Quaderni FinTech

Dimensionality reduction techniques to support insider trading detection

Consob - Scuola Normale Superiore di Pisa

12

febbraio 2024

1974 - 2024
CONSOB
COMMISSIONE NAZIONALE PER LE SOCIETÀ E LA BORSA

Nella collana dei Quaderni FinTech
sono raccolti lavori di ricerca relativi
al fenomeno «FinTech» nei suoi molteplici aspetti
al fine di promuovere la riflessione e
stimolare il dibattito su temi attinenti
all'economia e alla regolamentazione
del sistema finanziario.

CONSOB

00198 Roma - Via G.B. Martini, 3
t +39.06.84771 centralino
f +39.06.8477612

20121 Milano - Via Broletto, 7
t +39.02.724201 centralino
f +39.02.89010696

h www.consob.it
e studi_analisi@consob.it

Tecniche per la riduzione dimensionale dei dati a supporto del rilevamento dei casi di insider trading

Consob - Scuola Normale Superiore di Pisa

Sintesi del lavoro

L'identificazione degli abusi di mercato è un'attività estremamente complessa che richiede l'analisi di insiemi di dati grandi e complessi. Lo studio propone un metodo di apprendimento automatico non supervisionato per il rilevamento di anomalie contestuali, che fornisce un supporto alla vigilanza sui mercati finalizzata all'identificazione di potenziali attività di insider trading. Nello specifico, lo studio - basato su un data set anonimizzato - affronta il problema di identificazione di potenziali casi di insider trading e propone un metodo diverso rispetto ai precedenti studi che hanno fatto uso di tecniche di *unsupervised machine learning*: in questo caso, infatti, viene applicata la tecnica di decomposizione e successiva ricostruzione di una serie temporale di dati attraverso l'analisi delle "componenti principali" (PCA, *Principal Component Analysis*) e l'uso di *autoencoders*, in relazione alle posizioni assunte da gruppi di investitori in un determinato titolo azionario in prossimità di un evento price sensitive. L'unico input del metodo è la posizione di trading di ciascun investitore attivo sull'asset per il quale si è verificato un evento price sensitive (PSE). Dopo aver determinato gli errori di ricostruzione relativi ai profili di profili di trading, vengono imposte diverse condizioni al fine di identificare gli investitori il cui comportamento potrebbe essere sospetto di insider trading in relazione al PSE. In termini intuitivi, la logica che viene seguita nella procedura di identificazione di comportamenti anomali da parte degli investitori considera la posizione media ricostruita attraverso la tecnica PCA come rappresentativa di un'operatività normale. Qualsiasi scostamento nell'operatività di un singolo investitore dal comportamento medio ricostruito nel periodo di osservazione (che sia superiore ad una certa soglia di sensitività) viene segnalato dall'algoritmo come anomalo e potenzialmente meritevole di approfondimenti ulteriori attraverso tecniche di indagine "tradizionali". Un risultato particolarmente significativo di questo studio è la soddisfacente convergenza dei risultati ottenuti con quelli derivati dall'applicazione delle tecniche di *unsupervised machine learning* descritte nel precedente *paper* "*A machine learning approach to support decision in insider trading detection*", anch'esso frutto della collaborazione tra l'Istituto e la Scuola Normale Superiore di Pisa.

Il presente lavoro, frutto della collaborazione tra la Consob e la Scuola Normale Superiore di Pisa, è stato curato da:

- Adele **Ravagnani** - *SNS Pisa*
- Fabrizio **Lillo** - *SNS Pisa, Dipartimento di Matematica, Università di Bologna*
- Paola **Deriu** – *Consob*
- Piero **Mazzarisi** - *SNS Pisa, Dipartimento di Economia Politica e Statistica, Università di Siena*
- Francesca **Medda** - *Consob*
- Antonio **Russo** - *Consob*

Dimensionality reduction techniques to support insider trading detection[1]

Adele Ravagnani[2], Fabrizio Lillo[3], Paola Deriu[4],
Piero Mazzarisi[5], Francesca Medda[6], Antonio Russo[7]

January 8, 2024

Abstract

Identification of market abuse is an extremely complicated activity that requires the analysis of large and complex datasets. We propose an unsupervised machine learning method for contextual anomaly detection, which allows to support market surveillance aimed at identifying potential insider trading activities. This method lies in the reconstruction-based paradigm and employs principal component analysis and autoencoders as dimensionality reduction techniques. The only input of this method is the trading position of each investor active on the asset for which we have a price sensitive event (PSE). After determining reconstruction errors related to the trading profiles, several conditions are imposed in order to identify investors whose behavior could be suspicious of insider trading related to the PSE. As a case study, we apply our method to investor resolved data of Italian stocks around takeover bids.

Keywords: Dimensionality reduction, Principal component analysis, Autoencoder, Insider trading, Market abuse, Unsupervised learning.

[1]This paper represents the personal opinions of the authors and does not bind the membership organization in any way.

[2]adele.ravagnani@sns.it, Scuola Normale Superiore, Pisa, Italy.

[3]fabrizio.lillo@sns.it, Scuola Normale Superiore, Pisa and Dipartimento di Matematica, Università di Bologna, Italy.

[4]p.deriu@consob.it, Commissione Nazionale per le Società e la Borsa, Italy.

[5]piero.mazzarisi@unisi.it, Dipartimento di Economia Politica e Statistica, Università di Siena.

[6]f.medda@ucl.ac.uk, University College London, London, UK and Commissione Nazionale per le Società e la Borsa, Italy.

[7]a.russo@consob.it, Commissione Nazionale per le Società e la Borsa, Italy.

Contents

1 Introduction

Insider trading is the unlawful practice of trading by exploiting nonpublic confidential information about a listed company. It is a type of market abuse: it prevents full and effective market integrity, it violates natural demand-supply dynamics, it compromises public confidence. Knowing in advance how the price will likely move in response to the release of confidential information to the market, i.e. price sensitive event (PSE), such as, for example, the announcement of a takeover bid, can be easily exploited to make a profit. Such a type of practice is prohibited or criminalized in most jurisdictions around the world [4]. However, rules are specific of each country and efforts in persecuting insider trading vary considerably. In the European Union, it is expressly prohibited and administratively sanctioned. The member states are left with the possibility of also imposing criminal sanctions.

The "proof" and the subsequent imposition of a sanction (either administrative or criminal) to an investor that has operated as an *insider* is however a complex process, involving many steps: (i) the detection of alerts pointing to anomalies that appear attributable to abusive behaviors, (ii) the concrete assessment of the allegedly suspicious conduct with respect to possible rationale that may have supported the strategy under analysis, (iii) the investigation phase aimed at gathering evidence and clues of the abusive conduct, and (iv) the subsequent legal trial to confirm the fact that the unlawful conduct was committed.

In [19], we focus on the first two steps. We propose a methodology, based on unsupervised machine learning techniques, that is capable of providing an indication on whether the trading behavior of an investor or a group of investors is anomalous or not, thus supporting the monitoring and surveillance processes by the competent Authority and the assessment of the conduct. Our previous approach combines two methods, which are both based on well-known techniques, the k-means clustering algorithm [12] and the statistically validated co-occurrence networks [24]. With this new work, we want to provide an extension of the first method of [19]. The latter aims at identifying investors with suspicious behavior related to a price-sensitive event, by means of a dynamic clustering approach. This is done by focusing both on the discontinuities in the trading activity of single investors with respect to their normal activity and to the behavior of their peers, i.e., investors with similar activity. The first step of this method is the characterization of the trading activity of each investor in several time windows. Focusing on a time window, each investor is associated with a point in a three-dimensional space, which corresponds to three trading features (*signed turnover*, *magnitudo*, *maximum exposure*) that are relevant to our insider trading detection task and summarize the activity of the investor in that time window. The choice of three features was motivated by explanability reasons, since we can have a graphical representation and more easily interpret our findings. In this new work, we aim to overcome this choice of the three trading features by employing a dimensionality reduction approach. The idea is that the model should identify the features more relevant to investors' characterization by itself.

1.1 Literature review

Our work suits into the framework of anomaly detection. This field has been widely explored in the literature, especially in the last years, when its developments have been

going at the same pace as machine learning. The applications in the field of financial fraud detection are numerous [25] and, among them, some works are related to the detection of market abuse such as [21, 8, 17, 22]. However, insider trading detection in stock markets is a fairly unexplored topic.

Anomaly detection basically consists in identifying data instances that cannot be associated with normal behavior and that are rare in the data set. The goal of anomaly detection is to define a region of the features' space where normal observations lie; observations that do not lie in this region are defined as anomalies [5]. Identifying this normality region is not straightforward: the boundary between normal and anomalous behavior is not always sharp, behaviors that are actually anomalous could be disguised in order not to be identified, the definition of normal behavior could be time varying and it is strongly dependent on the application domain, it is difficult to distinguish noise from anomalous behavior [5]. From a practical point of view, there are four main aspects which determine the formulation of the anomaly detection method: availability of data labels, the desired output of the technique, the nature of the input data, the type of anomaly. A different type of anomaly detection approach is employed depending on the availability of data labels: supervised when each observation is labeled as normal or anomalous, semi-supervised when training data do not contain any anomalies and unsupervised when no labels are provided as in our interest case. Typically, the output of the anomaly detection algorithms associates with each observation a score, which quantifies the magnitude of its anomalous character. Setting a suited threshold, the ranked list of anomalies can provide labels for each data instance. Data instances can be of various type (binary/categorical/continuous, univariate/multivariate) and independent among them or related to each other, as it is the case of time series and sequences, spatial data and graph data, for which ad hoc methodologies have to be employed [2, 1]. Concerning the type of anomalies, the standard case is represented by *point anomalies*, which are single elements identified as anomalous; they could be *global* or *local* depending on whether the entire feature space or a specific region of it is considered [10]. Interestingly, there are cases where an element can be seen as normal, but when a given context is taken into account, it turns out to be an anomaly. We refer to this type as *contextual anomalies* [5], also termed *conditional anomalies* [23]. It may happen, for instance, that an investor has operated on a stock and, without a context, such an operation looks similar to other operations in the market. However, when compared to the own past behavior of the investor or to the operations of other investors, some discontinuity or synchronization patterns may be revealed and the operation could turn out to be identified as anomalous. Contextual anomalies problems can be tackled by algorithms for point anomaly detection once the context is included as a new feature. Finally, we could have data instances that are normal if considered individually, while they are anomalous together: they are the so-called *collective anomalies* [5] and can occur in a data set where data instances are dependent.

It is evident that anomaly detection problems are challenging, especially in the unsupervised setting. A variety of different approaches have been developed to address them. In particular, the main paradigms in time-series are: *clustering-based, distance-based, reconstruction-based* and *forecasting-based* [20]. Among them, the methods are multiple and their formulations are case-by-case dependent.

In this work, we develop an approach which aims at identifying *contextual anomalies* and lies in the *reconstruction-based* paradigm [20]. This framework aims at training

models that reconstruct normal data instances well. In this way, we expect that anomalous data will be reconstructed with a large error. An anomaly is detected when the reconstruction error is greater than a threshold i.e.

$$||X - \hat{X}|| > \delta$$

where X is an original data sample, $\hat{X}$ is its reconstructed counterpart and δ a suitable threshold. Models' performances are usually compared in terms of the most common metrics e.g. precision, recall and F1-scores [20].

Our case is even more complicated since we are not provided with labels which allow to compute the metrics to assess the models' performances. Therefore, in order to compare the results of different reconstruction models, we have to rely on qualitative inspections.

The standard model employed in *reconstruction-based* approaches is Principal Component Analysis (PCA) [14]. Its goal is to obtain a compressed representation of data, retaining the most important features. Data are mapped to a lower dimensional space by orthogonal transformations that aim at maximizing data variance or equivalently, minimizing reconstruction error.

The nonlinear counterpart of PCA is an autoencoder [11]. As PCA, autoencoders's goal is to minimize reconstruction errors but, in this case, the compression and decompression steps are made by means of neural network layers. Complex autoencoder architectures can be devised, as deep, convolutional, LSTM, variational autoencoders [11]. Also, generative adversarial networks have been used in *reconstruction-based* methods [3]. Moreover, combined approaches as in [6] can be employed.

If an autoencoder is provided with one hidden layer and linear activation functions, the analogy with PCA is evident and in the literature, it has been investigated in several works. In particular, in [16], the authors characterize the loss landscapes of linear autoencoders (LAEs), prove that LAEs with L_2 regularization learn the PCA's principal directions and provide an algorithm to recover them from LAEs' results.

Contributions of the paper and outline[1] The main contributions of the paper can be summarized as follows:

- We devise a method to support decision in insider trading detection which is not based on the definition of specific trading features;
- Our method is an unsupervised approach for contextual anomaly detection, without any labels to check results and compare performances;
- We apply principal component analysis and autoencoders for the reconstruction of trading profiles.

The paper is organized as follows. In Section 2, the proposed method is described. Section 3 presents the data set we use in our empirical analysis and Section 4 presents the results obtained by our method, with a special focus on one PSE. Finally, conclusions are drawn in Section 5. In the appendix, some figures, which are explained in the main text, are reported, and other collateral issues are investigated.

[1]The methodology presented in the paper was conceived in 2023 for the purpose of developing a proof of concept. It is, in no way, a tool used in the analysis and investigations carried out by Consob. The methodology may possibly constitute the future one of the tools to help and support the preliminary analysis and detection activities more efficiently. Any subsequent enforcement activity will, in any case, be based on the broader set of information that is gathered in the course of investigations and other possible types of analysis.

2 Method

2.1 Overview

As in [19], we are tackling an unsupervised problem without any availability of labels and we consider a specific class of price sensitive events (PSEs), namely announcements of takeover bids. A takeover bid is a public offer made by a physical person or a legal entity who is willing to buy other shareholders' shares at a price higher than the stock market value. If investors know in advance when the announcement of the takeover bid will occur, they can exploit their information by buying before the PSE. Indeed, when the takeover bid occurs, the shares' price goes up aligning with the offer price and thus, the informed investors can sell by making a no-risk profit.

