International Journal of Approximate Reasoning 138 (2021) 1–11

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

On the classification of financial data with domain agnostic features

João A. Bastos *, Jorge Caiado

REM/CEMAPRE, ISEG, Lisbon School of Economics and Management, Universidade de Lisboa, Rua do Quelhas 6, 1200-781 Lisboa, Portugal

ARTICLE INFO

Article history:
Received 31 March 2021
Received in revised form 20 July 2021
Accepted 20 July 2021
Available online 4 August 2021

Keywords:
Financial economics
Time series
Clustering
Classification
Machine learning

ABSTRACT

We compare a data-driven domain agnostic set of canonical features with a smaller collection of features that capture well-known stylized facts about financial asset returns. We show that these facts discriminate better different asset types than general-purpose features. Therefore, financial time series analysis is a domain where well-informed expert knowledge may not be disregarded in favor of agnostic representations of the data.

1. Introduction

It is a well-known fact that evaluating the similarity of temporal sequences using measures based on raw values, such as Minkowsky-type or dynamic warping distances, has several shortcomings when used in clustering and classification problems. For instance, because Euclidean distances are invariant to permutations of the coordinates they do not take into account the information contained in the autocorrelation structure of a time series. Parametric (Piccolo [31], Otranto [29], Caiado and Crato [9], D'Urso et al. [14], Cerqueti et al. [10]), nonparametric (Galeano and Peña [18], Caiado et al. [5,6], Alonso and Maharaj [1], Maharaj and D'Urso [25], Maharaj and D'Urso [26], Bastos and Caiado [2], Caiado et al. [8], D'Urso et al. [15]) and semi-parametric [9] methods have been proposed to address some of these limitations. Typically, these approaches use distance measures based on structural characteristics, or "features", derived from time series analysis algorithms. Feature-based representations of time series capture complex time-varying properties that can be used as inputs for statistical and machine learning models. Classification and clustering models based on features generally discriminate better different types of time series, and are less sensitive to missing or noisy data than those based on raw data (see Caiado et al. [7] and Maharaj et al. [27] for recent reviews on this topic).

Of course, nothing prevents feature based representations capturing different aspects of the data from being used as inputs for classification and clustering tasks. For instance, Wang et al. [35] evaluated the performance of features derived from a large variety of times series characteristics, ranging from distributional properties to nonlinearity and chaos. They also proposed a data-driven forward selection algorithm to select the best collection of features for a given task. Fulcher and Jones [17] also suggested an automated method for producing feature-based representations of temporal sequences derived from a large catalog of algorithms. They obtained over 9,000 features from time series properties commonly used

* Corresponding author.
E-mail addresses: jbastos@iseg.ulisboa.pt (J.A. Bastos), jcaiado@iseg.ulisboa.pt (J. Caiado).

https://doi.org/10.1016/j.ijar.2021.07.008
0888-613X/

in several scientific fields. Then, they derived an automated method, also based on forward selection, to select from the available pool of features those that are most appropriate to a specific problem, thereby eliminating the requirement of domain knowledge. Lubba et al. [24] proposed a somewhat different data-driven feature selection approach. Starting from the large pool of features analyzed in Fulcher and Jones [17], they selected a collection of features with strong individual and joint classification performance across a diverse set of problems, and that are minimally redundant. They obtained a final set of 22 "canonical features", named *catch22*: Canonical Time-series Characteristics.

Naturally, these data-driven methods for automatically selecting the best features for a specific task are valuable across a wide range of scientific fields, such as medicine, economics, physics, climate science, and many more. They certainly save valuable time spent on the manual and sometimes inefficient effort spent on searching for optimal feature sets. Still, we may ask if there are domains where well-informed expert knowledge may not be completely disregarded in favor of automated representations of the data. For instance, financial time series are known to exhibit a collection of rather consistent stylized facts, such as short-term dependence, conditional heteroskedasticity, long memory, and asymmetric reactions to shocks [12,34]. Also, the distribution of returns usually has fat tails consistent with a Pareto-Lévy stable distribution, which means that extreme events occur more often than would be predicted from a normal distribution [28]. In this paper, we formulate a small collection of 10 features that capture these well-known empirical facts. Using both supervised and unsupervised techniques, we show that this set of features discriminates better different asset types than the 22 canonical features of Lubba et al. [24].

Our empirical results are based on two datasets. The first dataset consists of a collection of international equity market indices constructed and maintained by Morgan Stanley Capital International (MSCI). Among these markets, 23 are classified by MSCI as "developed", and 23 are classified as "emerging". It is well known that emerging markets have higher mean returns than their developed counterparts, reflecting greater investment opportunities. On the other hand, emerging markets are more volatile, and have significant barriers to free capital flows [22]. Furthermore, asset returns in emerging markets are typically more predictable [2]. Our objective is to evaluate how well our stylized facts discriminate emerging markets from developed markets, and compare their performance with that of the canonical features.

In a second illustration, we study how well the two feature collections discriminate a set of large capitalization stock indices from a set of foreign exchange rates. Stock prices are determined by expected future cash flows and discount factors. In asset pricing theory, they are usually modeled by a geometric Brownian motion, possibly including generalized conditional heteroskedasticity. On the other hand, foreign exchange rates are determined by expected interest rate differentials between countries and currency risk premia [11]. In turn, interest rates are usually modeled with mean-reverting Vasicek or Cox-Ingersoll-Ross processes. Exchange rate returns exhibit some statistical regularities such as fat tails and volatility clustering [see, e.g., 33].

The paper is organized as follows. The next section provides a description of the 10 features based on stylized facts, as well as a brief description of the canonical features. The third section provides an overview of the models used for discriminating asset returns. Section 4 shows the results of the empirical analysis. Some concluding remarks are given in the final section.

