Journal of Economic Dynamics & Control 139 (2022) 104434

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

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

Directed acyclic graph based information shares for price discovery☆

Sebastiano Michele Zema

Institute of Economics, Scuola Superiore Sant'Anna, Piazza Martiri della Libertá Pisa, 56127, Italy

ARTICLE INFO

Article history:
Received 7 August 2021
Revised 22 April 2022
Accepted 7 May 2022
Available online 10 May 2022

JEL classification:
C32
C58
G14

Keywords:
Structural VECM
Information shares
Microstructure noise
Independent component analysis
Directed acyclic graphs

ABSTRACT

The possibility to measure the contribution of agents and exchanges to the price formation process in financial markets acquired increasing importance in the literature. In this paper I propose to exploit a data-driven approach to identify structural vector error correction models (SVECM) typically used for price discovery. Exploiting the non-Normal distributions of the variables under consideration, I propose a variant of the widespread Information Share measure, which I will refer to as the *Directed Acyclic Graph based-Information Shares*(DAG-IS), which can identify the leaders and the followers in the price formation process through the exploitation of a causal discovery algorithm well established in the area of machine learning. The approach will be illustrated from a semi-parametric perspective, solving the identification problem with no need to increase the computational complexity which usually arises when working at incredibly short time scales. Finally, an empirical application on IBM intraday data will be provided.

1. Introduction

The past decades have been characterized by dramatic changes in financial markets, where the proliferation of algorithmic trading strategies put aside the intervention of human agents in the price formation process. These algorithms execute orders at incredibly short time scales and there is no doubt anymore they account for most of the trading volumes in developed markets. In addition, processes of market fragmentation have been carried out jointly with the rising of high-frequency trading. This doubly increased the complexity of financial markets, since quotes and trades might be dispersed across different listing venues and at heterogeneous time scales which mix the slower dynamic of humans with the faster dynamic of machines.

The possible benefits of fragmented versus consolidated markets have been object of debates for both economists and regulators also in recent times (Hatheway et al., 2017; Kwan et al., 2015; O'Hara and Ye, 2011). As a consequence, the possibility to measure the relative contribution of different exchanges, agents, and financial instruments to the price formation process acquired increasing importance in the research environment.

☆ I am grateful to Alessio Moneta, Giacomo Bormetti, Fulvio Corsi, and Mattia Guerini for their precious comments. I gratefully acknowledge Joel Hasbrouck for sharing his data and useful clarifications. I am also particularly grateful to Marcelo Fernandes for the valuable suggestions, as well as to the participants of the 13th Annual SoFiE Conference for Young Scholars. I also thank two anonymous referees for their constructive comments which helped me to improve the paper. All errors are my own.

E-mail address: sebastianomichele.zema@santannapisa.it

https://doi.org/10.1016/j.jedc.2022.104434
0165-1889/

In this article I propose to adopt a completely data driven strategy based on *Independent Component Analysis* (ICA) to identify the structural vector error correction models (SVECM) widely adopted in the price discovery context, proposing a solution for the identification problem of the Information Share (IS) measures (Hasbrouck, 1995). The proposed methodology exploits the non-Normal distributions of the variables to identify the transitory shocks, and the associated mixing matrix according to which the observed model residuals correlate across markets. In particular, in presence of non-Normal latent shocks and assuming the existence of a causal chain in the system, it will be shown how to choose among all the potential Choleski orderings the one which is compatible with independent shocks.

Another popular measure widely adopted in price discovery analyses that it is worth mentioning is the Component Share (CS) based on the permanent-transitory (PT) decomposition introduced in Booth et al. (1999); Gonzalo and Granger (1995); Hansen and Lunde (2006); Harris et al. (1995). Both the IS and CS measures build their fundamentals upon the modeling of price changes through VECMs, with the substantial difference that while the CS is defined only in terms of speeds of adjustment toward the common trend (i.e. markets with lower cointegration loadings rapidly adjust and are thus more informative), the IS measure is more concerned with variations in the prices and seeks to measure the amount of information generated by each market. Both approaches have their merits and limits which have been documented by comprehensive discussions in the literature (Baillie et al., 2002; De Jong, 2002; Harris et al., 2002a; 2002b; Hasbrouck, 2002b; Lehmann, 2002). The IS approach, compared to the CS one, has a richer specification since it considers the speed of adjustment together with the relative share of variance of the efficient price process accounted by each market.

Still, from a microstructural modeling point of view, the IS can be uniquely determined only when the VECM residuals are uncorrelated given that the presence of substantial contemporaneous correlations hampers the correct identification of the shocks occurred in each market. Hasbrouck's suggested solution was, in absence of an established theory providing the causal chain to correctly order the variables in the model, to identify the SVECM using the Choleski decomposition and going through all the possible permutations of the variables to get upper and lower bounds for the IS. In empirical applications upper and lower bounds are often very wide giving rise to interpretative ambiguities about the real allocation of information between the analyzed variables, making impossible to distinguish between the exchanges which lead the price formation process and exchanges that follow it.

From a recent data-driven perspective instead, Hasbrouck (2021) proposed to exploit the high frequency at which quotes and trades occur, modeling thus in natural time to drastically reduce the distance between the upper and lower bounds obtained by permuting the variables. Sampling prices at very short time scales, even from microseconds to nanoseconds precision, heavily reduces contemporaneous cross correlations, which by construction leads to narrower IS bounds and allow to discard any interpretive ambiguity. To deal with the enormous amount of coefficients to be estimated in such a natural time framework, the author handled the problem by adopting the heterogeneous autoregressive approach (HAR) proposed by Corsi (2009). Nevertheless, this modeling approach raised interesting and useful comments and discussions in the literature, in some cases controversial, directly related to the econometric model specification, treatment of the high level of data sparsity in natural time, and subsequent identification of where price discovery occurs (Brugler and Comerton-Forde, 2021; Buccheri et al., 2021; Ghysels, 2021; de Jong, 2021).

Despite the identification issue above mentioned and even if other measures of price discovery have been proposed in the literature (see De Jong and Schotman, 2010; Lien and Shrestha, 2009; Putniņš, 2013; Yan and Zivot, 2010), the IS is still one of the most widely used measures for price discovery as documented by its adoption in recent works as well (Ahn et al., 2019; Baur and Dimpfl, 2019; Brogaard et al., 2019; Chen and Tsai, 2017; Entrop et al., 2020; Hagströmer and Menkveld, 2019; Kryzanowski et al., 2017; Lin et al., 2018). The idea to exploit the non-Normal distribution of financial returns to identify the IS measure, directly arises from the possibility of introducing a purely data-driven technique in a research field in which is very hard to provide general and robust theory-driven identification strategies. This will lead to the introduction of the *Directed Acyclic Graph based-Information Shares*(DAG-IS).

The idea of identifying the IS by means of the distributional properties of the variables was firstly introduced by Grammig and Peter (2013). The authors exploited the concept of tail dependence through the adoption in the modeling procedure of different variance regimes, inspired by Rigobon (2003), to identify the contribution of each market to the price discovery process. The intuition was that differences between tail and center correlations, caused by the occurrence of extreme price changes, could be exploited to reach full identification. In particular, following Lanne and Lütkepohl (2010), they assume price innovations to emerge as a mixture of two serially uncorrelated Normal random vectors with different covariance matrices, where one is the identity and the other is a diagonal matrix associated to different variance regimes. Still providing a solution based on the exploitation of the statistical properties of the variables of interest, the methodology proposed in this article differs under many aspects. First, the methodology which I am going to propose can work in principle under any non-Normal distribution, with no need of introducing different volatility regimes to identify the model. Second, keeping Hasbrouck (2021) as a clear benchmark, the strategy proposed in this article is found to provide coherent empirical results under different time specifications when identifying the leaders and the followers in the price formation process.

For all of these reasons the solution proposed in this article can be appealing, at the cost of introducing the assumption of independent shocks in place of uncorrelated ones. Together with the assumption of the presence of an acyclic contemporaneous causal structure (Hyvärinen, 2013; Shimizu et al., 2006), I show we can consistently identify the causal chain in the system and thus the correct permutation of the variables in the VECM with subsequent unique identification of the IS measures.

Recent developments about the ICA approach can be found particularly in macroeconometrics where the identification issue of structural VAR (SVAR) models is pervasive (Gouriéroux et al., 2017; Lanne et al., 2017; Moneta et al., 2013) but applications can be found also in financial econometric and forecasting studies (Audrino et al., 2005; Fabozzi et al., 2016; García-Ferrer et al., 2012; Hafner et al., 2020) as well as in the empirical validation of simulated models (Guerini and Moneta, 2017). Here its potential effectiveness in the identification of SVECM models for price discovery purposes will be addressed. The article is organized as follows. In Section 2 the general setting is provided, showing the baseline model with its identification issues for price discovery. In Section 3, the model and assumptions are illustrated explaining the identification scheme and a simulation exercise is provided to clarify the methodology. Section 4 provides an empirical application on IBM 3 October 2016 intraday data, in order to have the results of Hasbrouck (2021) as a clear benchmark to compare with. Conclusion and discussions are provided in Section 5.

