frac{1}{2}\sum_{t=1}^{T} 2\left(\{\kappa' \otimes (Dx_{t-1})\}\,\Sigma^{-1}\left\{\kappa \otimes (Dx_{t-1})'\right\}\right) vec(\beta_0)
\end{aligned}$$

Finally, based on the following property:

$$\begin{aligned}
\Sigma^{-1}Ax_t &= \operatorname{vec}(\Sigma^{-1}Ax_t) = \operatorname{vec}((\Sigma^{-1}Ax_t)') \\
\Sigma^{-1}\Gamma^* B^* x_{t-1} &= \operatorname{vec}((\Sigma^{-1}\Gamma^* B^* x_{t-1}) = \operatorname{vec}((\Sigma^{-1}\Gamma^* B^* x_{t-1})') \\
\operatorname{vec}(\mathbf{AXB}) &= (\mathbf{B}' \otimes \mathbf{A})\operatorname{vec}(\mathbf{X}) \\
(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) &= \mathbf{AC}\otimes\mathbf{BD} \\
\operatorname{vec}(\mathbf{A}+\mathbf{B}) &= \operatorname{vec}(\mathbf{A}) + \operatorname{vec}(\mathbf{B})
\end{aligned}$$

we thus can rewrite the previous expression as:

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \operatorname{vec}(\beta_0)} &= \operatorname{vec}\left\{\sum_{t=1}^{T} Dx_{t-1}x_t'A'\Sigma^{-1}\kappa\right\} \\
&\quad - \operatorname{vec}\left\{\sum_{t=1}^{T} Dx_{t-1}x_{t-1}'(\Gamma^* B^*)'\Sigma^{-1}\kappa\right\} \\
&\quad - \left[(\kappa'\Sigma^{-1}\kappa)\otimes\sum_{t=1}^{T}\left\{Dx_{t-1}x_{t-1}'D'\right\}\right]\operatorname{vec}(\beta_0)
\end{aligned}$$

Then, using again the following property of the vectorization of a matrix:

$$\operatorname{vec}(\mathbf{A}+\mathbf{B}) = \operatorname{vec}(\mathbf{A}) + \operatorname{vec}(\mathbf{B})$$

we can write:

$$
\begin{aligned}
&2\sum_{t=1}^{T}\Big[\operatorname{vec}\left\{Dx_{t-1}\left(x_t'A'\Sigma^{-1}\right)\kappa\right\}-\operatorname{vec}\left\{Dx_{t-1}\left(x_{t-1}'B^{*\prime}\Gamma^{*\prime}\Sigma^{-1}\right)\kappa\right\}\\
&-\kappa'\Sigma^{-1}\kappa\otimes\left(Dx_{t-1}x_{t-1}'D'\right)\Big)\operatorname{vec}\left(\beta_0\right)\Big]=0\\
&\operatorname{vec}\left\{DM_{10}A'\Sigma^{-1}\kappa-DM_{11}B^{*\prime}\Gamma^{*\prime}\Sigma^{-1}\kappa\right\}=\kappa'\Sigma^{-1}\kappa\otimes DM_{11}D'\operatorname{vec}(\beta_0)\\
&\operatorname{vec}(\beta_0)=(\kappa'\Sigma^{-1}\kappa\otimes DM_{11}D')^{-1}\operatorname{vec}\left\{DM_{10}A'\Sigma^{-1}\kappa-DM_{11}B^{*\prime}\Gamma^{*\prime}\Sigma^{-1}\kappa\right\}\\
&\operatorname{vec}(\beta_0)=(\kappa'\Sigma^{-1}\kappa)^{-1}\otimes(DM_{11}D')^{-1}\operatorname{vec}\left\{DM_{10}A'\Sigma^{-1}\kappa-DM_{11}B^{*\prime}\Gamma^{*\prime}\Sigma^{-1}\kappa\right\}
\end{aligned}
$$

