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

A non-Normal framework for price discovery: The independent component based information shares measure

Sebastiano Michele Zema [1]
Francesco Cordoni [2]

[1] Scuola Normale Superiore, Pisa, Italy
[2] Royal Holloway College, University of London,UK

2023/03 **August 2024**
ISSN(ONLINE) 2284-0400

A non-Normal framework for price discovery: The independent component based information shares measure*

Sebastiano Michele Zema[1] and Francesco Cordoni[2]

[1]Scuola Normale Superiore, Pisa
[2]University of London, Royal Holloway College

December 2023

Abstract

We propose a new measure of price discovery, which we will refer to as the *Independent Component based Information Share* (IC-IS). This measure constitutes a variant of the widespread Information Share, with the main difference being it does not suffer the same identification issues. Under the assumptions of non-normality of the shocks, a rather general theoretical framework leading to the definition of the IC-IS and its estimation via a pseudo maximum likelihood (PML) approach is illustrated. After testing the robustness of the measure in a Montecarlo simulation environment, we illustrate two separate empirical analyses encompassing different price discovery applications.

Keywords: vector error correction models (VECMs); information shares; market microstructure; independent component analysis; pseudo maximum likelihood; price discovery

JEL classification: C32, C58, G14.

*The contents of this article are reflective and express the personal opinions of the authors only. Sebastiano Michele Zema acknowledges the financial support granted by the MIUR, with the PRO3 program "Network Analysis of Economic and Financial Resilience", during his period spent at the Sant'Anna School of Advanced Studies. Francesco Cordoni acknowledges the financial support from the Leverhulme Trust Grant Award RPG-2021-359.

1 Introduction

The quantification of the contribution of agents and exchanges to the price formation process acquired increasing importance in the literature. Processes of market fragmentation, carried out together with the proliferation of algorithmic trading strategies and the introduction of complex financial products, dramatically increased the complexity of financial markets and made the possibility to measure their informativeness a concrete challenge in the financial environment. In this respect, the information share (IS) measure of Hasbrouck (1995) represents a milestone in the literature, being one of the most widely adopted measures for price discovery as documented by its large 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).

From a market microstructure modeling perspective, the IS build its fundamentals upon the modeling of price changes through vector error correction models (VECM). The main shortfall of the IS measure is it can be uniquely determined only when the VECM residuals are not contemporaneously correlated. Hasbrouck's suggested solution was, in absence of a sound financial theory suggesting appropriate causal relationships, to identify the model by performing the Choleski decomposition on the covariance matrix of the residuals for all the possible permutations of the variables, which leads to upper and lower bounds for the IS. In empirical applications upper and lower bounds are often very wide because of substantial cross-correlations in the model residuals, raising interpretative ambiguities about the real allocation of information between the analyzed variables.

In this paper we propose a solution by defining a variant of the widespread IS measure, which we refer to as the *Independent Component based Information Shares* (IC-IS), and for which the identification issues related to the well-established IS are tremendously alleviated[1]. The proposed measure builds its fundamentals on the exploitation of non-normality for both the estimation and identification of the mixing matrix through which the shocks are revealed in the market. We thus provide a rather general estimation framework for the IC-IS based on pseudo maximum likelihood (PML) (Gourieroux et al., 1984; Gouriéroux et al., 2017), to let the true non-gaussian probability density functions be unknown.

As explained also in Hasbrouck (2003), the upper and lower bounds of the IS measure cannot be interpreted as confidence bands, nor can they be used for statistically testing the contribution of each variable, but rather as an attempt to solve the identification problem. Being the newly introduced IC-IS a precise point estimate for the contribution of each variable to price discovery, not only the identification problem is strongly alleviated, but a simple testing framework based on the asymptotic properties of the estimator of the mixing matrix can be provided as well.

The article is organized as follows. Section 2 reviews the state of the art, recalling the most recent updates on the topic and the need for further progress. Section 3 set up the general market microstructure framework on which the newly proposed measure will be based on. Section 4 introduces the new price discovery measure. In section 5 we show how

[1]We have to stress that our measure should not be either confused with the *Component-share* measure (see Harris et al., 1995; Booth et al., 1999) or mentioned as an alternative to the latter.

to estimate the proposed measure, with details on theory, simulations, and estimation. In section 6 we finally show a variety of price discovery applications: After using the same IBM data of Hasbrouck (2021) to have a sound benchmark to compare with, we move to analyze the price discovery process at different levels of the order book for four stocks listed in the S&P500. Section 7 concludes.

2 State of the art

The present work is not the first one dealing with such a long-standing issue. Even if several attempts have been made to solve the identification problem associated with the IS measure, a general strategy which explicitly deals with the identification problem is not available yet. The idea of estimating unique IS measures by exploiting the distributional properties of the variables was first introduced by Grammig and Peter (2013). The authors, inspired by Rigobon (2003), introduced different volatility regimes to identify the IS. The intuition was that the occurrence of extreme price changes causes differences between tail and center correlations, information which can be exploited to reach full identification of the model. Subject to the condition of observing different volatility regimes in the market, which might not always be the case, the above mentioned solution is effective.

A solution to the problem of obtaining unique information share measures can also be found in Lien and Shrestha (2009) and Fernandes and Scherrer (2018). Both authors, handled the problem by computing the IS on the spectral decomposition of either the correlation or the covariance matrix of the innovations. Even if these approaches are effective in getting unique measures, the problem at the origin of the impossibility of obtaining a precise quantification of the IS is the lack of an identification procedure commonly accepted. Computing the IS on the factor structure associated with the spectral decomposition of the covariance(correlation) matrix does not provide a solution to the identification issue which constitutes the real problem.

From a more recent data-driven perspective Hasbrouck (2021) proposed to exploit the high-frequency at which quotes and trades occur, modeling in natural time to drastically reduce the range obtained by permuting the variables. The idea is that sampling prices at incredibly short time scales, even at micro or nanoseconds precision, inevitably and drastically reduce the presence of contemporaneous cross-correlations (see also Dias et al., 2021), which consequently leads to narrower IS bounds and discards any possible interpretative ambiguity. Still, modeling in this natural time framework requires to estimate an enormous amount of coefficients. 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; de Jong, 2021; Ghysels, 2021).

Another recent contribution which tried to provide a solution to the identification problem of the IS, by fixing the permutation indeterminacy of the variables in the model, can be found in Zema (2022). The author proposed the adoption of a causal discovery model, well-

established in the machine learning literature, which exploits the non-Normal distributions of the variables to recover the directed acyclic graph (DAG) structure which is more likely to be true given the data. Given the obtained DAG, the associated causal chain was finally used to pick the corresponding permutation of the variable and compute the associated and unique IS measure. However, this approach works if and only if the assumption of the existence of a recursive causal structure in the system holds true.

In this respect, the present work tries to make a step forward in the literature by providing a rather generalized and practical framework for the estimation of unique market information shares when the shocks are non-normally distributed. This will lead to the introduction of the previously mentioned IC-IS, for which it is not necessary to assume the presence of either different volatility regimes or causal chains in the system. Moreover, the measure has been found to provide consistent results, even if in a limited sample, with no need to increase the model and computational complexities introduced when working at incredibly high resolutions in natural time.

While the scope of this work is to provide a solution to a long-standing issue in the context of price discovery through the IS measures of Hasbrouck (1995), it should be noted that a variety of other measures and approaches have been proposed in the literature for price discovery (Harris et al., 1995; Booth et al., 1999; De Jong and Schotman, 2010; Putniņš, 2013, see for instance). For the readers interested in having a general overview, comprehensive reviews of different price discovery measures and how they relate to each other can be found in Baillie et al. (2002), Lehmann (2002), and Yan and Zivot (2010).

3 Measuring price discovery: The general framework

The general market microstructure setting is the one of Hasbrouck (1995). Let $p_t = \{p_{1t}, p_{2t}, \ldots, p_{nt}\}$ be a vector of time series log-prices observed in n different exchanges but pertaining the same security. Being the time-series arbitrage linked, their dynamic can be modeled by the vector error correction model (VECM) of Engle and Granger (1987), specified as

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

with $\beta \in \mathbb{R}^{n\times n-1}$ containing the $n-1$ cointegrating vectors $p_1 - p_2$, $p_1 - p_3$, $p_1 - p_n$ and $\alpha \in \mathbb{R}^{n\times n-1}$ being a matrix of loadings. The system in equation 1 is covariance stationary, with $\mathrm{Cov}(u_t) = \Omega$, and admits the common trend representation

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

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

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

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 $IS_j = \psi_j^2 \Omega_{jj} / \psi \Omega \psi^{'}$, with ψ being the common row of $\Psi(1)$ and ψ_j denoting the *j-th* element of ψ corresponding to market j. In many empirical applications Ω is non-diagonal and the information shares are not identified. A practical solution widely adopted in the literature is to consider the Choleski decomposition $\Omega = FF^{'}$ and compute

$$IS_j = \frac{\left([\psi F]_j \right)^2}{\psi \Omega \psi^{'}} \tag{4}$$

for each possible permutation of the variables in the model so to get upper and lower bounds for each IS. Zema (2022) proposed to identify the IS measure by means of a causal search algorithm which exploits the non-Normal distribution of the variables to pick a specific order and performing Choleski accordingly. Still, the proposed solution works only when the assumption of a causal recursive structure among the variables is not violated. In the next section a generalized framework to identify and test the IS measure in a non-Normal setting will be introduced, without imposing any recursive causal structure in the system (i.e., lower triangular contemporaneous coefficients matrix).