We focus on a single asset, the one for which we have a PSE, and on a time window with T trading days. The first part of this time window - e.g. 6 months - is a reference period and the second part - e.g. 1 month - is an investigation period i.e. a short time window preceding the PSE that will be defined as Δ in the following. As a first step, we compute the trading position of each investor on each day. Given N_0 investors and T trading days $\{t_0, t_1, \ldots, t_T\}$, the position of investor i on day t is defined as follows:

$$x_i(t) = \sum_{t_0 \leq t' \leq t} [V_b(i, t') - V_s(i, t')] \tag{1}$$

where $V_{b/s}(i, t)$ is the number of shares bought/sold by investor i on day t. Therefore, a vector $x_i = \big[x_i(t_0), \ldots, x_i(t_T)\big] \in \mathbb{R}^T$ is assigned to each investor. As usually done, we normalize data as

$$x_i(t) \to \frac{x_i(t)}{\max_t |x_i|} = \frac{x_i(t)}{||x_i||_\infty} \tag{2}$$

and investors with constant positions i.e. investors who do not trade or are strict daily investors (i.e. the number of shares purchased and sold on each day are equal), are discarded.

In the definition of Equation 1, we assume that investors' positions are null on t_0. Of course, this is not true in general. However, since information on the precise composition of the portfolio of each investor is not available, this sounds as the best proxy of asset positions. In the following, we will see that actually this is not an issue for this new method, contrary to what happens for our previous method [19].

We also observe that positions are computed using the number of shares and not Euro. The reason is that the monetary value of a portfolio fluctuates in response to the changing price and these fluctuations affect in the same direction positions with the same sign (e.g. long or short). Thus, spurious correlations between positions might be detected when using Euro as a unit of measurement.

Indicating with N the number of investors with non-constant positions, we end up with a data set $X \in \mathbb{R}^{N,T}$ with rows x_i. This data set is the input of a dimensionality reduction method (we are going to use PCA and autoencoders), which will allow to obtain a reconstructed representation of the data after a compression i.e.

$$X \to Z = f_1(X) \to \hat{X} = f_2(Z) \text{ where } X, \hat{X} \in \mathbb{R}^{N,T},\ Z \in \mathbb{R}^{N,K}$$

and such that the reconstruction error is minimized with respect to the transformations f_1 and f_2 i.e.

$$\hat{X} = \arg\min_{\hat{X}'} ||X - \hat{X}'||_F^2$$

where $||\cdot||_F$ is the Frobenius norm.

In the compression phase, observations are mapped to a lower dimensional space that captures common and essential characteristics. In our setting, the features which are subjected to compression are the positions of investors' on each day. This consists in identifying a subset of days or combinations of them with a major role in the characterization of agents' trading activity.

After the dimensionality reduction step, investors with anomalous activity (potential insiders) are identified following the *reconstruction-based* paradigm and assuming they are substantially less numerous than investors with normal behavior. The preparatory steps of the method we develop are the following:

- compute the reconstruction errors

$$\epsilon_i(t) = |x_i(t) - \hat{x_i}(t)| \quad i = 1, \ldots, N, \ t = 1, \ldots, T,$$

  we expect that normal observations have low ϵ_i, while anomalies have large ϵ_i;

- compute the anomaly scores i.e. the largest reconstruction errors for each investor:

$$s_i^* = \max_t \epsilon_i(t) \quad i = 1, \ldots, N;$$

- localize the largest reconstruction errors:

$$t_i^* = \arg\max_t \epsilon_i(t), \quad i = 1, \ldots, N$$

  which are such that

$$\epsilon_i(t_i^*) = s_i^*, \quad i = 1, \ldots, N;$$

- compute $n_t = card\{i : t_i^* = t\} \ \forall t = 1, \ldots, T$, where n_t is the number of investors having the largest reconstruction error on day t;

- compute d_i for $i = 1, \ldots, N$ i.e. the number of activity days of each investor.

Finally, in order to detect potential insiders, we devise a method which is based on the following idea. In order to be anomalous, an investor should satisfy the following conditions: (1) to have at least one day for which the reconstruction error is large, (2) the corresponding time lies in the investigation period, (3) she has either a small number of activity days or her identified anomalous score is on a day when not too many other investors do, and (4) she is in a net buying position on the day of the PSE. This last condition can be set by imposing that the difference between the position on the PSE and the position on the first day of the reference period is larger than a threshold, that we choose equal to 0.5.

Formalizing the above conditions, we say that an investor i is anomalous if

$$\left\{ \begin{array}{l} \Big(\epsilon_i(t), t, n_t\Big) : \epsilon_i(t) \geq \epsilon_\theta, \ t \in \Delta, \\ \quad n_t < n_\theta \text{ if } d_i > d_\theta \text{ or } \forall n_t \text{ if } d_i \leq d_\theta, \\ \quad i \text{ has a net buying position on the PSE} \end{array} \right\} \neq \emptyset. \tag{3}$$

This criterion depends on three threshold parameters d_θ, ϵ_θ, and n_θ. The parameter d_θ, the minimal number of days in item (3) of the criterion above, is set to 3. Instead, we choose ϵ_θ and n_θ in a data driven fashion: we estimate the probability density function of the anomaly scores s_i^* and of the times of the largest reconstruction errors t_i^*. Since we observe that the former distribution is bimodal (see the left panel of Figure 3), we expect that normal investor profiles are associated to small anomaly scores s_i^*, while anomalous ones to high scores. Thus ϵ_θ is chosen as the local minimum between the two modes of the distribution. In practice, we relied on the module *signal* of the Python library *scipy*. Finally, n_θ is chosen as the top decile of its distribution.

Supervising authorities are often interested in a ranking of potential insiders to identify the most suspicious investors. Our approach is able to deliver such a ranking. Investors are mapped in a two-dimensional space $(s_i^*, \bar{n}_{t_i^*})$, where

$$\bar{n}_{t_i^*} = n_{t_i^*} \mathbb{I}[d_i > d_\theta]$$

and then the two features are normalized to take values in $[0, 1]$. The Euclidean distance between each investor and the point $(1, 0)$ is the metric for our ranking. The smaller the distance the higher the ranking.

Finally, it is important to point out the extreme unsupervised nature of our problem. We are not provided with labels associated to each investor so, we train models for dimensionality reduction by using all data and we cannot compare models' performances in terms of accuracy. This and our previous work [19] tackle the same issue and we could be tempted to employ [19]'s results as ground truth. However, with this dimensionality approach, we aim to provide a new method that could give another tool to support regulators' investigations related to insider trading detection. Therefore, the results of [19] are not validation data: in this new work, they are employed for comparisons and robustness checks.

2.2 Dimensionality reduction methods

Principal Component Analysis

A standard method to apply dimensionality reduction is Principal Component Analysis (PCA) [14]. Starting from a feature scaled data set X, the goal of PCA is to obtain a compressed representation Z_K of data and then, a reconstructed version $\hat{X}$ by means of orthogonal transformations:

$$X \to Z_K = XP_K \to \hat{X} = Z_K P_K^T \text{ where } X, \hat{X} \in \mathbb{R}^{N,T},\ Z_K \in \mathbb{R}^{N,K},\ P_K \in \mathbb{R}^{T,K},$$

and the transformation matrix P_K is such that the reconstruction error is minimized with a rank constraint, i.e.

$$\hat{X} = \arg \min_{\hat{X}':\ rank(\hat{X}') \leq K} ||X - \hat{X}'||_F^2.$$

The solution is the truncated Singular Value Decomposition (SVD), as follows from the Eckart-Young theorem [9] or analogously, it is obtained by applying the spectral theorem on the covariance matrix of X which is mean-centered:

$$Cov(X) = \frac{1}{N} X^T X = P \Lambda P^T$$

where $P \in \mathbb{R}^{T,T}$ is orthogonal ($P^T = P^{-1}$), $\Lambda \in \mathbb{R}^{T,T}$ and $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_T)$ with $\lambda_1 > \lambda_2 > \ldots > \lambda_T$. The sum of the eigenvalues $\lambda_1, \ldots, \lambda_T$ of the covariance matrix is the total variance of the data. Thus, keeping more components means being able to explain more data variability. The eigenvectors matrix P is defined as $P = [p_1, p_2, \ldots, p_T]$ and p_i, $i = 1, \ldots, T$ are the loading vectors or principal components. The dimensionality reduction with K components is obtained as

$$Z_K = XP_K$$

where $P_K = [p_1, \ldots, p_K] \in \mathbb{R}^{T,K}$, the reconstructed data as $\hat{X} = Z_K P_K^T$ and so $\hat{X} = XP_K P_K^T$. More explicitly, we have

$$\hat{x}_i(t) = \sum_{k=1}^{K} (x_i \cdot p_k) p_k(t) \; i = 1, \ldots, N, \; t = 1, \ldots, T. \tag{4}$$

.

It is evident that PCA is a decorrelation transformation and that its solution is not unique. Indeed, the loss is invariant under the transformation $P \to PU$, where U is any orthogonal matrix. Under this transformation, the loading vectors are transformed into a different orthonormal basis for the same subspace. Moreover, according to the Eckart-Young theorem [9], the truncated SVD is the best low-rank approximation of a matrix. Therefore, PCA is the linear dimensionality reduction method that minimizes the least squares error in the distortion when we project back to the original space[2].

Finally, it is worth noticing that, compared to other insider detection methods, the one based on PCA does not depend on the knowledge of the initial position of the investors. As explained in Subsection 2.1, this information is indeed lacking in our dataset and in general it can be difficult to obtain because it requires the knowledge of the whole past history of an investor trading activity. Clustering based on k-means, adopted in [19] depends instead on the arbitrary choice of the initial position of the investors. To show that this is not the case for PCA, we prove that the PCA reconstructed position error ϵ is invariant under the addition of an arbitrary constant to the vector of positions of a given investor. *Proof.* Let us consider the vector x_i describing the position of investor i and let us add an arbitrary constant C to it, obtaining $x_i^C = x_i + C\mathbb{I}_{Tx1}$. Denoting with $\hat{x}_i$ and $\hat{x}_i^C$ the reconstructed positions, it is direct to show that $\hat{x_i}^C = \hat{x}_i + C\mathbb{I}_{Tx1}$. First, we observe that, given the high number of investors, performing PCA on a dataset where investor i has position x_i and then, on a dataset where investor i has position x_i^C leads to loading vectors which are basically the same.

[2]Observe that our starting data set is in the format $N \times T$, i.e. the trading days are the features that are subjected to compression. Alternatively, we could start with a data set $Y \in \mathbb{R}^{T,N}$ where the features are the investors. In this case, PCA consists in identifying a subset of investors, or combinations of them, with a major role in the characterization of agents' trading activity. However, this would lead to a more time consuming and computationally expensive procedure. Indeed, we would need to obtain the eigendecomposition of the covariance matrix $Cov(Y) \in \mathbb{R}^{N,N}$, which is much larger than $Cov(X) \in \mathbb{R}^{T,T}$ since $N \gg T$. Furthermore, as shown in Appendix B, if the same feature scaling is applied on the data, the results we obtain starting from the data set in two different formats are analogous.

Referring to Equation 4,

$$\begin{aligned}
\hat{x}_i^C(t) =& \sum_{k=1}^{K}(x_i^C \cdot p_k)p_k(t) = \sum_{k=1}^{K}[(x_i + C\mathbb{I}_{Tx1}) \cdot p_k]p_k(t) = \\
=& \sum_{k=1}^{K}(x_i \cdot p_k)p_k(t) + C\sum_{k=1}^{K}(\mathbb{I}_{Tx1} \cdot p_k)p_k(t) = \\
& = \hat{x_i}(t) + C\big\{[p_1(1)p_1(t) + p_2(1)p_2(t) + \ldots + p_K(1)p_K(t)] + \ldots + \\
& + [p_1(t)p_1(t) + p_2(t)p_2(t) + \ldots + p_K(t)p_K(t)] + \ldots + \\
& + [p_1(T)p_1(t) + p_2(T)p_2(t) \ldots + p_K(T)p_K(t)]\big\} = \\
& = \hat{x_i}(t) + C[(P_K P_K^T)_{1t} + \ldots + (P_K P_K^T)_{tt} + \ldots + (P_K P_K^T)_{Tt}] = \\
& = \hat{x_i}(t) + C \qquad \forall t = 1, \ldots, T
\end{aligned}$$

where $(P_K P_K^T)_{lm}$ is the element l, m of the matrix $P_K P_K^T$ and in the last step we exploit that the matrix P_K is orthogonal. As explained in Subsection 2.1, our anomaly detection approach is *reconstruction-based* and $\epsilon_i = \epsilon_i^C$ indeed,

$$\epsilon_i^C = ||\hat{x_i}^C - x_i^C|| = ||\hat{x}_i + C\mathbb{I}_{Tx1} - x_i - C\mathbb{I}_{Tx1}|| = \epsilon_i.$$

Therefore, the reconstruction error is independent of C and if the profile x_i is identified as anomalous, the same will be true for x_i^C. □

However, in our approach, after trading positions are computed, they are normalized according to Equation 2 and so, computing the position of an investor setting the zero of her portfolio on a different day means that the arbitrary constant C is added to her unnormalized position. Let us define as ψ_i the position of investor i before normalization. Then, we define $\psi_i^C = \psi + C\mathbb{I}_{Tx1}$ as the position of the same investor computed by setting the zero of the portfolio on another day. After normalization, the positions are

$$\begin{aligned}
x_i(t) &= \frac{\psi_i(t)}{||\psi_i||_\infty} \\
x_i^C(t) &= \frac{\psi_i^C(t)}{||\psi_i^C||_\infty} = \frac{\psi_i(t) + C}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty}.
\end{aligned}$$

Therefore, we have that the reconstructed positions of x_i^C are

$$\begin{aligned}
\hat{x}_i^C(t) =& \sum_{k=1}^{K}(x_i^C \cdot p_k)p_k(t) = \sum_{k=1}^{K}\left(\frac{\psi_i + C\mathbb{I}_{Tx1}}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty} \cdot p_k\right)p_k(t) = \\
=& \sum_{k=1}^{K}\left(\frac{\psi_i}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty} \cdot p_k\right)p_k(t) + C\sum_{k=1}^{K}\left(\frac{\mathbb{I}_{Tx1}}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty} \cdot p_k\right)p_k(t) = \\
=& \frac{||\psi_i||_\infty}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty}\sum_{k=1}^{K}\left(\frac{\psi_i}{||\psi_i||_\infty} \cdot p_k\right)p_k(t) + \frac{C}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty} = \\
=& \frac{||\psi_i||_\infty}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty}\hat{x}_i(t) + \frac{C}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty} \qquad \forall t = 1, \ldots, T,
\end{aligned}$$

and the reconstruction error is

$$\epsilon_i^C = \frac{||\psi_i||_\infty}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty}\epsilon_i.$$

This implies that $\epsilon_i^C = \epsilon_i$ if

$$\frac{||\psi_i||_\infty}{||\psi_i + C\mathbb{I}_{Tx1}||_\infty} = 1, \tag{5}$$

which holds if $\max_t |\psi_i(t)| = \max_t |\psi_i^C(t)|$ given that $\max_t |\psi_i^C(t)| \neq 0$. As we will show in Section 4, this last condition holds for the majority of the profiles in our dataset.