Considering that Σ is a covariance matrix satisfying $\Sigma^{-1}=\Sigma$, we have:

$$\operatorname{vec}(\beta_0)=\operatorname{vec}\left\{(DM_{11}D')^{-1}(DM_{10}A'\Sigma^{-1}\kappa-DM_{11}B^{*\prime}\Gamma^{*\prime}\Sigma^{-1}\kappa)[(\kappa'\Sigma^{-1}\kappa)^{-1}]'\right\}$$

where we can rewrite $[(\kappa'\Sigma^{-1}\kappa)^{-1}]'$ as follows:

$$[(\kappa'\Sigma^{-1}\kappa)^{-1}]'=[(\kappa'\Sigma^{-1}\kappa)']^{-1}=[\kappa'(\Sigma^{-1})'\kappa]^{-1}=[\kappa'\Sigma^{-1}\kappa]^{-1}$$

Accordingly, we obtain:

$$\operatorname{vec}(\beta_0)=\operatorname{vec}\left\{(DM_{11}D')^{-1}(DM_{10}A'\Sigma^{-1}\kappa-DM_{11}B^{*\prime}\Gamma^{*\prime}\Sigma^{-1}\kappa)[\kappa'\Sigma^{-1}\kappa]^{-1}\right\}$$

which eventually leads to the gradient formula that we use to update the parameter β_0, given the other parameters associated to the current iteration:

$$\beta_0^{(l+1)}=(DM_{11}D')^{-1}(DM_{10}A'\Sigma^{(l)-1}\kappa^{(l)}-DM_{11}B^{*\prime}\Gamma^{*(l)\prime}\Sigma^{(l)-1}\kappa^{(l)})[\kappa^{(l)\prime}\Sigma^{(l)-1}\kappa^{(l)}]^{-1}$$

C.2.2 Derivation with Respect to Γ

Considering the complete-data log-likelihood formula (44), we can write the derivative with respect to Γ as:

$$\frac{\partial\log\mathcal{L}}{\partial\Gamma}=-\frac{1}{2}\cdot\frac{\partial\sum_{t=1}^{T}W(\Gamma)}{\partial\Gamma}$$

with:

$$
\begin{aligned}
W(\Gamma)&=\left(x_t'A'-x_{t-1}'B'\Gamma'\right)\Sigma^{-1}\left(Ax_t-\Gamma Bx_{t-1}\right)\\
&=\left(x_t'A'\Sigma^{-1}-x_{t-1}'B'\Gamma'\Sigma^{-1}\right)\left(Ax_t-\Gamma Bx_{t-1}\right)\\
&=x_t'A'\Sigma^{-1}Ax_t-x_t'A'\Sigma^{-1}\Gamma Bx_{t-1}-x_{t-1}'B'\Gamma'\Sigma^{-1}Ax_t+x_{t-1}'B'\Gamma'\Sigma^{-1}\Gamma Bx_{t-1},
\end{aligned}
$$

where:

$$w1=x_t'A'\Sigma^{-1}\Gamma Bx_{t-1}$$

$$w2 = x'_{t-1}B'\Gamma'\Sigma^{-1}Ax_t$$

$$w3 = x'_{t-1}B'\Gamma'\Sigma^{-1}\Gamma Bx_{t-1}$$

Basing on the following properties:

$$\begin{aligned} Tr(\mathbf{ABC}) &= Tr(\mathbf{CAB}) = Tr(\mathbf{BCA}) \\ &= Tr(\mathbf{CAB}) = Tr(\mathbf{BCA}) \end{aligned}$$

we obtain:

$$\begin{aligned} \frac{\partial w_1}{\partial \Gamma} &= \frac{\partial}{\partial \Gamma} Tr\left[x'_t A'\Sigma^{-1}\Gamma Bx_{t-1}\right] \\ &= \frac{\partial}{\partial \Gamma} Tr\left[Bx_{t-1}x'_t A'\Sigma^{-1}\Gamma\right] \\ &= \left(Bx_{t-1}x'_t A'\Sigma^{-1}\right)' \\ &= \Sigma^{-1}Ax_t x'_{t-1}B' \end{aligned}$$

while:

$$\begin{aligned} \frac{\partial w_2}{\partial \Gamma} &= \frac{\partial}{\partial \Gamma} Tr\left[x'_{t-1}B'\Gamma'\Sigma^{-1}Ax_t\right] \\ &= \frac{\partial}{\partial \Gamma} Tr\left[\Sigma^{-1}Ax_t x'_{t-1}B'\Gamma'\right] \\ &= \Sigma^{-1}Ax_t x'_{t-1}B' \end{aligned} \tag{49}$$

