Unsupervised time-event probabilistic classification using large panels of time series *

Máximo Camacho† University of Murcia

Javier Palarea-Albaladejo University of Girona

Manuel Ruiz Marín Technical University of Cartagena

April 18, 2024

Abstract

This work presents a general framework for partitioning a time span into meaningful, non-overlapping segments using time series datasets characterized by a large cross-sectional dimension. These datasets commonly exhibit complexities and challenges such as non-linearities, structural breaks, asynchronicity, missing data or significant outliers that hamper data analysis and modelling. Aiming at accurate time-event classification and change or breakpoint detection in this setting, our proposal integrates three distinct components into a unified approach: symbolic analysis, compositional data analysis, and Markov-switching time series modeling. A comprehensive Monte Carlo simulation study is conducted to assess the performance of the method, demonstrating exceptional robustness across diverse scenarios. Moreover, its use in real-world applications is illustrated through two economic examples: (i) identifying recurrent recession and expansion regimes in the US economic cycle; and (ii) dating change points to high volatility episodes in the US stock market.

Keywords: Markov-switching models, event detection, change-point detection, symbolic dynamics, compositional data.

JEL Classification: E32, C22, E27.

*This work was partly supported by the grants PID2022-136547NB-I00 (MC), PID2022-136252NB-I00 (MR) and PID2021-123833OB-I00 (JPA) founded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE. JPA thanks grants ERDF A way of making Europe and 2021SGR01197 of the Department of Research and Universities of the Generalitat de Catalunya. Any possible errors are responsibility of the authors. Data and codes that replicate our results are available from the authors' websites.

†Corresponding author: University of Murcia, Department of Quantitative Methods for Economics and Business, 30100, Murcia, Spain. E-mail: mcamacho@um.es

1 Introduction

Event classification in time series involves partitioning time calendars into segments that represent distinct events or states, identifying critical change or breakpoints which correspond to shifts between different states. Determining the change points of the partition enables a deeper understanding of complex temporal patterns and assists in deploying a proactive mechanism to anticipate and respond to critical events and transitions.

Without aiming to be exhaustive, time-series event classification has been employed across various contexts. Thus, Liu et al. (2018) used it in clinical decision support systems, Reeves et al. (2007) for detecting climate changes, Küchenhoff et al. (2021) to monitor disease outbreaks in public health, or Lévy-Leduc and Roueff (2009) for network anomaly detection. Similarly, in economics and business, time-series event classification aids policymakers in monitoring key economic indicators such as exchange rates dynamics (Engel and Hamilton, 1990), real interest rates (García and Perron, 1998), and business cycles (Camacho et al., 2022). In finance, Jorda et al. (2011) and Barassi et al. (2020) identified episodes of global financial instability and change points in correlation structures, respectively. Additionally, Ranjan et al. (2018) showcased its utility in monitoring product quality in manufacturing, and Pourhabibi et al. (2020) in anomaly detection techniques for identifying fraudulent activities.

Given that practitioners often grapple with the challenge of event classification when labeled time series data are lacking, this work primarily focus on unsupervised classification. Time-event classification techniques involve algorithms that detect local minima and maxima within the time series data (Bry and Boschan, 1971), sequential change detection algorithms (Tartakovsky et al., 2014), minimizing cost functions over possible numbers and locations of change points (Killick et al., 2012), identifying latent discrete state variables that indicate regimes from which particular observations are drawn (Chib, 1988, and Hamilton, 1989), or handling kernel densities or rank statistics (Csörgö and Horváth, 1988).

These unsupervised methods encounter challenges that impact their effectiveness and reliability, many of which are outlined in Gupta et al. (2024). Key issues include sensitivity to assumptions such as stationarity, normality and independence of observations. *Ad hoc* constraints are sometimes required to ensure meaningful turning points, but slight deviations from assumptions can lead to inaccurate results. Furthermore, optimizing hyperparameters such as window size, penalty terms, or significance thresholds for some methods can be challenging; particularly in the absence of labeled data in at least part of the sample. Moreover, recent technological advancements have facilitated the analysis of vast collections of time series data, enhancing the richness of the inferences drawn from them. Nevertheless, the use of larger datasets introduces two additional challenges in empirical applications. Firstly, they notably increase the risk of issues such as non-linearities, structural breaks, missing data or outliers. Secondly, certain methods cannot cope with the increasing computational burden, then restricting scalability to real-time or high-dimensional data analysis.