Autoencoders

PCA is a linear method, consisting in applying the loading vectors' matrix to the starting data twice. Its nonlinear counterpart is an autoencoder (AE)[11]. Autoencoders' goal is analogous to PCA's: starting from a data set X, they aim to obtain a compressed representation of data Z and then, a reconstructed version $\hat{X}$:

$$X \to Z = f_1(X) \to \hat{X} = f_2(Z) \text{ where } X, \hat{X} \in \mathbb{R}^{N,T},\ Z \in \mathbb{R}^{N,K}.$$

The transformations f_1 and f_2 are such that the reconstruction error is minimized i.e.

$$\hat{X} = \arg\min_{\hat{X}'} ||X - \hat{X}'||_F^2$$

and, in this case, the compression and decompression steps are made by neural network layers. For an AE with one hidden layer, we have

$$\hat{x}_i(t) = g_2\Big(\sum_{k=1}^{K} g_1(x_i \cdot W_1(:,k))W_2(k,t)\Big)\ i = 1,\dots,N,\ t = 1,\dots,T, \tag{6}$$

where $W_1 \in \mathbb{R}^{T,K}$, $W_2 \in \mathbb{R}^{K,T}$ are layers' weight matrices, also called *encoder* and *decoder*, and g_1, g_2 are activation functions. The layers' weight matrices are determined by common gradient descent algorithms like Adam [15].

Above, we focus on an autoencoder with one hidden layer in order to highlight the analogy with PCA. Indeed, as it is shown in [16], if the activation functions are linear, the autoencoder with L_2-regularization learn PCA's principal directions. This issue is examined in depth in Appendix C. However, autoencoders can be deeper and can have complex architectures such as the well-know convolutional, LSTM, variational autoencoders [11].

As we saw above, the solution of PCA follows from the Eckart-Young theorem [9]. The latter states that the solution to the problem

$$\arg \min_{\hat{X}':\ rank(\hat{X}') \leq K} ||X - \hat{X}'||_F^2$$

is given by the truncated SVD. The latter approximates excellently data with linear relationships. On the other hand, concerning the reconstruction of nonlinear data, autoencoders outperform PCA, as empirical results in different fields show e.g. applications on image reconstruction. Theoretically, the difference between the problems tackled by PCA and AE is the constraint on the rank of $\hat{X}$. For PCA we impose $rank(\hat{X}) \leq K$, so there are no more than K columns of $\hat{X}$ which are linearly independent. Equivalently, we have no more than K independent features. On the other hand, AEs aim at minimizing the reconstruction error without any constraint on $rank(\hat{X})$. This means we could end up with $rank(\hat{X}) \in (K, T]$ so, more independent features.

Pros and cons of the different methods

PCA should be preferred against AEs if small datasets are considered, more interpretability and nested solutions are needed. PCA is also easier to implement than AEs. Moreover, it needs less computational resources and less training time.

The interpretability of the results which are provided by PCA, is due to the linearity of the method. However, this linearity can be a downside if data are nonlinear. On the other end, AEs' nonlinearity allows to capture complex relationships in the data. This last bright side of AEs leads to a better performance in the reconstruction of outliers, compared to PCA, and this could become a downside for our anomaly detection task, if anomalies are reconstructed such that they are indistinguishable from normal data instances.

Finally, PCA loading factors are ordered such that the associated eigenvalues are in decreasing order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_T$. Thus one can measure the importance of a factor in explaining the data and rank the factors according to this criterion. On the contrary, the weight vectors learned by AE are not constrained to form an orthonormal basis, nor to have a meaningful ordering.

Choice of K

The choice of the dimension K of the compressed representation of the input data should achieve a trade-off between capturing enough information and avoiding overfitting, which could lead to reconstruct profiles of investors with anomalous behavior well.

The relying assumption of our use of the dimensionality reduction approach is that the essential and common characteristics are captured by the lower dimensional space and that they explain a large fraction of data variance. Then, anomalous behavior cannot be reproduced given the compression and decompression, and anomalous observations have higher reconstruction errors than normal ones. However, the unsupervised nature of our case makes it extremely complicated, because our training data are anomaly-contaminated.

As a starting point we rely on standard methods to set the parameter K, like the percentage of explained variance and the Scree plot, that is the plot of the eigenvalues as a function of K [14]. However, given their erratic nature, we perform an analysis, which helps us in the choice. Let us define A_K as the set of investors identified as anomalous when K is used as dimension of the latent space. First, we determine the cardinality of A_K as a function of K. Then, we study the stability of this set by computing the Jaccard similarity [13] between each A_K and A_{K-1}. K is set to the lowest value of the interval in which we have stability in our results.

3 Data

3.1 Transaction reporting database

The analysis is based on transaction reports collected by Consob for the Italian stocks, according to the directive 2014/65 by European Union, also called MiFID II[3]. The

[3]In a nutshell, the MiFIDII/MiFIR regime has introduced new regulations for European financial markets and, among them, the transaction reporting obligation that requires investment firms or intermediaries

relevant dataset was built aggregating the daily transactions of all investors operating in any of the Italian stocks, in the period from January 1, 2019 to September 30, 2021. In details, the dataset was built according to the following rules: i) all the information related to the identity of individual investors have been anonymized; ii) with reference to each stock (identified by its ISIN code), each data point keeps a record of:

1. anonymous identifier of the investor;
2. type of investors (household: H, investment firm: IF, legal entity: L);
3. trading venue of the operation (Borsa Italiana - MTA, London Stock Exchange - LSE, off-exchange, etc.) for a total of 224 venues;
4. day of the operation;
5. buy and sell volumes (in shares);
6. buy and sell Euro volumes;
7. number of buy and sell contracts;
8. price of both the first and the last contracts (if there are more than one contract, otherwise they coincide);
9. minimum and max prices of contracts (if there are more than one contract, otherwise they coincide);
10. average price of buy (sell) contracts.

In the period covered by the data set, 2,253,707 investors were observed, operating in 286 Italian stocks. This is the same data set used in our previous paper about insider trading detection [19] and in another work related to the investigation of the trading behavior of Italian investors during the Covid pandemic [7].

3.2 Price sensitive events database

In addition to the transaction reporting database, a data set containing several price sensitive events (PSEs) was built; such events, obviously public, had all been analysed by the competent Authority with the aim of market abuse detection, by means of standard analytics methodologies. PSEs are events or a set of circumstances relating to listed companies which, when made public, had a significant impact on the price of the company's shares.

Our focus is on insider dealing in the Italian Stock Exchange. Investors who know in advance when a PSE will occur, can trade in a rewarding manner before the information spreads, thus closing their position after the PSE and making a profit. For instance, if a investor knows a few days before its public announcement that a takeover bid is going to occur for a given stock, they could exploit such information by buying shares of the stock considered. When the takeover bid occurs, the shares' price goes up aligning with the offer price and thus, the informed investor can sell by making a no-risk profit.

PSEs dataset contains a list of takeover bids for a number of stocks. As known, a takeover bid is a public offer made by a physical person or a legal entity who is willing

executing transactions in financial instruments to communicate "complete and accurate details of such transactions to the competent authority as quickly as possible, and no later than the close of the following working day".

Stock	PSE date	Investigation period (Δ)
IMA	July 28, 2020	June 29, 2020 - July 28, 2020
UBI	Feb 17, 2020	Jan 16, 2020 - Feb 17, 2020
PANARIAGROUP	Mar 31, 2021	Mar 1, 2021 - Mar 31, 2021
CARRARO	Mar 28, 2021	Jan 4, 2021 - Mar 28, 2021
MOLMED	Mar 17, 2020	Dec 2, 2019 - Mar 17, 2020

Table 1: Price sensitive events. The table reports the stock name, the date of the PSE, and the investigation period.

to buy other shareholders' shares at a price higher than the stock market value. As we saw, takeover bids can be exploited by an informed investor by buying before the event. It is worth mentioning that takeover bids have prolonged effects on the market, thus an insider can make a profit even without closing the position immediately after the announcement.

Our data report for each PSE the stock, its date, and the time window for insider trading investigation. This period varies depending on the type of PSE, which leads to different definitions of the time at which an information starts to be considered price sensitive. In Table 1, the PSEs database is displayed.

4 Results

As a first case study, we focus on the asset Industria Macchine Automatiche (IMA) whose takeover bid was announced on July 28, 2020. Figure 1 shows the price dynamics of this asset. The impact that the PSE had on the share price is evident: there is an increase of 13.16% on the day of the announcement and the takeover bid's price 68.0 Euro is reached. In analogy to [19], we identify the reference period as the time window going from January 2, 2020 to June 28, 2020. Instead, the business month preceding the PSE i.e. July 28, 2020, is the investigation period, as outlined in Table 1.

For each investor we extract from the database the asset position (in shares) at the end of each day. We assume that the position on January 2, 2020 is zero, but, as proved above, this arbitrary choice has no effect on the reconstruction error.

The trading days are $T = 149$, the investors active and with non-constant position are $N = 13,225$.

Anomaly detection with PCA

The first method we employ in order to perform the dimensionality reduction step, is PCA. In Appendix A we show the plot of the explained variance as a function of the number of the retained components and the Scree plot. Considering these figures we choose the latent space dimension $K = 16$, which allows to retain 97% of the explained variance. However, we perform a robustness analysis, investigating other choices of K in Appendix A.

After a feature scaling as pre-processing step, we run our method as illustrated in Section 2. As an example of the reconstructed trading position, in Figure 2 we show the position of an investor compared to its reconstructed counterpart obtained by PCA. Most of the days the reconstruction is quite close to the original trajectory and the

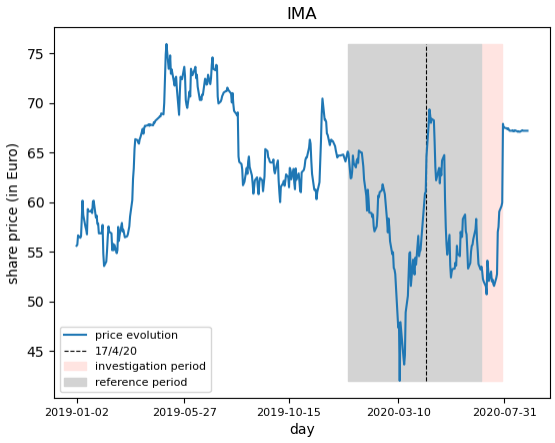

Figure 1: IMA price dynamics. The grey area is the reference period i.e. from January 2, 2020 to June 28, 2020. The pink area is the investigation period i.e. one business month before the PSE.

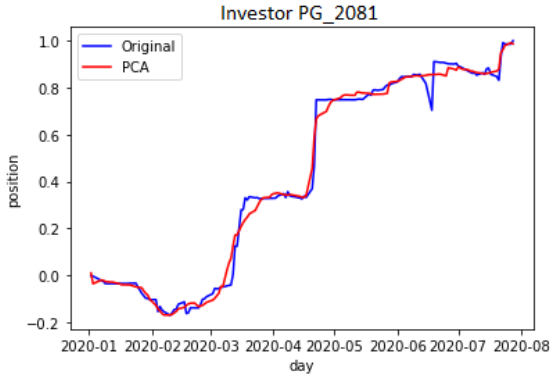

Figure 2: Comparison between the trading position of investor PG_2081 and the reconstructed one obtained by PCA with $K = 16$.

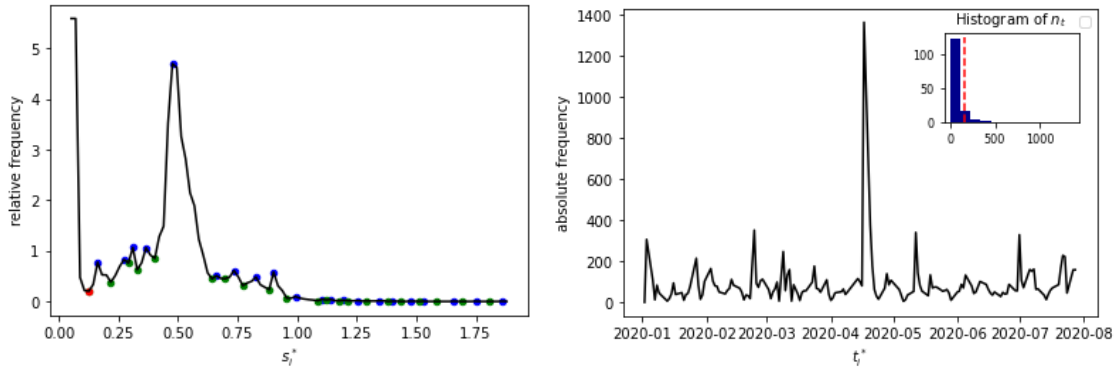

Figure 3: PCA on IMA. Left. Histogram of the anomaly scores. Blue points are local maxima, green local minima and the red point is $\epsilon_\theta \simeq 0.13$ that is the local minimum after the first peak. Right. Histogram of the times when anomaly scores are observed. The inset plot represents the histogram of the number of investors with given values of the times when their anomaly scores are observed. The vertical dashed line is n_θ i.e. the top decile of the distribution of n_t.

anomaly detection method identifies the days and investors for which the discrepancy, i.e. the reconstruction error, is large.

To identify the thresholds in the anomaly detection method, we plot in Figure 3 the histogram of the anomaly score s_i^* (left panel) and of the time of their occurrence t_i^* (right panel). As preannounced, a clear bimodal distribution in the former histogram is observed. The left mode (peak) contains investors with small anomaly score, thus "normal" investors, while the right mode contains potentially anomalous investors with a large maximal reconstruction error. Based on this empirical evidence, we set $\epsilon_\theta = 0.13$ as the first threshold parameter to identify potential insiders.

