INSTITUTE
OF ECONOMICS

Scuola Superiore
Sant'Anna

LEM | Laboratory of Economics and Management

Institute of Economics
Scuola Superiore Sant'Anna

Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy
ph. +39 050 88.33.43
institute.economics@sssup.it

LEM
WORKING PAPER SERIES

Directed Acyclic Graph based Information Shares for Price Discovery

Sebastiano Michele Zema [a]

[a] Institute of Economics, Scuola Superiore Sant'Anna, Pisa, Italy.

An updated version of this manuscript is published in the Journal of Economic Dynamics and Control.

2020/28 **October 2020**

ISSN(ONLINE) 2284-0400

Directed Acyclic Graph based Information Shares for Price Discovery

Sebastiano Michele Zema*
Institute of Economics, Scuola Superiore Sant'Anna
Piazza Martiri della Libertà 56127, Pisa
sebastianomichele.zema@santannapisa.it

August 7, 2021

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.

Keywords: Structural VECM; Information Shares; Microstructure noise; Independent Component Analysis; Directed acyclic graphs.

JEL classification: C32, C58, G14.

Declaration of interest: none

*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. All errors are my own.

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 (O'Hara and Ye, 2011; Kwan et al., 2015; Hatheway et al., 2017). As a consequence, the possibility to measure the relative contribution of each exchange in which the asset is listed, to the price formation process, acquired increasing importance in the research environment. In this article I propose to adopt a completely data driven strategy based on *Independent Component Analysis* (ICA) to identify the SVECM widely adopted in the price discovery context, proposing a solution for the identification problem of the Information Share 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.

Another popular measure widely adopted in price discovery analyses that worth to be mentioned is the Component Share (CS) based on the permanent-transitory (PT) decomposition introduced in Gonzalo and Granger (1995) (Harris et al., 1995; Booth et al., 1999; Hansen and Lunde, 2006). 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,b; 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 so to get upper and lower bounds for the IS of each market. 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 (2019) proposed to exploit the high frequency at which quotes and trades occur, modeling thus in natural time to drastically reduce the range obtained by permuting the variables. Sampling prices at very short time scales, even from microseconds to nanoseconds precision, heavily reduces contemporaneous cross correlations between the listing venues indeed, 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, 2019; Buccheri et al., 2019; De Jong, 2019; Ghysels, 2020). Despite the identification issue above mentioned and even if other measures of price discovery have been proposed in the literature (see Lien and Shrestha, 2009; De Jong and Schotman,

2010; Yan and Zivot, 2010; Putniņš, 2013), the IS is still one of the most widely used measures for price discovery as documented by its adoption in recent works as well (Chen and Tsai, 2017; Kryzanowski et al., 2017; Lin et al., 2018; Ahn et al., 2019; Baur and Dimpfl, 2019; Brogaard et al., 2019; Hagströmer and Menkveld, 2019; Entrop et al., 2020). The idea to exploit the non-Normal distribution of financial returns to identify the IS measure via machine-learning based causal search algorithms, 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 (2019) 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 structural shocks in place of uncorrelated ones. Together with the assumption of the presence of an acyclic contemporaneous causal structure (Shimizu et al., 2006; Hyvärinen, 2013), I show we can consistently identify the causal chain in the system and thus the correct permutation of

the variable 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 (Moneta et al., 2013; Gouriéroux et al., 2017; Lanne et al., 2017) but applications can be found also in financial econometric and forecasting studies (Audrino et al., 2005; García-Ferrer et al., 2012; Fabozzi et al., 2016; 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 (2019) as a clear benchmark to compare with. Conclusion and discussions are provided in section 5.

2 General setting

In this section I briefly go back to the microstructure setting introduced in Hasbrouck (1995), which exploits the vector error correction representation of Engle and Granger (1987), and repeated in Hasbrouck (2019). The starting point is to consider a vector of time series log-prices $p_t = \{p_{1t}, p_{2t}, \dots, p_{nt}\}$ 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 $p_1 - p_2$, $p_1 - p_3$, $p_1 - p_n$ since all price series naturally cointegrate each other, and $\alpha \in \mathbb{R}^{n\times n-1}$ is a loading matrix. The system in equation 1 is covariance stationary, with $\mathrm{Cov}(\epsilon_t) = \Omega$, and thus admits a VMA($\infty$) representation

$$\Delta p_t = \Psi(L)\epsilon_t \tag{2}$$

with

$$\Psi(L) = \sum_{i=0}^{\infty} \Psi_i L^i, \tag{3}$$

and also possess, as implied by the Granger representation theorem, a common trend representation given by