2. General setting

In this section I briefly review the microstructure setting introduced in Hasbrouck (1995), which exploits the vector error correction representation of Engle and Granger (1987), and reproposed in Hasbrouck (2021). The starting point is to consider a vector of time series log-prices $p_t = \{p_{1t}, p_{2t}, ..., p_{nt}\}$ observed in n different exchanges but pertaining the same security, thus all arbitrage linked and whose dynamic are modeled by VECM:

$$\Delta p_t = \alpha\beta' p_{t-1} + \sum_{i=1}^{k} \Phi_i \Delta p_{t-k} + \epsilon_t \tag{1}$$

where the matrix $\beta \in \mathbb{R}^{n\times n-1}$ contains the $n-1$ cointegrating vectors specified as $p_1 - p_2$, $p_1 - p_3$, $p_1 - p_n$ and $\alpha \in \mathbb{R}^{n\times n-1}$ is a loading matrix. The system in Eq. (1) is covariance stationary, with $\mathrm{Cov}(\epsilon_t) = \Omega$, and admits a common trend representation given by

$$p_t = p_0 + \Psi(1)\sum_{i=1}^{t} \epsilon_i + \Psi^*(L)\epsilon_t \tag{2}$$

where the decomposition $\Psi(L) = \Psi(1) + (1-L)\Psi^*(L)$ holds, with the matrix $\Psi(1)$ which can be computed as (Johansen, 1991):

$$\Psi(1) = \beta_\perp \left[\alpha_\perp' \left(I - \sum_{i=1}^{k} \Phi_i \right) \beta_\perp \right]^{-1} \alpha_\perp'. \tag{3}$$

Then, the information share measure for market j is the share of variance of the common component which is induced by the jth market, which means

$$IS_j = \frac{\psi_j^2 \Omega_{jj}}{\psi \Omega \psi'} \tag{4}$$

with ψ being the common row of $\Psi(1)$ and ψ_j denoting the jth element of ψ corresponding to market j. To deal with a non-diagonal Ω two practical solutions have been proposed. The first is to rewrite ϵ_t in terms of orthogonal innovations u_t through the Choleski decomposition C of Ω, then computing the IS as

$$IS_j = \frac{\left([\psi C]_j\right)^2}{\psi \Omega \psi'}. \tag{5}$$

However this allocation mechanism depends on the particular order in which the variables are inserted in the VECM, thus the heuristic solution was to consider upper and lower bounds for the IS by considering all the possible variable permutations.

The second practical solution consists in drastically reducing the gap between upper and lower bounds estimating the model in natural time at very high resolutions, since non zero cross correlations in Ω naturally arise as the sampling interval increases indeed (Dias et al., 2021; Hasbrouck, 2021). This clearly comes at costs, including both the computational aspect of dealing with such a number of observations characterized by high level of sparsity and a suitable model specification to estimate the coefficients still considering a sufficiently long lag-structure in the data.

As explained also in Hasbrouck (2003), the upper and lower bounds of the IS measures cannot be interpreted as confidence intervals but rather as an attempt to solve the identification problem. In the next section I will propose a methodology to uniquely identify, under a few assumptions, the permutation of the variables in the system to recover the exchanges which lead the price discovery and the following ones. It is worth mentioning that the IS measures have been used in a variety of price discovery applications beyond the analysis of multiple stock exchanges and for which high-frequency data are typically not available (see for example Blanco et al., 2005; Guidolin et al., 2021, among others, for the study of price discovery in CDS and bonds markets). In this respect, the methodology proposed in this paper can be appealing since it does not require to sample the data at very high-frequencies, which broaden the range of possible empirical analyses to be performed in the context of price discovery.

3. Model and assumptions

Consider the n-dimensional vector of price innovations $\epsilon_t = [\epsilon_{1t}, \epsilon_{2t}, \dots, \epsilon_{nt}]$ characterized by the non-diagonal covariance matrix Ω. Assume these observed signals to be a linear mixture of hidden components η_t, which can be modeled as

$$\epsilon_t = A_0 \eta_t, \tag{6}$$

where A_0 is a $n \times n$ mixing matrix through which the latent shocks η_t are revealed in each market. The Eq. (6) can be estimated up to permutation, sign, and scaling under some assumptions (Comon, 1994).

Assumption 3.1. The sequence of hidden sources, with finite and non-zero variance, of market microstructure noise η_t possesses at most one Normal marginal distribution,

Assumption 3.2. Independence of the latent shocks: $p(\eta_1, \eta_2, \dots, \eta_n) = \prod_i^n p(\eta_i)$.

Market microstructure noises embed a variety of frictions in the trading process, inherent not only to investment scheme strategies but also to market and asset specific factors and fundamentals. As evidenced by Aït-Sahalia and Yu (2009) market liquidity risk can lead to further adjustments, not explainable by asset specific fundamentals, in the asset bid-ask spread of the assets. Then, from a price discovery perspective the independence assumption in 3.2 would imply market microstructure noise to be market specific and independent from the efficient price process of the asset. Still, observed price innovations are allowed to correlate with each other by means of the mixing matrix A_0 (for example as a consequence of the time aggregation in the sampling process previously mentioned). However, since we directly observe only the mixtures, the independence of the hidden sources cannot be tested and has to be assumed. The non-normality assumption of financial returns is more a stylized fact rather than an assumption. The independence of the non-Normal structural shocks η_t is a stronger concept than uncorrelatedness which is not sufficient alone to get independent variables when non-Normally distributed. This additional information is what will allow to reach full identification of the model if there exists a contemporaneous causal chain between the variables in the system, leading to the third and last assumption.

Assumption 3.3. The observed price innovations ϵ_t can be arranged in a causal chain, meaning that their data generating process possesses a *directed acyclic graph structure* (DAG) (Spirtes et al., 2000).

Under Assumption 3.3 we can model the system in Eq. (6) as the following structural model,

$$\epsilon_t = B_0 \epsilon_t + \eta_t \tag{7}$$

where $A_0 = (I - B_0)^{-1}$ and the assumption of acyclical contemporaneous causal structure implies there exists an appropriate column ordering according to which B_0 is strictly lower triangular. It is worth noticing that the assumption of a causal chain structure is already implicit in the Choleski decomposition used to compute the standard IS measure in the literature. Thus, the value added of the Assumption 3.3 comes when taken jointly with Assumptions 3.1 and 3.2. We refer to this model as the *Linear Non-Gaussian Acyclic Model* (LiNGAM) firstly introduced by Shimizu et al. (2006) in the research field of non-Normal Bayesian networks.

To understand why non-normality is fundamental in the above specified model let us consider the normal case as the baseline to compare with. Consider then the two-dimensional case, for the sake of simplicity, with two innovations $\epsilon = [\epsilon_1, \epsilon_2]$ and two latent shocks $\eta = [\eta_1, \eta_2]$. The first meaningful condition to be fulfilled to reach identification is the orthogonality of the shocks, we thus need to proceed with the orthogonalization of the innovations

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, u_1 \perp u_2.$$

However orthogonal innovations are not sufficient especially in the Normal case. All possible representations fulfilling the meaningful orthogonality condition would be (Blanchard and Quah, 1989)

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \left[\begin{pmatrix} 1 & 0 \\ \lambda_1 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \right] \left[\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right],$$

meaning the models can be identified up to an arbitrary rotation of the space through a rotation matrix R, where the new model $\epsilon = A\eta$, with $A = \Lambda R$ and $\eta = R'u$, is observationally equivalent to $\epsilon = \Lambda u$. However, non-normality heavily reduce the difficulties in identifying the models and the shocks consequently (comprehensive and recent explanations can be found in Gouriéroux et al., 2017; 2020, among others), making possible to estimate $\epsilon_t = A_0 \eta_t$ and recover the latent shocks up to sign, scaling and permutations of the columns of A_0 only. In Fig. 1, a graphical illustration is also provided to clarify how the rotation indeterminacy is resolved by exploiting the ICA approach and why this would not be possible in the Normal case. Since for Normal random variables being uncorrelated coincide with being independent, the spherical symmetry of the joint density makes impossible to distinguish panel (c) from panel (a). Since any rotation would lead to an observationally equivalent density, independent shocks are recovered without knowing whether they are the true ones in panel (a) or a linear mixture of them as in (c). With non-Normal distributions this is not the case and the two densities after orthogonalization of the innovations can be distinguished, the variables in panel (c) are still statistically dependents. This makes possible to