We contribute to this literature with a novel time series event classification method that builds on latent class models, introducing an unobservable event indicator variable that represents temporal segments in a timeline. This is coupled with a large set of observable time series characterized

by a common dynamics which captures the characteristics of those segments. The goal is to categorize the distinct events occurring in each segment and identify the transitions between them by inferring the event indicator variable based on the observed patterns in the time series.[1]

To address the data issues typically found in large datasets, we apply a symbolic representation so that the analysis of the dynamics of the time series is simplified to evaluating a model-free sequence of symbols which, as suggested by the findings of Collet and Eckmann (2009), comprehensively captures the behavior of the time series.[2] Specifically, assuming that each event can be associated with high relative frequencies of some disjoint group of the potential set of symbols, inferring the unobserved event indicator variable involves tracking the evolution of the compositional transformations of these frequencies.[3] In addition, the event indicator variable is modeled as a discrete-time first-order Markov chain with a specified number of distinct regimes, with transition probabilities that may be constrained to ensure specific transitions. To estimate the probabilities of time t belonging to a particular regime, we rely on the approach suggested by Hamilton (1989).

Through a comprehensive Monte Carlo simulation study, our method demonstrates outstanding performance and robustness across various scenarios. These include the cases of missing data, outliers, structural breaks and time series asynchronicity. Thus, the simulation exercise shows that varying proportions of missing data or outliers had minimal impact on classification accuracy, especially when regimes were clearly distinct. Furthermore, the method exhibits notable prowess in detecting structural breaks, particularly when there is a substantial difference between regimes and low data variability. Finally, asynchronicity surrounding transition periods poses a significant challenge, potentially hindering performance. However, this issue is easily addressed by introducing a dedicated regime for the transition periods, leading to significant improvements in both turning point detection and regime classification.

Leveraging two empirical cases in the economic context, we showcase the practical value of our unsupervised time-event probabilistic classification method for large panel time series analysis. In the first application, we demonstrate its effectiveness in dating the United States (US) national reference cycle using a rich dataset of coincident indexes for all 50 states. These are sourced from the Federal Reserve Bank of St. Louis and derived from the single-index dynamic factor model developed by Crone and Clayton-Matthews (2005). Despite the extremely large outliers observed in 2020, our method performs remarkedly well, both accurately establishing the in-sample historical recession periods and offering promising results for pseudo real-time detection of national business cycle turning points. In the second application, we delve into the occurrence of highly synchronized upward trends in US stock market volatility by examining the squared returns of each asset comprising the Standard & Poor's index (S&P 500), identifying four prominent break dates that signify episodes of high volatility risk. Notably, two of these break dates coincide with the occurrence of the highest ever volatile index (VIX) levels during the COVID-19 virus outbreak and the financial crisis. The remaining two break dates align with the crash in October 1987 and the

[1] It is worth noting that the shifts between events can manifest as either recurrent or permanent phenomena.

[2] This approach has demonstrated remarkable robustness in dealing with data challenges when analyzing time-dependent systems, as shown for instance by Camacho, Romeu, and Ruiz (2021).

[3] These are formally required due to the non-negative nature and constant-sum properties of the vector of relative frequencies of the different groups of symbols.

1973-75 oil crisis.

In our view, the proposed method for time-event classification represents a significant advance with respect to traditional unsupervised algorithms, allowing to address key challenges inherent to the field as those mentioned above. Firstly, unlike many unsupervised methods, our approach does not rely on labeled training data, thereby alleviating the challenge of lacking ground truth labels. By leveraging innovative techniques from symbolic analysis and compositional data, our method does not require the determination of the optimal number of change points, overcoming the challenge of manually specifying required parameters. It is also invariant under monotonic transformations in the datasets and robust to the presence of missing data, outliers and structural breaks. Additionally, it is computationally cheap, thus facilitating real-time analysis of large-scale time series without compromising accuracy or scalability.

The rest of the paper is organized as follows. Section 2 introduces the formal background, including basic definitions and notation to be used throughout the paper, along with the proposed algorithm for change-point detection. Section 3 contains the Monte Carlo study assessing the performance of the method over a range of different scenarios and settings. Section 4 is devoted to the empirical examples using real data, namely with the purpose of dating the US reference economic cycle and detecting periods of high volatility in the S&P 500 index. Section 5 concludes with some final remarks.

2 Time-event classification

2.1 Preliminaries and notation

To represent latent classes in time series, we assume the existence of s^* distinct regimes each characterizing particular events. The regime evolution partitions the sample calendar time period $\{1, \ldots, T\}$ into non-overlapping segments. To characterize the occurrence of an event, we introduce an unobserved discrete event indicator variable, denoted by s_t, which takes values in the set $\{1, .., s^*\}$, such that $s_t = s$ indicates event s occurring at time t. For any given event $s \in \{1, .., s^*\}$, we define the support of s as $supp(s) = \{t|\ s_t = s\}$. Consequently, the sequence $S = \{s_1, \ldots, s_T\}$ encompasses the complete array of all conceivable classification outcomes.