2. Time-series features

2.1. Features based on stylized factors

2.1.1. Distributional properties

Standard univariate descriptive statistics of asset returns include the mean, the standard deviation, the skewness, and the excess kurtosis. The mean is a measure of the compounded return of the asset. The standard deviation, or unconditional volatility, measures the dispersion around the mean return and is a proxy for idiosyncratic asset risk. All else being equal, these quantities are positively correlated; a rational investor will only invest in a riskier asset if the expected return is higher.

The skewness measures the asymmetry of the distribution of returns. Investors are attracted by positive skewness (or long tail on right side) of the return distribution because it means a greater chance of extremely positive outcomes [23].

Finally, the kurtosis measures the "fatness" of the tails of the return distribution. If the data are normally distributed, the skewness and excess kurtosis should be close to zero. A distribution with positive excess kurtosis has heavy tails, whereas a distribution with negative excess kurtosis has short tails. In many empirical studies, the distribution of log returns usually has fatter tails than the normal distribution, which means that extreme events occur more often than would be predicted from a normal distribution. For instance, it is well known that emerging market returns depart from the normal distribution [22,3].

2.1.2. Short-term dependence

The short-term serial dependence describes the low-order correlation structure of a time-series. We examine the presence of short-term linear dependence in financial data using the autocorrelation of the returns. These autocorrelations are typically zero or very close to zero, in consonance with the random walk or martingale hypothesis. However, some returns often do exhibit serial correlation [2]. The presence of nonlinear dependence, and possible autoregressive heteroskedasticity effects is measured by the autocorrelations of squared or absolute returns. These are generally positive and significant for a

substantial number of lags. This stylized fact is known as volatility clustering, meaning that large (small) volatility is often followed by large (small) volatility. We characterize the low-order correlation structure in returns and squared returns of a time-series by the value of the Ljung-Box statistic,

$$Q(m) = T(T+2)\sum_{l=1}^{m}\frac{\hat{\rho}_l^2}{T-l}, \tag{1}$$

where $\hat{\rho}_l$ is the sample autocorrelation of returns, or squared returns, at lag l, and T is the length of the series. For better power properties a value $m \approx \ln(T)$ is typically chosen [34].

2.1.3. Long-memory

Some financial time-series exhibit long-memory or long-range dependence behavior [e.g., 20]. Of particular interest in financial economics is the long memory behavior of absolute stock returns and squared returns. Many empirical studies have noticed very slowly decaying autocorrelations for absolute (or squared) returns. As noted by Ding et al. [13] and Granger and Ding [21], the evidence of long memory is stronger for absolute returns than for squared returns. Using price series from various stock markets and commodity prices, Granger and Ding [21] showed that the absolute returns have the properties of an $I(d)$ process with memory parameter d around 0.45.

A stationary process exhibits long-memory with memory parameter d if its spectral density function $f(\omega)$ satisfies:

$$f(\omega) \sim C\omega^{-2d} \text{ as } \omega \to +\infty, \tag{2}$$

where C is a positive finite constant, and ω denotes the frequency. When $d < 0.5$ its autocorrelation function ρ_k decays at a hyperbolic rate,

$$\rho_k \sim C_\rho k^{2d-1}, \tag{3}$$

where C_ρ is a constant with respect to k. If $0 < d < 0.5$ the process has long memory. If $d = 0.5$ the process has no memory. If $-0.5 < d < 0$, the process has intermediate memory. For $d > 0.5$, the process is no longer covariance stationary.

2.1.4. Asymmetric volatility

An important stylized fact in finance is the conditional variance or volatility of asset returns. Volatility is a measure of the intensity of unpredictable changes in asset returns and it is commonly time varying dependent. The volatility clustering often seen in financial markets has increased the interest of researchers in applying good models that describe the historical pattern of asset volatility, and possibly use it to forecast future volatility. The univariate volatility models available in the literature include the autoregressive conditional heteroskedasticity (ARCH) model of Engle [16], the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev [4], and its various extensions.

Many time-varying volatility models have been proposed to capture the so-called "asymmetric volatility" effect, where volatility tends to be higher after a negative return shock than a positive shock of the same magnitude. An univariate volatility model commonly used to allow for asymmetric shocks to volatility is the threshold GARCH model [36],

$$\varepsilon_t = \sigma_t z_t, \text{ with } \sigma_t^2 = \alpha_0 + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \gamma d_{t-1}\varepsilon_{t-1}^2, \tag{4}$$

where $\{z_t\}$ is a sequence of independent and identically distributed random variables with zero mean and unit variance, $d_t = 1$ if ε_t is negative, and 0 otherwise. The volatility may either diminish ($\gamma < 0$), rise ($\gamma > 0$), or not be affected ($\gamma = 0$) by negative shocks or "bad news" ($\varepsilon_{t-1} < 0$). Good news has an impact of α while bad news have an impact of $\alpha + \gamma$.

2.2. Canonical features

The canonical features of Lubba et al. [24] are a collection of features that exhibits strong classification performance across a diverse set of problems, and are minimally redundant. The starting point of their analysis is a pool of 4791 "highly comparative time-series analysis" candidate features from Fulcher and Jones [17]. An initial filtering of these features is performed to find those that individually show good discriminatory power across a diverse range of time series. Each feature is scored according to its ability to distinguish the labeled classes in 93 classification tasks. After finding a set of features with good performance across all tasks, the authors find a subset of these features with minimal redundancy using hierarchical clustering with complete linkage on the Pearson correlation distance. This procedure results in a collection of 22 canonical features with a classification performance of about 90% of that given by the original set of features. Furthermore, they are more computationally efficient than the forward selection approach of Fulcher and Jones [17].

These features span a diverse range of time series characteristics: distributional properties, simple temporal statistics, linear autocorrelation, nonlinear autocorrelation, successive differences, fluctuation analysis, and "others". The complete list of features and their descriptions can be found in Table 1 of Lubba et al. [24]. In terms of classification performance, they compare well to other small feature sets proposed in the literature.