4 Non-Normal identification: The independent component based information shares

Let consider the n-dimensional vector of price innovations $u_t = [u_{1t}, u_{2t}, ..., u_{nt}]$, with non-diagonal covariance matrix Ω, to be a linear combination of n unobserved shocks $\epsilon_t = [\epsilon_{1t}, \epsilon_{2t}, ..., \epsilon_{nt}]$:

$$u_t = B\epsilon_t \tag{5}$$

Where $B \in \mathbb{R}^{n \times n}$ is an invertible mixing matrix through which the unobserved shocks ϵ_t are revealed in each market. If ϵ is normally distributed, the knowledge of u makes $BB^{'}$ identifiable but B itself cannot be identified. For any non-singular matrix Q, the matrix B and the shocks ϵ_t could be replaced respectively by $B^* = BQ$ and $\epsilon_t^* = Q^{-1}\epsilon_t$ leading to an observationally equivalent model. However, when ϵ_t is not Normal the identification problem almost disappears and B can be identified under a few conditions. This follows from well established results (see Comon, 1994; Eriksson and Koivunen, 2004) which lead to the following theorem[2]

Theorem 4.1. *Let $u_t = B\epsilon_t$ and the following conditions hold true:*

i The latent shocks $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are mutually independent: $p(\epsilon_1, \epsilon_2, ..., \epsilon_n) = \prod_i^n p(\epsilon_i)$.

ii The sequence $\epsilon_1, \epsilon_2, ..., \epsilon_n$ contains at most one Normal distribution.

then, B can be identified by column permutation, sign and scaling.

[2]The interested readers might refer to Moneta et al. (2013) for more intuitive explanations, and graphical representations, about the advantages of exploiting the non-normality assumption for identification purposes. Comprehensive and rigorous explanations also given by Gouriéroux et al. (2020).

This brings important implications to our price discovery framework. If the conditions in Theorem 4.1 are satisfied it is possible to introduce the following new information share measure, which we will refer to as the *Independent Component based Information Shares* (IC-IS), for which the identification problem is strongly alleviated:

$$\text{IC-IS}_j = \frac{\left([\psi B^{(p)}]_j\right)^2}{\psi(B^{(p)}B^{(p)\prime})\psi'}. \tag{6}$$

Where $B^{(p)}$ is matrix B after a specific permutation of its columns has been picked, that is $B^{(p)} = BP$. The scaling indeterminacy (i.e., local lack of identification) can be easily removed by pre-whitening the price innovations u_t (Hyvärinen and Oja, 1998, 2000; Moneta et al., 2013; Gouriéroux et al., 2017). Permutation and change in signs of the columns in B imply global lack of identification instead, since once B is estimated, the order in which the shocks are returned and the signs of their impact are unknown. The sign indeterminacy in the columns of B is not an issue in our price discovery framework. Being the newly defined IC-IS still a ratio between two quadratic forms, the sign of the columns of B does not influence the variance allocation mechanism for the efficient price process.

The only remaining cause of lack of identification is the column permutation indeterminacy. We thus need a criterion to fix an appropriate permutation of the column of the mixing matrix used to compute the IC-IS in equation 6, which leads us to the following proposition.

Proposition 4.1. *Let P be a permutation matrix such that the matrix $B^{(p)} = BP$ satisfies $|b_{ii}| \geq |b_{ij}| \ \forall \ \ i \neq j$ and assume the conditions stated in Theorem 4.1 are met. Then, the IC-IS measures defined in equation 6 are invariant to arbitrary permutations of the variable in the model.*

Proof. See Appendix A. □

The identification criteria is simply that of choosing a permutation matrix P so that $B^{(p)}$ satisfies $|b_{ii}^{(p)}| > |b_{ij}^{(p)}| \ \forall \ \ i \neq j$. Such permutation convention is quite standard in the literature (see for instance Ilmonen et al., 2011; Hallin and Mehta, 2015; Lanne et al., 2017) and it was interestingly adopted also by Grammig and Peter (2013) in their price discovery analysis[3]. In the current price discovery framework, such permutation criterion imposes nothing more than the following: each price series reacts to its own shock more than what other price series do. Hence, the measure is now identified since it is invariant to arbitrary permutation of the column of B which was the only source of indeterminacy left in our framework.

It is worth noticing the flexibility of such approach. When the permutation criterion just mentioned is not plausible from an economic standpoint, alternative permutation strategies for the columns of B can be implemented accordingly. This could be the case, for instance,

[3]It is worth noticing the authors reasonably assumed also $b_{ii}^{(p)} > 0$, together with the presence of two different variance regimes with one of them being strictly greater than one, to solve the sign and scaling indeterminacy respectively. As previously illustrated, such conditions, even if reasonable, are not necessary in our framework to reach identification of the model.

of price discovery trough derivative instruments (Blanco et al., 2005; Guidolin et al., 2021; Ahn et al., 2019). Imagine we want to quantify the contribution to the price formation process of leveraged exchange traded funds (ETFs). Leveraged ETFs are synthetic products which do not hold the underlying assets taken as benchmarks for their investment strategy. Still, these products replicate the benchmark's dynamic amplifying its return by opening derivative positions on the underlying (Leung et al., 2017; Shum et al., 2016). The spillover effect on these leveraged ETFs, originating from a shock on the underlying assets, would have a higher magnitude than the shock itself. As a consequence, considering for simplicity a system of two variables only, the permutation matrix P might be such that $B^{(p)} = BP$ satisfies the condition $b_{11}^{(p)} < b_{21}^{(p)}$ instead, where 1 = underlying and 2 = leveraged ETF.

Then, what is left is to choose an appropriate estimator for B so that we can compute the IC-IS previously introduced. This will be the object of the next section.

5 Estimation and testing

We set up an estimation framework for equation 5 based on the pseudo maximum likelihood (PML) framework of Gouriéroux et al. (2017). It should be remarked that other estimation strategies, different from the PML approach, could be implemented to get an estimate of the mixing matrix B (see Hyvarinen, 1999; Hyvärinen et al., 2010; Lanne and Lütkepohl, 2010; Lanne et al., 2017; Lanne and Luoto, 2021, among others), computing the IC-IS consequently. For instance, one could opt for a simpler but less general framework in which the non-Normal distributions are assumed to be known in a standard maximum likelihood estimation (MLE) as in Lanne et al. (2017).

The IC-IS measure we propose is rather flexible since it does not require, to be identified, the knowledge of the specific non-Normal distributions governing the shocks' behavior. For this reason, the PML approach is appealing since it allows to estimate B when the true probability density functions (p.d.f.) of the unobserved shocks ϵ_t are unknown. A comparison between different estimation strategies is not the objective of this work[3]. Still, the IC-IS should be robust to different sources of miss-specifications in the pseudo log-likelihood. We will show, via Montecarlo simulations, the PML to be a robust estimator for our IC-IS measure not only with respect to misspecification of the likelihood function in terms of non-Normal distributions chosen, but to time-varying variance in the error terms as well.

5.1 A pseudo maximum likelihood approach (PML)

For a correct estimation and identification of equation 5 we need first to import a set of assumptions as in Gouriéroux et al. (2017).

Assumption 5.1.

i The process $\epsilon_t = (\epsilon_{1t}, \dots, \epsilon_{nt})$ is an i.i.d. process with $E(\epsilon_t) = 0$ and $V(\epsilon_t) = Id$.

[3] An interesting evaluation study encompassing some of these different approaches can be found, among others, in Moneta and Pallante (2022).

ii The shocks $\epsilon_i, \ldots, \epsilon_n$ are mutually independent.

Such assumptions allow us to locally identify model by pre-whitening the price innovations u_t and set up its pseudo log-likelihood function. Being Σ the covariance matrix of u_t and S any matrix satisfying $SS^{'} = \Sigma$, we express $u_t = B\epsilon_t$ as $u_t = SC\epsilon_t$ so that $\tilde{u}_t = S^{-1}u_t$ are the pre-whitened innovations and C is orthogonal ($CC^{'} = Id$).

We work then with $\tilde{u}_t = C\epsilon_t$ and implement the PML estimator of matrix C. Let us consider a set of unknown p.d.f. $g_i(\epsilon_i)$, where $i = 1, \ldots, n$, and consider the pseudo log-likelihood function

$$\ln \mathcal{L}_T(C) = \sum_{t=1}^{T} \sum_{i=1}^{n} \ln g_i(c_i^{'} \tilde{u}_t) \tag{7}$$

where $c_i^{'}$ is the i-th row of the orthogonal matrix C, and $c_i^{'}\tilde{u}_t = \epsilon_i$. The problem to be solved is

$$\begin{aligned} &\hat{C} = \underset{C}{\text{argmax}} \ \ln \mathcal{L}_T(C) \\ &\text{s.t. } C'C = Id. \end{aligned} \tag{8}$$

The problem (8) is a typical constrained optimization where the constraints needed for local identification consist in the orthogonality condition for C, that is $c_i^{'}c_j = 0$ for $i < j$, and $c_i^{'}c_i = 1 \ \forall \ \ i$. The finite-sample first order conditions (FOCs) to be solved can be written as

$$\begin{cases} \sum_{t}^{T} \tilde{u}_t \dfrac{\partial \ln g_i(\hat{c}_i^{'}\tilde{u}_t)}{\partial \epsilon_i} - \sum_{j=1}^{n} \hat{\lambda}_{ij}\hat{c_j} = 0, \ \ \forall \ i, \\ \hat{c}_i^{'}\hat{c}_i = 1, \ \ \forall i \\ \hat{c}_i^{'}\hat{c}_j = 0, \ \ i < j. \end{cases} \tag{9}$$

The consistency of the estimates resulting from the system of equations (9) have been proven, under a set of necessary assumptions in addition to Assumption 5.1, by Gouriéroux et al. (2017). The assumptions are illustrated in the following.

Assumption 5.2.

i The functions $\ln g_i$ are twice continuously differentiable.

ii Uniform integrability of the (pseudo)likelihood function: $\sup_C |\sum_i^n \ln g_i(c_i^{'}\tilde{u}_t^{'})| \leq \infty$.

Assumptions 5.1 and 5.2 are needed to guarantee the uniform convergence of the finite-sample objective function $Q_T(C) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} \ln g_i(c_i^{'}\tilde{u}_t)$ to the asymptotic one $Q_\infty(C) = E[\sum_i^n \ln g_i(c_i^{'}\tilde{u}_t^{'})]$ (Gourieroux et al., 1984; Gourieroux and Monfort, 1995).