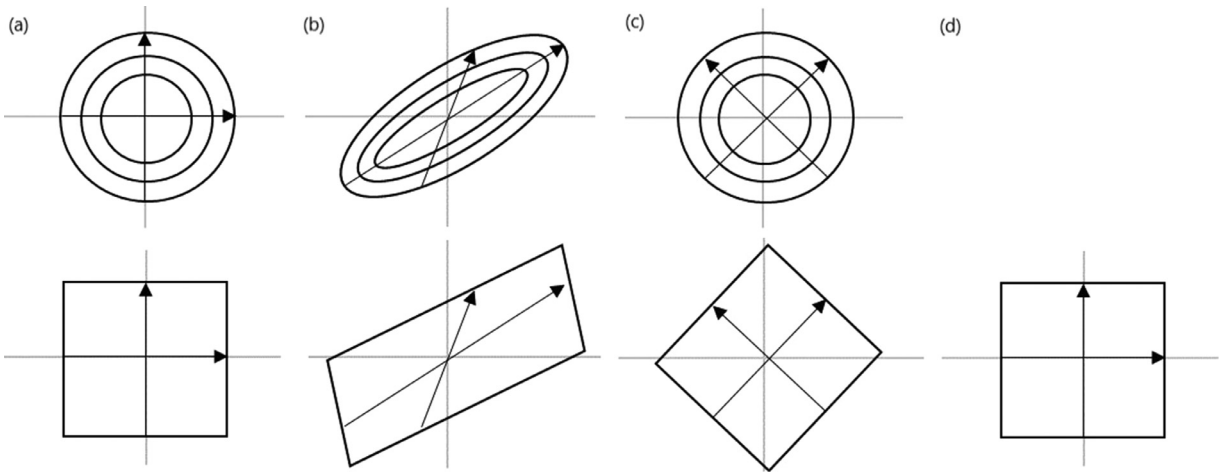

Fig. 1. The role of non-Normality and Independent Component Analysis. (a) The joint density of two standardized and independent random variables, representing our shocks in η_t, when they are Normally distributed (top) and when uniformly distributed (bottom). (b) The joint densities of the Normally and non-Normally distributed random variables after a linear transformation of the space. These would be the price innovations in ϵ_t arising as a linear mixture of the shocks in η_t. (c) Joint density after orthogonalization of the innovations. Despite getting orthogonal innovations, only for the non-Normal distributions panel (c) can be distinguished from panel (a) since they are uncorrelated but still statistically dependents. ICA performs an additional rotation of the space to minimize statistical dependencies, as shown in panel (d), recovering the latent shocks η_t still under arbitrary sign and columns permutations. Note that the Normal density is the only one yielding a spherical symmetric density for standardized and independent random variables.

perform an additional rotation of the space aimed at minimizing statistical dependencies to recover the latent shocks η_t. Searching for this additional rotation of the space, which is performed through the ICA procedure, eliminate the rotation indeterminacy above illustrated greatly alleviating the identification problem.

As a consequence the possible structural models are not perfectly symmetric anymore and can be reformulated either as

$$\begin{pmatrix}\epsilon_1\\ \epsilon_2\end{pmatrix} = \begin{pmatrix}1 & 0\\ \lambda_1 & 1\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix}$$

or

$$\begin{pmatrix}\epsilon_1\\ \epsilon_2\end{pmatrix} = \begin{pmatrix}1 & \lambda_2\\ 0 & 1\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix},$$

which means the model selection reduce to the choice of one of the two models above which are not equivalent anymore. Until now, only Assumptions 3.1 and 3.2 on the non-Normality and independence of the shocks η_t were necessary to illustrate the methodology. However, the identification is not over since the mixing matrix A_0 estimated with ICA is identified only up to column permutations meaning that we do not know in which order the latent shocks have been returned. This explain the necessity to introduce the third assumption on the existence of a directed acyclic graph structure in the model. If we assume there exist a causal chain in the model, than the only remaining step is to find the permutation such that the matrix $B_0 = I - A_0^{-1}$ in Eq. (7) is as close as possible to strictly lower triangular. This leads to perfect identifiability of the causal chain in the model to the extent the assumptions of non-normality, independence of the shocks and presence of a causal chain in the systems hold true.

Thus, the methodology proposed to solve the indeterminacy of the IS approach builds on two separate steps. In the first step ICA is performed to estimate the latent independent shocks, solving the rotation indeterminacy. In the second step, since we do not know the order in which the shocks are returned, the search for the column permutation consistent with a causal chain structure is performed. The two steps are necessarily interconnected, given that the second step building on Assumption 3.3 disentangles the permutation indeterminacy which remains after the implementation of the first step building on Assumptions 3.1 and 3.2. I thus proceed with the details and explanations on the implementation of both steps. After illustrating the ICA methodology, consisting in both the quantification of non-Normality and consequent estimation of the independent latent shocks, the entire set of procedures aimed at inferring the DAG structure will be illustrated discussing also potential limitations and caveats.

3.1. Quantifying non-normality and recovering the independent components: the ICA approach

The first step necessary for the identification process is to perform ICA to recover the non-Normal and statistically independent sources η_t from the observed price innovations ϵ_t. In particular, the ICA approach requires the adoption of suitable measure which quantify the non-Normality of a random variable. The estimation can thus proceed by estimating the

mixing matrix A_0 such that the non-normality of η_t is maximized. There are many approaches to estimate the model in 6 based for instance on the maximization of the kurtosis, negentropy, or minimization of the mutual information between the random variables. All methodologies are closely related and exploit the central limit theorem. The additive mixture ϵ_t of independent and non-normal components η_t, is always closer to a Normal distribution than the latter. Thus, maximizing the non-normality of η_t directly relates to finding a rotation through the inverse of A_0 such that their mutual dependence is minimized.

Going to the maximization schemes implemented so far in the literature, in this work the FastICA algorithm of Hyvärinen and Oja (2000) is adopted being one of the most popular estimators whose performances have been assessed theoretically and empirically, and for which efficient variants of the related algorithm have been also provided (Koldovsky et al., 2006; Miettinen et al., 2017; Reyhani et al., 2012). The optimization problem is solved quantifying the non-normality in terms of *approximated negentropy*. The entropy (amount of information) for a continuous random variables x is defined as

$$H(x) = -\int f(x)\ \log\ f(x)dx. \tag{8}$$

Given that a Normal variable has the largest entropy among random variables of equal variance (Cover and Thomas, 1991), one could optimally quantify, at least theoretically, the non-normality of a random variable by looking at the difference between its entropy and the one of a Normal one with the same variance. The so called *negentropy* is thus defined as

$$J(x) = H(\mathcal{N}) - H(x). \tag{9}$$

However, this would require in practice the knowledge of the probability density function from which the data are generated. For this reason the algorithm deals with an useful approximation of the negentropy of a random variable which takes the form

$$J(x) \approx [E(g(x)) - E(g(\mathcal{Z}))]^2, \tag{10}$$

where $\mathcal{Z}$ is a standardized normal and $g(\cdot)$ is any suitable non-quadratic function used to approximate the negentropy given the data (Hyvärinen and Oja, 1998), here $g(x) = -e^{-x^2/2}$. What is important is to choose $g(\cdot)$ in a way that important regularity conditions, here briefly discussed, are satisfied to guarantee the convergence of the algorithm and related asymptotic properties of the estimates. First, the mixtures are centered to be zero mean and whitened (i.e. uncorrelated and with their variances equal to one) which means I work with the quantities $z = PD^{-1/2}\epsilon$ as done also by Fernandes and Scherrer (2018), where PDP^t is the spectral decomposition of the covariance matrix of the mixtures Ω. Whitening the innovations allow us to get rid of the scaling indeterminacy. The algorithm searches for a vector w, being the rows of the inverse of A_0, which maximizes the non-normality of $w^t z$ measured as shown by Eq. (10), that is

$$\hat{w} = argmax_w E(J(w^t z)). \tag{11}$$

Proposition 3.1. *Suppose that Assumptions 3.1 and 3.2 hold true and that the following regularity conditions are satisfied:*

i $E(z) = \mathbf{0}$;
ii *All moments of z up to the fourth exist;*
iii *Both* $g'(\cdot)$ *and* $g''(\cdot)$ *are Lipschitz continuous. That is, there exist* $\delta_1, \delta_2 < \infty$ *such that* $||g'(x_1) - g'(x_2)|| \leq \delta_1||x_1 - x_2||$ *and* $||g''(x_1) - g''(x_2)|| \leq \delta_2||x_1 - x_2||$;
iv $g''(\cdot)$ *is bounded;*

Then, being $E(zg(w^t z) = \mathbf{0}$ *the first order optimality condition of the maximization problem in (11), the estimator* $\hat{w} = \{w : E(zg^{'}(w^t z) = \mathbf{0}))\}$ *is consistent and asymptotically normal, that is* $\sqrt{n}(\hat{w} - w) \xrightarrow{d} \mathcal{N}(0, \Omega)$.

Proposition 3.1 summarizes the regularity conditions needed to establish the asymptotical properties of the estimates. The asymptotic normality of the ICA estimates have been already proven for a variety of different optimization procedures. A comprehensive theoretical discussion on the statistical properties of the FastICA estimator can be found in Reyhani et al. (2012). It should be mentioned that also other studied proposed to use non-Normal distributions to identify structural shocks in SVAR models (Gouriéroux et al., 2017; Lanne and Lütkepohl, 2010; Lanne et al., 2017) by assuming specific density functions for the shocks.

3.2. Identifying the acyclical causal structure

Until now I used Assumptions 3.1 and 3.2 to estimate η_t and the mixing matrix A_0 up to sign and permutation. The permutation indeterminacy in particular prevent the possibility to determine an appropriate order for the variables. I thus introduce at this point the acyclicity assumption in 3.3 which implies the correct permutation to be the one yielding a strictly lower triangular matrix B_0 encoding the DAG structure.

In the low-dimensional case, the search over all possible permutations is feasible and can be performed also heuristically. However, general optimization procedures should be performed for large-dimensional systems in the form of *linear assignment problems*. In this respect, I refer the readers interested in the general methodology to the original contribution

Algorithm 1 VECM-LiNGAM algorithm for IS measures.