$$p_t = p_0 + \Psi(1) \sum_{i=1}^{t} \epsilon_i + \Psi^*(L)\epsilon_t \tag{4}$$

where holds the decomposition $\Psi(L) = \Psi(1) + (1-L)\Psi^*(L)$, which can be seen as the multivariate generalization of the decomposition introduced in Beveridge and Nelson (1981). The second term on the right hand side of equation 4 is the random walk component driving all prices in the system, and thus can be identified as the latent efficient price process, while the last term is the transitory component admitting the VMA(∞). The matrix $\Psi(1)$ can be computed as (Johansen, 1991):

$$\Psi(1) = \beta_\perp \left[\alpha_\perp^{'} \left(I - \sum_{i=1}^{k} \Phi_i \right) \beta_\perp \right]^{-1} \alpha_\perp^{'} \tag{5}$$

and has rank equal to one which is the dimension of the efficient price process behind the observed series, thus all rows of $\Psi(1)$ are identical. 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

$$IS_j = \frac{\psi_j^2 \Omega_{jj}}{\psi \Omega \psi'} \tag{6}$$

with ψ being the common row of $\Psi(1)$ and ψ_j denoting the *j-th* element of ψ corresponding to market *j*. The above definition uniquely allocate the total variance across markets only if the covariance matrix of the innovations Ω is diagonal, while an identification issue arises when price innovations are correlated. 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 as

$$\epsilon_t = C u_t \tag{7}$$

where C is the Choleski decomposition of Ω. The IS thus can be computed in terms of the new orthogonal innovations u_t, that is

$$IS_j = \frac{\left([\psi C]_j \right)^2}{\psi \Omega \psi'}. \tag{8}$$

This allocation mechanism defined through the causal chain implied by the lower triangular structure of C 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 of each market by considering all the possible variable permutations.

The second practical solution consists in drastically reducing the gap between upper and lower bounds, in order to eliminate interpretative ambiguities, estimating the model in natural time at very high resolutions. Non zero cross correlations in Ω naturally arise as the sampling interval increases indeed (Hasbrouck, 2019; Dias et al., 2020), thus they can be minimized by sampling at higher frequencies. 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 interval but rather as an identification problem. In the next section I will propose a methodology to uniquely identify, under few assumptions, the permutation of the variables in the system to recover the exchanges which lead the price discovery and the following ones.

3 Model and assumptions

Consider the n-dimensional vector of price innovations $\epsilon_t = [\epsilon_{1t}, \epsilon_{2t}, ..., \epsilon_{nt}]$ 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

$$\epsilon_t = A_0 \eta_t, \tag{9}$$

where A_0 is a $n \times n$ mixing matrix through which the latent structural shocks η_t are revealed in each market. The equation 9 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 possess at most one Normal marginal distribution;*

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

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 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. Concerning the non-normality assumption of financial returns it is more a stylized fact rather than assumption. The independence of the non-Normal structural shocks η_t is a stronger concept than uncorrelatedness which is not sufficient alone to get rid of all the dependence information in the data. 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 equation 9 as the following structural model,

$$\epsilon_t = B_0 \epsilon_t + \eta_t \tag{10}$$

where $A_0 = (I - B_0)^{-1}$ and the assumption of acyclical contemporaneous causal structure implies there exists an appropriate ordering of the variables according to which B_0 is strictly lower triangular. 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, consider for simplicity the two-dimensional case with ϵ_1 and ϵ_2. The structural model should

identify one variable as the exogenous and the other as the dependent one, which trivially means we should be able to choose between $\epsilon_1 = b_2\epsilon_2 + \eta_1$ and $\epsilon_2 = b_1\epsilon_1 + \eta_2$. In the gaussian case independence and uncorrelatedness coincide and there would be no way to understand whether $\epsilon_1 \to \epsilon_2$, or vice versa, relying on the covariance matrix of η. This because the spherical symmetry of the joint normal distribution of the random vector η_t makes impossible to uniquely identify the matrix A_0, which could be estimated only up to an arbitrary orthogonal transformation as usually done with the PCA approach in the Gaussian case (Hyvärinen, 2013; Moneta et al., 2013). This directly relates on the necessity to rely on a suitable economic theory which allow us to decide what the causal structure and subsequent shock propagation would be. However, under non-normal distributions of the structural shocks in η, the structural models will not be perfectly symmetric anymore and can be in principle distinguished and reformulated either as

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a_1 & 1 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

or

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} = \begin{pmatrix} a_2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix},$$

which means the model selection simply reduce to the choice of one of the two models above which will be now theoretically distinguishable.