Inference about the event indicator variable is made by analyzing the dynamics of an extensive collection of time series $\{X_{it}\}_{t=1}^{T}$, with $i = 1, 2, \ldots, N$. These time series are assumed to evolve uniquely within each event identified by s_t and distinctly across events. Importantly, these time series are synchronized, moving in unison whenever an event transition occurs. Therefore, the process of event classification from observed time series involves assigning a label s_t by drawing on the shared patterns embedded in the dataset $\{X_{it}\}$ at time point t.

2.2 Symbolic representation of time series

Addressing common issues in observed data is facilitated by a symbolic dynamic representation of the time series. In our context, symbolic dynamics involves partitioning the phase space into a finite number of regions and assigning a symbol to each partition. Subsequently, each time series

X_{it} at time t is transformed into a symbol, denoted by π, if, at time t, the corresponding trajectory of X_i in the phase space falls within the region associated with the symbol π. Specifically, we use the symbolic representation approach based on the phase space reconstruction method suggested by Takens (1981).

For each time series X_{it}, where $i = 1, \ldots, N$ and $t = m, \ldots, T$, we embed it into an m-dimensional space, so that it is expressed as

$$X_{it}(m) = (X_{it-m+1}, X_{it-m+2}, \ldots, X_{it}). \tag{1}$$

The vector $X_{it}(m)$ is referred to as the m-history, with the integer m representing the embedding dimension. For every cross-section $i = 1, 2, \ldots, N$, these m-histories encapsulate the dynamic behavior of the time series in the neighborhood of time t. While different symbolic representations are conceivable, we here follow the approach of Bandt and Pompe (2002). Specifically, we consider ordinal symbolic representations, which focus on the ordinal patterns within the m-histories. It has been shown that these representations are highly effective in the analysis of complex dynamics, as evidenced by Amigó (2010).

For this purpose, let Γ_m be the symmetric group of order $m!$ comprising all permutations of length m.[4] An element within this group, denoted by $\pi = (i_1, i_2, ..., i_m) \in \Gamma_m$, with $i_j \in \{0, 1, \ldots, m-1\}$ and $j = 1, 2, \ldots, m$, is referred to as a symbol. Subsequently, for each $i = 1, \ldots, N$ and $t = m, \ldots, T$, we map the m-histories $X_{it}(m)$ to a unique symbol $\pi_{it} = \pi \in \Gamma_m$ based on the relative value of the elements in the m-history. In this context, we say that $X_{it}(m)$ is of π_{it}-type if and only if $\pi_{it} = (i_1, i_2, ..., i_m)$ is the unique symbol in Γ_m satisfying the following conditions:

(a) $X_{it-i_1} \leq X_{it-i_2} \leq ... \leq X_{it-i_m}$,

(b) $i_{k-1} < i_k$ if $X_{it-i_{k-1}} = X_{it-i_k}$.

Condition (a) enforces the ordinal pattern, while condition (b) serves as a technical requirement ensuring the uniqueness of the symbol π_{it}.[5]

As a basic example, consider the time series $\{5, 3, 6, 1, 8, 9, 0, 2, 4\}$ of length $T = 9$. An embedding dimension $m = 3$ generates the $T - m + 1 = 7$ m-histories $\{(5, 3, 6), (3, 6, 1), (6, 1, 8), (1, 8, 9), (8, 9, 0), (9, 0, 2), (0, 2, 4)\}$. Now, using the integers $\{0, 1, 2\}$ to construct the symbols, these histories map to the set of symbols Γ_3 as follows: $\{(1, 2, 0), (0, 2, 1), (1, 2, 0), (2, 1, 0), (0, 2, 1), (1, 0, 2), (2, 1, 0)\}$. It is crucial to emphasize that the presence of specific symbols in particular time periods offers valuable insights into the dynamic behavior of the time series. For instance, in a negative trend period, the most common symbol will be $(0, 1, 2)$; whereas in periods of positive trend the symbol $(2, 1, 0)$ will tend to happen most frequently.

As s_t governs the dynamics of X_{it}, it naturally extends its influence to $X_{it}(m)$ and, by extension, to π_{it}. Consequently, the occurrence of a given regime can be associated with specific symbols. Thus, we assemble a collection $\Pi = \{\Pi_1, \ldots, \Pi_{s^*}\}$ comprising disjoint subsets of symbols within

[4] In both simulations and empirical applications, we have found that $m = 3$ provides a suitable choice for capturing the dynamics of time series.