By focusing on the histogram of the times t_i^* corresponding to the largest reconstruction errors (right panel of Figure 3), we observe the presence of several large peaks, i.e. days when a large number of investors displayed a large reconstruction error. We can understand the origin of these peaks by focusing on the largest one, happened on April 17, 2020 when more than $1,300$ investors display the largest reconstruction error. Looking at the price dynamics in Figure 1, we observe that on April 17, 2020 there was a high increase (7.4%) of the share price. The large number of investors having the largest reconstruction error on that day is likely due to their reaction to this very volatile day[4]. Clearly these peaks and the corresponding investors are not insiders and this explains why we impose the condition on n_t, that are the heights of the peaks in the histogram of t_i^*, in our methodology to identify anomalous investors - see Equation 3. The parameter n_θ is set equal to 158, which corresponds to the top decile of the distribution of n_t, which cuts off the investors whose anomaly score falls in a peak of t_i^*.

Applying our anomaly detection method, we obtain $1,246$ potential insiders out of the $13,225$ total investors. Given the high percentage of anomalous investors that we

[4]We remind that positions of investors are measured in shares and not in Euro, so the peaks are not associated with change in value of a position, due to the large price variation, but to a genuine trading activity.

obtain, their ranking, following the procedure explained at the end of Section 2.1, is fundamental to provide more insight.

We compare our results with the findings of a method based on k-means similar to that of our previous paper [19]. For each investor we extract the *signed turnover* and the *maximum exposure* in the period[5] and we use them as coordinates in a 2D space[6]. Then, we apply k-means to the set of points to identify clusters of investors and we label as anomalous an agent who, in the investigation period belongs to a different cluster than the ones in the reference period and the new cluster is the most rewarding one with respect to the PSE. If the PSE is the announcement of a takeover bid, the cluster with the most rewarding position is the closest to the point $(1, 1)$ i.e. both *signed turnover* and *maximum exposure* equal to 1. We distinguish two types of anomalous investors. They are *soft* if they are active in the reference period but with a different position than the one in the investigation period, while they are *hard* if they are only active in the investigation period[7]. In summary, the PCA (and later the AE) method acts directly on the whole trading profile of each investor (thus a vector of dimension T), while the method based on k-means considers two features, which are functions of the trading profile.

If we run the method based on k-means in 2D on IMA, 152 investors are identified as *soft* and 705 as *hard*. Among the $1,246$ potential insiders identified by the method based on PCA, 134 are *soft* and 671 are *hard*. From the comparison we find that the first 10 ranked anomalous investors, according to the method based on PCA, are all identified as suspicious by the method based on k-means. If we consider up to rank $50, 100, 150, 200, 300, 500$, investors who are also suspicious in the framework of the clustering method of [19] are $49, 99, 148, 185, 253, 451$ respectively.

The compatibility of the two methods is a positive sign of their robustness. However it is natural to ask what are the characteristics of the investors identified as anomalous only by one of the two methods. Of the first 500 ranked anomalous investors, 451 are also identified by the method based on k-means. Among the remaining ones, 46 are of the type represented in the top left panel of Figure 4 and 3 are of the type represented in the top right panel of Figure 4. The former performs one transaction on June 26, 2020 i.e. three days before the starting day of the investigation period. This investor could be suspicious given her aggressive buying position just in the vicinity of the PSE. However, this investor is not identified by the method based on k-means since the transaction is outside the investigation period. Therefore, contrary to the method based on k-means, the new method based on a dimensionality approach is not strictly dependent on an arbitrary choice of the investigation period.

On the other hand, the investor in the top right panel of Figure 4 sells a portion of her position on the day before the PSE, but still maintains a net buying position. Given this investor was not active in the reference period, this behavior of buying and then, selling in the investigation period, could be a strategy in order not to be identified

[5]The *signed turnover* is the aggregated Euro turnover of operations within the period, with positive (negative) sign for a net buying (selling) volume. The *maximum exposure* is the maximum of the absolute value of the position in Euro turnover within the period, with positive (negative) sign if the maximum is reached for a buying (selling) position. We refer to our previous paper [19] for a precise definition.

[6]In [19] we consider another feature, namely the *magnitudo/portfolio concentration* which represents the fraction of wealth in the investigated asset. Since our dimensionality reduction method considers only data related to the investigated asset, in the comparison we use a k-means approach in a 2D space.

[7]See our previous paper [19] for further details.

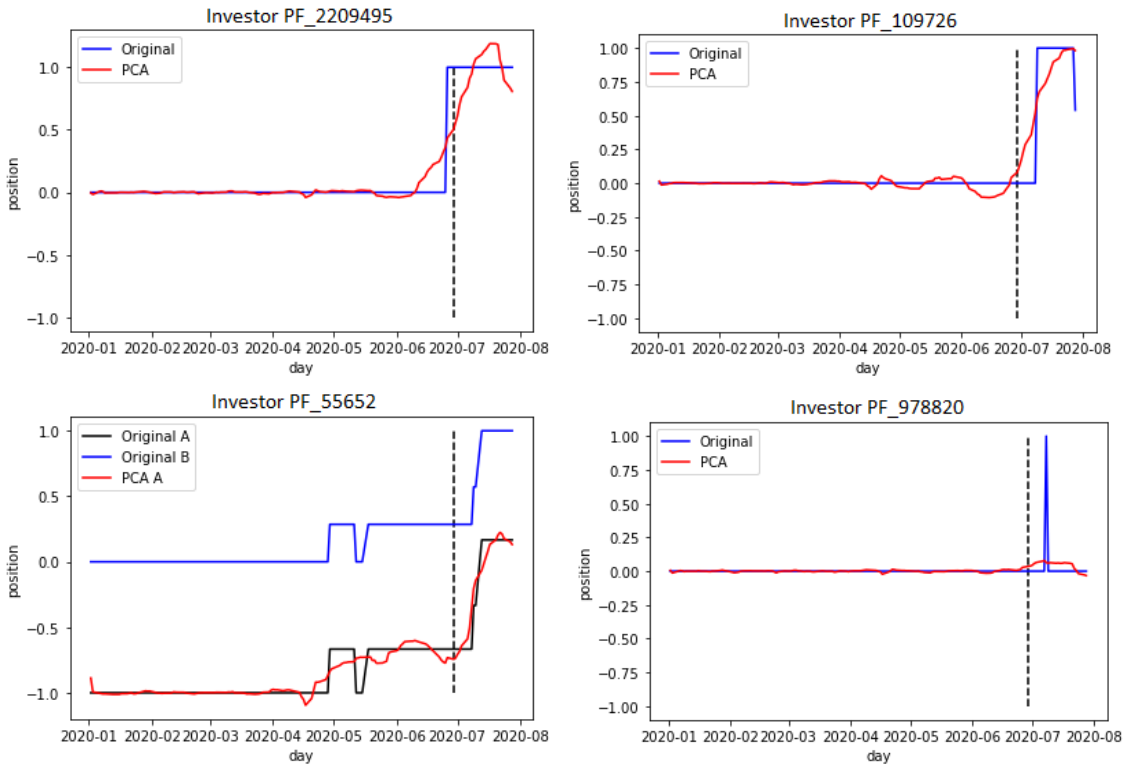

Figure 4: PCA on IMA. Top panels. Two anomalous investors identified by PCA but not by k-means. Bottom left. A refers to the trading position computed setting t_0 = January 2, 2019. B refers to the trading position computed setting t_0 = January 2, 2020.. Bottom right. Investor detected by the method based on k-means and not by the method based on PCA. In all panels the vertical dashed line is the day corresponding to the beginning of the investigation period i.e. June 29, 2020.

as suspicious by the regulator. This investor is not identified by the method based on k-means since in the last time window there is a drop in her *signed turnover* which leads the corresponding point to move away from the most rewarding cluster.

Another positive aspect of the new method based on a dimensionality reduction approach is its ability to detect given investors as anomalous, independently of the starting point chosen for the computation of her position (see Subsection 2.2). We observe that the condition in Equation 5 - which states when the reconstruction error of a given profile is the same of its shifted counterpart - holds for about 63% of investors in our dataset, if January 2, 2020 and January 2, 2019 are two choices of zeros for the computation of the positions. In the bottom left panel of Figure 4 we show the position of one investor when setting to zero the position on January 2, 2020 (blue) or on January 2, 2019 (black). Interestingly, this investor is identified as anomalous by the method based on PCA, but not by the method based on k-means. The fact that with the choice t_0 = January 2, 2019 this investor is identified as anomalous could surprise given we are focusing on detecting insider trading related to the announcement of a takeover bid. We recall that after this kind of PSE, the price increases and so, insiders are likely to have positive positions before the PSE. Indeed, this investor is not identified as anomalous

Name	Neurons of the hidden layers in the encoder	Encoding dimension	Neurons of the hidden layers in the decoder
AE-1	-	16	-
AE-2	32	16	32
AE-3	64, 32	16	32, 64
AE-4	128, 64, 32	16	32, 64, 128

Table 2: Autoencoders' architectures.

by the method based on k-means given her negative value of *signed turnover*. On the other hand, if the profile is computed with the choice $t_0 =$ January 2, 2020, she has a net buying position before the PSE and *signed turnover* equal to 1. Therefore, only with this choice of t_0, she is identified as *soft discontinuous* by the method based on k-means.

It is also interesting to investigate why some investors are detected by the method based on k-means and not by the method based on PCA. The total number of these investors is 79; among them, 27 are investors with constant position in terms of shares and so, they are not included in the analysis based on PCA. The majority of the remaining profiles are of the type displayed in the bottom right part of Figure 4. They are investors who have a null position in the investigation period, if positions are computed in shares. Instead, if positions are computed in Euros, as in the method based on k-means, the *signed turnover* and the *maximum exposure* of these investors are equal to 1 in the investigation period. They are in the best rewarding position and this makes them extremely suspicious according to the clustering approach of [19]. On the contrary, these investors are not identified by the method based on PCA, since the condition (4) relative to Equation 3 does not hold. That condition requires that the difference between the position on the PSE and the position on the first day of the reference period is larger than 0.5. For investors like PF_978820 (bottom right part of Figure 4), this difference is null since the investor closes her position before the PSE.

Finally, in Appendix B, we provide a comparison between the results obtained starting with the data set in the formats $N \times T$ and $T \times N$. As shown in Figure 9 and 10, if data are feature scaled in the same way, there is no difference between these results. In Appendix C, the relation between PCA and L_2-regularized autoencoders is tested on our data set.

Going nonlinear: the Autoencoder

As extensively proved in other research fields such as image reconstruction, adopting nonlinear and deep autoencoders can lead to a gain in performances, giving their ability to capture more complex relations in data. However, it is important to stress that our ultimate goal is not to best reconstruct our data. We wish to achieve a trade-off and to avoid overfitting. The idea is to obtain a lower dimensional space which captures the common and essential data characteristics; in this way, normal trading profiles will be well described while anomalous ones will not.

Bearing this in mind, we investigate the use of nonlinear autoencoders for our problem. The symmetric architectures we employ are schematized in Table 2. The number of neurons is chosen according to the geometric pyramid rule [18] and for

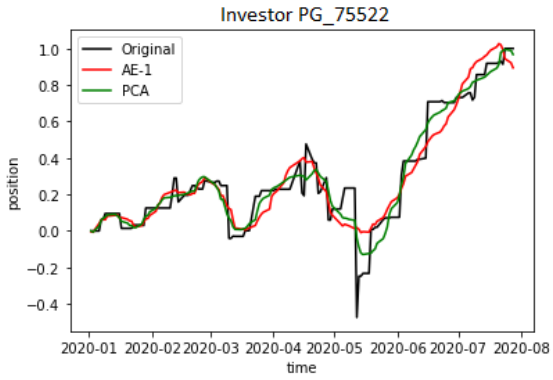

Figure 5: IMA. Comparison between the trading position of investor PG_75522 and the reconstructed ones obtained by PCA and AE-1 with $K = 16$.

Model	$\|\|X - \hat{X}\|\|_F$	EVS	$\bar{s^*}$	$\bar{s^*}_{anomalous}$	$\bar{s^*}_{normal}$	M_1	M_2
PCA	132.0	97.66	0.4525	0.5051	0.4489	0.1251	1.116
AE-1	168.7	96.18	0.5100	0.6088	0.5031	0.2099	1.193
AE-2	137.5	97.46	0.4574	0.5128	0.4536	0.1308	1.121
AE-3	129.3	97.76	0.4404	0.4997	0.4363	0.1453	1.134
AE-4	121.1	98.00	0.4228	0.4620	0.4201	0.0999	1.093

Table 3: Metrics for different dimensionality reductions of the IMA dataset. EVS is the explained variance score and it is defined as $100\left(1 - \frac{Var(X-\hat{X})}{Var(X)}\right)$; $\bar{s^*}$ is the mean anomaly score; $\bar{s^*}_{anomalous}$ is the mean anomaly score for investors who are identified as *hard/soft* by the method based on k-means; $\bar{s^*}_{normal}$ is the mean anomaly score for all investors except the ones identified as *hard/soft* by the method based on k-means; $M_1 = \frac{\bar{s^*}_{anomalous} - \bar{s^*}_{normal}}{\bar{s^*}_{normal}}$; $M_2 = \frac{\bar{s^*}_{anomalous}}{\bar{s^*}}$. The results related to the autoencoders are averaged over 10 runs.

all architectures the activation function of the hidden layers is the ReLU, while the activation function of the output layer is the hyperbolic tangent. This last choice allows to exploit the nonlinearity of the neural networks and yet to produce outputs with values in the interval $[-1, 1]$, the domain of the normalized trading positions. The loss function is the mean squared error (MSE) and Adam [15] is used as optimization algorithm.

The PCA and the 4 autoencoder architectures of Table 2 are run on our data set of trading positions for IMA. A comparison between the reconstructed profile of an investor, obtained with PCA and a type of autoencoder is represented in Figure 5. Greater smoothness is associated to the profile obtained by AE-1 however, the overall similarity between the two different profile reconstructions is evident. Table 3 summarizes the main results for all autoencoders' architectures in terms of several metrics. Due to their nonlinear character, we expect deep autoencoders can capture more complex features in data. This could lead to the identification of characteristics for the data compression which are more relevant than the ones identified by PCA, which is a linear model. However, we need at least 5 hidden layers to outperform PCA in terms of the loss function $||X - \hat{X}||_F$. The lower MSE is accompanied by greater

explained variance score (EVS) which is defined as[8]

$$EVS = 100\left(1 - \frac{Var(X - \hat{X})}{Var(X)}\right).$$

This means that deeper autoencoders can explain a larger variance of the data.