3. Models

3.1. Unsupervised classification

Unsupervised classification, or clustering, does not incorporate any information about the labels to train the algorithm. If the actual labels are known, they may, of course, be used for evaluating the quality of the clustering outcome. Our analysis is based on k-means clustering; possibly the most simple non-hierarchical clustering algorithm. Because we know beforehand that there are two classes in our datasets, we imposed a prior of $k = 2$ clusters. The dissimilarity between time-series was measured by the Euclidean distance between features. Because the k-means algorithm may be stuck in a local optimum in terms of the within sum of squares, we ran the algorithm 10,000 times with different centroid seeds and chose one of the several seeds that gave the solution with best within sum of squares. All features were z-score standardized in order to have zero mean and unit variance.

3.2. Supervised classification

Many supervised models have been suggested for classifying time series. Caiado et al. [7] and Maharaj et al. [27] provide recent reviews on this topic. Supervised classification uses information about the labels to train the classifier. Let $Y \in \{0, 1\}$ denote a Bernoulli random variable that codifies the two classes in the following binary classification problems, and $\boldsymbol{X}$ denote a vector of features. We consider the following classifiers:

- Logistic regression. This is a parametric binary choice model with a logit link function:

$$\Pr(Y = 1|\boldsymbol{X}) = \left(1 + \exp(-\boldsymbol{X}^T\boldsymbol{\beta})\right)^{-1}, \tag{5}$$

  where $\boldsymbol{\beta}$ is a vector of coefficients obtained by minimization of the regularized cost function

$$J(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} Y_i \log p(\boldsymbol{X}_i) + (1 - Y_i)\log(1 - p(\boldsymbol{X}_i)) + \frac{C}{2n}\sum_{j=1}^{p}\beta_j^2. \tag{6}$$

  The first term is proportional to the negative of the Bernoulli log-likelihood function, whereas the second term is a L2 shrinkage penalty, with cost parameter C, that prevents overfitting the training data. A cutoff value of 0.5 for $\Pr(Y = 1|\boldsymbol{X})$ is used to classify observations.
- Decision trees. These are non-parametric models represented by a sequence of if-then-else tests on the features. Starting from the "root node", an observation follows one of the tree branches according to the outcome of these tests on its features, eventually ending its path in a terminal or "leaf" node. This observation is classified according to the most common class in the training observations that formed that leaf.
- Random Forest. This is a "committee" of decision trees, in which each tree is trained using bootstrap samples of the original data. Furthermore, in the tree growing procedure, random subsets of the available features are considered when selecting the optimal way to split the nodes. An observation is classified according to the most frequent classification of each tree in the committee.

4. Empirical analysis

4.1. Case I: developed vs emerging markets

First, we revisit the stock market data used in [2]. It consists of free float-adjusted market capitalization equity indices constructed and maintained by Morgan Stanley Capital International (MSCI). In order to avoid effects due to exchange rates, all indices are specified in local currency. Securities included in the indices are subject to minimum requirements in terms of market capitalization, free-float, liquidity, availability to foreign investors, and length of trading. The data includes 23 markets that were classified by MSCI as "developed", and 23 markets that were classified as "emerging". A market is classified as developed if i) the country GNI per capita is 25% above the World Bank high income threshold for 3 consecutive years, ii) there is a minimum number of companies satisfying minimum size and liquidity requirements, and iii) there is a very high openness to foreign ownership, ease of capital inflows/outflows, efficiency of the operational framework, and stability of the institutional framework.

Table 1 lists the markets in the dataset according to their MSCI classification.[1] The data cover a period from January 1995 to December 2009, corresponding to 3,914 daily observations.

For each market, we calculated the 10 features corresponding to the stylized facts of asset prices discussed in section 2.1. Table 2 lists feature names and their descriptions. The estimator of the long-memory parameter (d) is based on the

[1] Israel's stock market is currently classified by MSCI as developed. However, we adhere to the original classifications used in [2].

Table 1
List of global equity markets according to Morgan Stanley Capital International (MSCI) classification.

Developed markets	Emerging markets
Australia (AUST), Austria (AUS), Belgium (BEL), Canada (CAN), Denmark (DEN), Finland (FIN), France (FRA), Germany (GER), Greece (GRE), Hong Kong (HK), Ireland (IRE), Italy (ITA), Japan (JAP), Netherlands (NET), New Zealand (NZ), Norway (NOR), Portugal (POR), Singapore (SING), Spain (SPA), Sweden (SWE), Switzerland (SWI), United Kingdom (UK), United States (US)	Argentina (ARG), Brazil (BRA), Chile (CHI), China (CHI), Czech Republic (CR), Colombia (COL), Egypt (EGY), Hungary (HUN), India (IND), Indonesia (INDO), Israel (ISR), Korea (KOR), Malaysia (MAL), Mexico (MEX), Morocco (MOR), Peru (PER), Philippines (PHI), Poland (POL), Russia (RUS), South Africa (SA), Taiwan (TAI), Thailand (THA), Turkey (TUR)

Table 2
List of features capturing the stylized facts about asset prices.

Name	Description
mean	Mean value of returns
stdev	Standard deviations of returns
skew	Skewness of returns
kurt	Kurtosis of returns
qstat	Ljung-Box statistic for short-term dependence in the returns
qstat2	Ljung-Box statistic for short-term dependence in the squared returns
arch	ARCH parameter α in Equation (4)
garch	GARCH parameter β in Equation (4)
lever	Parameter for asymmetric shocks γ in Equation (4)
d	Long-memory parameter in Equation (2)

Table 3
Model out-of-sample accuracy for discriminating developed markets from emerging markets when the inputs are: i) canonical features; ii) features derived from stylized facts.

Model	Canonical features	Stylized facts
K-means	71.7%	91.3%
Logistic regression	73.9%	76.1%
Decision tree	76.1%	82.6%
Random forest	78.3%	91.3%

frequency domain Gaussian approach of [32]. The estimated parameters of the threshold GARCH(1,1) model (arch, garch and lever) assume t-student error innovations.