1: Estimate the VECM equation by equation given the known cointegrating relationships, and perform the ICA estimation on the model residuals (any suitable ICA estimator) to recover A_0 and η_t.
2: Given the unmixing matrix $W = A_0^{-1}$, find the permutation of the rows of W such that the permuted version W^* minimize $\sum_i^n 1/|W_{ii}^*|$.
3: Divide each row of W^* by its diagonal element so to get a matrix $\tilde{W}$ with ones in the main diagonal.
4: Let $\tilde{B_0} = I - \tilde{W}$ be the estimate of B_0. Find a permutation matrix Z such that $Z\tilde{B}_0Z'$ as close as possible to be strictly lower triangular. Set the upper triangular elements to zero and permute back to get the matrix $\widehat{B}_0$ containing the directed acyclical graphical structure (DAG). A non zero element b_{ij} in matrix $\widehat{B}_0$ indicates the variable in position j to cause the variable in position i.
5: Thus, order the variable in the VECM according to the DAG structure obtained and perform Choleski on the estimated price innovations. Compute the IS measures.

It is useful to note that a test of statistical significance for the non zero elements of $\tilde{B}_0$ can be performed following if a sufficiently long time series is available, which is the case for high-frequency data. The code implementation of the pruning edges method is available publicly in the online ICA-based LiNGAM code repository.

of Shimizu et al. (2006) where the LiNGAM discovery algorithm was introduced. Here, I illustrate through Algorithm 1 the whole procedure to finally get the IS measures without permutation indeterminacy. I refer to it as the VECM-LiNGAM algorithm, leading to the DAG-IS measures. After the first step in which the non-Normal shocks are estimated by performing ICA on the VECM innovations, the steps in Algorithm 1 from 2 to 5 deals with the scaling, sign, and permutation indeterminacy. Note that step 3 deals in principle with the scaling indeterminacy by forcing the shocks to have unit variance. This is rather standard and the scaling indeterminacy has been already addressed here by performing ICA on the whitened VECM innovations in the first step. Steps 2 and 4 are more critical instead. Given the permutation indeterminacy of ICA, the columns of A_0 will be in random order, meaning we do not have a correct correspondence between the innovations ϵ_t, corresponding to the rows of A_0, and the shocks η_t corresponding to the columns. The procedure in step 2, given the DAG assumption which is crucial to solve the permutation indeterminacy (remember that $B_0 = I - A_0^{-1}$ must be as close as possible to strictly lower triangular), search for the column permutation by minimizing a cost function which penalizes small absolute values in the main diagonal. From an economic point of view this implies that each price series responds to its own shock more than what other price series do. The objective function minimized in step 2 can be also derived from a maximum likelihood approach assuming a generalized normal distribution for the errors (see Shimizu et al., 2006). Given the correspondence between rows and columns, the only remaining step is to order the variables consistently with a DAG structure. That is, we need to permute both rows and columns of $\tilde{B}_0$ through a permutation matrix Z such that $Z\tilde{B}_0Z'$ as close as possible to be strictly lower triangular. Once I get to the permuted matrix encoding the DAG structure, I only need to check what the causal chain is and compute the IS measure with the Choleski decomposition corresponding to the causal chain obtained via the methodology implemented. This leads to Proposition 3.2.

Proposition 3.2. *Suppose that Assumptions 3.1, 3.2 and 3.3 hold true. Then the Information Shares computed by following Algorithm 1 are uniquely identified.*

Proof. See Appendix A. □

The identification scheme proposed ensures the uniqueness of the permutation according to which the price innovations in ϵ_t are mapped in a one-to-one correspondence with the shocks η_t.

Assuming a causal chain among the variables, searching for the implied DAG structure through the algorithm, clearly comes at cost. In principle the matrix $Z\tilde{B}_0Z'$ might be such that no lower triangular matrix can be obtained by permutation. In that case the assumption of a recursive structure would not be adequate, and forcing the algorithm to find the permutation such that $Z\tilde{B}_0Z'$ is as close as possible to lower triangular would lead to biased results.

Rejecting the assumption of a recursive structure would have much severe consequences that go beyond the identification of the IS through the DAG structure. When no recursive structure is detected in the data the Choleski decomposition itself would not be reliable consequently, intrinsically hampering the validity of the IS approach whenever the assumption of a diagonal covariance matrix of the error is violated. A first heuristic check for the matrix to be close to a lower triangular one is to fulfill the condition $\sum_{i\leq j} \widehat{b_{ij}}^2 < 0.2$, however the null hypothesis for the coefficients to be zero can be statistically tested by bootstrap (Shimizu et al., 2006).

In the next section, a simulation exercise is provided to clarify the methodology. An empirical application will follow afterward.

3.2.1. An illustrative simulation exercise

Here I present the proposed identification mechanism on simulated data. In light of Assumptions 3.1 and 3.2 I generate samples of $T = 5000$ observations of independent sources η_t from an *Exponential Power Distribution* (EPD) whose density

function is defined as

$$f(\eta|p,\mu,\sigma_p) = \frac{p}{2\sigma_p p^{1/p}\Gamma(1+1/p)}\exp\left(-\frac{1}{p}\left|\frac{\eta-\mu}{\sigma_p}\right|^p\right) \tag{12}$$

where

$$\begin{aligned}\Gamma(1+1/p) &= \int_0^\infty \eta^{1/p}e^{-\eta}dx \\ &= (1/p)!\end{aligned} \tag{13}$$

is the gamma function. The variances are governed through the scale parameter σ_p according to

$$\sigma^2 = \frac{\sigma_p^2\Gamma(3/p)}{\Gamma(1/p)} \tag{14}$$

Since we need η_t to be non-Normal, I choose to simulate from the EPD density (see Kalke and Richter, 2013; Nardon and Pianca, 2009, for extensive discussions about simulation methodologies) to have flexibility in modeling through the additional shape parameter p. The EPD become a normal when $p=2$ and allows for fat tails by setting $p<2$ (DiCiccio and Monti, 2004; Nadarajah, 2005), which is useful in the present setting to simulate data displaying excess kurtosis as financial price changes do. When $p=1$ the distribution converges to a Laplace, I start with a shape parameter $p=1.2$ which implies an excess kurtosis of 1.8 according to

$$k = \frac{\Gamma(1/p)\Gamma(5/p)}{\Gamma(3/p)^2} - 3. \tag{15}$$

Given that the EPD encompasses different distributions, being a Normal when $p=2$, it represents a convenient choice in a setting where being non-Normal is crucial, ensuring to do not approach to a Normal distribution by controlling the shape parameter.[1] Typically, intraday financial returns display higher levels of volatility both at the beginning and at the end of the trading day, and lower levels of volatility in the middle. For this reason I let the variance of the distributions from which I simulate η_t vary over time, modelling it through the diurnal U-shape pattern (Andersen et al., 2012; Bollerslev et al., 2016; Hasbrouck, 2002a).

$$\sigma_{\eta t} = C + Ae^{-at} + Be^{-b(1-t)} \tag{16}$$

where parameters are set as in Andersen et al. (2012), that is $C=0.88929198$, $A=0.75$, $B=0.25$, $a=10$, and $b=10$. In the light of the empirical application provided in the next section, in which no more than 4-variables will be contemporaneously considered, I simulate a 4-dimensional VECM process driven by only one common stochastic trend. The signals ϵ_t are obtained by mixing the simulated non-Normal and independent shocks η_t through the matrix

$$A_0 = \begin{pmatrix} 0.9 & 0 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0.5 & 0.2 & 0.7 & 0 \\ 0.3 & 0.5 & 0.3 & 0.1 \end{pmatrix}, \tag{17}$$

whose lower triangular structure implies a causal chain from the first to the forth variables passing through the second and the third ones. The shocks in η_t are set to be independent and such that $Cov(\eta_t) = \Sigma_t$ is diagonal with equal variances, the information shares of the two markets are affected by the speed of adjustments in $\boldsymbol{\alpha}$ as well. Details about the simulation setting and parameters can be found in Appendix B.

With the specified parameters and the imposed causal chain, the true IS measures are $IS_1 = 0.58$, $IS_2 = 0.01$, $IS_3 = 0.39$, and $IS_4 = 0.02$. The identification procedure yields the following acyclic structure.

$$\widehat{B_0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.44 & 0 & 0 & 0 \\ 0.42 & 0.43 & 0 & 0 \\ 0.2 & 0.68 & 0.43 & 0 \end{pmatrix}, \tag{18}$$

which means the estimated DAG structure consistently recover the causal chain from the first variable to the fourth, passing before through the second and third variables. Fig. 2 shows the scatter plots for the residuals ϵ_t, clearly correlated as imposed in the data generating process (DGP), and the recovered independent structural sources η_t. Note that the estimated mixing matrix, upon which the causal search algorithm 1 is performed, closely resemble the true A_0 up to sign indeterminacy as shown below

$$\widehat{A_0} = \begin{pmatrix} -1 & 0.01 & 0.03 & 0.004 \\ -0.43 & 0.69 & 0.04 & 0.01 \\ -0.59 & 0.26 & -0.75 & 0.005 \\ -0.34 & 0.58 & -0.3 & 0.1 \end{pmatrix}. \tag{19}$$

[1] Additional simulations have been done, as a robustness check, using the Student's t-distribution. Results have been found to be robust and are available upon request.