For w_3 we consider the following property:

$$(\Sigma^{-1})' = \Sigma^{-1} \qquad (\Sigma \text{ is a covariance matrix})$$

$$Bx_{t-1}x'_{t-1}B' = Bx_{t-1}(Bx_{t-1})' = \left(Bx_{t-1}x'_{t-1}B'\right)'$$

$$\frac{\partial}{\partial \mathbf{X}} Tr(\mathbf{X}'\mathbf{AXB}) = \mathbf{AXB} + \mathbf{A}'\mathbf{XB}'$$

to write its derivative with respect to Γ as follows:

$$\begin{aligned} \frac{\partial w_3}{\partial \Gamma} &= \frac{\partial}{\partial \Gamma} Tr\left[x'_{t-1}B'\Gamma'\Sigma^{-1}\Gamma Bx_{t-1}\right] \\ &= \frac{\partial}{\partial \Gamma} Tr\left[\Gamma'\Sigma^{-1}\Gamma Bx_{t-1}x'_{t-1}B'\right] \\ &= 2\Sigma^{-1}\Gamma Bx_{t-1}x'_{t-1}B' \end{aligned}$$

Therefore, we obtain:

$$\begin{aligned} \frac{\partial \log \mathcal{L}}{\partial \Gamma} &= -\frac{1}{2}(0 - \frac{\partial w_1}{\partial \Gamma} - \frac{\partial w_2}{\partial \Gamma} + \frac{\partial w_3}{\partial \Gamma}) \\ &= -\frac{1}{2}(0 - x_t Ax'_{t-1}B'\Sigma^{-1} - \Sigma^{-1}Ax_t x'_{t-1}B' + \Gamma Bx_{t-1}x'_{t-1}B'\Sigma^{-1} + \Gamma\Sigma^{-1}Bx_{t-1}x'_{t-1}B') \\ &= AM_{00}B'\Sigma^{-1} - \Gamma BM_{11}B'\Sigma^{-1} \end{aligned}$$

which, following the first-order condition $\frac{\partial \log \mathcal{L}}{\partial \Gamma} = 0$, eventually leads to the gradient formula that we use to update the parameter matrix Γ, given the value of the parameter $\beta_0^{(l+1)}$:

$$\Gamma^{(l+1)} = (AM_{01}B^{(l+1)\prime})(B^{(l+1)}M_{11}B^{(l+1)\prime})^{-1}$$

C.2.3 Derivation with Respect to Σ

To obtain the derivative of the complete-data log-likelihood formula (44) with respect to Σ, we note:

$$\begin{aligned} V\left(\Sigma^{-1}\right) &= \left(Ax_t - \Gamma B x_{t-1}\right)' \Sigma^{-1} \left(Ax_t - \Gamma B x_{t-1}\right) \\ &= \left(x_t'A' - x_{t-1}'\Gamma' B'\right) \Sigma^{-1} \left(Ax_t - \Gamma B x_{t-1}\right) \\ &= x_t'A'\Sigma^{-1}Ax_t - x_t'A'\Sigma^{-1}\Gamma B x_{t-1} - x_{t-1}'\Gamma' B'\Sigma^{-1}Ax_t + x_{t-1}'\Gamma' B\prime\Sigma^{-1}\Gamma B x_{t-1} \end{aligned}$$

Then using the following property:

$$\frac{\partial \mathbf{a}'\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}'$$

we get:

$$\frac{\partial \left(B\Gamma x_{t-1}\right)' \sum' Ax_t}{\partial \Sigma^{-1}} = \left(B\Gamma x_{t-1}\right)\left(Ax_t\right)' = B\Gamma x_{t-1}x_t'A'$$