Fig. 1 shows box-plots for the distribution of these features for developed and emerging markets. As expected, emerging markets offer investors higher average returns at the cost of higher volatility, when measured by the standard deviation of the returns. Emerging markets also have higher short-term dependence in the returns, possibly reflecting a lower informational efficiency. In terms of reaction to negative shocks and persistence of long memory there are no significant differences, on average, between developed and emerging markets. However, emerging markets have a higher dispersion of these features possibly reflecting their lower integration due to greater barriers to free capital inflows and outflows, and to foreign ownership.

Let y denote the true labels of the data and $\hat{y}$ denote the predicted labels. We evaluate the classification performance of the methods using the proportion of misclassified observations, or accuracy:

$$\text{acc} = \frac{1}{n}\sum_{i=1}^{n} \text{I}(y_i = \hat{y}_i), \tag{7}$$

where n is the number of time series in the data, and $\text{I}(\cdot)$ is an indicator function that is equal to 1 if the argument evaluates to true, and 0 otherwise. Because the number of time series is rather small, we evaluate out-of-sample accuracy using leave-one-out cross-validation. The analysis was performed in Python 3 using `scikit-learn` and its default model hyper-parameters. Intensive hyper-parameter tuning in small datasets may lead to information leakage to the validation data.

Table 3 shows the model out-of-sample accuracies for discriminating developed markets from emerging markets when the inputs are the canonical features and the features that we derived from domain knowledge of the data (stylized facts). Despite k-means being an unsupervised technique, knowing the actual labels allows us to calculate the accuracy from the observations that belong to the "wrong" cluster. The styled facts discriminate better than the canonical features-based methods, with emphasis on k-means and random forest.

Fig. 2 shows the first two principal components of the k-means solution for the MSCI markets. The plot on the left corresponds to the canonical features, while the plot on the right corresponds to the features from stylized facts. Comparing

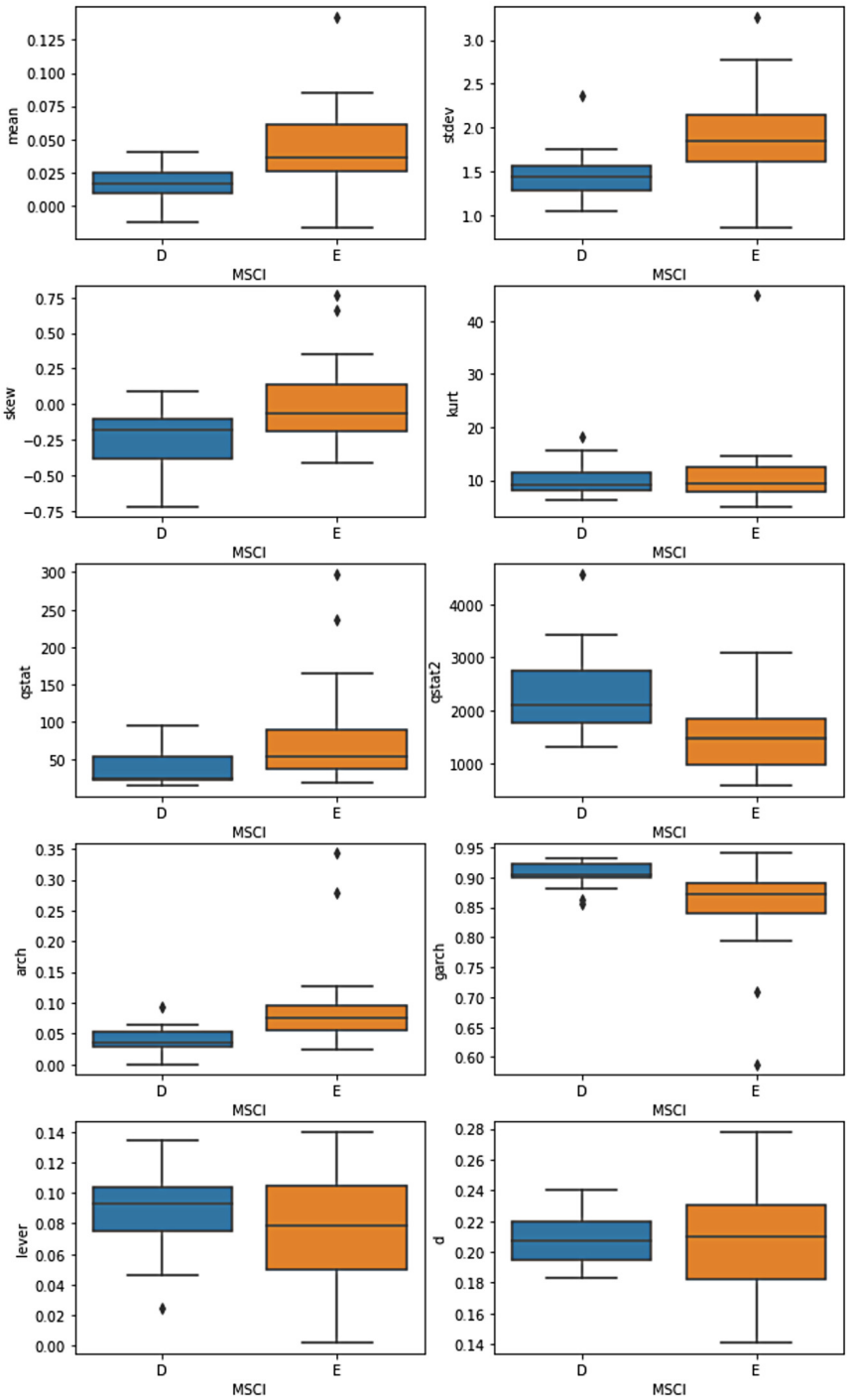

Fig. 1. Box-plots for the features representing stylized facts about asset prices, for developed stock markets (D) and emerging markets (E).

these solutions with the MSCI market classifications in Table 1, we can see that the features from stylized facts produce fewer misclassifications.