Assumption 5.3. *(Identification). The system of equations in 9 admits as only solutions the elements of the set $\mathcal{P}(C)$ obtained by permuting and multiplying by -1 the columns of C.*

Assumption 5.3 pertains to the identification problems already discussed, meaning we assume the output of the PML estimator $\hat{C}$ to be identified up to permutation and sign

change of its columns. This assumption is not violated as long as the conditions outlined in Theorem 4.1 hold true, that is, ϵ_t have at most one Normal distribution, and its components are mutually independent. This implies restrictions on the true distribution of ϵ_T and, consequently, on the choice we can make on the definition of the pseudo probability density functions in the PML procedure. Being Assumption 5.3 satisfied, we are capable of estimating C up to the column's sign and permutations. As evidenced by Gouriéroux et al. (2017), even if Assumption 5.3 is satisfied, we need two additional assumptions to ensure that asymptotically the PML yield us an estimate $\hat{C}$ which correspond to a maximum of the optimization problem in (8). The assumptions are the following

Assumption 5.4. *(Local concavity). The (pseudo)log-likelihood function ln-$\mathcal{L}_T(C)$ is locally concave in a neighbourhood of a matrix C of $\mathcal{P}(C)$.*

Assumption 5.5. *(Distinct distributions). The pseudo distributions g_i, as well as the true distributions of the ϵ_{it}, are different and asymmetric.*

Assumption 5.4 guarantees the matrix C to be a local maximum of the optimization problem. Moreover, if Assumption 5.5 is also satisfied, the values at the local maxima for C will be all different so that the global maximum will be reached by a unique element C of $\mathcal{P}(C)$. Under Assumptions 5.1-5.5 the PML estimator $\hat{C}_T$ of C is asymptotically normal with the speed of convergence $\sqrt{T}$, that is $vec\sqrt{T}(\hat{C}_T - C_0) \sim N(0, \Sigma_C)$[4].

Given the PML estimate of C, we can proceed to compute the IC-IS measure defined in equation 6. Recalling that we pre-whitened the price innovations, being the column permutation indeterminacy on C, we have that $B^{(p)} = SC^{(p)}$ where S is the matrix used for pre-whitening. We thus compute the IC-IS for market/variable j as $(\psi SC^{(p)}]_j)/\psi(SC^{(p)}C^{(p)\prime}S')\psi'$.

As already illustrated, S helps us remove the scaling indeterminacy in the estimates. The sign indeterminacy of the columns of C is not relevant for our measure, while the permutation indeterminacy is dealt with by choosing a permutation matrix P so that the permutation criterion for $B^{(p)}$ mentioned in section 4 is satisfied. Finally, we can exploit the asymptotic properties of $\hat{C}_T$ to derive the asymptotic properties of the IC-IS measure. This leads us to the following proposition

Proposition 5.1. *Given $vec\sqrt{T}(\hat{C}_T - C_0) \sim N(0, \Sigma_c)$ for $T \to \infty$, and given $\hat{B}_T = S\hat{C}_T$, it holds $B_T \sim N(SC_0, S\Sigma_c S')$. Being ψ the common row of $\Psi(1)$, let consider the contribution $\psi\hat{b}_j$, with $\hat{b}_j = s_j'\hat{c}_j$, of market j to the variance of the efficient price process. Then, the second central moment of $\psi\hat{b}_j$ is distributed according to a Gamma distribution with shape parameter $\lambda = 1/2$ and scale parameter $k = 2\psi s_j \Sigma^c_{jj} s_j'\psi'$. That is, $(\psi\hat{b}_j - \psi b_j)^2 \sim \Gamma(1/2, 2\psi s_j \Sigma^c_{jj} s_j'\psi')$.*

Proof. See Appendix B. □

[4]For the sake of clarity and readability, we refer the reader interested in the proof of the asymptotic normality of the PML estimator under the mentioned assumptions, and computational details, to the Appendix B of Gouriéroux et al. (2017).

Remark 1. The Gamma distribution simply arise as a scaled-χ^2 with scaling parameter equal to $\psi s_j \Sigma^c_{jj} s_j^{'} \psi^{'}$, that is $(\psi \hat{b}_j - \psi b_j)^2 / \psi s_j \Sigma^c_{jj} s_j^{'} \psi^{'} \sim \chi^2_1$. Then $(\psi \hat{b}_j)^2 \sim \sigma \chi^{'2}(m)$, where $\sigma \chi^{'2}(m)$ is a non-central $\chi^{'2}(m)$ with non-centrality parameter $m = \psi b_j$ multiplied by a scaling factor $\sigma = \psi s_j \Sigma^c_{jj} s_j^{'} \psi^{'}$.

The above results imply that typical Wald testing procedures can be easily implemented to test whether the contribution of each market to the price discovery process is significant or not. That this we can simply test $H_0 : \psi \hat{b}_j = 0$ through the Wald statistics $\psi \hat{b}_j / \psi s_j \Sigma^c_{jj} s_j^{'} \psi^{'}$, which under the null it follows asymptotically a χ^2_1 distribution as usual in a single parameter testing framework.

All of these results can be summarized as follows. First, when the transitory shocks generating market microstructure noise are not normally distributed, the historical identification problem of the IS measure is tremendously alleviated since it is not necessary to compute all the possible permutations in the model to get a heuristic range of solutions for the IS measure. A unique IC-IS measure, consisting of a precise point estimate for the contribution of each market/variable to the price formation process, can be defined and implemented. Second, this measure can be statistically tested starting from the asymptotic distribution of the estimated matrix C of contemporaneous coefficients. This represents in a sense a change of perspective for the interpretation of the measure, since it is not required anymore to get upper and lower bounds which are often very large thus raising interpretative ambiguities.

What remains is to assess the robustness of the IC-IS measure to both misspecifications of the pseudo log-likelihood and heteroskedasticity. As a matter of fact, heteroskedasticity is a stylized fact of financial time series. Still, in the PML framework formulated in equation 7, ϵ_t is assumed to be an $i.i.d.$ process which is incompatible with a heteroskedastic behaviour. Hence, it is mandatory to show that such an assumption in the PML estimation of the mixing matrix C, is not harmful in our price discovery framework, having no detrimental effect in practice for a correct quantification of the IC-IS measure. To this purpose, we set up a comprehensive Montecarlo simulation where we show our measure to be robust to different misspecifications including heteroskedasticity in the error term.

Before moving to the Montecarlo simulations, it is worth mentioning that the $i.i.d.$ assumption for ϵ_t could be relaxed following the generalized method of moments (GMM) approach introduced by Lanne and Luoto (2021) to estimate the mixing matrix B. The GMM framework, despite being less restrictive in terms of assumptions, it is highly computationally intensive and time consuming. In addition, as shown by Lanne and Lütkepohl (2010) themselves, yields estimates which are inferior in terms of accuracy compared to the PML. For the sake of completeness, we nevertheless provide an adaptation of the GMM approach to our price discovery framework to have less restrictive assumptions. Still, for the reasons just mentioned and for the sake of readability, we limit the GMM discussion for price discovery to a supplemental appendix [5].

[5] Codes for both the PML and GMM will accompany the supporting material.

5.2 Montecarlo simulations

In this subsection the robustness of the IC-IS measure to different sources of misspecifications in the density functions $g_i(\epsilon_i)$ is investigated. The IC-IS measure, computed starting from the PML estimates, should not be too sensitive to the specification of the pseudo log-likelihood function. In other words, the allocation mechanism for the variance of the common trend should be consistent across different suitable non-Normal density functions chosen. The results will be compared with the standard and well established IS computed with upper and lower bounds associated with different Choleski decompositions.

We simulate N = 500 samples each of length T = 5000. The shocks ϵ_t are drawn from Student distributions with time-varying degree of freedoms v_t to let the variance change over time with a U-shape pattern to take into account heteroskedasticity in the error terms (Andersen et al., 2012; Bollerslev et al., 2016; Hasbrouck, 2002)

$$\sigma_\epsilon(t) = M + De^{-dt} + We^{-w(1-t)} \tag{10}$$

where parameters are set as $M = 1$, $D = 0.75$, $W = 0.25$, $d = 10$, and $w = 10$ (see Appendix B). Time-varying variance is introduced for two main reasons. First, the U-shape patterns allow us to simulate the data more realistically. Intraday financial returns typically display higher levels of volatility both at the beginning and at the end of the trading day, with lower levels of volatility in the middle. Second, being the variance of a Student equal to $v/(v-2)$, the time-varying variance is obtained by letting the degrees of freedoms $v(t)$ of the Student distributions vary accordingly. This introduces an additional source of misspecification with respect to which the robustness of the IC-IS measure is evaluated, being the true distributions not known to the econometrician.

After having specified the 4-dimensional orthogonal mixing matrix C to be estimated, and a matrix S which is used to cross-correlate the shocks ϵ_t to get the correlated price innovations $u_t = SC\epsilon_t$, we simulate 4-dimensional VECMs using the price innovations u_t. For each simulated sample, the VECM is estimated equation by equation given the known cointegration vectors. The price innovations u_t are recovered as residuals and jointly prewhitened using a Choleski decomposition to both remove the scaling indeterminacy and be compliant with the orthogonality conditions. Then, the PML procedure is performed on the whitened innovations, obtaining the estimates $\hat{C}$ needed to compute the IC-IS as illustrated in section 3.

Being $\mathcal{D}_i$ the true Students with time-varying $v_i(t)$, we set as pseudo distributions $g_i = t(v_i)$, meaning we use Student distributions with different but fixed and predetermined degrees of freedom v_i. Different combinations of degrees of freedom v_i will be considered. The obtained IC-IS are compared with the true IS implied by the simulated model parameters [4], but also with the well-known upper and lower bounds of the IS we would obtain by performing Choleski decompositions over all the possible variable permutations in the model. As an additional robustness check the IC-IS will be computed under an additional source of misspecification: We do not limit the misspecification to the degrees of freedom

[4]Parameters are shown in Appendix C, all codes for both the empirical and simulation analyses are available publicly in a supplemental Appendix.