Nonetheless, we are not solely interested in better reconstruction errors. A larger gap between the errors of the anomalous investors and the ones who are not anomalous is desired. Given our unavailability of labels, we compare the mean anomaly score of investors who are detected as *hard/soft* by the method based on k-means [19] i.e. $\bar{s^*}_{anomalous}$, and the mean anomaly score of the other investors i.e. $\bar{s^*}_{normal}$.

The two metrics M_1 and M_2, that we introduce, shed light on this issue. They are defined as

$$M_1 = \frac{\bar{s^*}_{anomalous} - \bar{s^*}_{normal}}{\bar{s^*}_{normal}}$$

and

$$M_2 = \frac{\bar{s^*}_{anomalous}}{\bar{s^*}}.$$

We obtain that the model AE-1 has the greatest values of M_1 and M_2 i.e. it leads to a greater gap between the anomaly scores of our proxy of anomalous investors and the others. On the other hand, AE-4, which allows to obtain the lowest error in the data reconstruction, leads to the lowest value of the two quantities. It is important to stress that the comparison between the values of M_1 and M_2 that are obtained with different models, provide information which could be useful to our insiders detection task. However, the anomaly score is not the only quantity which determines whether an investor is identified as anomalous; the distribution of the times corresponding to the largest reconstruction errors also plays a fundamental role, both on the identification and on the ranking. We will investigate this issue in the following by employing both AE-1 and AE-4 for the anomaly detection step.

Anomaly detection with autoencoders

We perform our anomaly detection task by employing two different architectures of autoencoders i.e AE-1 and AE-4. The main results are summarized in Table 4 and compared with the results of PCA.

First of all we notice that when considering the first 150 ranked investors, the different methods provides almost identical set of anomalous cases. This, once more, indicates that machine learning methods (k-means, PCA, autoencoders) essentially

[8]Notice that the EVS is different from a common metric that is usually employed for PCA, that is the explained variance ratio (EVR), defined as

$$EVR = \frac{\sum_{k=1}^{K} var(z_k)}{\sum_{t=1}^{T} var(x_t)},$$

where $x_t \in \mathbb{R}^N$ are the columns of $X \in \mathbb{R}^{N,T}$, that is the data matrix, and $z_k \in \mathbb{R}^N$ are the columns of $Z_K \in \mathbb{R}^{N,K}$, that is X's representation in the K-dimensional latent space. The EVR is not a good indicator for autoencoders since it strictly depends on the activation functions we choose. Therefore, EVS is preferred to EVR, thus allowing a fair comparison with the PCA results.

Method	$\lvert A\rvert$	$I/\lvert A_{KM}\rvert$	I_{10}	I_{50}	I_{100}	I_{150}	I_{200}	I_{300}	I_{500}
PCA	1,246	805/857	10	49	99	148	185	253	451
AE-1	1,502	812/857	10	49	99	148	186	193	337
AE-4	1,325	807/857	10	49	99	148	166	226	424

Table 4: IMA: anomaly detection. A is the set of investors identified as anomalous. I is defined as $I = |A \cap A^{KM}|$ where A_{KM} is set of investors identified as *hard/soft* by the method based on k-means. I_n is defined as $I_n = |A_n \cap A^{KM}|$ where A_n is the set of the first n ranked anomalous investors.

agree in the identification of the most suspicious investors, demonstrating an overall robustness of the adopted methodologies.

When we focus on the first 500 ranked investors, PCA is the method with the largest overlap with the method based on k-means. This cannot be explained by the anomaly score values since, as shown in Table 3, the value of the metrics M_1, M_2 obtained by PCA are lower than the one obtained by AE-1. The two different dimensionality reduction methods lead to different histograms of the times corresponding to the largest reconstruction errors, which are shown in the right panels of Figures 3 and 14. While the histogram obtained with PCA has a maximum on April 17, 2020, the one obtained with AE-1 has the highest peak on the day of the PSE. Contrary to PCA, the autoencoder AE-1 is able to provide a dimensionality reduction where the trading activity on April 17, 2020 is treated as "normal" for a large fraction of investors. Therefore, the decrease of the overlapping could be ascribed to the change in the distribution of the times corresponding to the largest anomaly scores.

Among the first 500 ranked investors identified by AE-1 or AE-4, investors who are detected by AE-1 or AE-4 and that are not detected by the method based on k-means are analogous to the ones which were identified by relying on PCA and not by the method based on k-means i.e. profiles like the one in the top left panel of Figure 4, with one buying transaction just before the beginning of the investigation period.

Instead, among the first 500 ranked investors identified by AE-1, 62 are not identified by PCA. Among them, the investors ranked 15 (Figure 15), 114, 146, 396 are identified as *hard* by the method based on k-means. The others are of the type represented in the top left panel of Figure 4. In a similar way, among the first 500 ranked identified by AE-4, investors who are not detected by PCA are 16. Among them, 3 are *hard* and are like the investor in Figure 17, the others are like the one in the top left panel of Figure 4. It is evident the ability of the autoencoders to capture as anomalous a type of profile like the ones in Figure 15 - 17, that were not identified by PCA and that are *hard* according to the method based on k-means. Moreover, again, the method based on dimensionality reduction approaches shows its independence of the choice of the investigation period.

If we compare the results obtained by employing different architectures of autoencoders, among the first 500 ranked by AE-4, only 1 was not detected by AE-1.

4.1 Other case studies

While we have extensively covered the case study related to the asset IMA, we now focus on the other PSEs shown in Table 1. Table 5 summarizes the main results obtained

Asset	N	T	K	$\lvert A\rvert$	$I/\lvert A^{KM}\rvert$	I_{10}	I_{50}	I_{100}	I_{150}	I_{200}	I_{300}	I_{500}
IMA	13,225	149	16	1,246	805/857	10	49	99	148	185	253	451
UBI	31,970	118	16	1,801	1,255/1,432	10	50	100	150	200	300	499
PANARIAGROUP	1,068	56	12	232	178/188	10	42	91	125	150	-	-
CARRARO	4,500	317	24	537	431/500	9	49	99	149	199	283	401
MOLMED	11,976	307	38	1121	465/1,264	1	10	41	60	62	86	286

Table 5: Anomaly detection on all assets, obtained by employing PCA. N and T are the numbers of investors and days in the data set. K is the encoding dimension. A is the set of investors identified as anomalous. I is defined as $I = |A \cap A^{KM}|$ where A_{KM} is set of investors identified as *hard/soft* by the method based on k-means. I_n is defined as $I_n = |A_n \cap A^{KM}|$ where A_n is the set of the first n ranked anomalous investors.

Model	$\lvert\lvert X - \hat{X}\rvert\rvert_F$	EVS	$\bar{s^*}$	$\bar{s^*}_{anomalous}$	$\bar{s^*}_{normal}$	M_1	M_2
PCA	205.8	97.07	0.4953	0.5982	0.4905	0.2196	1.208
AE-1	234.0	96.20	0.5225	0.6123	0.5183	0.1816	1.172
AE-2	185.6	97.61	0.4417	0.5267	0.4377	0.2034	1.192
AE-3	168.3	98.04	0.4003	0.5015	0.3955	0.2684	1.253

Table 6: UBI: dimensionality reduction. EVS is the explained variance score and it is defined as $100\Big(1 - \frac{Var(X-\hat{X})}{Var(X)}\Big)$; $\bar{s^*}$ is the mean anomaly score; $\bar{s^*}_{anomalous}$ is the mean anomaly score for investors who are identified as *hard/soft* by the method based on k-means; $\bar{s^*}_{normal}$ is the mean anomaly score for all investors except the ones identified as *hard/soft* by the method based on k-means; $M_1 = \frac{\bar{s^*}_{anomalous} - \bar{s^*}_{normal}}{s^*_{normal}}$; $M_2 = \frac{\bar{s^*}_{anomalous}}{s^*}$. The results related to the autoencoders are averaged over 10 runs.

Method	$\lvert A\rvert$	$I/\lvert A_{KM}\rvert$	I_{10}	I_{50}	I_{100}	I_{150}	I_{200}	I_{300}	I_{500}
PCA	1,801	1,255/1,432	10	50	100	150	200	300	499
AE-3	2,106	1,348/1,432	10	50	100	150	200	300	457

Table 7: UBI: anomaly detection. The encoding dimension is 16. A is the set of investors identified as anomalous. I is defined as $I = |A \cap A^{KM}|$ where A_{KM} is set of investors identified as *hard/soft* by the method based on k-means. I_n is defined as $I_n = |A_n \cap A^{KM}|$ where A_n is the set of the first n ranked anomalous investors.

by using PCA. The overlapping with the results of the method based on k-means are analogous to what is obtained for IMA, except for MOLMED. For this asset, the small overlapping is due to the choice of the investigation period, thus highlighting the ability of our method based on a dimensionality reduction approach to be independent of the choice of the investigation period. Moreover, the value added by our new method to the insider trading detection task, is analogous to what is obtained for IMA.

Now, let us deepen the main results related to the asset UBI. We employ both PCA and autoencoders. In Table 6, a comparison between the reconstruction results obtained by employing different architectures is shown. A trend different from IMA can be observed. In this case, we need at least 3 hidden layers to outperform PCA in the reconstruction of trading profiles and AE-3 shows the lowest MSE. Contrary to IMA, the architecture which leads to the greatest values of M_1 and M_2 is still AE-3. If we rely on this autoencoder and apply our anomaly detection method, the results we obtain are provided in Table 7, compared with the ones of PCA.

If the first 500 ranked by PCA are considered, our new method does not provide new information compared to the one based on k-means. This is in contrast with the autoencoder. The profiles detected by AE-3 and not by k-means are analogous to the one in the top left panel of Figure 4. On the other hand, the investors detected by AE-3 and not by PCA are *hard* according to the method based on k-means and analogous to the profile in Figure 17. The ability of autoencoders to capture this type of investors which are not detected by PCA, was already shown in the study related to IMA.

5 Conclusion

We proposed a novel unsupervised approach for contextual anomaly detection, to support decision in insider trading detection. This method tackles the same issue of our previous paper [19] with a different point of view. In particular, the method based on k-means, that we develop in [19], is based on the definition of three features i.e. *signed turnover*, *magnitudo*, *maximum exposure*. With this new method, we aim at overcoming the features' choice: our only input is the trading position of each investor for a given asset and the model learns the relevant characteristics by itself.

This new approach lies in the *reconstruction-based* paradigm of anomaly detection and it involves several steps. First, we employ PCA or autoencoders and we obtain the reconstruction errors for the trading profiles of each investor active on a given asset for which we have a takeover bid. Then, we localize the largest errors and impose several conditions in order to detect anomalous investors, who could be suspicious of insider trading related to the PSE.

We observe a consistent overlapping with the results of the method based on k-means. However, the value added of this new method is evident. If PCA is employed as dimensionality reduction approach, the method is extremely fast and easy to implement. Both with PCA and autoencoders, we do not longer have to choose the trading features which allow to characterize the trading activity of each investor. The method is not strictly dependent on the choice of the beginning of the investigation period and actually, it could provide insight on whether this time window should be fixed. The method is also independent of the choice of the initial time for the computation of the trading positions.

The differences between the performances of PCA and autoencoders are case-by-case dependent. We showed that autoencoders allow to identify as anomalous, profiles that are not detected by PCA and are actually *hard* according to the method based on k-means. We think that for small data sets, PCA is a sufficient method to perform the dimensionality reduction step. Instead, for larger data sets, a coupled use of PCA and autoencoders should be preferred. This conclusion is also motivated by the extreme complexity of our problem, that is also strengthened by the unavailability of labels, which force us to evaluate the performance of our method without a systematic procedure.

A natural extension of this work is the employment of more complex architectures of autoencoders.

List of abbreviations

- AE: Autoencoder
- CARRARO: Carraro S.p.A.
- EVS: Explained Variance Score
- IMA: Industria Macchine Automatiche S.p.A.
- LAE: Linear Autoencoder
- LSTM: Long Short-Term Memory
- MOLMED: MolMed S.p.A.
- MSE: Mean Squared Error
- PANARIAGROUP: Panariagroup Industria Ceramiche S.p.A.
- PCA: Principal Component Analysis
- PSE: Price Sensitive Event
- ReLU: Rectified Linear Unit
- SVD: Singular Value Decomposition
- UBI: UBI Banca - Unione di Banche Italiane S.p.A.

Declarations

Availability of data and materials

The data that support the findings of this study are from the Commissione Nazionale per le Società e la Borsa (CONSOB), that is the public authority responsible for regulating the Italian financial markets. Restrictions apply to the availability of these

data, because of the severe privacy policy related to the data collected within the MiFIDII/MiFIR regime and are not publicly available.

Competing interests

The authors declare that they have no competing interests.

Funding

This paper is funded by the project "Percorso di formazione su data analysis, network science, artificial intelligence e big data per gli abusi di mercato", a research collaboration between Scuola Normale Superiore and CONSOB.

Authors' contributions

Adele Ravagnani and Fabrizio Lillo contributed to the conception and design of the work and the interpretation of the results. Adele Ravagnani developed the methodology, analyzed the data, and was the major contributor to writing the manuscript. All authors read and approved the final manuscript.

References

[1] Aggarwal CC. Outlier Analysis. Springer-Verlag New York; 2013.

[2] Akoglu, L., Tong, H. & Koutra, D. Graph based anomaly detection and description: a survey. Data Mining and Knowledge Discovery 29, 626-688 (2015).

[3] Audibert Julien, Unsupervised anomaly detection in time-series (PhD thesis), 2022.

[4] Bhattacharya, U., & Daouk, H. (2002). The world price of insider trading. The Journal of Finance, 57(1), 75-108.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. ACM Comput. Surv. 41, 3, Article 15 (July 2009), 58 pages, doi.org/10.1145/1541880.1541882.

[6] Stéphane Crépey, Lehdili Noureddine, Nisrine Madhar, Maud Thomas, Anomaly Detection on Financial Time Series by Principal Component Analysis and Neural Networks, arXiv:2209.11686, 2022.