[5] For continuous distributions, where single values have a theoretical probability of occurrence equal to zero, condition (b) becomes redundant.

Γ_m.[6] If all N m-histories are mapped into Π_s at time t, then time t is classified as regime s. However, recognizing the uncertainty in symbol mapping in real-world applications, the proposed event classification method involves assigning time t to the event labeled as $s_t = s$ if Π_s is the set of symbols more likely to occur across the N time series at t.

Specifically, let $\mathcal{P}_t = \{P_t(\Pi_1), \ldots, P_t(\Pi_{s^*})\}$ be the set of probabilities referred to the occurrence of the subsets of symbols in Π at time t. Thus, we classify $s_t = s$ when there exists a subset of symbols $\Pi_s \subset \Gamma_m$ for which $P_t(\Pi_s) > P_t(\overline{\Pi}_s)$, where $\overline{\Pi}_s$ refers to the complementary set of Π_s in Γ_m. To ensure a one-to-one mapping between time series and their symbolic representation in terms of event classification, the following two conditions must be satisfied for each $s \in \{1, 2, \ldots, s^*\}$:

1. There should always exist a subset of symbols associated to label s, $\Pi_s \subsetneq \Gamma_m$, such that $P_t(\Pi_s) > P_t(\overline{\Pi}_s)$ for all $t \in supp(s)$. In other words, when $s_t = s$, we must consistently identify a set of symbols Π_s as predominant among the entire set of possible symbols within the support of s at that time.

2. If there are two events $s_i, s_j \in \{1, 2, \ldots, s^*\}$ with $s_i \neq s_j$, then it should be the case that $\Pi_{s_i} \cap \Pi_{s_j} = \emptyset$. This means that the partition of Γ_m should result in disjoint subsets of symbols. Therefore, if a symbol $\pi \in \Pi_{s_i}$ this means that such symbol will continue to characterize any time t where $s_t = s_i$. Importantly, this symbol will never be part of the set of symbols Π_{s_j}.

In practice, it is necessary to estimate the probability of occurrence of each subset of symbols. To accomplish this, we establish that an m-history that maps to the symbol π is called of Π_j-type if and only if $\pi \in \Pi_j$. By considering all m-histories ending at a given time point t for all cross sections $i = 1, 2, \ldots, N$, the probability associated with the set of symbols Π_j can be estimated by the proportion of m-histories of Π_j-type at time t, written as

$$\hat{P}_t(\Pi_j) = \frac{\#\{i \mid X_{it}(m) \text{ is of } \Pi_j\text{-type}\}}{N}. \tag{2}$$

Here, # denotes the cardinality, $j = 1, \ldots, s^*$, and $t = m, \ldots, T$.

The feasibility of using the observed proportions as estimators of the probability of occurrence of the symbols will relate to the number of series N and the number of symbols ($m!$) considered. Drawing on the ordinary considerations for the common statistical test on multinomial proportions, based as an approximation to the χ^2 distribution, a practical rule of thumb is to ensure that the expected frequencies are at least 5 (see e.g. Rohatgi, 1976, Chapter 10). Applying this to our context, we recommend that the number of time series should be at least five times the total number of symbols. In any case, this is not anticipated to be an issue in general when dealing with moderate to large time series panels.

[6] The selection of subsets of symbols depends on the specific event classification. We illustrate the methodology for designing these subsets in the simulation and empirical sections of the manuscript.

2.3 Compositional characterization of symbol frequencies

By construction, the distribution of frequencies of the symbols, $\mathcal{P}_t = (P_t(\Pi_1), \ldots, P_t(\Pi_{s^*}))$, defines a data point within an s^*-part unit simplex

$$\mathcal{S} = \{\mathcal{P} = (P(\Pi_1), \ldots, P(\Pi_{s^*})) : P(\Pi_j) \geq 0; \sum_{j=1}^{s^*} P(\Pi_j) = 1\}. \tag{3}$$

The simplex corresponds to the sample space of so-called compositional data, i.e. multivariate data accounting for parts of a whole and then carrying relative information, subject to non-negativity and constant-sum constraints. This implies that conventional statistical methodology and models, designed for data obeying the ordinary (unconstrained) real Euclidean geometry, are incompatible with the unique characteristics of data defined on the simplex; which can lead to misleading or spurious results[7].