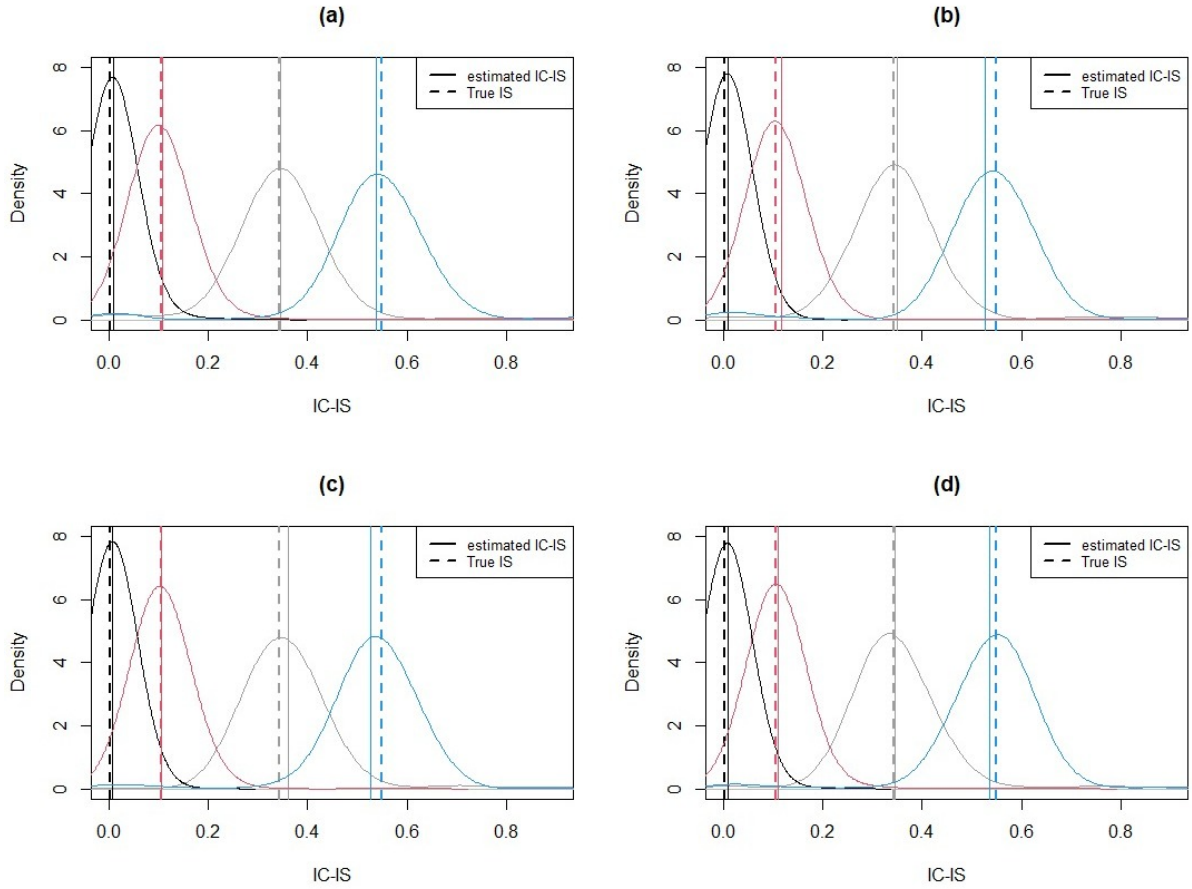


Figure 1: Comparison between the estimated IC-IS (continuous vertical lines) and true IS (dashed lines) measures for each simulated variable, together with the underline distributions of the IC-IS obtained from the N=500 Montecarlo samples. Each color corresponds to a different variable.

of the Student but we estimate the IC-IS using other non-Normal distributions as well, namely the Laplace and the Hyperbolic secant distributions.

Table 1 reports the average values of the IC-IS obtained from the 500 simulated samples. It is common practice in price discovery applications to get a 'mid-point' of the IS range obtained by implementing all the possible Choleski permutations in the model. For this reason we also display that mid-point under the field '(Mean range)', which simply is the average of the upper and lower bounds. Results can be summarized as follows. Independently from the non-Normal distributions chosen in the specification of the pseudo log-likelihood, the proposed IC-IS always allocates the variances across the variables consistently (i.e., the ranking in terms of informativeness is respected and the magnitude of the different shares is close to the true ones). Figure 1 graphically shows some of the scenarios illustrated in table 1, comparing the estimated IC-IS with the true values and showing also the distribution of the IC-IS measure which belongs to the closed interval [0,1].

Remark 2. Given $(\psi\hat{b}_j - \psi b_j)^2 \sim \Gamma(1/2, 2\psi s_j \Sigma^c_{jj} s_j' \psi')$, the quantity X= $(\psi\hat{b}_j$-$\psi b_j)^2 / \sum_{j=1}^{N} (\psi\hat{b}_j$

Table 1: Montecarlo simulation results.

True IS = $[0.0025, 0.1042, 0.5500, 0.3433]$				
	(1) $\epsilon_{1,3} \sim t(3), \epsilon_{2,4} \sim t(4)$			
IC-IS	0.0103	0.1194	0.5186	0.3515
All permutations	[0.0011, 0.4563]	[0.0000, 0.3018]	[0.1242, 0.9168]	[0.0151, 0.4194]
(Mean)	(0.2287)	(0.1509)	(0.5205)	(0.2173)
	(2) $\epsilon_{1,3} \sim t(5), \epsilon_{2,4} \sim t(6)$			
IC-IS	0.0084	0.1133	0.5313	0.3469
All permutations	[0.0046, 0.5027]	[0.0035, 0.3125]	[0.1191, 0.8695]	[0.0163, 0.3716]
(Mean)	(0.2537)	(0.158)	(0.4943)	(0.1940)
	(3) $\epsilon_1 \sim t(4), \epsilon_2 \sim t(5), \epsilon_3 \sim t(7), \epsilon_4 \sim t(12)$			
IC-IS	0.0072	0.1042	0.5371	0.3515
All permutations	[0.0000, 0.4280]	[0.0016, 0.2800]	[0.1316, 0.9381]	[0.0180, 0.4387]
(Mean range)	(0.2140)	(0.1408)	(0.5349)	(0.2284)
	(4) $\epsilon_{1,2,3,4} \sim$ Laplace$(\mu = 0, p = 1)$			
IC-IS	0.0358	0.1473	0.4624	0.3545
All permutations	[0.0035, 0.5276]	[0.0009, 0.2597]	[0.1199, 0.8755]	[0.0211, 0.3604]
(Mean range)	(0.2656)	(0.1304)	(0.4976)	(0.1908)
	(5) $\epsilon_{1,2,3,4} \sim$ Hyperbolic secant			
IC-IS	0.0102	0.1066	0.5405	0.3427
All permutations	[0.0009, 0.5170]	[0.0000, 0.2557]	[0.1265, 0.8969]	[0.022, 0.3776]
(Mean range)	(0.2590)	(0.1278)	(0.5117)	(0.1998)

Notes: Average IC-IS obtained in each Montecarlo simulation. Five different simulation settings from (1) to (5) have been implemented, each of them corresponding to a different specification of the pseudo log-likelihood used to estimate the matrix C and needed to compute the IC-IS. For each specification, the results are compared with the IS obtained performing the standard Choleski procedure with upper and lower bounds and their average. The true IS implied by the data generating process are shown in the first row of the table.

- $\psi b_j)^2$ follows a Beta distribution being a ratio of independent Gamma distributions of the form $\Gamma_j / \sum_{j=1}^{N} \Gamma_j$. Then, the non-central version of this quantity is exactly the IC-IS measure which arises as the ratio of non-central chi-square distributions, which follows the non-central Beta distribution (see Johnson et al., 1995).

From such Montecarlo simulation experiments we can conclude that our price discovery measure, which bases its fundamentals on the identification via non-gaussian distributions, is capable of correctly disentangling market information shares accurately even under different sources of misspecifications in the log-likelihood. Such results are also in line with those of Herwartz et al. (2022), where the authors show that statistical identification schemes which exploit non-gaussianity to recover independent components are generally robust under very different data structures.

6 Empirical Applications

In this section, we present two empirical applications which are different in their scope. The first analysis consists of a replication of Hasbrouck (2021)'s results on IBM data, so that we can compare the results of the classical IS, based on all Choleski permutations, with the proposed IC-IS measure. We thus move to a second empirical application in which we investigate, for the first time to our knowledge, the information content of quotes at different levels of the order-book. By exploiting intraday order-book data from the LOBSTER database, we try to disentangle whether limit order placed at deeper level of the order book contribute to the price discovery process as those placed in the first level (i.e., best bid and ask). We perform such empirical investigation for 4 different US stocks, which are Amazon, Microsoft, BlackRock and Abiomed.

6.1 Comparison of ISs on IBM data: replication of Hasbrouck (2021)

To evaluate the goodness of the proposed measures, we perform the first empirical application 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 [6]. This allows to keep detailed analyses already established in the literature as a benchmark to compare with, making clearer the interpretation and validity of the obtained results. The recent results of Hasbrouck (2021) have already been reproduced also by Zema (2022), for this reason, they will be simply reported here with no need to recompute them. The econometric analysis is then performed on IBM's trades and quotes recorded on the day 3 October 2016, with each record reporting both participants and Securities Information Processor (SIP) timestamps, with a sample for the day consisting of around 35.000 observations for each variable.

The empirical analysis follows three main lines of investigation. The first study focuses on the analysis of participants and SIP timestamps, quantifying the impact of time reporting differentials on the measurement of price discovery. SIP data are needed to establish a consolidated and transparent way to disseminate market data to the public audience.

[6] We deeply thank Joel Hasbrouck for sharing the data

Starting from participants' trades and quotes, the SIP compute and publicly disseminates national best bids (NBBs) and offers (NBOs) at which brokers are required to trade, by the regulation, when acting on behalf of their customers. Since SIP data are by construction delayed signals of the participants ones, one expects to attribute leadership in the price discovery process to the participants-based data almost entirely. To perform the analysis, a 4-variables VECM will be estimated including both SIP national NBBs and NBOs plus participants bid and ask prices.