[7] Deriu, Paola and Lillo, Fabrizio and Mazzarisi, Piero and Medda, Francesca and Ravagnani, Adele and Russo, Antonio, How Covid mobility restrictions modified the population of investors in Italian stock markets (July 29, 2022). Available at SSRN: ssrn.com/abstract=4176182 or dx.doi.org/10.2139/ssrn.4176182.

[8] Donoho, S. (2004, August). Early detection of insider trading in option markets. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 420-429).

[9] Eckart, C. and Young, G., The Approximation of One Matrix by Another of Lower Rank. Psychometrika, 1(3):211-218, 1936.

[10] Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one, 11(4), e0152173.

[11] Goodfellow, I., Bengio, Y., and Courville, A. Deep Learning. MIT Press, 2016, deeplearningbook.org.

[12] Hartigan, J.A., 1975. Clustering algorithms. John Wiley & Sons, Inc..

[13] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2). New York: Springer.

[14] I. T. Jolliffe (2002), Principal Component Analysis, Springer New York, NY.

[15] Kingma, D. and Ba J. (2014), Adam: A method for stochastic optimization. Preprint: arxiv.org/abs/1412.6980.

[16] Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, Cotton Seed, Loss Landscapes of Regularized Linear Autoencoders, Proceedings of the 36th International Conference on Machine Learning, PMLR 97:3560-3569, 2019.

[17] Li, A., Wu, J., & Liu, Z. (2017). Market manipulation detection based on classification methods. Procedia Computer Science, 122, 788-795.

[18] Masters, T. 1993. Practical neural network recipes in C++. New York: Academic Press.

[19] Mazzarisi, Piero and Ravagnani, Adele and Deriu, Paola and Lillo, Fabrizio and Medda, Francesca and Russo, Antonio, A Machine Learning Approach to Support Decision in Insider Trading Detection (December 6, 2022). Available at SSRN: `https://ssrn.com/abstract=4294752` or `http://dx.doi.org/10.2139/ssrn.4294752`.

[20] Mejri et al. (2022), Unsupervised Anomaly Detection in Time-series: An Extensive Evaluation and Analysis of State-of-the-art Methods, arXiv:2212.03637.

[21] Minenna, M. (2003). The detection of market abuse on financial markets: A quantitative approach. Quaderni di finanza, (54).

[22] Morgia, M. L., Mei, A., Sassi, F., & Stefa, J. (2021). The doge of wall street: Analysis and detection of pump and dump cryptocurrency manipulations. ACM Transactions on Internet Technology (TOIT).

[23] Song, X., Wu, M., Jermaine, C., & Ranka, S. (2007). Conditional anomaly detection. IEEE Transactions on knowledge and Data Engineering, 19(5), 631-645.

[24] Tumminello, M., Micciche, S., Lillo, F., Piilo, J., & Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. PloS one, 6(3), e17994.

[25] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. Computers & security, 57, 47-66.

[26] Zhou, Y. and Liang, Y., Critical Points of Linear Neural Networks: Analytical Forms and Landscape Properties. In ICLR, 2018.

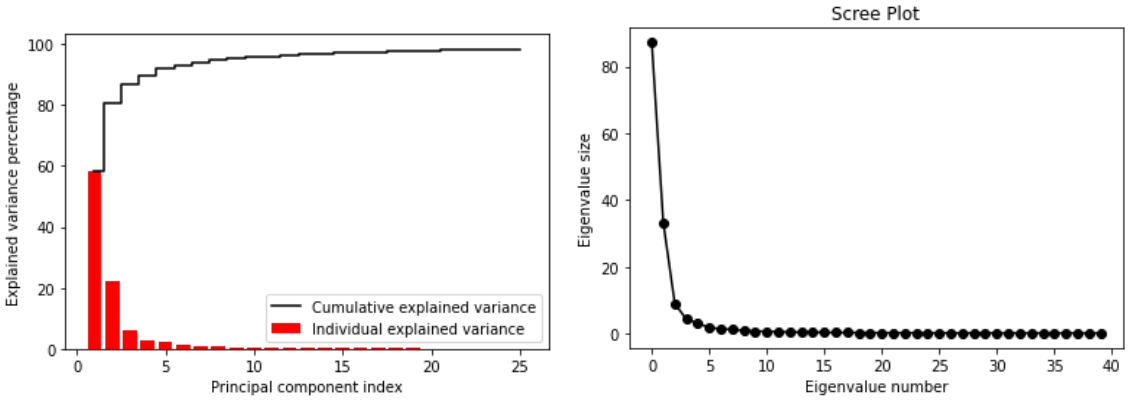

Figure 6: PCA on IMA. Explained variance percentage as a function of the number of retained components (left) and Scree plot i.e. eigenvalue size as a function of the corresponding component index (right).

APPENDICES

A The choice of K in PCA

In this Appendix we investigate the optimal dimension K in PCA analysis and we perform a robustness check. The results refer to IMA.

The explained variance percentage and Scree plots are shown in Figure 6. Keeping only 1 component allows to retain 58.5% of the data variance and we only need 5 components to reach 90%.

In Figure 7, the trading profile of an investor is compared with the reconstructed ones obtained by PCA for several values of K. It is evident that increasing the latent space dimension leads to an improvement in the reconstruction of the profile. Moreover, $K = 16$ allows to obtain reconstruction errors which are comparable to the ones related to higher values of K. Indeed, the choice $K = 16$ is such that 97% of the data variance can be explained. However, as we illustrated in Subsection 2.2, giving the unsupervised nature of our problem and our complete lack of labels, *a priori* we do not know the best suited value of this parameter. This motivates us to perform an analysis to understand how different choices of K could impact our results.

For several choices of K, we run our methodology using PCA for the dimensionality reduction step. Then, we identify a set of anomalous investors A_K and in order to test the stability of this set, we compute the Jaccard similarity [13] between each A_K and A_{K-1}. Results are reported in the left panel of Figure 8, which shows that, especially for small values of K, the similarity is unstable and oscillates. On the other hand, in an interval between $K = 16$ and $K = 18$, we have a more stable trend with very high values of the metric. This stability is also evident by looking at the right panel of Figure 8, which shows the cardinality of A_K for several values of K. This motivates us to set $K = 16$ i.e. the lowest value of the interval in which we have stability in our findings.

B PCA on data in two different formats

Our starting data set is in the format $N \times T$ i.e. the trading days are the features which are subjected to compression. We could start with a data set $Y \in \mathbb{R}^{T,N}$ where the features are the investors. As we explained in Subsection 2.2, this would lead to a more time consuming and computationally expensive procedure since in our dataset $N \gg T$. However, it is legitimate to ask whether there is a difference in the results obtained with these two approaches.

Before running PCA, it is fundamental to perform feature scaling. This preprocessing step consists in rescaling each feature such that it has unit standard deviation and null mean. Our input data are investors' positions which are normalized as explained in the main text. This first normalization is such that the activity of each investor is normalized compared to her own trading history. If the features are the trading days, the feature scaling before PCA leads to a data set where

$$x_i(t) \to \frac{x_i(t) - mean(x_t)}{std(x_t)} \tag{7}$$

where x_t are the columns of $X \in \mathbb{R}^{N,T}$. Therefore, this second normalization step consists in normalizing the position of each investor on a day with respect to the positions of all other investors on that day.

If instead, the feature scaling is performed with respect to investors, it would lead to a data set where

$$x_i(t) \to \frac{x_i(t) - mean(x_i)}{std(x_i)}.$$

where x_i are the rows of $X \in \mathbb{R}^{N,T}$. This means we are normalizing the position of each investor on a day with respect to the positions of the same investor on other days. We remind that also the normalization used in the main text, although different, uses the whole history of an investor's position.

Therefore, we adopt the feature scaling of Equation 7 and we apply PCA using as input the data set in the format $N \times T$ and then, in the format $T \times N$. The eigenvalues that we obtain in the two cases are the same, as it is represented in Figure 9. Analogously, Figure 10 shows the equivalency between the anomaly scores distributions.

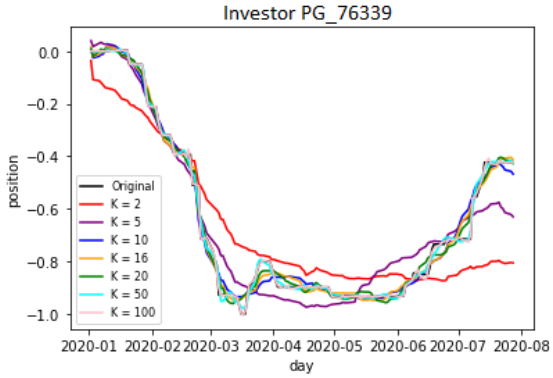

Figure 7: Comparison between the trading position of investor PG_76339 and the reconstructed ones obtained by PCA with several values of K.

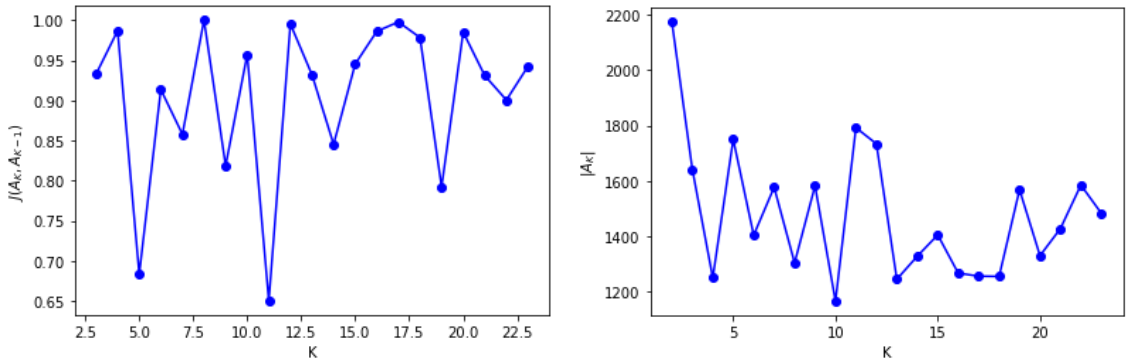

Figure 8: Left. Jaccard similarity between the set of anomalous investors identified with K and $K-1$. Right. Cardinality of the set of anomalous investors as a function of K.

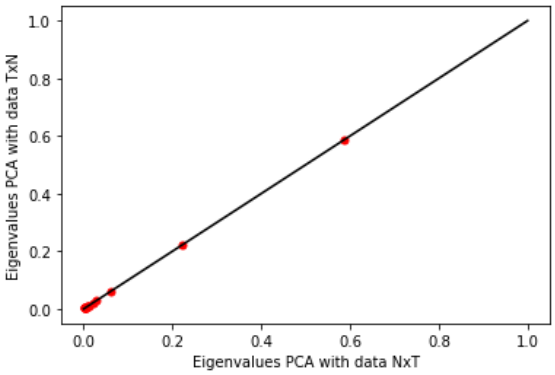

Figure 9: Comparison between the eigenvalues obtained by running PCA on data in the formats $N \times T$ and $T \times N$. The dark line is the bisector.

Formally, this can be explained by observing that PCA identifies the eigenvalues of the data covariance matrix. This means:

$$(X^T X)p = \lambda p$$

where λ is an eigenvalue and p is the corresponding eigenvector. If we multiply by X, we obtain

$$(XX^T)(Xp) = \lambda(Xp) \iff Cov(Y)(Xp) = \lambda(Xp)$$

where $Y = X^T$. Therefore, the eigenvalues of $Cov(X)$ and $Cov(Y)$ are the same while the eigenvectors are p and Xp.

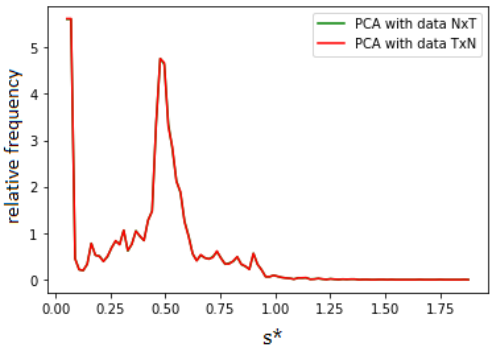

Figure 10: Comparison between the anomaly scores histograms obtained by running PCA on data in the formats $N \times T$ and $T \times N$.

C Relation between linear autoencoders and PCA

If we compare Equations 4 and 6, it is pretty evident they are analogous if the activation functions g_2 and g_1 are the identity functions and $W_1 = W_2^T$. Indeed, as illustrated in [16], if the activation functions are linear, the autoencoder with L_2-regularization learn PCA's principal directions.

Given the data $X \in \mathbb{R}^{N,T}$, linear autoencoders' (LAEs) goal is to obtain the following transformations:

$$X \to Z = XW_1 \to \hat{X} = XW_1W_2$$

where $Z \in \mathbb{R}^{N,K}$, $W_1 \in \mathbb{R}^{T,K}$, $W_2 \in \mathbb{R}^{K,T}$, and such that the loss function is minimized i.e.

$$W_{1,2} = \arg\min_{W_1,W_2} \mathcal{L}(W_1, W_2) = \arg\min_{W_1,W_2} ||X - XW_1W_2||_F^2.$$

As for standard autoencoders, W_1 is called *encoder* and W_2 *decoder*.

By the Eckart-Young theorem [9], the optimal rank-K solution is the truncated Singular Value Decomposition (SVD) i.e.

$$XW_1W_2 = U_K S_K V_K^T = USI_{TxK}V_K^T = USV^T V_K V_K^T = XV_K V_K^T.$$

Therefore, a LAE learns the principal subspace. However, it does not learn the principal directions indeed W_1, W_2 are optimal under the following transformations:

$$W_1 \to W_1 G$$
$$W_2 \to G^{-1} W_2$$
$$\forall G \in GL_K(\mathbb{R})$$

where $GL_K(\mathbb{R})$ is the general linear group i.e. matrices which are invertible.

Contrary to traditional PCA loading factors, the weight vectors learned by a LAE are not constrained to form an orthonormal basis, nor to have a meaningful ordering. However, they span the same subspace.