Fig. 3 shows how important the features derived from stylized facts were at discriminating developed markets from their emerging counterparts. The importance of a given feature is measured as the cumulative reduction in the Gini coefficient of the data that it has achieved across all nodes and trees in the random forest committee. We normalized the features such that the most important feature had a score of 100.

We identified five features with strong importance for the random forest model (the mean and the standard deviation of returns, the Ljung-Box statistic for squared returns, and the autoregressive conditional heteroskedasticity parameters). The

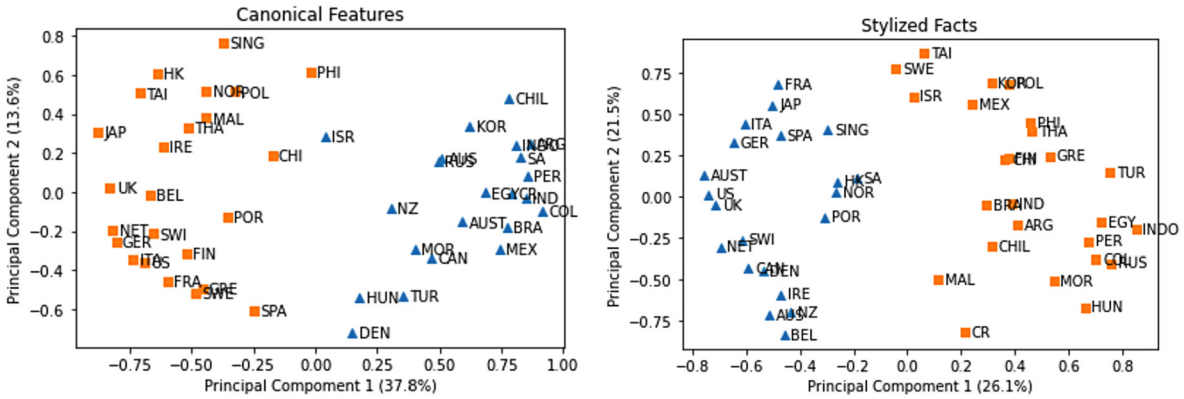

Fig. 2. Projection in the first two principal components of the K-means solution for the MSCI markets. The plot on the left corresponds to the canonical features, while the plot on the right corresponds to the stylized facts.

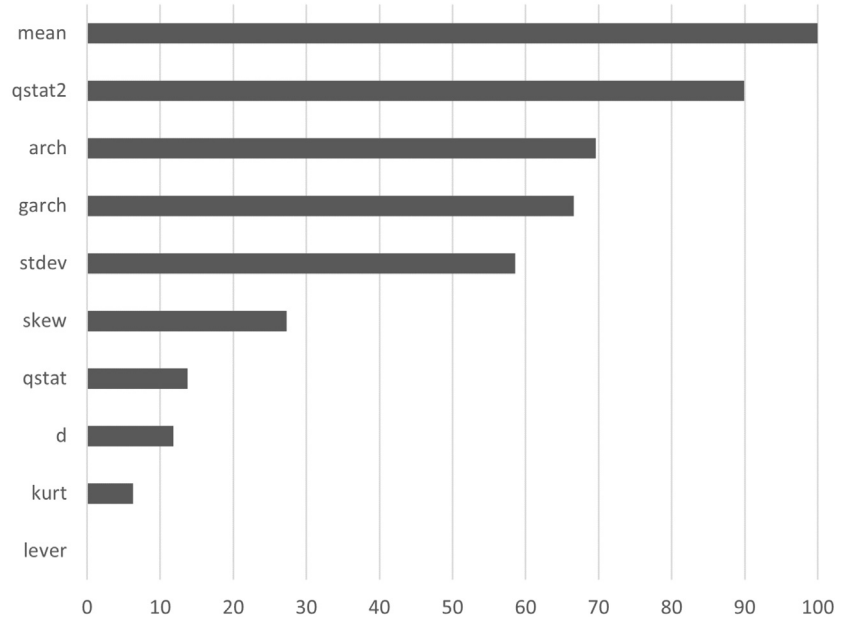

Fig. 3. Feature importance for discriminating developed markets from emerging markets. The scores were normalized such that the most important feature has a score of 100.

average log return (mean) is positively correlated with the unconditional volatility (stdev), which is simply a materialization of the risk-return trade-off. The presence of short-term non-linear dependence (qstat2) and conditional heteroskedasticity (garch) effects is more salient in developed markets. This can be explained by the fact that the volatility in emerging markets is primarily driven by local factors [3]. The asymmetric shock effect (lever) is negligible in the discrimination between developed and emerging markets.

4.2. Case II: stock indices vs foreign exchange rates

We have collected a dataset with stock indices and foreign exchange rates reported by Yahoo Finance. The time-series consist of daily close prices covering the period from January, 2005, to December, 2019. Table 4 lists the Yahoo ticker symbols of the stock indices, and the currency pairs used in the analysis.

Fig. 4 shows box-plots for the distribution of these features for stock indices (I) and currency pairs (C). In the period covered by the data, stocks had higher average returns than foreign exchange rates, in accordance with the mean-reverting nature of exchange rate dynamics. Stock indices exhibit more non-linear dependency with a greater interquartile range than exchange rates. In contrast, the midspread of the non-linear dependence statistic is much more noticeable in exchange rates. In general, there are substantially more outliers in the exchange rate series than in the stock indices.

Table 5 shows the model out-of-sample accuracies for discriminating stocks indices from foreign exchange rates when the inputs are the canonical features and the features that we derived from domain knowledge of the data. All models suggest

Table 4
List of stock indices (ticker symbols), and currency exchange rates used in the analysis.