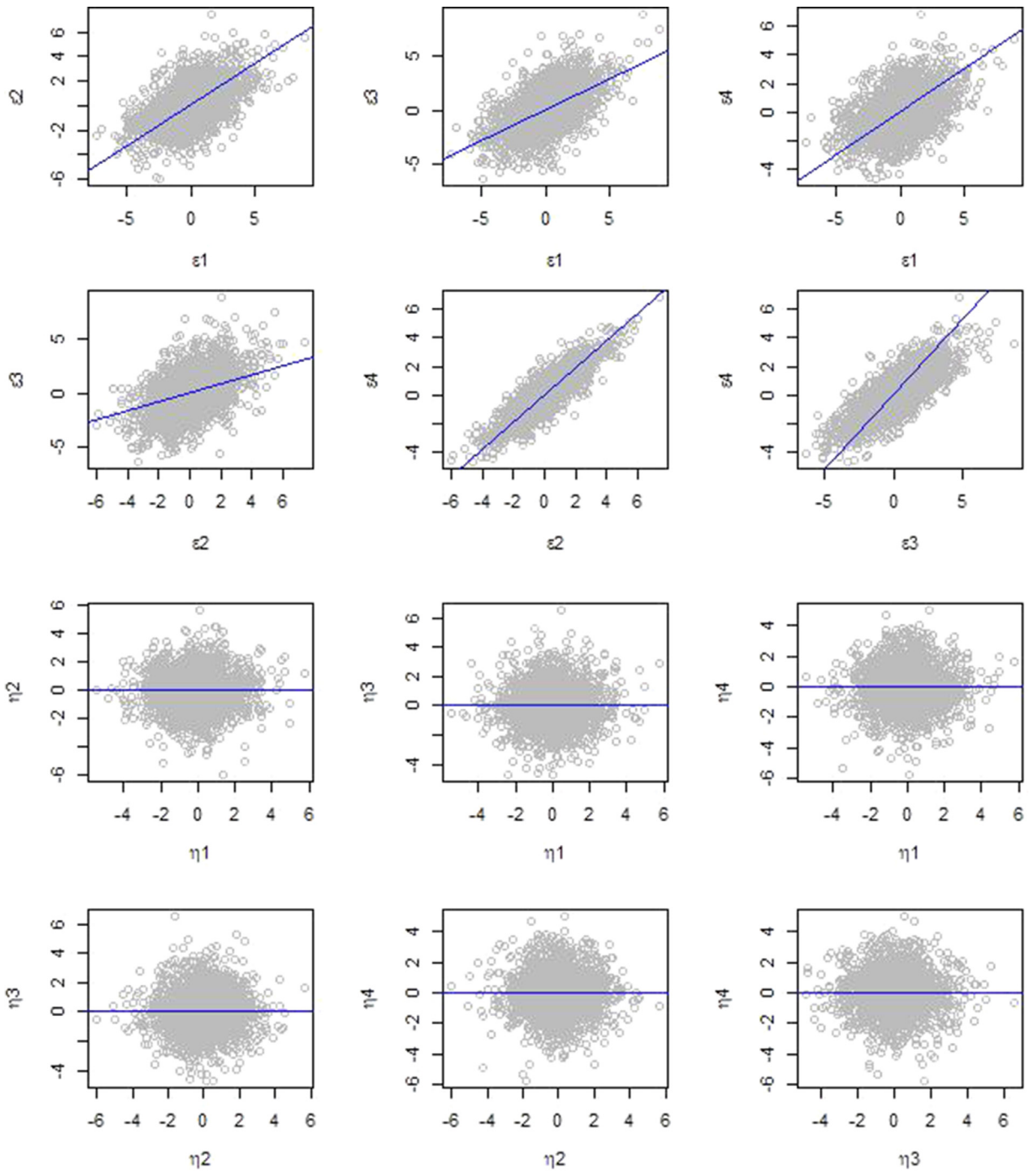

Fig. 2. Scatter plots for the simulated residuals (top half) and estimated latent structural shocks (bottom half).

The computation of the ISs going through all the possible permutations would provide us with $IS_1 = [0.1, 0.58], IS_2 = [0.01, 0.32], IS_3 = [0.1, 0.6]$, and $IS_4 = [0.01, 0.31]$, which make impossible to correctly disentangle the contribution of each market to the variance of the efficient price process. However, recovering the correct causal chain by means of the proposed identification strategy makes possible to correctly permute the variables to get the right causal ordering and, consequently, the true IS measures implied in the simulation setting. In the next section, an empirical application based on IBM data keeping previous results in the literature as a benchmark will be provided.

4. Empirical application

4.1. Benchmarking the model

Bringing the procedure on high-frequency data exposes to several caveats, mostly related to the sparsity of the data and to model specification issues. To have a benchmark to compare with, I empirically test the proposed methodologies on the same IBM data adopted by Hasbrouck (2021), for the day 3 October 2016, which have been shared under the authorization of the NYSE making this analysis possible. I thus try to disentangle the relative contribution to the price discovery process of primary listing and non-primary listing exchanges, participant-based and SIP-based quotes, trades and quotes. As previously illustrated, the main power of the approaches relies in the exploitation of the non-Normal distributions to separate the sources of noise in each variable. In this respect it becomes interesting to test the model stability both in natural and event time, adopting a relatively low level of resolution (i.e. second precision) in the data for the natural time specification. This to eventually check the consistency of the obtained results in both time specifications without increasing the computational complexity and data sparsity introduced when working at very high frequencies.

4.2. IBM, 3 October 2016

The empirical application focuses on some detailed analyses already conducted in the literature in order to have a direct comparison which makes clearer the interpretation of the obtained results. The econometric analysis is performed on IBM's quotes and trades for the day 3 October 2016, with each record reporting both participant-based and SIP-based timestamps. The final whole sample for the day consists of around 30.000 observations. VECM models are thus estimated both in natural-time and event-time with a maximum lag $k = 10$, and then the data-driven identification strategies for the IS measures are implemented.

The first study disentangles the impact of time reporting differentials on the quantification of price discovery measures, through the estimation of a 4-variables VECM including national best bids (NBBs) and offers (NBOs) constructed from both participant and Securities Information Processor (SIP) timestamps. The purpose of the SIP is to establish a consolidated and transparent way to view the market activity for all US equities. Starting from the participant trades and quotes, the Security Information Processor compute and publicly disseminate national best bids and offers at which broker are required to trade, by the regulation, when acting in the interest of their customers. Given that the SIP timestamps are by construction delayed signals of the participant ones, one expects to attribute the price discovery to the participant-based data.

I then proceed with the second analysis which consists in quantifying the price discovery in both the primary listing and other exchanges. The VECM will include bids and offers placed on the primary listing, plus best bids and offers taken from all the exchanges except the primary one.

Finally, the third study is aimed at determining the relative contributions of trades and quotes. I thus insert in the model trades occurred on lit and dark pools separately, plus NBBs and NBOs quotes from participant timestamps. Dark pools are private trading venues, alternative to public accessible exchanges which are defined here as lit pools (examples are the NYSE, NASDAQ, or LSE among others), with no regulatory transparency requirements. This allows institutional investors to trade large securities volume without making their hands visible, thus avoiding possible adverse price effects for their trades when huge volumes are involved since there is no order book visible to the public.

To schematically summarize the empirical application, three separate VECMs will be estimated and identified by the proposed methodology containing respectively:

1. $\mathrm{p}_t^{\mathrm{Model1}} = \left[\mathrm{NBB}_t^{\mathrm{Participants}}, \mathrm{NBO}_t^{\mathrm{Participants}}, \mathrm{NBB}_t^{\mathrm{SIP}}, \mathrm{NBO}_t^{\mathrm{SIP}}\right]$;
2. $\mathrm{p}_t^{\mathrm{Model2}} = \left[\mathrm{NBB}_t^{\mathrm{OtherExchanges}}, \mathrm{NBO}_t^{\mathrm{OtherExchanges}}, \mathrm{Bid}_t^{\mathrm{Primary}}, \mathrm{Ask}_t^{\mathrm{Primary}}\right]$;
3. $\mathrm{p}_t^{\mathrm{Model3}} = \left[\mathrm{NBB}_t^{\mathrm{Participants}}, \mathrm{NBO}_t^{\mathrm{Participants}}, \mathrm{Trade}_t^{\mathrm{LitPools}}, \mathrm{Trade}_t^{\mathrm{DarkPools}}\right]$.

In Fig. 3, the quantile-quantile plots for the VECM residuals are displayed. It can be noticed they are visibly leptokurtic as expected (the normality hypothesis was soundly rejected at the 1% by different tests usually adopted as the Jarque–Bera and the Shapiro–Wilk tests). The residuals of the models estimated for the participant versus SIP timestamps are not reported in the quantile-quantile plots to avoid useless redundancies, given that the variables would be again NBBs and NBOs with just the time-delays differentials in reporting them.

For each model related to a given price discovery analysis, the identification procedure leading to the DAG-IS measures is performed and compared with the approach in which upper and lower bounds are computed by going through all the possible permutations and applying the Choleski decomposition. While Table 1 shows the estimated coefficients of the structural matrix A_0 for each analysis, Table 2 summarizes the information shares estimated for each variable. The autoregressive and loading coefficients, for each estimated VECM, are not reported here for the sake of brevity. However, as also reported in Hasbrouck (2021), estimates are mostly insignificant at the 1-second resolution while they are very significant in the event-time specification. As illustrated in the previous section the underlying acyclical causal structure is encoded in the instantaneous effect matrix A_0, where non-zero elements represent the links among the variables involved.