If instead of $\mathcal{L}(W_1, W_2)$, we consider

$$\mathcal{L}_\sigma(W_1, W_2) = \mathcal{L}(W_1, W_2) + \lambda(||W_1||_F^2 + ||W_2||_F^2),\ \lambda > 0,$$

the penalization term $\lambda(||W_1||_F^2 + ||W_2||_F^2)$ is not invariant under the general linear group indeed

$$||\alpha W_1||_F^2 = \alpha^2 ||W_1||_F^2 \neq ||W_1||_F^2.$$

On the other hand, it is invariant under the orthogonal group indeed

$$||W_1 O||_F^2 = ||W_1||_F^2\ \forall O \in O_K(\mathbb{R})$$

and we recall $O_K(\mathbb{R}) \subset GL_K(\mathbb{R})$. So, $\mathcal{L}(W_1, W_2)$ is invariant under the general linear group while $\mathcal{L}_\sigma(W_1, W_2)$ under the orthogonal group (the invariance is considered with respect to the transformation applied to W_1 and W_2).

As we said above, if W_1 is optimal, so does $W_1 G\ \forall G \in GL_K(\mathbb{R})$ and we observe that

$$W_1 G = U S V^T G$$

i.e. it is not in SVD form. On the other hand, we have that

$$W_1 O = U S V^T O$$

i.e. $W_1 O$ is in SVD form.

In [16], after this reasoning, authors provide an algorithm to recover the principal directions of PCA from LAE weight matrices. This is as follows:

- train a L_2-regularized LAE with loss function $\mathcal{L}_\sigma$ (input data can be not mean-scaled). The optimal W_1 and W_2 are W_1^* and W_2^*;
- apply SVD on W_2^{*T} $(T \times K)$: $W_2^{*T} = U \Sigma V^T$;
- the loading vectors are the columns of U i.e. the left singular vectors of the decoder.

This algorithm is a consequence of the *Landscape Theorem* of the paper [16]. Indeed, according to this Theorem, we have that the optimal value of the decoder and the encoder matrices for $\mathcal{L}_\sigma$ are defined up to an orthogonal map $O \in O_K(\mathbb{R})$:

$$W_2^T = U_K (I - \lambda \Sigma_K^{-2})^{\frac{1}{2}} O = W_1$$

where $X = U \Sigma V^T$ and $\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_K^2 > \lambda$. In the last equality, the *Transpose Theorem* [16] has been employed: it states that all critical points of $\mathcal{L}_\sigma$ satisfy $W_1 = W_2^T$.

To sum up, the L_2-regularized LAEs are transposes at all critical points and learn the principal directions as the left singular vectors of the decoder. Given this relation between LAE and PCA and the algorithm above, using LAE instead of PCA could be useful for large datasets. Indeed, SVD will be performed on a smaller matrix W_2^* which is $K \times T$, instead of X that is $N \times T$. Moreover, having a PCA-like solution allows to exploit nested solutions easily. Indeed, if results are obtained for a given K then, we can obtain the solution for $K' \neq K$, by truncating the loading vectors' matrix U at K' instead of K.

Finally, we recall that as proved in [26], the loss function for linear networks has no spurious local minimum, while such point does exist for nonlinear networks with ReLU activation.

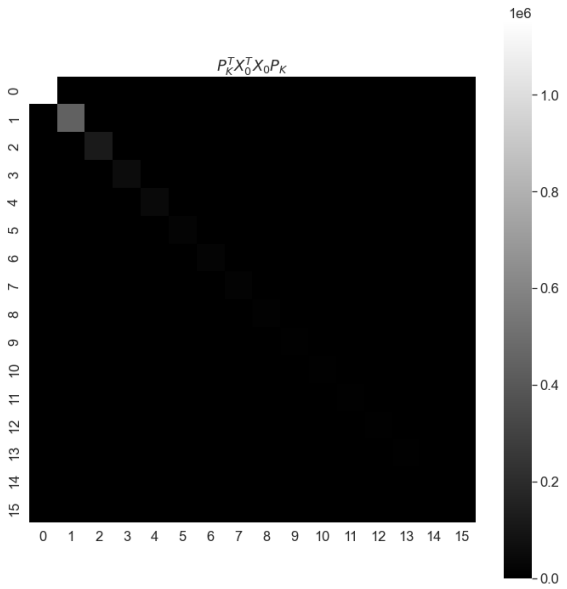

Figure 11: Covariance matrix of $X_0 P_K$ i.e. PCA compressed representation of the mean-centered data.

Results

Let us consider the case study related to IMA using $K = 16$ and performing the dimensionality reduction step with a LAE. We would like to test the analogy between LAE and PCA, by relying on the results of [16] and as explained in the previous paragraph. Therefore, our architecture is a L_2-regularized LAE with one hidden layer with K neurons. We apply three different transformations to the mean-centered data X_0 i.e. $X_0 P_K$, $X_0 W_1$, $X_0 U_K$ where P_K is the loading vectors' matrix obtained by PCA, W_1 is the encoder of the LAE and U_K is the loading vectors' matrix obtained by the LAE. In Figures 11-13, the covariance matrices of these transformed data are represented. As expected according to [16], the covariance matrix is diagonal and with descending diagonal elements for $X_0 P_K$ and $X_0 U_K$; this is not the case for the covariance matrix of $X_0 W_1$.

D Anomaly detection with autoencoders

In this section, figures concerning the results related to our method based on the employment of autoencoders and applied on the asset IMA, are provided. Explanations of these results are inserted in the main text.

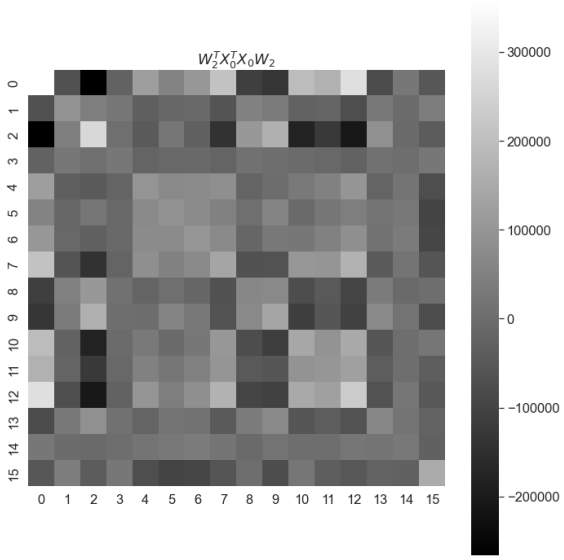

Figure 12: Covariance matrix of $X_0 W_1$ i.e. LAE compressed representation of the mean-centered data.

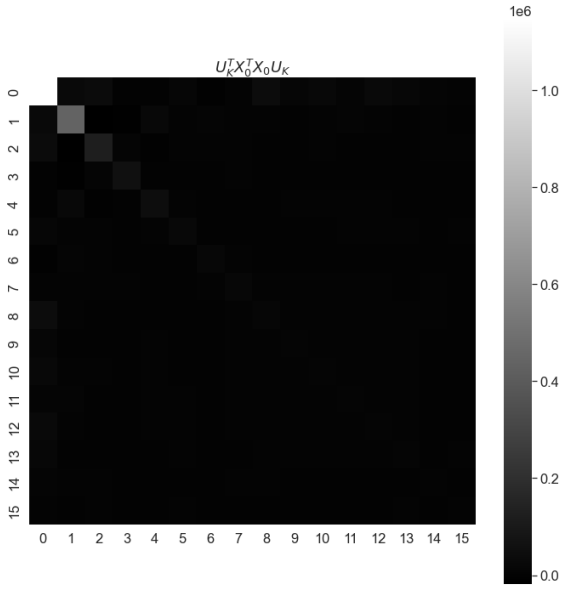

Figure 13: Covariance matrix of $X_0 U_K$ i.e. compressed representation of the mean-centered data, using the loading vectors obtained by the LAE.

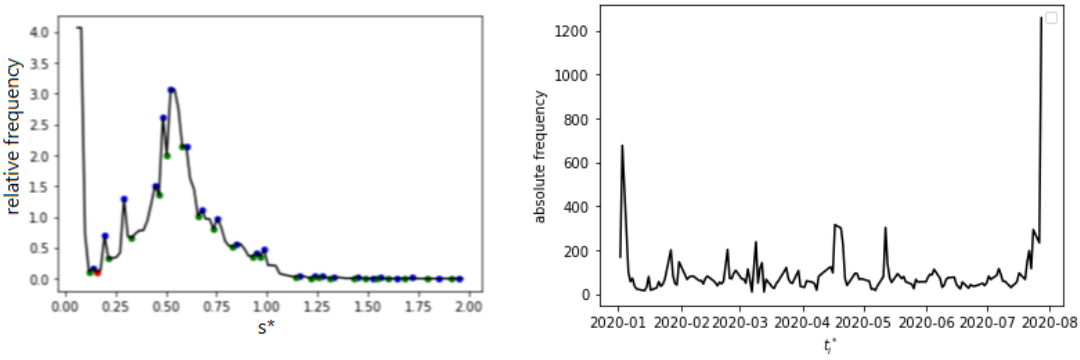

Figure 14: AE-1 on IMA. Left. Histogram of the anomaly scores; blue points are local maxima, green local minima and the red point is $\epsilon_\theta \simeq 0.156$ that is the local minimum after the first peak. Right. Histogram of the times corresponding to the anomaly scores.

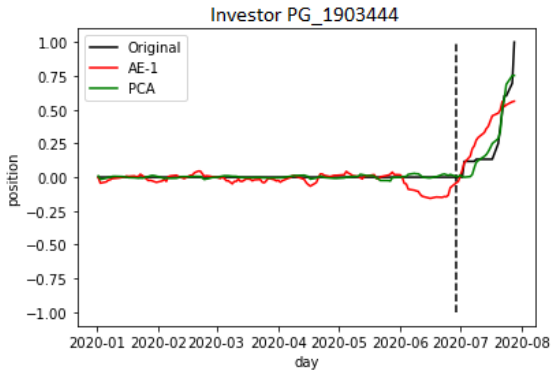

Figure 15: IMA. Position of an investor detected by the method based on AE-1 and not by the method based on PCA. The vertical dashed line is the day corresponding to the beginning of the investigation period i.e. June 29, 2020.

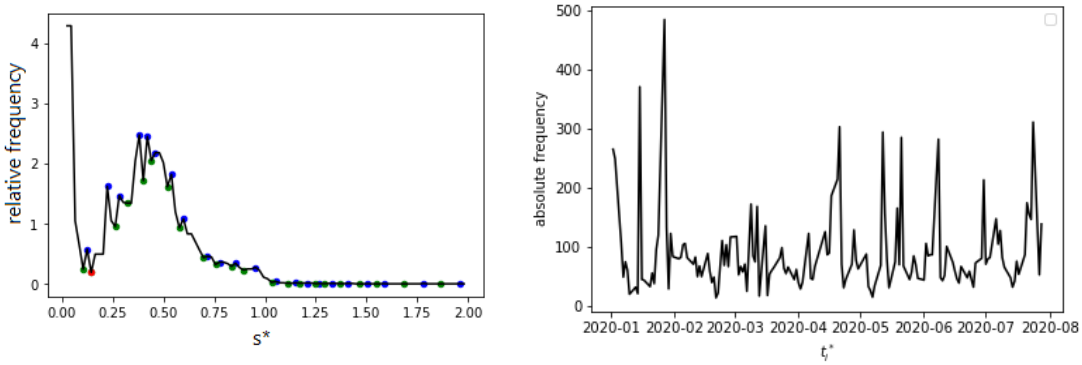

Figure 16: AE-4 on IMA. Left. Histogram of the anomaly scores; blue points are local maxima, green local minima and the red point is $\epsilon_\theta \simeq 0.14$ that is the local minimum after the first peak. Right. Histogram of the times corresponding to the anomaly scores.

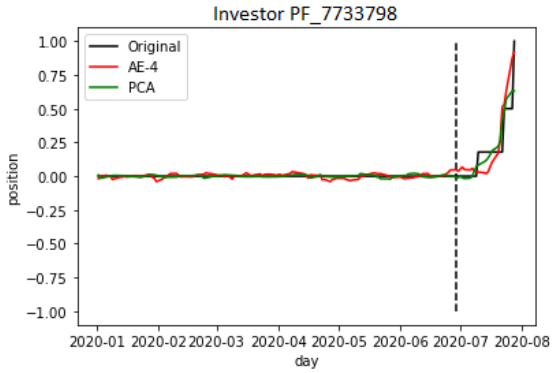

Figure 17: IMA. Investor detected by the method based on AE-4 and not by the method based on PCA. The vertical dashed line is the day corresponding to the beginning of the investigation period i.e. June 29, 2020.

	IMA		UBI	
Investors' type	Households	Firms	Households	Firms
Group 1 latent space	$9,734$	453	$12,829$	306
Group 2 latent space	$2,846$	192	$18,027$	808
A: Investors with $s_i^* \geq \epsilon_\theta$	$11,011$	606	$26,789$	686
B: Investors with $s_i^* < \epsilon_\theta$	$1,569$	39	$4,067$	428

Table 8: IMA and UBI. Composition of the two groups of investors in the latent space and in the anomaly score distribution.

E Households versus firms

As we explained in Section 3, the data set we are provided with, contains information about each investor type. We consider two broad categories of investors: households, that include individual households and joint accounts of several households, and firms, that include investment firms and other legal entities. The goal of this section is to investigate whether it could be more advantageous to run our whole methodology separately for households and firms. This is motivated by the fact that, in principle, these two classes of investors have different behavior.

E.1 PCA using all data: household-firm composition

Let us consider our case study related to the asset IMA. We have $12,580$ households and 645 firms. As expected, the dataset is highly imbalanced towards households who constitute 95.1% of investors. However, their corresponding exchanged volume is less than 10% of the total.

We consider the representation in the latent space obtained by PCA which uses all data. A clustering method (the k-means) for two groups is run and we check whether one of the group is mainly composed of households and the other one of firms. The groups' composition is reported in Table 8. We perform a Fisher test with the null hypothesis that there is not association between groups in the latent space and investor

types. The p-value turns out to be 4.3e-5 so, we need to reject the null hypothesis: there is a relation between investor type and group. Analogously, we test whether a given investor type is over/under-expressed in a group, as in [24]. For this test, the null hypothesis is defined by assuming the random co-occurrence of a given investor type and her belonging to a given group. The hypergeometric distribution is used as a benchmark for randomness. It results that in group 1 (2), the investor type *household* (*firm*) is over-expressed and the investor type *firm* (*household*) is under-expressed.