Stock indices	Currency exchange rates
AEX, AORD, ATHEX, ATX, AXJO, BFX, BSESN, BVSP, FCHI, FTSE, GDAXI, GSPTSE, HSI, IBEX, IPSA, IXIC, JKSE, JSE, KLSE, KS11, MERV, MIB, MXX, N100, N225, NSEI, NYA, NZ50, OMX, OMXH25, PSEI, RUT, SP500, SSE, SSMI, STI, STOXX50, SZ, TA125, TWII, XAX	AUD/USD, USD/CNY, EUR/CAD, EUR/CHF, EUR/GBP, EUR/HUF, EUR/JPY, EUR/SEK, EUR/USD, GBP/JPY, GBP/USD, USD/HKD, USD/IDR, USD/INR, USD/JPY, USD/MXN, USD/MYR, NZD/USD, USD/PHP, USD/RUB, USD/SGD, USD/THB, USD/ZAR

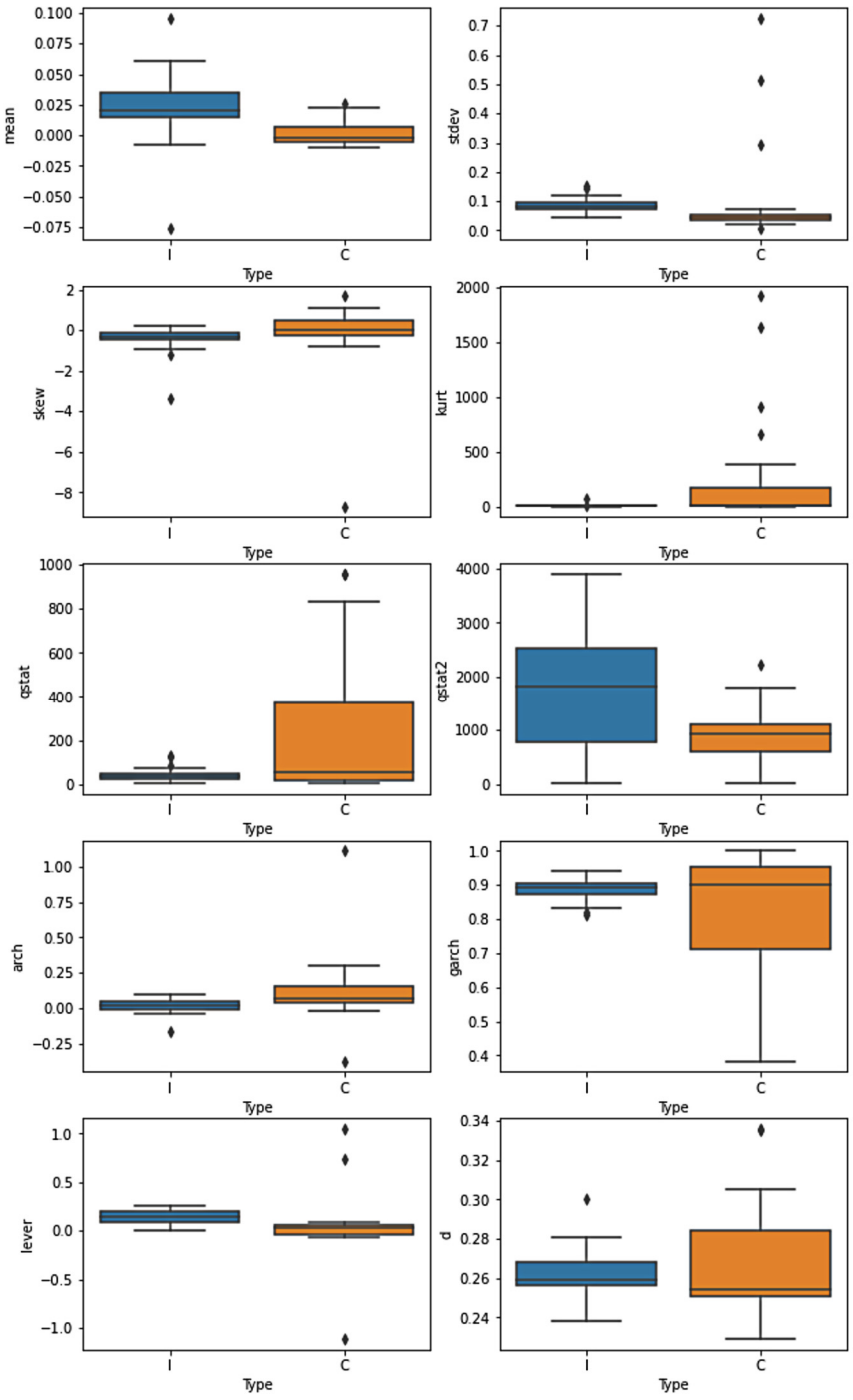

Fig. 4. Box-plots for the features representing stylized facts of asset prices, for stock indices (I) and foreign exchange rates (C).

Table 5
Model out-of-sample accuracy for discriminating stocks indices from foreign exchange rates when the inputs are: i) canonical features; ii) features derived from stylized facts.

Model	Canonical features	Stylized facts
K-means	51.6%	71.9%
Logistic regression	71.9%	95.3%
Decision tree	81.2%	89.1%
Random forest	85.9%	96.9%