The LiNGAM procedure exactly aims at inferring the underlying causal network structure which is more consistent with the statistical dependencies observed in the data, and the assumption 3.3 of an acyclical causal structure is introduced exactly to identify the model. In the Normal case the identification of the ISs would come through the imposition of particular Choleski ordering restrictions which in turn imply both independence and an acyclical causal structure among the variables (either one variable causes the other or viceversa according to the economic theory we rely on). Thus, in this context, the value added of assuming independence and acyclical causal relations arises indeed by taking them jointly with the assumption of non-Normality. In what follows, the estimation steps necessary to achieve the structural identification of the VECM with associated information shares

are illustrated. The first step is the estimation of the hidden non-normal and independent components η_t.

3.1 Quantifying non-normality and recovering the independent components

The first step necessary for the identification process is to recover the non-Normal and statistically independent sources η_t from the observed price innovations ϵ_t. This requires the adoption of suitable measures 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 9 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 direction of the space through the inverse of A_0 such that their mutual dependence is minimized. Going to the optimization 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 algorithm whose performances have been assessed theoretically and empirically, and for which efficient variants of the algorithm have been also provided (Reyhani et al., 2012; Koldovsky et al., 2006; Miettinen et al., 2017). 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

$$H(x) = -\int f(x)\, log\, f(x)\, dx. \tag{11}$$

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 variable with the same variance. The so called *negentropy* is thus

defined as

$$J(x) = H(\mathcal{N}) - H(x). \tag{12}$$

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

$$J(x) \approx \left[E\left(g(x)\right) - E\left(g(\mathcal{Z})\right)\right]^2, \tag{13}$$

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. 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. The algorithm searches for a vector w which maximizes the non-normality of $w^t z$ measured as shown by equation 13, that is

$$\hat{w} = argmax_w E(J(w^t z)). \tag{14}$$

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 (14), 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 (Lanne and Lütkepohl, 2010; Lanne et al., 2017; Gouriéroux et al., 2017) by assuming specific density functions for the structural shocks.

3.2 Identifying the acyclical causal structure

Until now I made use only of assumptions 3.1 and 3.2 to estimate η_t and the mixing matrix A_0 up to permutation, sign, and scaling. The permutation indeterminacy in particular prevent the possibility to determine an appropriate order for the variables. I thus introduce at this point assumption 3.3 to identify the structural model by adapting to our context an heuristic causal search algorithm, well established in the machine learning research area (Shimizu et al., 2006; Hyvärinen et al., 2010), in which the acyclicity assumption makes possible to exploit statistical dependencies to recover a unique causal chain between price innovations in the SVECM. As a consequence I will be able to impose a specific order of the variables in the Choleski decomposition. In algorithm 1, the whole procedure to finally get the IS measure for each market without permutation indeterminacy is illustrated. While step 3 deals with the scaling indeterminacy of the ICA estimation, steps 2 and 4 deal with the sign and permutation indeterminacy which is the crucial problem we have when we want to identify the IS measures for each market, leading to proposition 3.1.

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

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 $\sum_i^n 1/|W_{ii}^*|$. The objective function to minimize in this steps can be derived from maximum likelihood approach assuming a generalized normal distribution for the errors (see Shimizu et al., 2006) .

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 $Z\tilde{B}_0 Z'$ 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. Code implementation of the pruning edges method publicly in the online LiNGAM code repository.

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 closest 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

$$f(\eta \mid p,\ \mu,\ \sigma_p) = \frac{p}{2\sigma_p p^{1/p}\Gamma(1+1/p)} exp\left(-\frac{1}{p}\left|\frac{\eta-\mu}{\sigma_p}\right|^p\right) \tag{15}$$

where

$$\begin{aligned}\Gamma(1+1/p) &= \int_0^\infty \eta^{1/p}e^{-\eta}dx \\ &= (1/p)!\end{aligned} \tag{16}$$

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

$$\sigma^2 = \frac{\sigma_p^2\Gamma(3/p)}{\Gamma(1/p)} \tag{17}$$

Since we need η_t to be non-Normal, I choose to simulate from the EPD density (see Nardon and Pianca, 2009; Kalke and Richter, 2013, 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

$$k = \frac{\Gamma(1/p)\Gamma(5/p)}{\Gamma(3/p)^2} - 3. \tag{18}$$

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 (Hasbrouck, 2002a; Andersen et al., 2012; Bollerslev et al., 2016).