However, given that our method relies on the computation of the reconstruction errors, we further investigate the household-firm composition in the anomaly score distribution obtained by running PCA using the whole dataset. We split investors in two categories: investors with anomaly score s_i^* greater than or equal to the threshold ϵ_θ (group A) and investors with anomaly score lower than the threshold (group B). The groups' composition is shown in Table 8. Also in this case, the Fisher test points out that there is association between the investor type and the group in the anomaly score distribution. We test the over/under-expression of investor types in the two groups, as in [24]. It results that in group A (B), the investor type *firm* (*household*) is over-expressed and the investor type *household* (*firm*) is under-expressed. We can conclude that, basically, firms are associated with higher values of anomaly score.

The results reported so far are related to the asset IMA, which is illiquid and, as shown in [19], exhibits strong synchronization signals related to the PSE under investigation. If we consider the asset UBI, which is much more liquid than IMA, the findings related to the latent space representation, are analogous. On the other hand, if we focus on the composition household-firm in the anomaly score distribution, we find that in group A (B), the investor type *household* (*firm*) is over-expressed and the investor type *firm* (*household*) is under-expressed. Therefore, contrary to IMA, higher scores are associated with households. This difference between IMA and UBI could be explained by the fact that, as mentioned above, in [19] we showed that investors trading IMA were having strong synchronization signals related to the PSE. Indeed, in the second clustering approach of [19], based on the statistically validated co-occurrence networks and aimed at identifying groups of investors with coordinated suspicious behavior related to the PSE, we identify an highly synchronized cluster made up of more than $2,000$ investors, who are mainly households and with the portfolios managed by the same entity. This issue together with the fact that IMA's data set is small, could have make easier reconstructing households' profiles.

E.2 PCA using households' and firms' data separately

We perform PCA using the datasets made up of the two categories of investors separately. We compare the results between them and with the results obtained by running PCA with the whole dataset.

Let us start to focus on IMA. Given the extremely high fraction of households (95.1%), the difference between PCA results obtained by considering only households and by considering all the investors is negligible. The differences between PCA results obtained by considering only households or only firms are not substantial, especially for the first components, i.e. the components which retain more data variability. This is shown in Figure 18 where the first 6 components and the twentieth component are shown (let us recall that each principal component is a vector of dimensionality T). Moreover, in Figure 19, a comparison between the eigenvalues obtained is provided.

Investors' type	$\lvert A^c \rvert$	$\lvert A^{all} \cap C \rvert$	$\lvert A^c \cap A^{all} \rvert$	$\lvert A^{all}_{500} \cap C \rvert$	$\lvert A^c \cap A^{all}_{500} \rvert$	$\lvert A^c_{500} \cap A^{all}_{500} \rvert$
Households	1,580	1,722	1,580	491	483	432
Firms	95	79	78	9	9	9

Table 9: UBI. A_c with $c = \{\text{households, firms}\}$ is the set of potential insiders obtained by using only the data related to investors of type c. A^{all} is the set of potential insiders obtained by using all data. C is the set of households/firms in the data set. A^{all}_{500} and A^c_{500} are the set of the first 500 ranked potential insiders obtained by using all data or only the data of investors of type c respectively.

Analogous results are obtained for UBI.

E.3 Insider trading detection using households' and firms' data separately

Now, let us tackle our major goal of this section, that is investigating whether it could be more advantageous to run our whole methodology for insider trading detection, separately for households and firms. As we illustrated in the previous subsection, for IMA (UBI), firms (households) are associated with higher values of anomaly score and households (firms) with lower values of anomaly score. For IMA, this issue together with the small number of firms $(606 + 39)$ imply that, if we perform PCA using only the data related to firms, the anomaly score distribution we obtain, does not show the bimodality we want to exploit in order to set the threshold ϵ_θ, which has a major role in the criterion of Equation 3. On the other hand, for UBI, the higher number of firms (in absolute value) allows to preserve the bimodality of the anomaly score distribution obtained by running PCA with only the data related to firms, as shown in Figure 20. Therefore, we focus on UBI for the subsequent analysis.

We apply our whole methodology to identify potential insiders, for households and firms separately: results are shown in Table 9. The method which uses only the data related to firms identifies 16 more anomalous investors than the method which uses the whole dataset. Some of them, like the profile in Figure 21, could be interesting for our scope. However, if we focus on the first 500 ranked potential insiders, there is no difference. On the other hand, if the methodology is run by using only the data related to households, a consistent number of potential insiders is not identified with respect to the method which uses all the data and, among the first 500 ranked potential insiders, 59 investors are not detected. These households are actually extremely suspicious since they are all just active in the investigation period with a net buying position, similarly to the profile in Figure 21.

The difference between the results obtained by using all data and the data only related to households, could be surprising: in subsection E.2, we observed that the difference between the principal components and the eigenvalues obtained in the two cases is negligible. However, it is important to remember that in the criterion of Equation 3, also the times corresponding to the largest reconstruction errors t^*_i have a role and in fact, using the data only related to households, causes a change in the t^*_i histogram.

To conclude, we verified that if PCA using all data is performed, there is a split be-

tween households and firms, both in the latent space representation and in the anomaly score distribution. However, for small assets as IMA, the anomaly score distribution loses its bimodality once PCA is applied using the data related to only firms. Thus, setting the threshold to run our *reconstruction-based* approach, is problematic. This does not occur for more liquid assets as UBI, for which the number of firms, even if it is less than 5% of the total number of investors, is greater. We find that for UBI, performing our method using the data related to the two investors' classes separately leads to an improvement for firms if we go beyond the first 500 ranked anomalous investors. On the other hand, for households, it leads to a deterioration in our results. Investors who have a net suspicious activity related to the PSE, are missed. Therefore, running our whole methodology separately for households and firms does not seem to be consistently more advantageous.

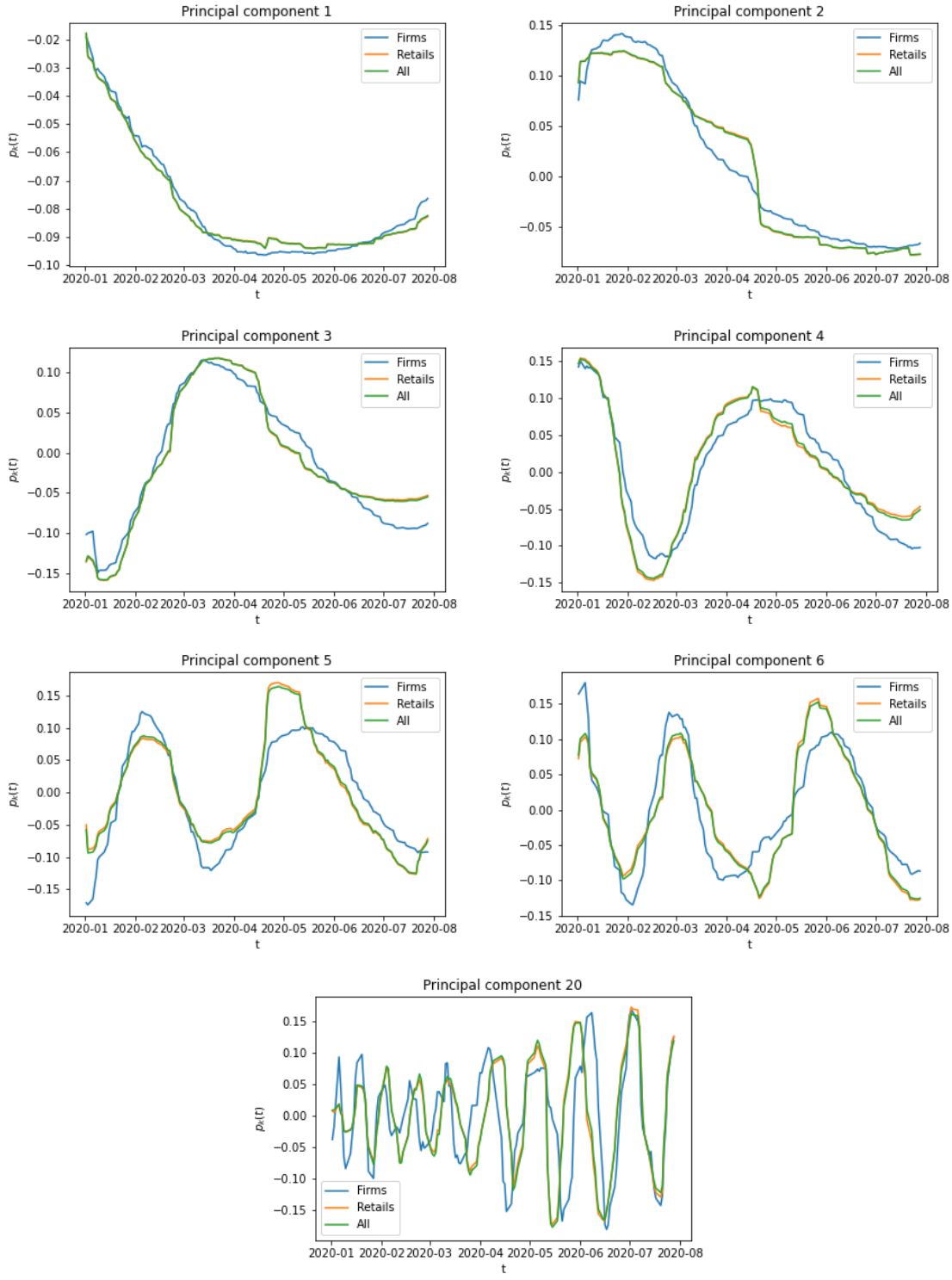

Figure 18: IMA. Representation of some of the principal components obtained by performing PCA using the whole dataset (*All*), the dataset made up of households (*Retails*) and the dataset made up of firms (*Firms*).

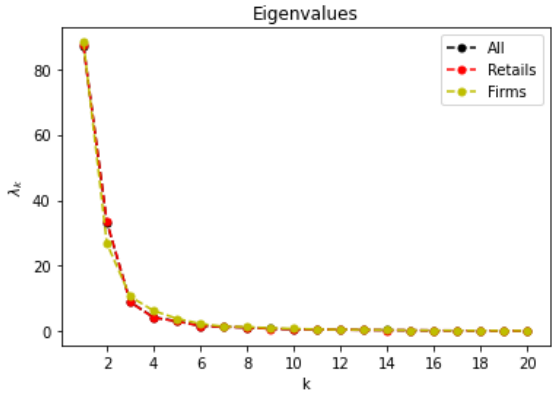

Figure 19: IMA. Eigenvalues obtained by performing PCA using the whole dataset (*All*), the dataset made up of households (*Retails*) and the dataset made up of firms (*Firms*).

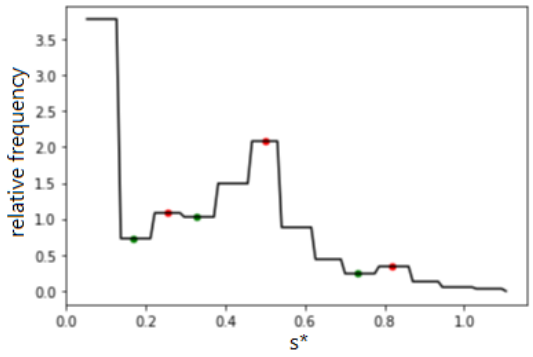

Figure 20: UBI. Histogram of the anomaly scores obtained by using PCA with $K = 16$; red points are local maxima, green local minima.

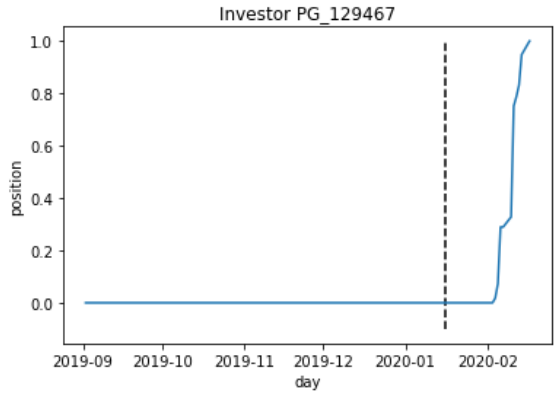

Figure 21: UBI. Profile of an investor identified as anomalous by using only the dataset made up of firms and not by using the whole dataset. The dotted black line is the beginning of the investigation period.

Quaderni FinTech

12 – febbraio 2024
Dimensionality reduction techniques to support insider trading detection
Consob - Scuola Normale Superiore di Pisa

11 – dicembre 2022
A machine learning approach to support decision in insider trading detection
Consob - Scuola Normale Superiore di Pisa

10 – luglio 2022
How Covid mobility restrictions modified the population of investors in Italian stock markets
Consob - Scuola Normale Superiore di Pisa

9 – giugno 2022
L'intelligenza artificiale nell'*asset* e nel *wealth management*
N. Linciano, V. Caivano, D. Costa, P. Soccorso, T.N. Poli, G. Trovatore; in collaborazione con Assogestioni

8 – aprile 2021
La portabilità dei dati in ambito finanziario
A cura di A. Genovese e V. Falce

7 – settembre 2020
Do investors rely on robots?
Evidence from an experimental study
B. Alemanni, A. Angelovski, D. Di Cagno, A. Galliera, N. Linciano, F. Marazzi, P. Soccorso

6 – dicembre 2019
Valore della consulenza finanziaria e *robo advice* nella percezione degli investitori
Evidenze da un'analisi qualitative
M. Caratelli, C. Giannotti, N. Linciano, P. Soccorso

5 – luglio 2019
Marketplace lending
Verso nuove forme di intermediazione finanziaria?
A. Sciarrone Alibrandi, G. Borello, R. Ferretti, F. Lenoci, E. Macchiavello, F. Mattassoglio, F. Panisi

4 – marzo 2019
Financial Data Aggregation e Account Information Services
Questioni regolamentari e profili di business
A. Burchi, S. Mezzacapo, P. Musile Tanzi, V. Troiano

3 – gennaio 2019
La digitalizzazione della consulenza in materia di investimenti finanziari
Gruppo di lavoro Consob, Scuola Superiore Sant'Anna di Pisa, Università Bocconi, Università di Pavia, Università di Roma 'Tor Vergata', Università di Verona