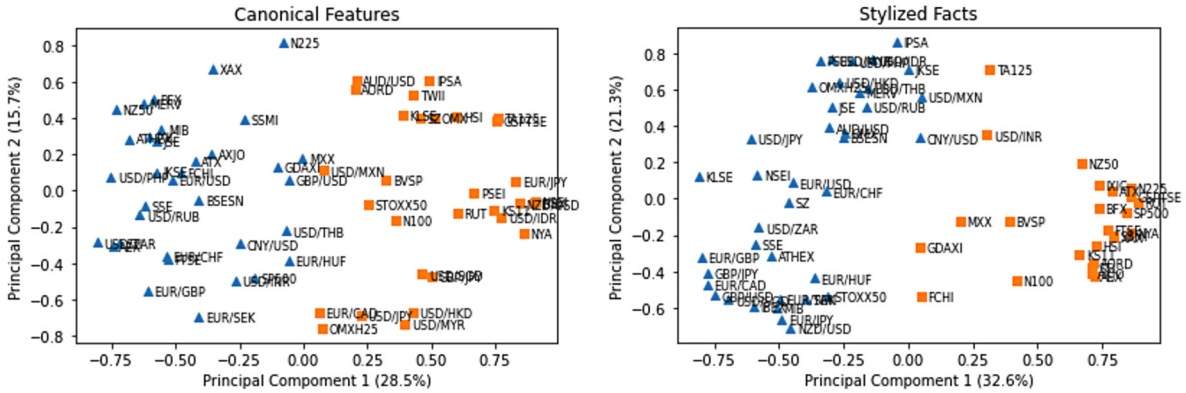

Fig. 5. Projection in the first two principal components of the K-means solution for discriminating stocks indices from foreign exchange rates. The plot on the left corresponds to the canonical features, while the plot on the right corresponds to the stylized facts.

that the stylized facts are better at discriminating stocks from foreign exchange rates. Surprisingly, using the canonical variables as inputs to K-means gives a classification accuracy just slightly above that of a random classifier.

Fig. 5 shows the first two principal components of the K-means solution for discriminating stocks indices from foreign exchange rates. The plot on the left corresponds to the canonical features, while the plot on the right corresponds to the features from stylized facts. Again, we can confirm that the features from stylized facts produce fewer misclassifications. In particular, the features-based map derived from stylized facts indicates that all currencies except one (USD/IMR) are "correctly" clustered together in the same cluster. In contrast, the K-means solution based on the canonical features is not able to discriminate between stock indices and exchange rates.

Fig. 6 shows how important the features derived from stylized facts were at discriminating stock indices from foreign exchange rates. The most important features were the non-linear dependence, the kurtosis of returns and the average log return. The least important feature was the coefficient of asymmetry of the distribution. The empirical results indicate that currencies exhibit higher excess kurtosis than stock indices (273.4 and 12.8, respectively). In contrast, the average rate of return for stock indices (0.0023) is higher than that for exchange rates (0.002) but the difference between their standard deviations (or unconditional volatilities) is negligible.

5. Conclusions

This study compares data-driven and knowledge-driven sets of features for clustering and classification of financial time series. We use the set of 22 canonical features selected by Lubba et al. [24] to capture the dynamic properties of time series across diverse applications. These features include time series characteristics such as distribution, linear and non-linear autocorrelation, successive differences, temporal statistics, and fluctuation analysis properties. The second feature set consists of 10 features that capture the well-known stylized facts of financial returns including distributional properties, short-term dependence, long-memory, and autoregressive conditional volatilities.

Both sets of time series features (canonical and stylized facts) were used as input variables for supervised and unsupervised learning of two financial application examples. The first one is concerned with discrimination between emerging and developed equity indices constructed by Morgan Stanley Capital International (MSCI). The second one uses world's majors stock indices and currencies from 2005 to 2019. In both studies, the feature-based learning methods extracted from stylized facts performed better than canonical feature-based learning methods to distinguish and separate between the two distinct financial asset classes. In addition, our smaller knowledge-based feature set provides a more meaningful economical interpretation. In the case of the logistic regression, the lower performance of the canonical features may be due to the inclusion of irrelevant variables that use the precious degrees of freedom in our small samples without contributing to the discrimination of the two classes. On the other hand, decision trees and random forests are rather robust to variables with low discrimination power since these will not be selected in the node splitting process. Therefore, the canonical features perform worse because they miss important variables included in our set of 10 features.

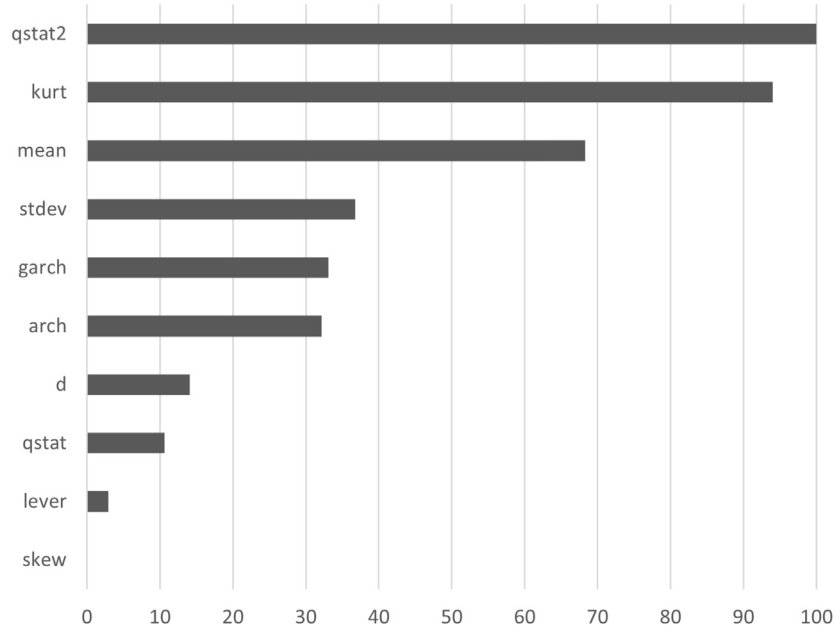

Fig. 6. Feature importance for discriminating stock indices from foreign exchange rates. The scores were normalized such that the most important feature has a score of 100.

Future research on clustering and classification of financial time series should explore other methods for making groups, associating each financial price or return series with a vector of parameters or features (autoregressive estimates, autocorrelations, spectrum densities, volatilities, distributions, or other features extracted from time series). Assuming that parameters or features are generated by a mixture of normal distributions, the objective is to find the number of distributions and the probability of each series coming from each distribution. This means we will have to define the vector of parameters or features and to have an approach for fitting mixtures of normal variables. The projection pursuit approach (see Pena and Prieto [30] and Galeano et al. [19]) of projecting the points onto certain directions according to some optimal criterion would be a good way to find these groups.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia [grant number UIDB/05069/2020]. We would like to thank the comments of two reviewers.