The second study will quantify price discovery across different exchanges instead. Financial instruments are often traded on multiple markets. In particular, public companies can have their stocks traded contemporaneously both in the primary listing exchanges (i.e., where the initial public offering occurred) and other exchanges indeed (same examples include cross-listing, dark pools, OTC markets among others). The VECM here will include IBM bids and offers placed on the primary listing exchange, plus the best bids and offers that were placed in all the other exchanges in which IBM was traded except the primary one.

Finally, the third study analyzes the contribution to price discovery of trades and quotes. Here, the model will include trades that occurred on lit and dark pools plus NBBs and NBOs quotes from participant timestamps. Differently from lit pools characterized by stricter regulatory requirements (such as NYSE, NASDAQ, or LSE among others), dark pools are alternative private trading venues with no regulatory transparency requirements. The rationale behind the existence of these dark pools is to let institutional investors trade large volumes of securities without making their hands visible. The analysis of the benefits of lit versus dark pools is not the objective of OUR analysis from a regulatory perspective, we rather want to quantify the informational content of trades versus that of all quotes in the book. For this reason, information shares of lit and dark trades have been summed up together and compared with quotes' information shares[7].

The three VECM models implemented to perform the three empirical analyses mentioned are thus separately estimated on the following variables

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. $\mathrm{p}_{\mathrm{t}}^{\mathrm{Model3}} = \left[\mathrm{NBB}_{\mathrm{t}}^{\mathrm{Participants}}, \mathrm{NBO}_{\mathrm{t}}^{\mathrm{Participants}}, \mathrm{Trade}_{\mathrm{t}}^{\mathrm{LitPools}}, \mathrm{Trade}_{\mathrm{t}}^{\mathrm{DarkPools}}\right]$.

The results of the analyses are then displayed and commented in the next subsection[8].

6.1.1 Empirical results

For each model, related to a given price discovery analysis, the IC-IS measure is computed and compared with the standard Choleski based IS in which upper and lower bounds are

[7]However, as expected and in line with Hasbrouck (2021), dark trades are found by the IC-IS to have a contribution to price discovery which is not statistically different from zero.

[8]Having Hasbrouck (2021) as a direct comparison for this section, all VECM models have been estimated using the same specification adopted by the authors with a maximum number of lags $k = 10$

Table 2: IC-IS and IS comparison: Summary results.

IC-IS		All permutations			
participants	SIP	participants		SIP	
		Min	Max	Min	Max
0.997	0.002	0.943	0.999	0.001	0.057
primary	non-primary	primary		non-primary	
		Min	Max	Min	Max
0.78	0.22	0.46	0.56	0.44	0.54
Quotes	Trades	Quotes		Trades	
		Min	Max	Min	Max
0.49	0.51	0.61	0.67	0.33	0.39

Notes: Information shares measures for each price discovery analysis, comparing the IC-IS with the classical IS based on all permutations. Trades include both lit and dark trades, given that the contribution of the latter is negligible. The all permutations approach yielded results consistent with Hasbrouck (2021). For the specification of the pseudo log-likelihood pairs of Student distributions with 3 and 4 d.o.f. have been used, but results have been found to be consistent with the adoption of other heavy tail distributions such as the Laplace or the Hyperbolic secant.

computed by going through all the possible permutations of the variables in the model. The results are shown in table 2 and can be summarized as follows.

In the participant versus SIP timestamps analysis, the IC-IS attributed importance to the participants in the price discovery process almost entirely, as obviously expected, with a 99 percent share of price discovery which is consistent with the results of Hasbrouck (2021) based on the classical IS measure. Al already mentioned, SIP timestamps are simply delayed signals of the participant timestamps. For this reason, quantifying price discovery in such framework is rather easy and serves as the starting point to evaluate the possible goodness of the measure we are proposing.

In the trade versus quotes analysis, the IC-IS measure and the classical IS based on all permutations interestingly gave results which are in contrast with each other instead. While the range based on Choleski-permutations tells us that quotes are more informative than trades, the IC-IS measure yields very balanced results in which trades are as informative as quotes or even more (51 percent against 49 percent). Being such an analysis meant to be a comparative one, based on one trading only and for just a single listed company (IBM), it is not possible to draw general conclusions about the intraday information content of quotes and trades. Nevertheless, it is worth stressing what follows to have a plausible sense of the discrepancies across measures and estimators. The Choleski-based IS measure entirely is based on the second moment of the distributions of ϵ_t only, hence huge loss of information might happen in the quantification of the measures. This is not the case of the IC-IS, where higher moments than the second one are taken into account. As shown in the previous section, even in a simulated environment the permutation-based approach can fail in detecting the right IS under non-gaussianity and time-varying variances, (under)overestimating the contribution of variables with (higher)lower levels of informativeness. Thus, it is reasonable to expect prices at which the trades actually occurred to be more informative, compared to what the Choleski based approach would quantify, than all other quotes entered in the book.

More interestingly, and consistently with the above considerations, the IC-IS measure allowed us to get rid of some quantification ambiguities related to the price discovery analysis of primary versus non-primary listing exchanges. While the IC-IS attributed a clear 78 percent share to the primary listing market, the standard IS based on all permutations was not capable of clearly distinguishing between the primary and non-primary exchanges. Such results have been achieved without either increasing the modeling and computational complexity which arise when working at incredibly short time scales (i.e., microsecond precision as in Hasbrouck (2021), or introducing the rather restrictive directed acyclic graph structure assumption of Zema (2022).

6.2 Price discovery across order-book levels

In financial markets, large blocks of shares are commonly split by traders into smaller orders which are subsequently placed at different levels of the order book over a designated time interval. In this section we focus on quantifying the information content of those orders placed at different and deeper levels of the order-book, comparing their information shares ratios with those of the best bid and ask orders. As we will show, while our newly proposed

IC-IS measure is remarkably capable of clearly distinguishing the best bid and ask orders from limit orders placed at different levels of the book, the same does not hold for the classical IS measure based on the midpoint of the min-max range.

A correct quantification of the informational content of orders placed at different levels of the order-book is relevant for multiple reasons. First, the limit order book can be informative about future price changes. Having the possibility to see deeper levels of the book or even the entire system of limit orders might offer a unique advantage when competing with other traders in the market, making it possible to strategically place the orders over specific price levels at which the trade should be executed (Harris and Hasbrouck, 1996; Harris and Panchapagesan, 2005; Cao et al., 2009). Second, a significant challenge when large orders are executed is deciding between executing orders by crossing the spread or placing passive orders at subsequent levels of the order book. Although passive orders may reduce market impact, they do not guarantee immediate execution. However, empirical studies suggest that passive orders can also influence market impact by distorting the limit order book (Donnelly, 2022; Cordoni and Sancetta, 2023). Indeed, the distribution of orders over different book levels reflects the market's ability to absorb new orders (i.e., liquidity) without incurring in large price variations, since gaps in the limit-order book trigger large midpoint price changes (see Farmer et al., 2004; Toth et al., 2012, among others). Then, traders need to find the optimal proportion of wealth to be allocated over the different levels of the book so that adverse market impact feedbacks are minimized (Alfonsi et al., 2010; Predoiu et al., 2011; Chen et al., 2018).

We proceed considering four stocks included in the S&P500 with still different market capitalization and belonging to different industrial sectors: Amazon (AMZN), Abiomed (ABMD), BlackRock (BLK) and Monster Beverage Corporation (MNST). For each stock we collect tick-by-tick data on the sample period from 3-September-2019 to 31-August-2020, from 9:30am until 4:30pm on every trading day. The sample period was specifically chosen to include the market crash caused by the COVID-19 pandemic in the analysis. Data are collected from LOBSTER database[9] which is a Level 3 dataset, meaning that it contains all limit orders for the first 10 levels of the order-book, all in sequential order. On each day we consider three levels, which are the first level (top Bid and top Ask), the 5th level (intermediate level) and the 10th level (deepest Bid and Ask level at the bottom of the book). We thus have six variables at each timestamp containing the bid and ask quotes observed at these three levels. Anytime a limit-order is executed the book gets modified, we thus update the time-counter and move to the next order-book observations. On each day we estimate a VECM model and we compute both the IC-IS and the classical IS based on all Choleski permutations.

Analyzing the IS directly on individual ask and bid levels is challenging because the contribution of each level can vary significantly depending on the day. For example, during days characterized by selling activities, the ask side may have a more pronounced impact, leading to considerable oscillation in the analysis due to its inherent nature[10]. Since we are

[9]Huang and Polak (2011) *https://lobsterdata.com/*.

[10]We postpone the analysis at the individual level to future research, which could introduce additional complexity due to daily variability and the inherently oscillatory nature of daily IS at the individual bid ask level.

interested in understanding how different levels contribute to the price discovery process, once we obtain the information measures for both bids and asks, we aggregate the measures at their respective levels in the same spirit as in the previous empirical application in section 6.1. This involves summing the IS values of ask and bid quotes placed at the same level of the order book.

6.2.1 Empirical Results

In this section we show the results concerning the informativeness of the different levels of the order-book. Since we work with intraday book data, we have the possibility to exploit thousands of observations in each trading day to get daily estimations of the IC-IS over the period 3-September-2019 to 31-August-2020. We thus provide a summary of the results. In Table 3, we show the average IC-IS measure obtained over the entire sample for each stock at different levels of the order-book. In the same table, the IC-IS is compared with the min-max intervals and their midpoints associated to the classical IS measure. The results show that the IC-IS is capable of disentangling the informativeness of the different levels. As expected and in line with the literature, the best Bid and Ask placed at the first level of the book are the most informative ones (see for instance Cao et al., 2009, among others), followed by the 5th and 10th levels in order in the ranking. Such rankings of importance in terms of informativeness are remarkably consistent across all the four stocks analyzed. Even if the absolute magnitudes can differ between stocks (as one would reasonably expect given that we are considering different companies), the ranking of the order-book levels is consistent across them. On the contrary, the IS based on min-max ranges and midpoints is not capable of disentangling any of the three order-book levels, thus making it impossible to rank the best Bid and Ask, 5th and 10th levels in terms of contribution to the price formation process. Moreover, as already acknowledged in the literature, the midpoints of such wide ranges do not add-up to one, inconsistently signalling less than 50% around of informativeness for all levels of the order-book for each stock.