Accordingly, we can write:

$$\begin{aligned} \frac{\sum_{t=1}^{T}\left(V\left(\Sigma^{-1}\right)\right)}{\partial \sum^{-1}} &= \sum_{t=1}^{T}\left(x_t'A'Ax_t - x_t'A\Gamma B x_{t-1} - B\Gamma x_{t-1}x_t'A + x_{t-1}'B'\Gamma'\Gamma B x_{t-1}\right) \\ &= AM_{00}A' - \Gamma BM_{10}A' - B\Gamma M_{10}A' + \Gamma BM_{11}B'\Gamma' \end{aligned}$$

Since we know that:

$$\begin{aligned} \Gamma(BM_{11}B') &= AM_{01}B' \\ \Gamma(BM_{11}B') &= (BM_{10}A')' \\ [(\Gamma BM_{11}B')']' &= (BM_{10}A')' \\ (BM_{11}B'\Gamma')' &= (BM_{10}A')' \\ BM_{11}B'\Gamma' &= BM_{10}A' \end{aligned}$$

we can rewrite (50) as:

$$\begin{aligned} \frac{\sum_{t=1}^{T}\left(V\left(\Sigma^{-1}\right)\right)}{\partial \sum^{-1}} &= AM_{00}A' - \Gamma BM_{10}A' - B\Gamma M_{10}A' + B\Gamma M_{10}A' \\ &= AM_{00}A' - \Gamma BM_{10}A' \end{aligned}$$

Therefore, we have:

$$\begin{aligned}\frac{\partial \log \mathcal{L}}{\partial \Sigma^{-1}} &= \frac{T}{2}\log\left|\Sigma^{-1}\right| - \frac{1}{2}\cdot\frac{\sum_{t=1}^{T}\left(V\left(\Sigma^{-1}\right)\right)}{\partial \sum^{-1}} \\ &= \frac{T}{2}\Sigma - \frac{1}{2}\left(AM_{00}A' - \Gamma B M_{10}A'\right)\end{aligned}$$

which, following the first-order condition $\frac{\partial \log \mathcal{L}}{\partial \Sigma^{-1}} = 0$, eventually leads to the gradient formula that we use to update the parameters matrix Σ, given the value of the parameters $\beta_0^{(l+1)}$ and $\Gamma^{(l+1)}$:

$$\Sigma^{(l+1)} = T^{-1}(AM_{00}A' - \Gamma^{(l+1)}B^{(l+1)}M_{10}A')$$

C.2.4 Update of the Parameter Matrix R

Using the set of updated parameters $\theta^{(l+1)}$ to run the Kalman filter and Kalman smoother, we can then update the matrix R as follows:

$$R^{(l+1)} = T^{-1}\sum_{t=1}^{T}\left[\left(y_t - H_t x_t^T\right)\left(y_t - H_t x_t^T\right)' + H_t P_t^T H_t'\right] \tag{50}$$

However, following (Shumway & Stoffer 1982), when, for a given t, there is missing data within y_t, we then consider that there is no contribution to $R^{(l+1)}$ and just add the associated covariance estimate from the previous iteration. More precisely, let's define a partition of the $q \times 1$ vector $y_t = (y_t^{(1)}, y_t^{(2)})$, where $y_t^{(1)}$ corresponds to the $q_1 \times 1$ observed elements and $y_t^{(2)}$ to the $q_2 \times 1$ unobserved portion. We can then rewrite the observation equation (20) as:

$$\begin{bmatrix} y_t^{(1)} \\ y_t^{(2)} \end{bmatrix} = \begin{bmatrix} H_t^{(1)} \\ H_t^{(2)} \end{bmatrix} x_t + \begin{bmatrix} w_t^{(1)} \\ w_t^{(2)} \end{bmatrix}$$

where $H_t^{(1)}$ and $H_t^{(2)}$ are respectively $q_1 \times n$ and $q_2 \times n$ matrices, and we assume uncorrelated errors:

$$cov\begin{bmatrix} w_t^{(1)} \\ w_t^{(2)} \end{bmatrix} = \begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix}$$