References

[1] A.M. Alonso, E.A. Maharaj, Comparison of time series using subsampling, Comput. Stat. Data Anal. 50 (2006) 2589–2599.
[2] J.A. Bastos, J. Caiado, Clustering financial time series with variance ratio statistics, Quant. Finance 14 (2014) 2121–2133.
[3] G. Bekaert, C. Harvey, Emerging equity market volatility, J. Financ. Econ. 43 (1997) 29–77.
[4] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econom. 31 (1986) 307–327.
[5] J. Caiado, N. Crato, D. Peña, A periodogram-based metric for time series classification, Comput. Stat. Data Anal. 50 (2006) 2668–2684.
[6] J. Caiado, N. Crato, D. Peña, Comparison of time series with unequal length in the frequency domain, Commun. Stat., Simul. Comput. 38 (2009) 527–540.
[7] J. Caiado, E.A. Maharaj, P. D'Urso, Time series clustering, in: C. Henning, M. Meila, F. Murtagh, R. Rocci (Eds.), Handbook of Cluster Analysis, CRC Press, Taylor & Francis Group, 2015, pp. 241–263.
[8] J. Caiado, N. Crato, P. Poncela, A fragmented-periodogram approach for clustering big data time series, Adv. Data Anal. Classif. 14 (2020) 117–146.
[9] J. Caiado, N. Crato, Identifying common dynamic features in stock returns, Quant. Finance 10 (2010) 797–807.
[10] R. Cerqueti, M. Giacalone, R. Mattera, Model-based fuzzy time series clustering of conditional higher moments, Int. J. Approx. Reason. 134 (2021) 34–52.
[11] R. Clarida, J. Gali, Sources of real exchange-rate fluctuations: how important are nominal shocks?, Carnegie-Rochester Conf. Ser. Public Policy 41 (1994) 1–56.
[12] R. Cont, Empirical properties of asset returns: stylized facts and statistical issues, Quant. Finance 1 (2001) 223–236.
[13] Z. Ding, C.W.J. Granger, R.F. Engle, A long memory property of stock market returns and a new model, J. Empir. Finance 1 (1993) 83–106.
[14] P. D'Urso, L. De Giovanni, R. Massari, GARCH-based robust clustering of time series, Fuzzy Sets Syst. 305 (2016) 1–28.

[15] P. D'Urso, L.A. Garcia-Escudero, L. De Giovanni, V. Vitale, A. Mayo-Iscar, Robust fuzzy clustering of time series based on B-splines, Int. J. Approx. Reason. 136 (2021) 223–246.
[16] R.F. Engle, Autoregressive conditional hetroskedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50 (1982) 987–1008.
[17] B.D. Fulcher, N.S. Jones, Highly comparative feature-based time-series classification, IEEE Trans. Knowl. Data Eng. 26 (2014) 3026–3037.
[18] P. Galeano, D. Peña, Multivariate analysis in vector time series, Resen. Inst. Mat. Estat. Univ. Sao Paulo 4 (2000) 383–404.
[19] P. Galeano, D. Peña, R.S. Tsay, Outlier detection in multivariate time series by projection pursuit, J. Am. Stat. Assoc. 101 (2006) 654–669.
[20] S. Granero, J.T. Segovia, J.G. Perez, Some comments on Hurst exponent and the long memory processes on capital markets, Phys. A, Stat. Mech. Appl. 387 (22) (2008) 5543–5551.
[21] C.W.J. Granger, Z. Ding, Varieties of long-memory models, J. Econom. 73 (1996) 61–77.
[22] C. Harvey, Predictable risk and returns in emerging markets, Rev. Financ. Stud. 8 (1995) 773–816.
[23] A. Kraus, R. Litzenberger, Skewness preference and the valuation of risky assets, J. Finance 21 (1976) 1085–1094.
[24] C.H. Lubba, S.S. Sethi, P. Knaute, S.R. Schultz, B.D. Fulcher, N.S. Jones, catch22: CAnonical Time-series CHaracteristics, Data Min. Knowl. Discov. 33 (2019) 1821–1852.
[25] E.A. Maharaj, P. D'Urso, A coherence-based approach for the pattern recognition of time series, Phys. A, Stat. Mech. Appl. 389 (17) (2010) 3516–3537.
[26] E.A. Maharaj, P. D'Urso, Fuzzy clustering of time series in the frequency domain, Inf. Sci. 181 (2011) 1187–1211.
[27] E.A. Maharaj, P. D'Urso, J. Caiado, Time Series Classification and Clustering, CRC Press, Taylor & Francis Group, United States, 2019.
[28] B. Mandelbrot, The variation of certain speculative prices, J. Bus. 36 (1963) 394–419.
[29] E. Otranto, Clustering heteroskedastic time series by model-based procedures, Comput. Stat. Data Anal. 52 (2008) 4685–4698.
[30] D. Peña, F.J. Prieto, Cluster identification using projections, J. Am. Stat. Assoc. 96 (2001) 1433–1445.
[31] D. Piccolo, A distance measure for classifying ARIMA models, J. Time Ser. Anal. 11 (1990) 152–164.
[32] P.M. Robinson, Gaussian semiparametric estimation of long-range dependence, Ann. Stat. 23 (1995) 1630–1661.
[33] S.J. Taylor, Modelling Financial Time Series, World Scientific, Singapore, 2008.
[34] R.S. Tsay, Analysis of Financial Time Series, 3rd edition, Wiley, 2010.
[35] X. Wang, K. Smith, R.J. Hyndman, Characteristic-based clustering for time series data, Data Min. Knowl. Discov. 13 (2006) 335–364.
[36] J.-M. Zakoian, Threshold heteroskedastic models, J. Econ. Dyn. Control 18 (1994) 931–955.