$$\sigma_{\eta t} = C + Ae^{-at} + Be^{-b(1-t)} \tag{19}$$

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

$$A_0 = \begin{pmatrix} 0.9 & 0 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0.5 & 0.2 & 0.7 & 0 \\ 0.3 & 0.5 & 0.3 & 0.1 \end{pmatrix}, \tag{20}$$

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, 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.

$$\widehat{B_0} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.44 & 0 & 0 & 0 \\ 0.42 & 0.43 & 0 & 0 \\ 0.2 & 0.68 & 0.43 & 0 \end{pmatrix}, \tag{21}$$

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. Figure 1 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

$$\widehat{A_0} = \begin{pmatrix} -1 & 0.01 & 0.03 & 0.004 \\ -0.43 & 0.69 & 0.04 & 0.01 \\ -0.59 & 0.26 & -0.75 & 0.005 \\ -0.34 & 0.58 & -0.3 & 0.1 \end{pmatrix}. \tag{22}$$

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 we are able to correctly permute the variables to get the true IS measures. 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 (2019), 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. $\mathrm{p}_{\mathrm{t}}^{\mathrm{Model1}} = \left[\mathrm{NBB}_{\mathrm{t}}^{\mathrm{Participants}}, \mathrm{NBO}_{\mathrm{t}}^{\mathrm{Participants}}, \mathrm{NBB}_{\mathrm{t}}^{\mathrm{SIP}}, \mathrm{NBO}_{\mathrm{t}}^{\mathrm{SIP}}\right]$;
2. $\mathrm{p}_{\mathrm{t}}^{\mathrm{Model2}} = \left[\mathrm{NBB}_{\mathrm{t}}^{\mathrm{OtherExchanges}}, \mathrm{NBO}_{\mathrm{t}}^{\mathrm{OtherExchanges}}, \mathrm{Bid}_{\mathrm{t}}^{\mathrm{Primary}}, \mathrm{Ask}_{\mathrm{t}}^{\mathrm{Primary}}\right]$;
3. $\mathrm{p}_{\mathrm{t}}^{\mathrm{Model3}} = \left[\mathrm{NBB}_{\mathrm{t}}^{\mathrm{Participants}}, \mathrm{NBO}_{\mathrm{t}}^{\mathrm{Participants}}, \mathrm{Trade}_{\mathrm{t}}^{\mathrm{LitPools}}, \mathrm{Trade}_{\mathrm{t}}^{\mathrm{DarkPools}}\right]$.

In figure 2, 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 and can be found in

the supplemental online appendix. However, as also reported in Hasbrouck (2019), 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. 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 second);
2. $NBO_{participants} \longrightarrow NBB_{SIP} \longrightarrow NBO_{SIP} \longrightarrow NBO_{participants}$ in event time
3. $Bid_{primary} \longrightarrow Ask_{primary} \longrightarrow NBB_{others} \rightsquigarrow NBO_{others}$;
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. Surprisingly, the DAG structures recovered in the primary versus non-primary listing exchanges analysis and quotes versus trades analysis is stable and consistent across the natural and event time settings. 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. 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. 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 (2019) 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 (2019) 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 $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_{n!}\}$, with

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

and $\sigma_i(\cdot) : \{1, ..., n\} \to \{1, ..., n\}$, 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 $\boldsymbol{C_{(\sigma)}} = \{C_{(\sigma_1)}, ..., C_{(\sigma_{n!})}\}$. 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

$$DAG - IS_j = \frac{\left(\left[\psi C_{(\sigma_i^*)}\right]_j\right)^2}{\psi \Omega \psi'} \tag{23}$$

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

$$\Pi(L)p_t = \epsilon_t \tag{24}$$

where

$$\Pi(L) \equiv I_n - \sum_i^k \Pi_i L^i \tag{25}$$

$$\boldsymbol{\alpha\beta'} = (\sum_i^k \Pi_i - I_n) \tag{26}$$

$$\phi_s = -(\Pi_{s+1} + \Pi_{s+2} + ... + \Pi_k) \tag{27}$$

for $s = 1, 2, ..., k-1$, and such that $|I_n - \Pi_1 z - \Pi_2 z^2 - ... - \Pi_k z^k| = 0$ 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

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

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

while in the four-dimensional case are

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

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

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

References

Ahn, K., Y. Bi, and S. Sohn (2019). Price discovery among sse 50 index-based spot, futures, and options markets. *Journal of Futures Markets 39*(2), 238–259.