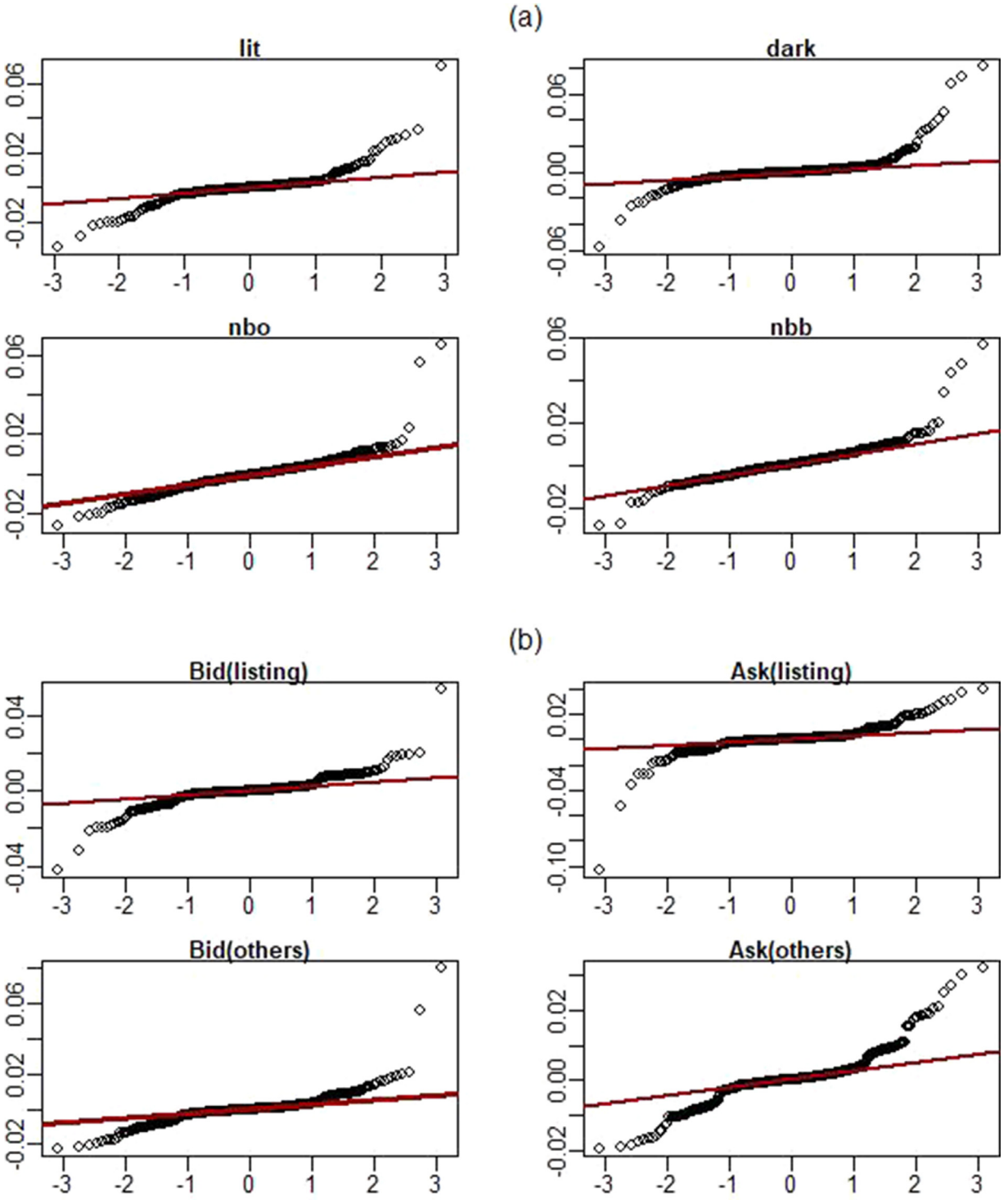

Fig. 3. Quantile-quantile plots of the VECM residuals. In Panel (a) are displayed the model residuals related to the price discovery analysis across trades and quotes, while in panel (b) the one across exchanges using quotes.

Table 1
Estimated instantaneous effect matrices A_0.

Participant VS SIP timestamps									
natural-time	1	0	0	0	event-time	1	−0.038	−0.05	−0.046
	0.34	1	**−0.36**	0		0	1	0	0
	−0.99	0	1	0		0	0.063	1	0
	0.016	**−1.001**	−0.016	1		0	**0.13**	**−0.12**	1
Non-primary VS Primary									
natural-time	1	0.026	**−0.45**	**−0.22**	event-time	1	0	**−0.33**	−0.012
	0	1	**−0.23**	**−0.45**		0.08	1	**−0.015**	**−0.034**
	0	0	1	0		0	0	1	0
	0	0	**−0.35**	1		0	0	−002	1
Quotes VS Trades									
natural-time	1	0	−0.0013	0	event-time	1	0	0	0
	0.012	1	0	**0.039**		−0.011	1	−0.0083	**0.019**
	−0.062	0	1	0		**−0.032**	0	1	0
	−0.051	0	**0.071**	1		**−0.033**	0	**−0.028**	1

Notes: Coefficients in bold are significants at the 1%. Statistical significance has been tested using standard errors from 1000 bootstrap samples.

Table 2
Information shares: summary results.

	DAG-IS		All permutations			
	Participants	SIP	Participants		SIP	
			Min	Max	Min	Max
1-sec	0.999	0.001	0.002	0.999	0.001	0.998
Event time	0.962	0.038	0.943	0.999	0.001	0.057
	Primary	Non-primary	Primary		Non-primary	
			Min	Max	Min	Max
1-sec	0.994	0.006	0.12	0.994	0.006	0.88
Event time	0.56	0.44	0.46	0.56	0.44	0.54
	Quotes	Trades	Quotes		Trades	
			Min	Max	Min	Max
1-sec	0.67	0.33	0.39	0.979	0.021	0.61
Event time	0.64	0.36	0.61	0.67	0.33	0.39

Notes: Information shares measures for each identification procedure and for each price discovery analysis across participants and SIP timestamps, trades and quotes, and exchanges. In the natural-time(1-sec) setting the most recent price observed in a given second interval is taken. In the event time specification, the time counter is incremented whenever there is an update to any variable in the system. Trades comprises both lit and dark trades, given that the contribution of the latter to the IS measure is negligible. The all permutations approach yielded results consistent with Hasbrouck (2021).

Given the estimated results, the following acyclical structures have been recovered

1. $NBB_{participants} \longrightarrow NBB_{SIP} \longrightarrow NBO_{participants} \longrightarrow NBO_{SIP}$ in natural time (1 s);
2. $NBO_{participants} \longrightarrow NBB_{SIP} \longrightarrow NBO_{SIP} \longrightarrow NBO_{participants}$ in event time
3. $Bid_{primary} \longrightarrow Ask_{primary} \rightsquigarrow NBB_{others} \rightsquigarrow NBO_{others}$;
4. $NBB_{participants} \longrightarrow Trades_{Lit} \longrightarrow Trades_{Dark} \longrightarrow NBO_{participants}$.

For the participant versus SIP timestamps the recovered acyclical structure changes with the time framework adopted, but most importantly participants are always placed in the first position and this is the reason why the DAG-IS is able to identify them as the leaders in both cases.

The DAG structures recovered in the primary versus non-primary listing exchanges analysis and quotes versus trades analysis are stable and consistent across the natural and event time settings instead. When the $\rightsquigarrow$ is present in place of the straight arrow $\rightarrow$ it simply means that the recovered coefficient associated to the causal relations is not statistically significant, meaning that the causal chain is interrupted in that specific point. This is the case for the primary versus non-primary listing exchange analysis for example, where no statistically significant relation is detected among shocks in different exchanges other than the primary one and the shocks propagate only from the primary listing to the others.

While the DAG-IS measure is able to identify the participant timestamps as the dominating ones, suggesting the correct variable's order in the system even in the low resolution case (1-second precision), the permutation approach would not

solve the identification issue given the very wide upper and lower bounds (min/max is 0.002/0.999 for participants and 0.001/0.998 for the SIP). There is no doubt in the event time specification instead, where also the approach based on all the possible permutations identify the participant timestamps as the variables leading the price formation process.

Also in the price discovery across exchange analysis, the DAG-IS consistently identify the primary listing exchange as the leader both in natural and event-time. This would not be possible using the heuristic solution with upper and lower bounds (min/max is 0.12/0.994 for the primary listing and 0.006/0.88 for the non-primary in the 1-second resolution). It has to be noticed, however, that the DAG-IS works by finding a permutation respecting the most the statistical dependencies of the data but does not solve the temporal aggregation issue we have when using low levels of resolution. This means that if we discard price variations in each market by aggregating over seconds, the measurement will be obviously overestimated or viceversa but we will still be able to correctly identify the leaders (primary listing) and the followers (other exchanges).

Finally, no sound difference has been detected, surprisingly, when measuring the informational content of quotes and trades in the natural and event time settings. Quotes are more informative than trades and the finding is consistently reported by the DAG-IS measure. Since the contribution of dark trades turns out to be negligible, their shares have been put together with the ones of lit trades differentiating only between trades and quotes.