A principled way to overcome this challenge is to work with log-ratios between parts. This establishes a one-to-one mapping between the simplex and the real space spanned by a log-ratio transformation (or, more generally, a coordinate representation) of the original data, enabling an equivalent application of ordinary statistical methods on the latter. Moreover, working with log-ratios guarantees desirable properties, such as the irrelevance of both the scale of the data (i.e., results not depending on the constant-sum constraint, e.g. either 1 or 100) and the number of parts forming the composition. Accordingly, we conduct inferential analysis regarding the unobserved discrete variable s_t, not using the probabilities of the sets of symbols, but rather their log-ratio representation.

Among the various alternatives discussed in the specialized literature, orthonormal log-ratio (olr) coordinates have been favored for statistical modeling due to their natural alignment with the geometrical structure of the simplex and some desirable metric properties (Barceló-Vidal and Martín-Fernández, 2016). However, for simplicity of the exposition we opt here for an additive log-ratio (alr) coordinate representation (Aitchison, 1986). This decision is actually irrelevant for our modelling as stressed below.[8] Specifically, an alr-coordinate vector $y_t = (y_{1t}, \ldots, y_{s^*-1t})$ consists of $s^* - 1$ log-ratio coordinates derived from the probabilities of the set of symbols $\mathcal{P}_t$ as follows

$$y_t = \left(\ln \frac{P_t(\Pi_1)}{P_t(\Pi_{s*})}, \ldots, \ln \frac{P_t(\Pi_{s^*-1})}{P_t(\Pi_{s*})} \right). \tag{4}$$

The alr representation requires one symbol frequency to be used as reference and placed in the denominator of the log-ratios. Without loss of generality, we use the frequency of the last symbol for this. Note that using a different part as denominator is equivalent to perform a permutation of the parts of the composition, which formally results in a matrix linear transformation between two alternative alr-coordinate vectors.

Importantly, assuming multivariate normality and applying maximum likelihood estimation as

[7] The interested reader is referred to the seminal work by Aitchison (1986) and more recent accounts of compositional data analysis such as Pawlowsky et al. (2015), Greenacre (2018), and Filzmoser et al. (2021).

[8] An alr-coordinate system however implies an oblique mapping between the simplex and the real space, potentially leading to distortion of results when using statistical methods based on distance measures (Pawlowsky et al., 2015).

ordinarily, it can be shown that the results are invariant to linear transformations between two alr-coordinates (Aitchison, 1986; Palarea-Albaladejo and Martín-Fernández, 2008). In addition, both alr and olr coordinates lead to the same probability law (Mateu-Figueras et al., 2013). Therefore, the difference between using alr or olr lies in how the information about the distribution of symbols is represented in the real space of coordinates.

Lastly, it is worth noting that the log-ratio operation is not defined when some parts take a value of zero. In practice, specialized imputation methods are commonly employed to replace these zeros by small, sensible values while preserving the relative structure of the data. This assumes that the zeros are not true zeros but rather the result of some form of left-censoring (Palarea-Albaladejo and Martín-Fernández, 2015). In our context, any zeros in the data are addressed through compositional imputation, assuming that limited sampling prevented the observation of a particular symbol at a certain time point, but that a positive frequency might have been obtained from another sample.

2.4 Inference on the latent variable

To perform inference on the unobserved discrete variable s_t, based on the dynamics of the resulting series of alr-coordinates of the composition of the probabilities of the sets of symbols, we turn to Markov-switching models (MSMs). In this framework, we assume that s_t changes its regime at unknown points in the calendar time period, referred to as turning points. These regime changes influence both the observed time series X_{it} and the series of alr-coordinates y_t representing the relative distribution of symbols.

Given the temporal dependence of events in time series contexts, there is typically some degree of inertia in the regimes. Therefore, the latent variable is modeled as a discrete-time, first-order, stationary, and ergodic Markov chain with s^* states. Let $Y_t = \{y_1, \ldots, y_t\}$ represent the information available at time t. In a first-order Markov chain, the conditional probability that s_t takes a specific value j, given $s_{t-1} = i$, depends on the past only through the most recent value, s_{t-1}:

$$Pr(s_t = j|s_{t-1} = i, s_{t-2} = k, ..., Y_{t-1}) = Pr(s_t = j|s_{t-1} = i), \tag{5}$$

which is abbreviated as p_{ij} in the following. The transition matrix, consisting of these transition probabilities, defines the characteristics of the Markov process. In certain scenarios, the matrix may be constrained to ensure that a regime can transition to only one particular regime, as proposed in Boldin (1996). In other cases, transition probabilities may be restricted to allow the state variable to either remain at its current value or progress to the next higher value, with no possibility of returning, as suggested in the multiple change-point model proposed by Chib (1998).