Aït-Sahalia, Y. and J. Yu (2009). High frequency market microstructure noise estimates and liquidity measures. *The Annals of Applied Statistics 3*(1), 422 – 457.

Andersen, T. G., D. Dobrev, and E. Schaumburg (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics 169*(1), 75–93.

Audrino, F., G. Barone-Adesi, and A. Mira (2005). The stability of factor models of interest rates. *Journal of Financial Econometrics 3*(3), 422–441.

Baillie, R. T., G. G. Booth, Y. Tse, and T. Zabotina (2002). Price discovery and common factor models. *Journal of financial markets 5*(3), 309–321.

Baur, D. G. and T. Dimpfl (2019). Price discovery in bitcoin spot or futures? *Journal of Futures Markets 39*(7), 803–817.

Beveridge, S. and C. R. Nelson (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary economics 7*(2), 151–174.

Bollerslev, T., A. J. Patton, and R. Quaedvlieg (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics 192*(1), 1 – 18.

Booth, G. G., R. W. So, and Y. Tse (1999). Price discovery in the german equity index derivatives markets. *Journal of Futures Markets: Futures, Options, and Other Derivative Products 19*(6), 619–643.

Brogaard, J., T. Hendershott, and R. Riordan (2019). Price discovery without trading: Evidence from limit orders. *The Journal of Finance 74*(4), 1621–1658.

Brugler, J. and C. Comerton-Forde (2019). Comment on: Price Discovery in High Resolution. *Journal of Financial Econometrics*. nbz005.

Buccheri, G., G. Bormetti, F. Corsi, and F. Lillo (2019). Comment on: Price Discovery in High Resolution. *Journal of Financial Econometrics*. nbz008.

Chen, Y.-L. and W.-C. Tsai (2017). Determinants of price discovery in the vix futures market. *Journal of Empirical Finance 43*, 59–73.

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

Corsi, F. (2009). A Simple Approximate Long-Memory Model of Realized Volatility. *Journal of Financial Econometrics 7*(2), 174–196.

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

De Jong, F. (2002). Measures of contributions to price discovery: A comparison. *Journal of Financial markets 5*(3), 323–327.

De Jong, F. (2019). Comment on: Price Discovery in High Resolution*. *Journal of Financial Econometrics*. nbz006.

De Jong, F. and P. C. Schotman (2010). Price discovery in fragmented markets. *Journal of Financial Econometrics 8*(1), 1–28.

Dias, G. F., M. Fernandes, and C. M. Scherrer (2020). Price Discovery in a Continuous-Time Setting. *Journal of Financial Econometrics*. nbz030.

DiCiccio, T. J. and A. C. Monti (2004). Inferential aspects of the skew exponential power distribution. *Journal of the American Statistical Association 99*(466), 439–450.

Engle, R. F. and C. W. Granger (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.

Entrop, O., B. Frijns, and M. Seruset (2020). The determinants of price discovery on bitcoin markets. *Journal of Futures Markets 40*(5), 816–837.

Fabozzi, F. J., R. Giacometti, and N. Tsuchida (2016). Factor decomposition of the euro-zone sovereign cds spreads. *Journal of International Money and Finance 65*, 1–23.

Fernandes, M. and C. M. Scherrer (2018). Price discovery in dual-class shares across multiple markets. *Journal of Futures Markets 38*(1), 129–155.

García-Ferrer, A., E. González-Prieto, and D. Peña (2012). A conditionally heteroskedastic independent factor model with an application to financial stock returns. *International Journal of Forecasting 28*(1), 70–93.

Ghysels, E. (2020). Comment on: Price Discovery in High Resolution and the Analysis of Mixed Frequency Data*. *Journal of Financial Econometrics*. nbz007.

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

Gouriéroux, C., A. Monfort, and J.-P. Renne (2017). Statistical inference for independent component analysis: Application to structural var models. *Journal of Econometrics 196*(1), 111–126.

Grammig, J. and F. J. Peter (2013). Telltale tails: A new approach to estimating unique market information shares. *Journal of Financial and Quantitative Analysis*, 459–488.

Guerini, M. and A. Moneta (2017). A method for agent-based models validation. *Journal of Economic Dynamics and Control 82*, 125–141.

Hafner, C. M., H. Herwartz, and S. Maxand (2020). Identification of structural multivariate garch models. *Journal of Econometrics*.

Hagströmer, B. and A. J. Menkveld (2019). Information revelation in decentralized markets. *The Journal of Finance 74*(6), 2751–2787.