Moreover, given that the informational content of the three levels of the order-book can in principle vary over time, we graphically show in Figure 1 the evolution of the IC-IS every two-weeks rather than its average value over the entire sample period as in Table 3. Such dynamics show some interesting but consistent patterns. The best Bids and Asks placed at the first level retain most of the information at almost every point in time and for each stock, clearly distinguishing themselves from the 5th and 10th levels which are closer to each other. Interestingly, while the 5th level is on average more informative than the 10th one, very heterogeneous patterns emerge both across companies and over time, with the 10th level occasionally jumping above the 5th level and, in some extreme cases, almost contributing as the 1st level to the price formation process. We might suspect such jumps in the IC-IS at the deepest levels of the book to be associated with high volatility in the market, since all stocks exhibit spikes in the measure during the Covid-19 turmoil and subsequent recovery.

The analysis of the causal forces driving changes in the IC-IS is beyond the scope of this paper, and would require a much wider sample of assets and control variables. However, it is important to notice how the measure is remarkably capable of both disentangling the

Table 3: Average daily information shares measures for each price discovery analysis, for Amazon, Abiomed, BlackRock and Monster Beverages Corp., comparing the IC-IS with the classical IS based on all permutations. For the standard IS, we report average min-max range and the related mid points.

	IC-IS	All permutations		
		min	mid	max
AMZN				
Book Level 1	0.6197	0.0059	0.2947	0.5835
Book Level 5	0.2262	0.0022	0.3490	0.6958
Book Level 10	0.1527	0.0021	0.3296	0.6572
ABMD				
Book Level 1	0.5616	0.0008	0.2893	0.5779
Book Level 5	0.2540	0.0005	0.3774	0.7542
Book Level 10	0.1600	0.0019	0.4033	0.8049
BLK				
Book Level 1	0.5962	0.0001	0.3339	0.6678
Book Level 5	0.2095	0.0001	0.4238	0.8475
Book Level 10	0.1880	0.0001	0.4656	0.9290
MNST				
Book Level 1	0.7322	0.0004	0.4030	0.8057
Book Level 5	0.1604	0.0003	0.4354	0.8706
Book Level 10	0.1090	0.0020	0.4033	0.8048

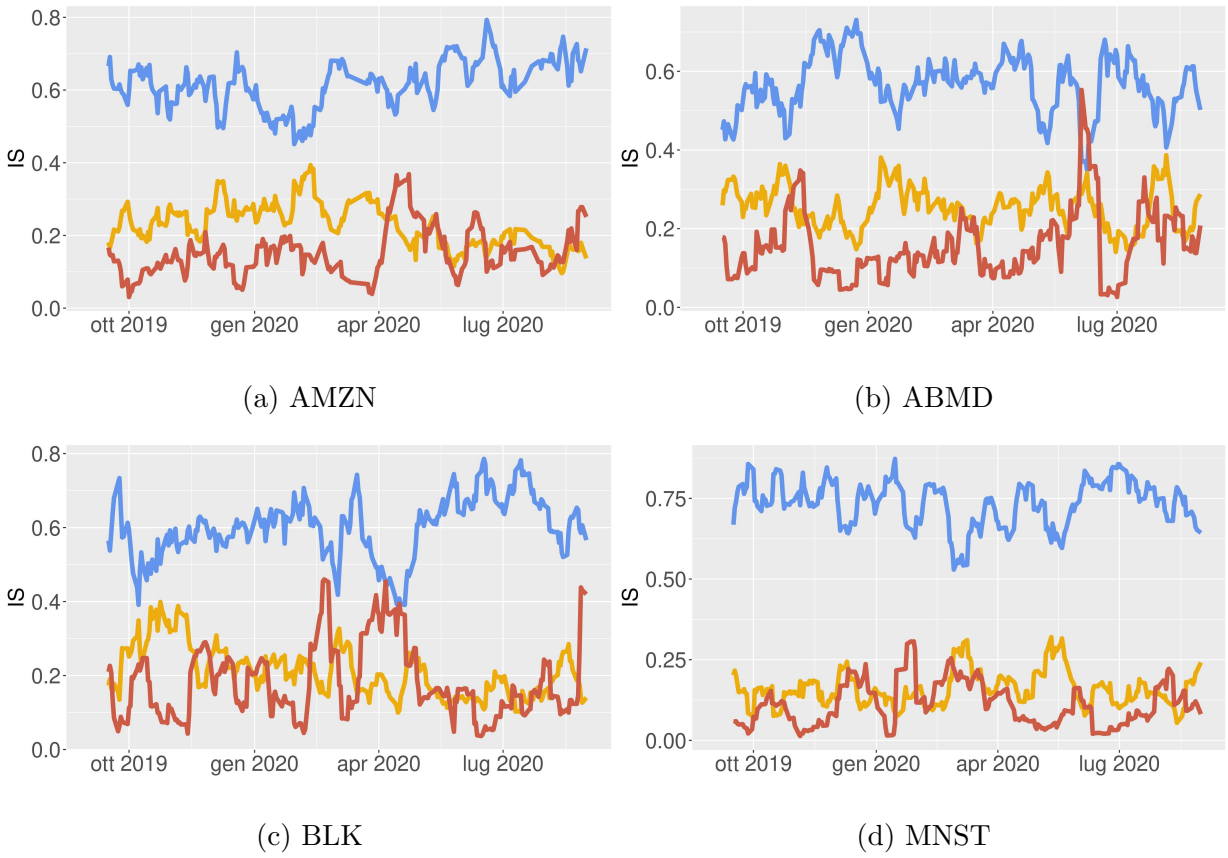
information content of the different levels of the order book and sensibly detecting changes, thus providing a new promising tool to understand how information propagates across the limit orders. This would not be possible with the classical IS measure based on min-max intervals and their midpoints, where the time-varying analysis, in line with the results in Table 3, turns out to be completely inconclusive as shown in Figure 3[11].

We can now move to the conclusions.

7 Conclusion

Measuring the contributions to the price formation process by estimating unique information share measures represents a long-standing issue in empirical finance and market

[11]For the sake of space and brevity, we show the min-max ranges and midpoints over time for Amazon only but the same holds for the other 3 stocks as well. Results are available upon request.



(a) AMZN (b) ABMD

(c) BLK (d) MNST

Figure 2: Biweekly IC-IS measures for Amazon (AMZN), Abiomed (ABMD), BlackRock (BLK) and Monster Beverage Corp. (MNST) at book level 1 (blue lines), level 5 (orange lines) and 10 (red lines) from 03 September 2019 to 31 August 2020 based on daily IC-IS measure.

microstructure modeling. While several attempts have been made in the literature, a general and well-established procedure to solve the issue from a data-driven perspective is not available yet. To this end, a new measure of price discovery, namely the IC-IS, has been introduced. The measure, does not suffer from the identification issues inherited by the historical IS measure.

Differently from both the historical IS and the DAG-IS defined by Zema (2022), for which a recursive structure is imposed in the system through the adoption of Choleski decompositions, the IC-IS provides a less restrictive framework in which no triangular structure assumptions are needed to resolve the identification issues. Moreover, this new measure neither requires the adoption of different volatility regimes as in Grammig and Peter (2013) nor requires to model in natural-time at incredibly short time-scales as done by Hasbrouck (2021). For these reasons, the IC-IS could bring new insights about the way in which the contributions to price discovery, through the variance of the efficient price process as historically proposed by Hasbrouck (1995), are both identified and quantified.

The empirical application on IBM data, performed keeping the results of Hasbrouck

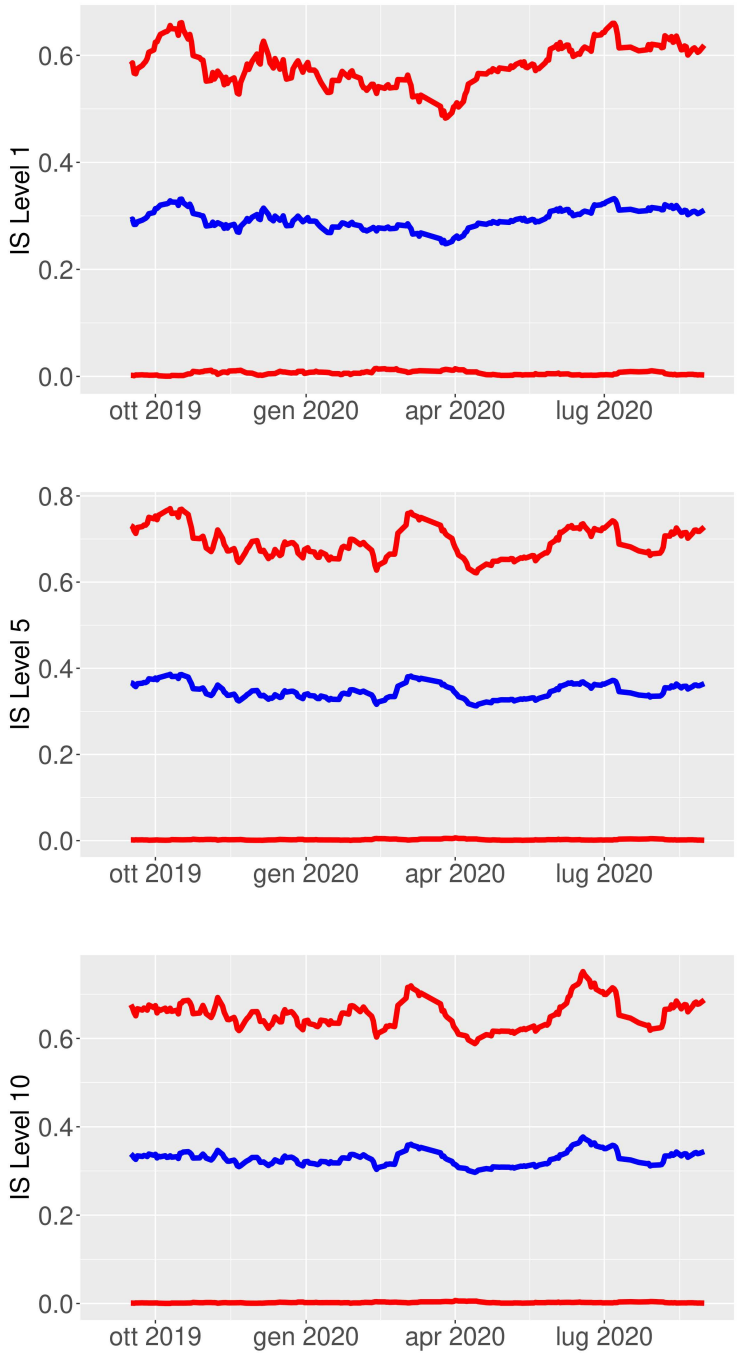


Figure 3: Biweekly IS measures for AMZN book level 1, level 5 and level 10 between 03Sep2019 to 31Aug2020 based on the classical IS measure with all permutations. The red lines exhibit the biweekly max and min IS interval, respectively, while the blue line illustrates the biweekly midpoint. The biweekly IC-IS measures are computed using a 10 days moving average.