To estimate the probability of time t belonging to a particular regime, we adopt the approach outlined by Hamilton (1989). Building on the observed alr-coordinates, we consider the simplified version of the linear MSM proposed by Camacho and Perez-Quiros (2007):

$$y_t = \mu_{s_t} + \epsilon_t. \tag{6}$$

Here, μ_{s_t} represents the expected value $E(y_t)$, and $\epsilon_t \sim N(0, \Sigma)$ denotes a normally distributed

random error term.[9] Gathering the model parameters together, including means $\mu_1, \ldots, \mu_{s^*}$, elements of the covariance matrix Σ, and transition probabilities p_{ij}, into a vector θ, the Bayes' rule and Markov chain properties lead to

$$Pr(s_t = i \mid \theta, Y_t) = \frac{f(y_t \mid \theta, Y_{t-1}, s_t = i) Pr(s_t = i \mid \theta, Y_{t-1})}{f(y_t \mid \theta, Y_{t-1})}. \tag{7}$$

Here, $f(y_t \mid \theta, Y_{t-1}, s_t)$ is the probability density function of a multivariate normal distribution with mean μ_{s_t} and variance Σ. Also, the denominator of this expression can be obtained by summing the elements of the numerator over $i = 1$ to s^*. Hamilton (1989) outlined a forward filter to store these filtered probabilities. Starting from an initial value $Pr(s_0 = i \mid \theta, Y_0)$, for $i = 1, \ldots, s^*$, the following two steps are carried out recursively from $t = m$ to $t = T$. First, the one-step-ahead prediction from information up to $t - 1$,

$$Pr(s_t = i \mid \theta, Y_{t-1}) = \sum_{i=1}^{s^*} p_{ij} Pr(s_{t-1} = i \mid \theta, Y_{t-1}), \tag{8}$$

is computed. Second, when the t-th observation of the alr-coordinates of $\mathcal{P}_t$ is added to this recursion, the filtered probabilities are updated using (7). In addition, the likelihood of the mixture $f(y_t \mid \theta, Y_{t-1})$ is obtained as a byproduct of this recursion, enabling parameter estimation by maximum likelihood.

3 Monte Carlo simulations

In this section, we present the results of a Monte Carlo study aimed at assessing the performance of the proposed time-event classification procedure when applied to large panels of time series data. We evaluate the procedure using common performance metrics, relying on lists of actual and predicted regimes over time. The study examines the performance and robustness of the proposed method across various scenarios defined for a range of model parameter settings. These scenarios include cases with missing data, outliers, structural breaks, and asynchronicity.

To provide a visual representation of the cases studied, we present illustrative time series panels in Figure 1, each of which referring to the different simulated scenarios. To facilitate the discussion, the event periods or regimes are indicated in the figures by white and grey background colors.[10] Different aspects are consider when assessing the classification performance of the proposal, including its ability to provide an accurate prediction, detect breakpoints, and indicate the need for adjustments to accommodate asynchronicities.

In this exercise, our primary focus is on a state variable that governs the first moment of the time series. However, it is worth noting that the analysis will remain largely analogous when the state variable governs the second moment. In such cases, the attention would shift to the squared time series, providing then insights into the evolution of the variance of the data. This versatility enables

[9] We found this simple model to be enough and useful in our empirical applications. However, if required, it could be readily extended to e.g. incorporate autoregressive parameters or a state-dependent covariance matrix.

[10] The implementation of the method was carried out using the R system for statistical computing version 4.2.3 (R Core Team, 2023), with the simulation pipelines set up in parallel using built-in R functions.

to adapt the method to varied characteristics of the data, ensuring a comprehensive assessment of their performance across different moments of the time series to detect event changes.

3.1 Missing and outlying observations

In this initial simulation study, denoted by DGP1, we generate a set of N time series. These time series exhibit a unit root in their levels but have stationary differences, following a Markov-switching autoregressive process with two regimes and recurrent shifts. To accomplish this, for each Monte Carlo simulation ($r = 1, 2, \ldots, R$), we create a random sequence of length T consisting of indicator variables, denoted by b_t^r. These indicators take values in $\{1, 2\}$ and represent the partition of the calendar time period of the sample into two distinct and recursive regimes. We assume that b_t^r follows a 2-state first-order Markov-switching model with transition probabilities p_{11} and p_{22}. Simultaneously, we generate N idiosyncratic components, denoted by u_{it}^r, of length T. These components are drawn from Gaussian random variables with mean zero and variance σ^2. Therefore, for each simulation, we obtain N nonlinear time series given by

$$X_{it}^r = \alpha_{b_t^r} + X_{it-1}^r + u_{it}^r. \tag{9}$$

We consider panels comprising $N = 30, 60, 120$ time series, each with a length of $T = 250$, and setting model parameters $p_{11} = 0.9$, $p_{22} = 0.3, 0.7$, and $\sigma^2 = 1, 2$. The two regimes associated with the underlying state b_t^r are defined by values $\alpha_1 < 0$ and $\alpha_2 > 0$, with $\alpha_2 - \alpha_1$ taking values 1 and 3 which determine the expected difference in the first difference of the time series between both regimes. The embedding dimension is fixed at $m = 3$, resulting in a set of symbols Γ_3 with a cardinality of 6.