Hansen, P. R. and A. Lunde (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics 24*(2), 127–161.

Harris, F. H. d., T. H. McInish, G. L. Shoesmith, and R. A. Wood (1995). Cointegration, error correction, and price discovery on informationally linked security markets. *The Journal of Financial and Quantitative Analysis 30*(4), 563–579.

Harris, F. H. d., T. H. McInish, and R. A. Wood (2002a). Common factor components versus information shares: a reply. *Journal of Financial Markets 5*(3), 341–348.

Harris, F. H. d., T. H. McInish, and R. A. Wood (2002b). Security price adjustment across exchanges: an investigation of common factor components for dow stocks. *Journal of financial markets 5*(3), 277–308.

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

Hasbrouck, J. (2002a). The dynamics of discrete bid and ask quotes. *The Journal of Finance 54*(6), 2109–2142.

Hasbrouck, J. (2002b). Stalking the "efficient price" in market microstructure specifications: an overview. *Journal of Financial Markets 5*(3), 329–339.

Hasbrouck, J. (2003). Intraday price formation in us equity index markets. *The Journal of Finance 58*(6), 2375–2400.

Hasbrouck, J. (2019). Price Discovery in High Resolution*. *Journal of Financial Econometrics*. nbz027.

Hatheway, F., A. Kwan, and H. Zheng (2017). An empirical analysis of market segmentation on us equity markets. *Journal of Financial and Quantitative Analysis 52*(6), 2399–2427.

Hyvärinen, A. (2013). Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371*(1984), 20110534.

Hyvärinen, A. and E. Oja (1998). Independent component analysis by general nonlinear hebbian-like learning rules. *signal processing 64*(3), 301–313.

Hyvärinen, A. and E. Oja (2000). Independent component analysis: algorithms and applications. *Neural networks 13*(4-5), 411–430.

Hyvärinen, A., K. Zhang, S. Shimizu, and P. O. Hoyer (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research 11*(5).

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica: journal of the Econometric Society*, 1551–1580.

Kalke, S. and W.-D. Richter (2013). Simulation of the p-generalized gaussian distribution. *Journal of Statistical Computation and Simulation 83*(4), 641–667.

Koldovsky, Z., P. Tichavsky, and E. Oja (2006). Efficient variant of algorithm fastica for independent component analysis attaining the cramér-rao lower bound. *IEEE Transactions on neural networks 17*(5), 1265–1277.

Kryzanowski, L., S. Perrakis, and R. Zhong (2017). Price discovery in equity and cds markets. *Journal of Financial Markets 35*, 21–46.

Kwan, A., R. Masulis, and T. H. McInish (2015). Trading rules, competition for order flow and market fragmentation. *Journal of Financial Economics 115*(2), 330–348.

Lanne, M. and H. Lütkepohl (2010). Structural vector autoregressions with nonnormal residuals. *Journal of Business & Economic Statistics 28*(1), 159–168.

Lanne, M., M. Meitz, and P. Saikkonen (2017). Identification and estimation of non-gaussian structural vector autoregressions. *Journal of Econometrics 196*(2), 288–304.

Lehmann, B. N. (2002). Some desiderata for the measurement of price discovery across markets. *Journal of Financial Markets 5*(3), 259 – 276. Price Discovery.

Lien, D. and K. Shrestha (2009). A new information share measure. *Journal of Futures Markets: Futures, Options, and Other Derivative Products 29*(4), 377–395.

Lin, C.-B., R. K. Chou, and G. H. Wang (2018). Investor sentiment and price discovery: Evidence from the pricing dynamics between the futures and spot markets. *Journal of Banking & Finance 90*, 17–31.

Miettinen, J., K. Nordhausen, H. Oja, S. Taskinen, and J. Virta (2017). The squared symmetric fastica estimator. *Signal Processing 131*, 402–411.

Moneta, A., D. Entner, P. O. Hoyer, and A. Coad (2013). Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics 75*(5), 705–730.

Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics 32*(7), 685–694.

Nardon, M. and P. Pianca (2009). Simulation techniques for generalized gaussian densities. *Journal of Statistical Computation and Simulation 79*(11), 1317–1329.

O'Hara, M. and M. Ye (2011). Is market fragmentation harming market quality? *Journal of Financial Economics 100*(3), 459 – 474.

Putniņš, T. J. (2013). What do price discovery metrics really measure? *Journal of Empirical Finance 23*, 68 – 83.