(2021) as a sound benchmark in the literature, provided consistent and reasonable results. Moreover, an additional empirical application on the limit order-book of four different stocks revealed the IC-IS to be capable, differently from the classical IS measure, of disentangling the informational content of the three different order-book levels analyzed. Such results raise the possibility of future applications in the field benefiting from the new measure.

Finally, even if the IC-IS can be adopted as a standalone measure, it is worth stressing that the greatest benefits might come with an adoption which is complementary to other established measures, especially when no sound prescription in favour of a specific approach is available.

Appendix A. Proof of proposition 1

Proof. Let $\hat{B}P^{(r)}$ be the estimated instantaneous coefficient matrix with randomly permuted columns (i.e., identified up to column permutation). Let then P be the permutation matrix such that $\hat{B}P = \hat{B}^{(p)}$ satisfies the identification condition $|b_{ii}| \geq |b_{ij}| \ \forall \ \ i \neq j$ stated in Proposition 1. Then, there always exists a permutation matrix $P^* = P^{(r)'}P$ such that $\hat{B}P^{(r)}P^* = \hat{B}P = \hat{B}^{(p)}$ is the estimate of the instantaneous coefficient matrix we are looking for. □

Exploiting the basic properties of permutation matrices, the proof is straightforward and simply shows that independently from the arbitrary column ordering obtained after estimating the mixing matrix B, it is always possible to find an appropriate permutation matrix, being the product of two permutation matrices a permutation matrix itself, such that the identification conditions in Proposition 4.1 are met.

Appendix B. Proof of proposition 2

The proof for the asymptotic distribution of $(\psi\hat{b}_j - \psi b_j)^2$, recalling $b_j = s_j^{'}c_j$ where c_j is the j-th column of C, is a trivial application of the delta method in the multivariate case knowing that asymptotically $vec\sqrt{T}(\hat{C}_T - C) \sim N(0, \Sigma_C)$.

Proof. Given the continuous differentiable function $f(\hat{c}_j) = \psi s_j^{'}\hat{c}_j$, consider the first order Taylor series expansion around the true value c_j (higher order terms are exactly zero being $f(\hat{c}_j)$ linear)

$$f(\hat{c}_{jT}) = f(c_j) + \nabla f(c_j)^{'}(\hat{c}_{jT} - c_j)$$

that is

$$\psi s_j^{'}\hat{c}_{jT} = \psi s_j^{'}c_j + \psi s_j^{'}(\hat{c}_{jT} - c_j),$$

then from the Slutsky's Theorem (Gut, 2005) follows that if $\sqrt{T}(\hat{c}_{jT} - c_j) \xrightarrow{d} N(0, \Sigma^{c_j})$, being Σ^{c_j} the covariance matrix of the j-th column of C, then

$$\sqrt{T}(f(\hat{c}_j) - f(c_j)) \xrightarrow{d} N(0, \nabla f(c_j)^{'}\Sigma^{c_j}\nabla f(c_j))$$

that is

$$\sqrt{T}(\psi s_j^{'}\hat{c}_j - \psi s_j^{'}c_j) \xrightarrow{d} N(0, \psi s_j \Sigma^{c_j} s_j^{'}\psi^{'}).$$

$$\sqrt{T}(\psi\hat{b}_j - \psi b_j) \xrightarrow{d} N(0, \psi s_j \Sigma^{c_j} s_j^{'}\psi^{'}).$$

It follows that $(\psi\hat{b}_j - \psi b_j)/\sqrt{\psi s_j \Sigma^{c_j} s_j^{'}\psi^{'}} \sim N(0,1)$ which implies $(\psi\hat{b}_j - \psi b_j)^2/\psi s_j \Sigma^{c_j} s_j^{'}\psi^{'} \sim \chi^2(1)$. Then, $(\psi\hat{b}_j - \psi b_j)^2 \sim \psi s_j \Sigma^{c_j} s_j^{'}\psi^{'} * \chi^2(1) \rightarrow (\psi\hat{b}_j - \psi b_j)^2 \sim \Gamma(1/2, 2\psi s_j \Sigma_{jj}^{c} s_j^{'}\psi^{'})$ being the χ^2 a special case of a Gamma with parameters $\lambda = \text{d.o.f.}/2$ and $k = 2$, where the scale parameter k absorbs the variance $\psi s_j \Sigma_{jj}^{c} s_j^{'}\psi^{'}$ and d.o.f = 1.

□

Appendix C: Simulation setting and parameters

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

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

where

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

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

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

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 cointegration vectors and has rank equal to *n-1*. We simulate the system with 1 lag only for simplicity, so the parameters are

$$\alpha = \begin{pmatrix} 0.025 & 0.05 & 0.03 \\ 0.08 & 0.07 & 0.06 \\ 0.1 & 0.01 & 0.04 \\ 0.09 & 0.06 & 0.09 \end{pmatrix}, \quad \beta = \begin{pmatrix} 1 & \\ \vdots & -\mathrm{I}_{n-1} \\ 1 & \end{pmatrix}, \Phi_1 = \begin{pmatrix} 0.4 & -0.9 & -0.25 & 0.3 \\ 0.6 & 0.35 & 0.55 & -0.1 \\ 0.2 & -0.2 & -0.7 & 0.4 \\ 0.1 & 0.35 & 0.6 & 0.1 \end{pmatrix},$$

and $\Phi_1 = -\Pi_2$, $\Pi_1 = \alpha\beta' + I - \Pi_2$. Finally, the matrix S used in $u_t = SC\epsilon_t$ is

$$S = \begin{pmatrix} 0.9 & 0 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0.5 & 0.2 & 0.7 & 0 \\ 0.3 & 0.5 & 0.3 & 0.1 \end{pmatrix}.$$

It must be mentioned that typically the diurnal U-shape pattern is quantified in the literature by setting $M \approx 0.89$. Here $M = 1$ simply to guarantee the existence of the variance of the Students from which shocks are generated, being the degrees of freedom of the Students mapped over time in a one-to-one relation with the time-varying variances. This useful shift still preserves a U-shape patterns and does not hamper in any way the IC-IS measure in the simulation exercise.

Appendix D: Relaxing the i.i.d assumption based on moment restrictions.

In this section we relax the assumption of ϵ_t being an i.i.d. process. We do so by exploiting the GMM estimation framework of Lanne and Luoto (2021) where the authors assume the shocks to be only uncorrelated rather than independent and identically distributed. Since

orthogonality does not suffice to reach full identification of the model, they do so by imposing an additional number of fourth-moment restrictions. We thus import their assumptions and discuss shortly thereafter the related moment conditions we need to impose in our price discovery framework.

Assumption 7.1.

i *The process* $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{nt})$ *is a strictly stationary random vector with each shock component* ϵ_{it} *having mean zero and unit variance.*

ii *The shocks* $\epsilon_{1t}, \ldots, \epsilon_{nt}$ *are serially uncorrelated, that is* $cov(\epsilon_{it}, \epsilon_{i,t+k}) = 0 \ \forall \ k \neq 0$.

iii *The shocks* $\epsilon_i, \ldots, \epsilon_n$ *are mutually orthogonal and at most one of them is normally distributed.*

iv $E(\epsilon_{it}^3 \epsilon_{jt}) = 0$ *for at least* $n(n-1)/2$ *combinations of* i *and* j *with* $i \neq j$.

As proven by Lanne and Luoto (2021) B can be identified up to columns' sign and permutations (i.e., local identification) under the above assumptions. This means that given the estimate of B, through appropriate moment conditions which will be discussed shortly, the IC-IS can be easily identified as in the scenario illustrated in section 4 starting from Theorem 4.1. Equivalently to that scenario, non-normality is still the key ingredient which heavily alleviates the identification problem. The only difference is we are now relaxing the independence assumption for the shocks through the imposition of some additional moment conditions. Such moment conditions, deserve now some discussion and clarifications.

Let us rewrite the VECM model for price discovery in equation 1 in a more compact form as follows

$$\Delta p_t = \Theta \boldsymbol{X}_{t-1} + B\epsilon_t \tag{11}$$

where $\boldsymbol{X}_{t-1} = (p_{t-1}, \Delta p_{t-1}, \ldots, \Delta p_{t-k})$ is a $(nk+1) \times 1$ vector and $\Theta = (\alpha\beta', \Phi_1, \ldots, \Phi_k)$, while B is the matrix we need to identify to compute the information shares. In our price discovery framework, the moment conditions to be imposed will pertain to B only. Being the cointegrating relationships derived from the theory (i.e., no arbitrage, one efficient price process only underlying the series) and imposed in the model, the parameters in Θ are easily estimated through OLS. This means we are left, as in the PML strategy, with n^2 unknown parameters over which the moment conditions are imposed as follows:

$$E(\epsilon_{it}^2) - 1 = 0, \quad i = 1, \ldots, n \tag{12}$$

$$E(\epsilon_{it}\epsilon_{jt}) = 0, \quad i \neq j \tag{13}$$

$$E(\epsilon_{it}^3\epsilon_{jt}) = 0, \quad i \neq j \tag{14}$$

$$E(\epsilon_{it}^2\epsilon_{jt}^2) - 1 = 0, \quad i \neq j. \tag{15}$$

The n moment conditions in 12 deal with the scaling indeterminacy restricting the shocks to be unit variance. The mutual orthogonality of the shocks in Assumption 7.1(iii) yield us the additional $n(n-1)/2$ moment conditions in equation 13. The equations in 12 and 13 provide us with $n(n+1)/2$ moment conditions which are not sufficient yet to identify the n^2 parameters in B. We thus need at least $n(n-1)/2$ additional moment conditions. To this purpose, we exploit non-Normality to provide additional co-kurtoses conditions which are the ones in equations 14 and 15.