Thus, in the r-th Monte Carlo simulation, 3-histories $X_{it}^r(3)$, for $i = 1, \ldots, N$ and $t = 3, \ldots, T$, are computed and subsequently mapped into Γ_3. The number of states for the subsets of symbols is fixed at $s^* = 2$, and these states govern the dynamics of two subsets of symbols, namely $\Pi = \{\Pi_1, \Pi_2\}$. Here, we are interested in performing inference on the regime that characterizes decreasing paths, occurring when $s_t = 1$, and associated with the set of symbols $\Pi_1 = (0, 1, 2)$. Consequently, we define a second set of symbols, Π_2, encompassing the complementary set $\overline{\Pi}_1 = \{(2, 1, 0), (0, 2, 1), (1, 2, 0), (1, 0, 2), (2, 0, 1)\}$. These two sets of symbols are sufficient to characterize the regime changes in the panel of time series for this particular simulation scenario.

The elements $P_t^r(\Pi_j)$ of $\mathcal{P}_t$ are then calculated as the proportion of times any of the symbols belonging to Π_j (with $j = 1, 2$) appear in the N 3-histories $X_{it}^r(3)$ observed at time point t. If any of these proportions is occasionally zero, we consistently impute them with a value below the minimum observed proportion for the affected symbol, using the log-ratio expectation-maximization (lrEM) method (Palarea-Albaladejo and Martín-Fernández, 2008, 2015). This method takes into account the interdependencies between the proportions of symbols and preserves the constraint that relative frequencies sum up to 1. Finally, the series of alr-coordinates $y_t^r = \ln(P_t^r(\Pi_1)/P_t^r(\Pi_2))$ leads to the estimation of a univariate 2-state MSM as described in (6), with filtered regime probabilities $P(s_t^r = i|\theta, Y_t^r)$, where $\theta = (\alpha_1, \alpha_2, \sigma^2, p_{11}, p_{22})$ and $i = 1, 2$. This process is repeated in $R = 500$ simulation runs.

The common values of the model parameters for this DGP1 study (N, p_{11}, p_{22}, σ^2, and $\alpha_2 - \alpha_1$) are combined with a parameter set that controls for the presence of missing data and outliers. In the case of missing data, we implement a random mechanism to simulate missingness in a fraction p_{miss} of the N time series, where $p_{miss} = \{0.2, 0.4, 0.6, 0.8\}$. For each time series within this fraction, missing values are randomly imposed at a rate of 20-60% of the time points. It is important to note that when translating the original time series into a symbolic representation, a symbol is considered missing if any element of the associated m-history $X_{it}^r(m)$ is missing (illustrated in Figure 1A). In practice, the computation of each proportion $P_t^r(\Pi_j)$ is therefore based only on the observed cases.

In the case of outlying observations, we randomly select fractions of time series with varying levels of outlying data, denoted by $p_{out} = \{0.2, 0.4, 0.6, 0.8\}$. For each selected time series, we introduce outlying observations at 10-30% of the time points. These outliers are generated by firstly calculating the first and third quartiles of the data distribution across time series at a specific time point. Then, subtracting 1.5 times the interquartile range from the first quartile and adding 1.5 times the interquartile range to the third quartile.[11] Given this, the selected time points in the time series are randomly transformed into outliers by assigning to them the result of subtracting (resp. adding) a percentage deviation from the tip of the lower (resp. upper) whisker, with such percentage being in the set $\{20\%, 40\%, 60\%\}$. Figure 1B provides a visual representation of the generated data sets.