Reyhani, N., J. Ylipaavalniemi, R. Vigário, and E. Oja (2012). Consistency and asymptotic normality of fastica and bootstrap fastica. *Signal processing 92*(8), 1767–1778.

Rigobon, R. (2003). Identification through heteroskedasticity. *Review of Economics and Statistics 85*(4), 777–792.

Shimizu, S., P. O. Hoyer, A. Hyvärinen, and A. Kerminen (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research 7*(Oct), 2003–2030.

Spirtes, P., C. N. Glymour, R. Scheines, and D. Heckerman (2000). *Causation, prediction, and search*. MIT press.

Yan, B. and E. Zivot (2010). A structural analysis of price discovery measures. *Journal of Financial Markets 13*(1), 1–19.

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% level in both LiNGAM and MLE t-student approaches. For the LiNGAM approach, 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 obtained 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 instead. Trades comprises both lit and dark trades, given that the contribution of the latter to the IS measure is negligible.

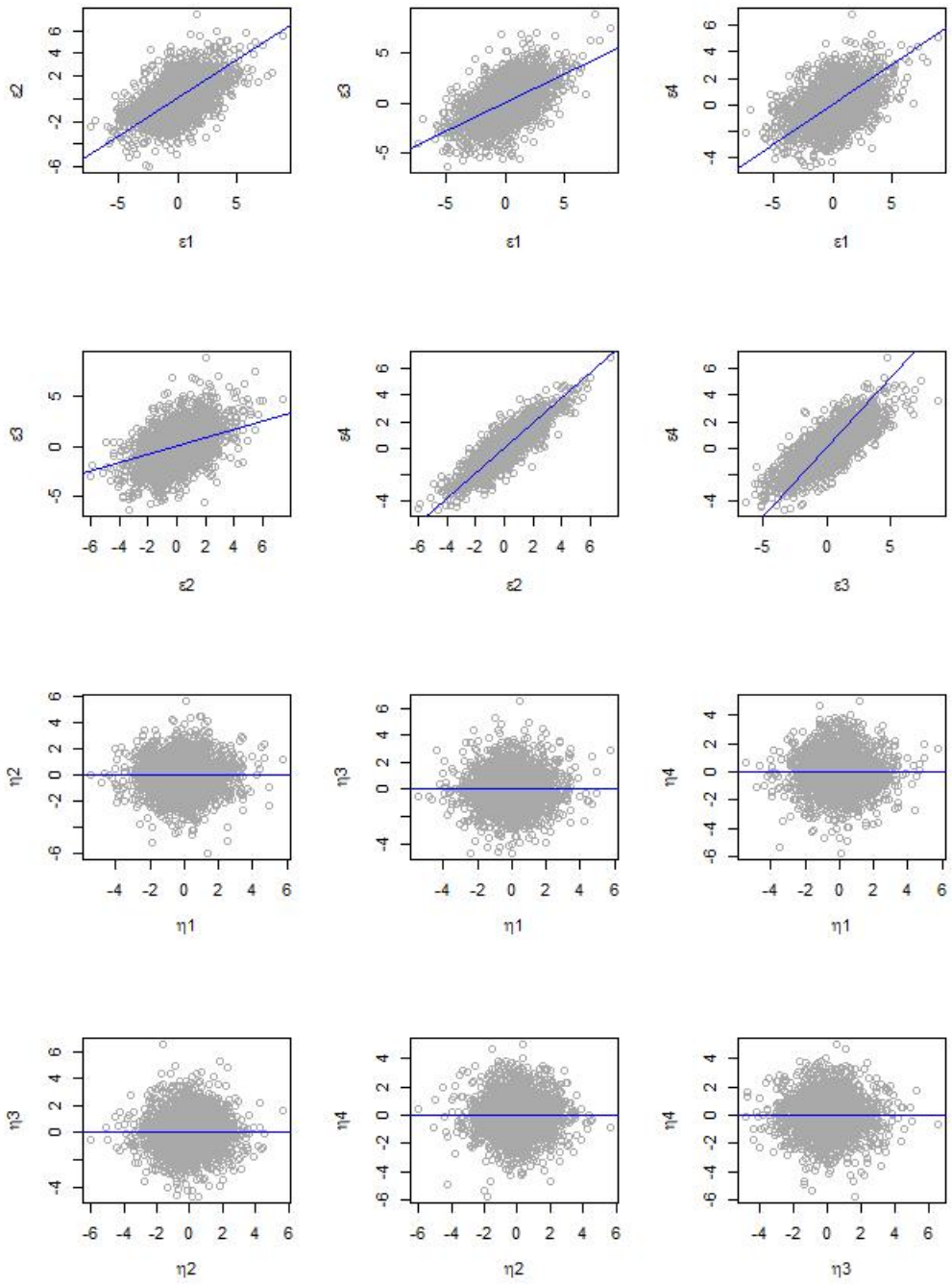

Figure 1: Scatter plots for the simulated residuals (top half) and estimated latent structural shocks (bottom half).