Overall, the results obtained in the empirical application just illustrated are coherent in choosing the leaders in the price formation process, and in line with the results of Hasbrouck (2021) but without increasing the modeling and computational complexity introduced by working at incredibly short time-scales.

5. Conclusion

Measuring the informational content of fragmented financial markets acquired increasing importance over time for both academics and practitioners. This article proposes a data-driven methodology with the roots in the machine learning research field, exploiting the typical non-Normal distributions of financial returns, to uniquely identify one of the most widely adopted measures for price discovery and for which no identification solutions had been proposed for almost twenty years until the first approach proposed by Grammig and Peter (2013). Differently from the cited approach, with this article I put forward an identification procedure in which the Information Shares measures can be always determined, under some statistical and structural assumptions, with no need of exploiting the possible presence of different volatility regimes caused by extreme price changes, thus providing a general identification framework for price discovery analyses. To this purpose, the DAG-IS measure is introduced. The new estimation procedure has been discussed both theoretically and empirically, with an illustrative simulation exercise. Keeping the empirical analysis of Hasbrouck (2021) as a direct benchmark to compare with, the proposed procedure is found to yield coherent results even across different time specifications, being able to correctly identify the leaders in the price formation process. Given the flexibility of the modeling strategy which can be assessed from a semiparametric perspective, future applications in the field might benefit from the revisited Information Share measures here introduced when the assumption of a causal structure among the data is plausible to exist but no sound theory is provided to decide the direction of causality *a-priori*.

Appendix A

Proof of Proposition 2.2.. Let $\boldsymbol{\sigma} = \{\sigma_1, \dots, \sigma_{n!}\}$, with

$$\sigma_i = \begin{pmatrix} 1 & 2 & \dots & n \\ \sigma_i(1) & \sigma_i(2) & \dots & \sigma_i(n) \end{pmatrix}$$

and $\sigma_i(\cdot) : \{1, \dots, n\} \to \{1, \dots, n\}$, be the set of all possible permutations of the n variables in the model. Consider the set of the Cholesky factors, of the covariance matrices, associated to each permutation of the variables $\boldsymbol{C}_{(\boldsymbol{\sigma})} = \{C_{(\sigma_1)}, \dots, C_{(\sigma_{n!})}\}$. The uniqueness of the Information Share follows directly from the fact that given the estimates of the independent components, there is only one permutation, among the possible ones, yielding a strictly lower triangular matrix $\hat{B_0}$ representing the DAG structure of the variables in the model (result proven in Shimizu et al., 2006). Then, being σ_i^* and $C_{(\sigma_i^*)}$ unique solutions, the identified Information Shares given the estimated DAG structure and computed as

$$DAG - IS_j = \frac{\left(\left[\psi C_{(\sigma_i^*)}\right]_j\right)^2}{\psi \Omega \psi'} \tag{A.1}$$

are unique. □

Appendix B

Data for the illustrative exercise are simulated from the equivalent VAR representation of the VECM adopted in the paper as follows

$$\Pi(L)\, p_t = \epsilon_t \tag{B.1}$$

where

$$\Pi(L) \equiv I_n - \sum_{i}^{k} \Pi_i L^i \tag{B.2}$$

$$\boldsymbol{\alpha}\boldsymbol{\beta}' = (\sum_{i}^{k} \Pi_i - I_n) \tag{B.3}$$

$$\phi_s = -(\Pi_{s+1} + \Pi_{s+2} + \ldots + \Pi_k) \tag{B.4}$$

for $s = 1, 2, \ldots, k-1$, and such that $|I_n - \Pi_1 z - \Pi_2 z^2 - \ldots - \Pi_k z^k| = 0$ has only one unit root since the system is driven by only one common stochastic trend. Consequently, the matrix $\boldsymbol{\beta}$ contains the known cointegrating vectors and has rank equal to *n-1*. In the two-dimensional case the parameters are

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.1 \\ 0.5 \end{pmatrix}, \boldsymbol{\Omega} = \begin{pmatrix} 1 & 0.45 \\ 0.45 & 0.32 \end{pmatrix}, \phi_1 = \begin{pmatrix} 0.6 & 0.3 \\ -0.7 & -0.9 \end{pmatrix}$$

$$\boldsymbol{\beta}' = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \Pi_2 = \begin{pmatrix} -0.6 & -0.3 \\ 0.7 & 0.9 \end{pmatrix} \Pi_1 = \begin{pmatrix} 1.7 & 0.2 \\ -0.2 & -0.4 \end{pmatrix},$$

while in the four-dimensional case are

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.025 & 0.05 & 0.03 \\ 0.08 & 0.07 & 0.06 \\ 0.1 & 0.01 & 0.04 \\ 0.09 & 0.06 & 0.09 \end{pmatrix}, \boldsymbol{\Omega} = \begin{pmatrix} 1 & 0.45 & 0.57 & 0.34 \\ 0.45 & 0.67 & 0.4 & 0.54 \\ 0.57 & 0.4 & 0.98 & 0.58 \\ 0.34 & 0.54 & 0.58 & 0.56 \end{pmatrix},$$

$$\phi_1 = \begin{pmatrix} 0.2 & -0.2 & -0.7 & 0.4 \\ 0.1 & 0.35 & 0.6 & 0.1 \\ 0.6 & 0.35 & 0.55 & -0.1 \\ 0.4 & -0.9 & -0.25 & 0.3 \end{pmatrix}, \quad \Pi_1 = \begin{pmatrix} 1.305 & -0.225 & -0.75 & 0.37 \\ 0.31 & 1.270 & 0.53 & 0.04 \\ 0.75 & 0.25 & 1.54 & -0.14 \\ 0.64 & -0.99 & 0-.31 & 1.21 \end{pmatrix},$$

$$\Pi_2 = \begin{pmatrix} -0.2 & 0.2 & 0.7 & -0.4 \\ -0.1 & -0.35 & -0.6 & -0.1 \\ -0.6 & -0.35 & -0.55 & 0.1 \\ -0.4 & 0.9 & 0.25 & -0.3 \end{pmatrix}, \boldsymbol{\beta}' = \begin{pmatrix} 1 & \\ \vdots & -\boldsymbol{I_{n-1}} \\ 1 & \end{pmatrix}.$$

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jedc.2022.104434

References

Ahn, K., Bi, Y., Sohn, S., 2019. Price discovery among SSE 50 index-based spot, futures, and options markets. J. Futures Mark. 39 (2), 238–259.

Aït-Sahalia, Y., Yu, J., 2009. High frequency market microstructure noise estimates and liquidity measures. Ann. Appl. Stat. 3 (1), 422–457. doi:10.1214/08-AOAS200.

Andersen, T.G., Dobrev, D., Schaumburg, E., 2012. Jump-robust volatility estimation using nearest neighbor truncation. J. Econom. 169 (1), 75–93.

Audrino, F., Barone-Adesi, G., Mira, A., 2005. The stability of factor models of interest rates. J. Financ. Econom. 3 (3), 422–441.

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

Baur, D.G., Dimpfl, T., 2019. Price discovery in bitcoin spot or futures? J. Futures Mark. 39 (7), 803–817.

Blanchard, O., Quah, D., 1989. The dynamic effects of aggregate demand and supply disturbances. Am. Econ. Rev. 79 (4), 655–673.

Blanco, R., Brennan, S., Marsh, I.W., 2005. An empirical analysis of the dynamic relation between investment-grade bonds and credit default swaps. J. Finance 60 (5), 2255–2281.

Bollerslev, T., Patton, A.J., Quaedvlieg, R., 2016. Exploiting the errors: a simple approach for improved volatility forecasting. J. Econom. 192 (1), 1–18. doi:10.1016/j.jeconom.2015.10.007. http://www.sciencedirect.com/science/article/pii/S0304407615002584

Booth, G.G., So, R.W., Tse, Y., 1999. Price discovery in the german equity index derivatives markets. J. Futures Mark. 19 (6), 619–643.

Brogaard, J., Hendershott, T., Riordan, R., 2019. Price discovery without trading: evidence from limit orders. J. Finance 74 (4), 1621–1658.

Brugler, J., Comerton-Forde, C., 2021. Comment on: price discovery in high resolution. J. Financ. Econom. 19 (3), 431–438.

Buccheri, G., Bormetti, G., Corsi, F., Lillo, F., 2021. Comment on: price discovery in high resolution. J. Financ. Econom. 19 (3), 439–451.

Chen, Y.-L., Tsai, W.-C., 2017. Determinants of price discovery in the VIX futures market. J. Empir. Finance 43, 59–73.

Comon, P., 1994. Independent component analysis, a new concept? Signal Process. 36 (3), 287–314.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. J. Financ. Econom. 7 (2), 174–196. doi:10.1093/jjfinec/nbp001. https://academic.oup.com/jfec/article-pdf/7/2/174/2543795/nbp001.pdf

Cover, T.M., Thomas, J.A., 1991. Information theory and statistics. Elem. Inf. Theory 1 (1), 279–335.

De Jong, F., 2002. Measures of contributions to price discovery: acomparison. J. Financ. Mark. 5 (3), 323–327.

De Jong, F., Schotman, P.C., 2010. Price discovery in fragmented markets. J. Financ. Econom. 8 (1), 1–28.