If ϵ_t were Normally distributed, only the first two moments would be informative and the moment conditions in 14 and 15 would be redundant, adding no information beyond the one provided by the mutual orthogonality condition in 13. On the contrary, when at most one of the variables in the model is Normally distributed as required by Assumption 7.1(iii), the co-kurtosis conditions taken for all pairs i, j with $i \neq j$ becomes informative. Such co-kurtosis conditions are implied by independence but they do not necessarily imply the latter. Thus, the idea is to impose those moment conditions that would prevail if the components in ϵ_t were independent without imposing independence, thus allowing for various forms of conditional heteroskedasticity in the data[12].

We can then estimate B in equation (11) by minimizing

$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} g(\epsilon_t, \theta) W_T \frac{1}{T}\sum_{t=1}^{T} g(\epsilon_t, \theta) \tag{16}$$

where $\theta = (vec(B))$ is the row vector of n^2 parameters to be estimated, while W is a positive semi-definite $q \times q$ matrix of weights for the $q \times 1$ vector of sample moment conditions in $g(.)$. The weighting matrix W_T can be efficiently estimated, as shown by Hansen (1982), by setting $W_T = S_T^{-1}$ with S_T being the inverse of the asymptotic covariance matrix of the moment conditions. Such matrix S_T can be estimated consistently, under some regularity conditions (Newey and West, 1987), by the following heteroskedasticity and auto-correlation (HAC) matrix estimator

$$\hat{S}_{HAC} = \hat{\Gamma}_0 + \sum_{t=1}^{T-1} w(i,T)[\hat{\Gamma}_i + \hat{\Gamma}_i{}'] \tag{17}$$

where $\Gamma_{i,T}(\theta)$ is the *i-th* sample autocovariance matrix of $g(\epsilon_t, \theta)$ evaluated at an initial consistent estimate θ^* of the true parameter θ, that is $\hat{\Gamma}_i \equiv \sum_{t=i+1}^{T} g(\epsilon_t, \theta_T^*) g(\epsilon_{t-i}, \theta_T^*)/T$, while $w(i,T)$ is the kernel controlling for the number of autocovariances included. The estimation procedure is based then a 3-step procedure. In the first step we estimate the VECM equation by equation via OLS given the known cointegrating relationship. In step 2, we extract the estimated model's residuals and implement a further two-step GMM estimation procedure Hansen (1982) where we first minimize equation (16) using a sub-optimal W_T to get an initial estimate for B, we then update such estimate based on the consistent heteroskedastic and auto-correlation matrix estimator $\hat{S}_{HAC}$.

The $n(n-1)/2$ asymmetric co-kurtosis conditions in 14 are already sufficient to reach full-identification. Nevertheless, the additional $n(n-1)/2$ symmetric conditions in 15 give

[12]Lanne and Luoto (2021) provide details and a useful example for this point.

us the possibility to test for over-identifying restrictions. This is useful to ensure that proper moment conditions are selected given the data, which might increase the accuracy of the GMM estimator. To this end, we rely on the widely known J-test of over-identifying restrictions by Hansen (1982). Given that several over-identifying sets of moment conditions typically turns out to pass the test, we need to adopt a selection criteria to pick the optimal set among the suitable ones. We thus follow the relevant moment selection criterion (RMSC) of Hall et al. (2007) to pick the optimal set of moment conditions and estimate the n^2 parameters in B.

As shown in Hall (2004) the GMM estimator is, under typical standard assumptions of strict stationarity and ergodicity, consistent and asymptotically normal. That is, as well as in the PML framework, the IC-IS asymptotic properties can be derived in a very analogous way as we did for the PML case. The only difference being the variance associated with the two difference estimates. As a result, Proposition 4.1 and remarks 1 and 2 hold true.

It should be evidenced, as also done by Lanne and Luoto (2021), that the GMM estimation procedure via the selection of the optimal moment restrictions, despite relaxing some theoretical assumptions, has the side-effect of being computationally burdensome, slowing down very significantly the estimation process. In our framework for instance, we need at least $n(n-1)/2$ co-kurtosis conditions to be selected, without repetition, out of $n(n-1)$ possible ones in total (i.e., asymmetric plus symmetric conditions). Considering all the possible over-identifying restrictions, yield us $\sum_{k=(n-1)/2+1}^{n(n-1)} \binom{n(n-1)}{k}$ combinations to be tested.

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.

Alfonsi, A., A. Fruth, and A. Schied (2010). Optimal execution strategies in limit order books with general shape functions. *Quantitative finance 10*(2), 143–157.

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

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.

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

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 (2021). Comment on: Price discovery in high resolution. *Journal of Financial Econometrics 19*(3), 431–438.

Buccheri, G., G. Bormetti, F. Corsi, and F. Lillo (2021). Comment on: Price discovery in high resolution. *Journal of Financial Econometrics 19*(3), 439–451.

Cao, C., O. Hansch, and X. Wang (2009). The information content of an open limit-order book. *Journal of Futures Markets: Futures, Options, and Other Derivative Products 29*(1), 16–41.

Chen, Y., X. Gao, and D. Li (2018). Optimal order execution using hidden orders. *Journal of Economic Dynamics and Control 94*, 89–116.

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.

Cordoni, F. and A. Sancetta (2023). Consistent causal inference for high-dimensional time series. *arXiv preprint arXiv:2307.03074*.

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

de Jong, F. (2021). Comment on: Price Discovery in High Resolution*. *Journal of Financial Econometrics 19*(3), 452–458.

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 (2021). Price discovery in a continuous-time setting. *Journal of Financial Econometrics 19*(5), 985–1008.

Donnelly, R. (2022). Optimal execution: A review. *Applied Mathematical Finance 29*(3), 181–212.

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.

Eriksson, J. and V. Koivunen (2004). Identifiability, separability, and uniqueness of linear ica models. *IEEE signal processing letters 11*(7), 601–604.

Farmer, J. D., L. Gillemot, F. Lillo, S. Mike, and A. Sen (2004). What really causes large price changes? *Quantitative finance 4*(4), 383–397.

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

Ghysels, E. (2021). Comment on: Price discovery in high resolution and the analysis of mixed frequency data. *Journal of Financial Econometrics 19*(3), 459–464.

Gourieroux, C. and A. Monfort (1995). *Statistics and econometric models*, Volume 1. Cambridge University Press.

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.

Gouriéroux, C., A. Monfort, and J.-P. Renne (2020). Identification and estimation in non-fundamental structural varma models. *The Review of Economic Studies 87*(4), 1915–1953.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica: journal of the Econometric Society*, 681–700.

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.

Guidolin, M., M. Pedio, and A. Tosi (2021). Time-varying price discovery in sovereign credit markets. *Finance Research Letters 38*, 101388.

Gut, A. (2005). *Probability: a graduate course*, Volume 200. Springer.

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

Hall, A. (2004). *Generalized method of moments*. Wiley Online Library.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007). Information in generalized method of moments estimation and entropy-based moment selection. *Journal of Econometrics 138*(2), 488–512.

Hallin, M. and C. Mehta (2015). R-estimation for asymmetric independent component analysis. *Journal of the American Statistical Association 110*(509), 218–232.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, 1029–1054.

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, L. and J. Hasbrouck (1996). Market vs. limit orders: the superdot evidence on order submission strategy. *Journal of Financial and Quantitative analysis 31*(2), 213–231.

Harris, L. E. and V. Panchapagesan (2005). The information content of the limit order book: evidence from nyse specialist trading decisions. *Journal of Financial Markets 8*(1), 25–67.

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

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

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

Hasbrouck, J. (2021). Price discovery in high resolution. *Journal of Financial Econometrics 19*(3), 395–430.

Herwartz, H., A. Lange, and S. Maxand (2022). Data-driven identification in svars—when and how can statistical characteristics be used to unravel causal relationships? *Economic Inquiry 60*(2), 668–693.

Huang, R. and T. Polak (2011). Lobster: Limit order book reconstruction system. *Available at SSRN 1977207*.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks 10*(3), 626–634.

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).

Ilmonen, P., D. Paindaveine, et al. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *the Annals of Statistics 39*(5), 2448–2476.

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

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous univariate distributions, volume 2*, Volume 289. John wiley & sons.

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

Lanne, M. and J. Luoto (2021). Gmm estimation of non-gaussian structural vector autoregression. *Journal of Business & Economic Statistics 39*(1), 69–81.

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.

Leung, T., M. Lorig, and A. Pascucci (2017). Leveraged etf implied volatilities from etf dynamics. *Mathematical Finance 27*(4), 1035–1068.

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.

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.

Moneta, A. and G. Pallante (2022). Identification of structural var models via independent component analysis: a performance evaluation study. *Journal of Economic Dynamics and Control 144*, 104530.

Newey, W. K. and K. D. West (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777–787.

Predoiu, S., G. Shaikhet, and S. Shreve (2011). Optimal execution in a general one-sided limit-order book. *SIAM Journal on Financial Mathematics 2*(1), 183–212.

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

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

Shum, P., W. Hejazi, E. Haryanto, and A. Rodier (2016). Intraday share price volatility and leveraged etf rebalancing. *Review of Finance 20*(6), 2379–2409.

Toth, B., Z. Eisler, F. Lillo, J. Kockelkoren, J.-P. Bouchaud, and J. D. Farmer (2012). How does the market react to your order flow? *Quantitative Finance 12*(7), 1015–1024.

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

Zema, S. M. (2022). Directed acyclic graph based information shares for price discovery. *Journal of Economic Dynamics and Control 139*, 104434.