Then, if, for a given t, there is missing data within y_t, the contribution C_t to the matrix $R^{(l+1)}$ defined by the equation (50) will be given by:

$$C_t = T^{-1}\left(\begin{bmatrix} y_t^{(1)} - H_t^{(1)}s_t^T \\ 0 \end{bmatrix}\begin{bmatrix} y_t^{(1)} - H_t^{(1)}s_t^T \\ 0 \end{bmatrix}' + \begin{bmatrix} H_t^{(1)} \\ 0 \end{bmatrix} P_t^T \begin{bmatrix} H_t^{(1)} \\ 0 \end{bmatrix}' + \begin{bmatrix} 0 & 0 \\ 0 & R_{22}^{(l)} \end{bmatrix}'\right)$$

D Relative Market Information Share

Following the Granger representation theorem and assuming the presence of $n - h$ unit roots associated to the polynomial characterising the cointegration equation (17), the vector error correction model can be rewritten as a VAR representation:

$$z_t = \Xi \sum_{i=1}^{t} e_t + \Xi^*(L) e_t \tag{51}$$

where:

$$\Xi = \beta_\perp \left[\kappa'_\perp \left(I_K - \sum_{j=1}^{p-1} \Gamma_j \right) \beta_\perp \right]^{-1} \kappa'_\perp \tag{52}$$

and $\Xi^*(L)$ is a matrix polynomial in the lag operator. The first element of eq. (51) corresponds to the permanent component of the marginal price dynamics, which is permanently impounded into the individual prices and embodies new information about the given asset. The second element is a stationary process with a zero mean, which represents a transient effect on the price. Gonzalo & Granger (1995) have demonstrated that this permanent-transient model is equivalent to a common-factor model, where the first part of the equation (51) can be expressed as:

$$\Xi \sum_{i=1}^{t} e_t = A_1 f_t \tag{53}$$

where A_1 is any basis of $\beta_\perp$, the null space associated to β, and where f_t, named the common factor, is equal to $\kappa'_\perp z_t$. Then, the ij relative market information share proposed by Hasbrouck (1995) can be defined as:

$$\frac{S_i}{S_j} = \frac{\gamma_i \Sigma_{ii}}{\gamma_j \Sigma_{jj}} \tag{54}$$

where γ_i corresponds to the i-th element of the first row in the matrix Ξ. This measurement of the relative contribution of each asset to the common factor has been demonstrated to be equivalent to considering the ratio of the respective components of the vector $\kappa'_\perp$ weighted by the variance-covariance matrix of the innovations Baillie et al. (2002).

E Robustness Tests

To confirm the validity and robustness of our results, we propose a set of robustness tests for the data pre-processing methods that we retained for the trading session and the rolling techniques, as well as tests regarding data frequency.

E.1 Trading Sessions and Rolling Techniques

We first verify in this section whether there is any diurnal effect with regard to cointegration. This phenomenon has been commonly observed in financial markets high-frequency data on price level (Harris 1986) and, especially, price volatility (Engle & Sokalska 2012, Christensen et al. 2018). In this literature, we generally assume that the stochastic dynamics is frictionless within the day and thus consistent, whereas markets' microstructures and associated frictions between days significantly perturb price dynamics, making its modelling much more hazardous. The tables below (table 7, table 8 and table 9) confirm this finding and show that cointegration is mainly playing during session 2 (market official trading hours) of a trading day and significantly less so during session 1 (electronic trading hours). In addition, prices in session 1 are so greatly affected by market noise that even the combining of both sessions yields no additional information about the joint dynamics of asset prices and, even worse, prevents identification of the cointegrated days. This result somewhat questions the relevance and informative content of the electronic trading sessions. The fact that only the official trading hours are informative about the joint dynamics of financial assets might be explained by the absence of manufacturers outside of the official trading hours and the prevalence of traders during electronic sessions. Therefore, to mitigate the impact of this diurnal effect on our study of the joint dynamics of asset prices, we mainly retained daily samples and aggregate volume and volatility measures at daily frequencies.