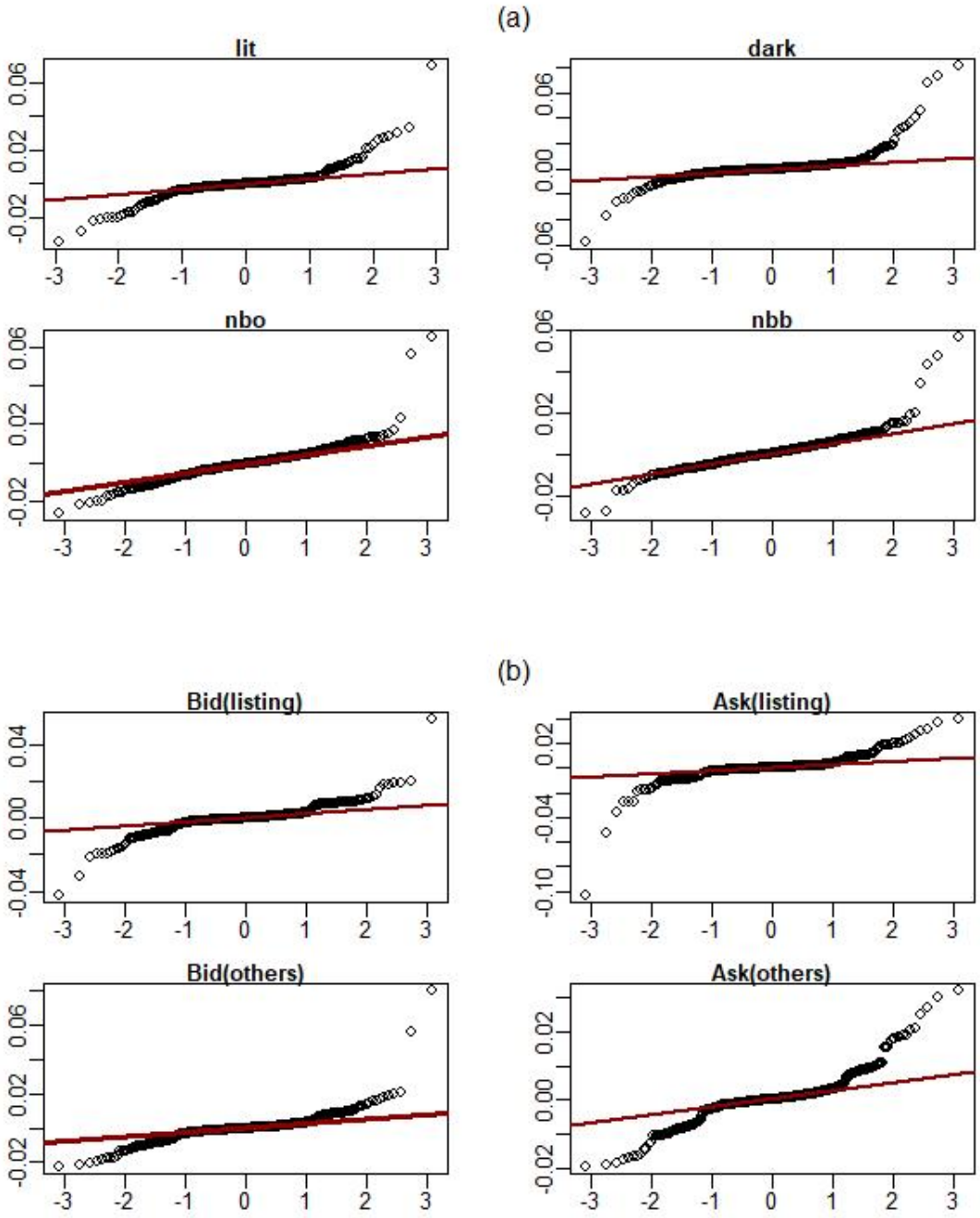

Figure 2: Quantile-quantile plots of the VECM residuals. In Panel (a) are displayed the model residuals related to the price discovery analysis across trades and quotes, while in panel (b) the one across exchanges using quotes.