Dias, G.F., Fernandes, M., Scherrer, C.M., 2021. Price discovery in a continuous-time setting. J. Financ. Econom. 19 (5), 985–1008.

DiCiccio, T.J., Monti, A.C., 2004. Inferential aspects of the skew exponential power distribution. J. Am. Stat. Assoc. 99 (466), 439–450. doi:10.1198/016214504000000359.
Engle, R.F., Granger, C.W., 1987. Co-integration and error correction: representation, estimation, and testing. Econometrica 251–276.
Entrop, O., Frijns, B., Seruset, M., 2020. The determinants of price discovery on bitcoin markets. J. Futures Mark. 40 (5), 816–837.
Fabozzi, F.J., Giacometti, R., Tsuchida, N., 2016. Factor decomposition of the Eurozone sovereign CDSspreads. J. Int. Money Finance 65, 1–23.
Fernandes, M., Scherrer, C.M., 2018. Price discovery in dual-class shares across multiple markets. J. Futures Mark. 38 (1), 129–155.
García-Ferrer, A., González-Prieto, E., Peña, D., 2012. A conditionally heteroskedastic independent factor model with an application to financial stock returns. Int. J. Forecast. 28 (1), 70–93.
Ghysels, E., 2021. Comment on: price discovery in high resolution and the analysis of mixed frequency data. J. Financ. Econom. 19 (3), 459–464.
Gonzalo, J., Granger, C., 1995. Estimation of common long-memory components in cointegrated systems. J. Bus. Econ. Stat. 13 (1), 27–35.
Gouriéroux, C., Monfort, A., Renne, J.-P., 2017. Statistical inference for independent component analysis: application to structural VAR models. J. Econom. 196 (1), 111–126.
Gouriéroux, C., Monfort, A., Renne, J.-P., 2020. Identification and estimation in non-fundamental structural VARMA models. Rev. Econ. Stud. 87 (4), 1915–1953.
Grammig, J., Peter, F.J., 2013. Telltale tails: a new approach to estimating unique market information shares. J. Financ. Quant. Anal. 48, 459–488.
Guerini, M., Moneta, A., 2017. A method for agent-based models validation. J. Econ. Dyn. Control 82, 125–141.
Guidolin, M., Pedio, M., Tosi, A., 2021. Time-varying price discovery in sovereign credit markets. Finance Res. Lett. 38, 101388.
Hafner, C.M., Herwartz, H., Maxand, S., 2020. Identification of structural multivariate GARCHmodels. J. Econom. doi:10.1016/j.jeconom.2020.07.019. http://www.sciencedirect.com/science/article/pii/S0304407620302098
Hagströmer, B., Menkveld, A.J., 2019. Information revelation in decentralized markets. J. Finance 74 (6), 2751–2787.
Hansen, P.R., Lunde, A., 2006. Realized variance and market microstructure noise. J. Bus. Econ. Stat. 24 (2), 127–161.
Harris, F.H.d., McInish, T.H., Shoesmith, G.L., Wood, R.A., 1995. Cointegration, error correction, and price discovery on informationally linked security markets. J. Financ. Quant. Anal. 30 (4), 563–579. http://www.jstor.org/stable/2331277
Harris, F.H.d., McInish, T.H., Wood, R.A., 2002. Common factor components versus information shares: a reply. J. Financ. Mark. 5 (3), 341–348.
Harris, F.H.d., McInish, T.H., Wood, R.A., 2002. Security price adjustment across exchanges: an investigation of common factor components for dow stocks. J. Financ. Mark. 5 (3), 277–308.
Hasbrouck, J., 1995. One security, many markets: determining the contributions to price discovery. J. Finance 50 (4), 1175–1199.
Hasbrouck, J., 2002. The dynamics of discrete bid and ask quotes. J. Finance 54 (6), 2109–2142. doi:10.1111/0022-1082.00183. https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-1082.00183
Hasbrouck, J., 2002. Stalking the “efficient price” in market microstructure specifications: an overview. J. Financ. Mark. 5 (3), 329–339.
Hasbrouck, J., 2003. Intraday price formation in us equity index markets. J. Finance 58 (6), 2375–2400.
Hasbrouck, J., 2021. Price discovery in high resolution. J. Financ. Econom. 19 (3), 395–430.
Hatheway, F., Kwan, A., Zheng, H., 2017. An empirical analysis of market segmentation on us equity markets. J. Financ. Quant. Anal. 52 (6), 2399–2427.
Hyvärinen, A., 2013. Independent component analysis: recent advances. Philos. Trans. R. Soc. A 371 (1984), 20110534.
Hyvärinen, A., Oja, E., 1998. Independent component analysis by general nonlinear Hebbian-like learning rules. Signal Process. 64 (3), 301–313.
Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. 13 (4–5), 411–430.
Johansen, S., 1991. Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica 1551–1580.
de Jong, F., 2021. Comment on: price discovery in high resolution*. J. Financ. Econom. 19 (3), 452–458. doi:10.1093/jjfinec/nbz006. https://academic.oup.com/jfec/article-pdf/19/3/452/40507808/nbz006.pdf
Kalke, S., Richter, W.-D., 2013. Simulation of the p-generalized Gaussiandistribution. J. Stat. Comput. Simul. 83 (4), 641–667. doi:10.1080/00949655.2011.631187.
Koldovsky, Z., Tichavsky, P., Oja, E., 2006. Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér–Rao lower bound. IEEE Trans. Neural Netw. 17 (5), 1265–1277.
Kryzanowski, L., Perrakis, S., Zhong, R., 2017. Price discovery in equity and CDS markets. J. Financ. Mark. 35, 21–46.
Kwan, A., Masulis, R., McInish, T.H., 2015. Trading rules, competition for order flow and market fragmentation. J. Financ. Econ. 115 (2), 330–348.
Lanne, M., Lütkepohl, H., 2010. Structural vector autoregressions with nonnormal residuals. J. Bus. Econ. Stat. 28 (1), 159–168.
Lanne, M., Meitz, M., Saikkonen, P., 2017. Identification and estimation of non-Gaussianstructural vector autoregressions. J. Econom. 196 (2), 288–304.
Lehmann, B.N., 2002. Some desiderata for the measurement of price discovery across markets. J. Financ. Mark. 5 (3), 259–276. doi:10.1016/S1386-4181(02)00025-3. Price Discovery. http://www.sciencedirect.com/science/article/pii/S1386418102000253
Lien, D., Shrestha, K., 2009. A new information share measure. J. Futures Mark. 29 (4), 377–395.
Lin, C.-B., Chou, R.K., Wang, G.H., 2018. Investor sentiment and price discovery: evidence from the pricing dynamics between the futures and spot markets. J. Bank. Finance 90, 17–31.
Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., Virta, J., 2017. The squared symmetric FastICA estimator. Signal Process. 131, 402–411.
Moneta, A., Entner, D., Hoyer, P.O., Coad, A., 2013. Causal inference by independent component analysis: theory and applications. Oxf. Bull. Econ. Stat. 75 (5), 705–730.
Nadarajah, S., 2005. A generalized normal distribution. J. Appl. Stat. 32 (7), 685–694. doi:10.1080/02664760500079464.
Nardon, M., Pianca, P., 2009. Simulation techniques for generalized Gaussiandensities. J. Stat. Comput. Simul. 79 (11), 1317–1329.
O'Hara, M., Ye, M., 2011. Is market fragmentation harming market quality? J. Financ. Econ. 100 (3), 459–474. doi:10.1016/j.jfineco.2011.02.006. http://www.sciencedirect.com/science/article/pii/S0304405X11000390
Putniņš, T.J., 2013. What do price discovery metrics really measure? J. Empir. Finance 23, 68–83. doi:10.1016/j.jempfin.2013.05.004. http://www.sciencedirect.com/science/article/pii/S0927539813000340
Reyhani, N., Ylipaavalniemi, J., Vigário, R., Oja, E., 2012. Consistency and asymptotic normality of FastICA and bootstrap FastICA. Signal Process. 92 (8), 1767–1778.
Rigobon, R., 2003. Identification through heteroskedasticity. Rev. Econ. Stat. 85 (4), 777–792.
Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A., 2006. A linear non-Gaussian acyclic model for causal discovery. J. Mach. Learn. Res. 7 (Oct), 2003–2030.
Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D., 2000. Causation, Prediction, and Search. MIT Press.
Yan, B., Zivot, E., 2010. A structural analysis of price discovery measures. J. Financ. Mark. 13 (1), 1–19.

Update

Journal of Economic Dynamics and Control
Volume 148, Issue , March 2023, Page

DOI: https://doi.org/10.1016/j.jedc.2023.104608

Journal of Economic Dynamics & Control 148 (2023) 104608

Contents lists available at ScienceDirect

Journal of Economic Dynamics & Control

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

Corrigendum

Corrigendum to 'Directed acyclic graph based information shares for price discovery' [Journal of Economic Dynamics and Control 139 (2022) 104434]

Sebastiano Michele Zema

Institute of Economics, Scuola Superiore Sant'Anna, Piazza Martiri della Libertá Pisa, 56127, Italy

The authors regret <'each price series respond' I think should be replaced by 'each prices series responds'>. The authors would like to apologise for any inconvenience caused.

DOI of original article: 10.1016/j.jedc.2022.104434
E-mail address: sebastianomichele.zema@santannapisa.it

https://doi.org/10.1016/j.jedc.2023.104608