We also study the extent to which the rolling techniques affect the cointegration relationship. According to table 8, we find that the rolling technique based on soybean open interest, which we retained for our study, is the most efficient technique for appreciating the cointegration process. This can be associated with the fact that some maturities on soft commodities are not even considered by manufacturers, who are thus not enforcing, through trades, the convergence of the crush spread towards its long-run equilibrium price. The independent rolling technique also performs well, although it is less informative at the monthly frequency.

Nevertheless, whatever rolling technique we consider, the traded volumes, particularly those associated with soymeal, are always statistically significant in explaining the rank

of matrix Π and thus the presence or absence of cointegration within intraday prices[30] Furthermore, the associated coefficient's sign stays the same, regardless of the rolling technique used.

Table 7: Cointegration Summary of Session 1

every 1 minute data[a]	day	month	quarter
ZS Open Interest Rollover	52	2	1
Independent Rollover	53	4	1
Monthly Rollover	49	4	1
30s Monthly Rollover	33	2	0

[a] This table records the occurrence of cointegration for different sample sizes. Day/month/quarter respectively indicates the number of cointegration occurrences when considering daily/monthly/quarterly samples of one-minute data from session 1 only.

Table 8: Cointegration Summary of Session 2

every 1 minute data[a]	day	month	quarter
ZS Open Interest Rollover	180	11	3
Independent Rollover	181	7	3
Monthly Rollover	132	6	0
30s Monthly Rollover	131	2	0

[a] This table records the occurrence of cointegration for different sample sizes. Day/month/quarter respectively indicates the number of cointegration occurrences when considering daily/monthly/quarterly samples of one-minute data from session 2 only.

[30]For the sake of clarity, table 3 only shows results for the rolling technique based on soybean open interest, but similar tables for the other rolling techniques are available on demand.

Table 9: Cointegration Summary of Session 1 + Session 2

every 1 minute data[a]	day	month	quarter
ZS Open Interest Rollover	52	2	1
Independent Rollover	54	3	0
Monthly Rollover	53	4	2
30s Monthly Rollover	48	0	1

[a] This table records the occurrence of cointegration for different sample sizes. Day/month/quarter respectively indicates the number of cointegration occurrences when considering daily/monthly/quarterly samples of one-minute data from both session 1 and session 2.

E.2 Data Frequency

In this section, we verify how the frequency of the data studied may impact the results. The market's regular trading hours (four hours and fifty minutes per day, labeled 'session 2' in this article) yield 290 data points per trading day using the 1-minute data set and only 145 when using a 2-minute data set. According to Hakkio & Rush (1991), the gain in the degrees of freedom as we increase the frequency of the observations for a given sample length is more apparent than real when it comes to testing and estimating a cointegration relationship. The authors indeed demonstrated that it is not so much the data frequency or the number of points within a given period of time which improve the power of the cointegration test as the length of the period of time under scrutiny. However, they only compared monthly, quarterly and annual simulated data, and hence did not consider high-frequency data and the associated microstructure noise. Furthermore, this conclusion does not hold when dealing with time-varying cointegration (Bierens & Martins 2010, Koop et al. 2011) and by extension, as in our case, volume-varying cointegration.

The two-minute daily data set allows us to detect 229 days of cointegration (versus 180 using the 1-minute daily data set), while the monthly sampled data set only displays 9 months with monthly cointegration (versus 11 using the 1-minute monthly data set).

The quarterly sampled data set reveals only 3 cointegrated quarters (versus 3 using the 1-minute quarterly data set), suggesting that there is still residual microstructure noise after filtering the 1-minute data set using our technique, the effect of which is smoothed when considering less frequent intraday data. This does not affect our conclusions, however, regarding the relationship between cointegration and traded volumes.