To evaluate the performance of the method in the setting above, we use four common metrics computed from the outputs generated in each simulated scenario, which are averaged across runs to provide a final estimate. These metrics include (1) Area Under the Receiver Operating Characteristic curve (AUROC), which jointly considers sensitivity and specificity to measure the ability of the classifier to separate between regimes as the binary discrimination threshold varies; (2) Brier score, which measures the mean squared difference between the actual outcome of the event at time point t (0 when $s_t = 2$ and 1 when $s_t = 1$) and the probability of this event estimated from the model; (3) F1 score, which assesses the ability to detect change of regime while accounting for the unbalanced occurrence of regimes over time; and (4) balanced accuracy, which is the arithmetic mean of sensitivity and specificity and provides an accuracy measure adjusted for unbalance occurrence of events. For all metrics except the Brier score, higher values closer to 1 indicate better performance (see e.g. Sammut and Webb, 2017).[12]

3.2 Structural breaks

In the second simulation study, referred to as DGP2, we simulate Markov-switching autoregressive processes with means subject to structural breaks or permanent shifts. These structural breaks in the means of the time series occur at specific time points, including switches upwards at $T/4$ and switches downwards at $T/2$ and $3T/2$. Within each of these four segments, the time series evolve stationarily around the corresponding mean.

The structural breaks are introduced using an indicator variable v_t that takes values $\{1, 2, 3, 4\}$

[11] Therefore, we are computing lower and upper whiskers commonly used as thresholds beyond which a data point is classed as outlier in boxplots.

[12] Note that the predicted regime at a given time point is that having the highest filtered probability as estimated by (7) for each scenario.

for each of the four segments respectively, and governs the dynamics of the time series as

$$X_{it}^r = \mu_{v_t} + u_{it}^r. \tag{10}$$

Here, u_{it}^r follows an independent distribution $N(0, \sigma^2)$, and $r = 1, \ldots, R$. Specifically, $\mu_1 = 1$ is set as the reference mean, and then the others are define as $\mu_2 = \mu_1 + 3\delta$, $\mu_3 = \mu_1 - \delta$ and $\mu_4 = \mu_1 - 2\delta$, with the parameter δ controlling the relative separation between them.

Similarly to DGP1, we compute 3-histories and their corresponding symbols for the DGP2 study. However, in this case, we allow for a 3-state regime switching variable to determine whether the time series are evolving within one of the four segments or transitioning between them. Specifically, the state variable s_t can take three values: $s_t = 1$ when the time series are switching upwards, $s_t = 2$ when the time series are switching downwards, and $s_t = 3$ when the time series are steadily evolving within one of the four segments.

This setup results in a collection of symbols $\Pi = \{\Pi_1, \Pi_2, \Pi_3\}$, where $\Pi_1 = (2, 1, 0)$ refers to upward switches (occurring around $T/4$), $\Pi_2 = (0, 1, 2)$ refers to downward switches (occurring around $T/2$ and $3T/2$), and $\Pi_3 = \{(0, 2, 1), (1, 2, 0), (1, 0, 2), (2, 0, 1)\}$ refers to any steady time period within any of the four segments, where the time series fluctuate randomly around the mean. We compute the proportions of symbols $P_t^r(\Pi_j)$, with $j = 1, 2, 3$, and subsequently obtain the alr-coordinates $y_{1t}^r = \ln(P_t^r(\Pi_1)/P_t^r(\Pi_3))$ and $y_{2t}^r = \ln(P_t^r(\Pi_2)/P_t^r(\Pi_3))$. The time series $y_t^r = (y_{1t}^r, y_{2t}^r)$ are then modeled by a 3-state bivariate MSM as in (6) to estimate the probabilities of each of the three regimes over time.

For the simulation in DGP2, we consider a range of scenarios primarily based on values for $\delta = \{0.5, 1, 2, 3, 4, 5\}$ and $\sigma^2 = \{1, 2, 3, 4\}$. The number of time series per panel and their length are fixed at $N = 250$ and $T = 250$, respectively. To specify the transition matrix of the underlying Markov chain governing the move from one state to another, we set the elements p_{ij} with $i, j = 1, 2, 3$ as follows. After a switch, we assume that the time series remain steady regardless of its direction. This implies that the transition probabilities are restricted to avoid two consecutive switches, resulting in $p_{12} = 0$ and $p_{21} = 0$. This case is illustrated in Figure 1C.

The assessment of the performance of the method focuses here on its ability to detect structural breaks. Thus, the prediction outcome is simplified to a binary split into steady regimes (series within any of the segments) or breaks (series switching either upwards or downwards). The true positive rate is used as performance metric, which represents the proportion of times an actual break is correctly detected by the model over $R = 500$ simulation runs. Since there are only three single time points corresponding to breaks over the time span defined by T, whereas the probability of a break as estimated by the model varies continuously, it is possible that a break is not precisely predicted at its exact predefined point. Therefore, to ensure a fair evaluation and prevent potential false negatives caused by slight misalignments, we allow for a window of width $\pm w$ around the actual break points. Success is then recorded whenever a break is predicted within a close vicinity of the break point. Here, we consider different values for such window width, namely $w = \{0, 1, 2, 3\}$ time points.

3.3 Asynchronous time series

To investigate scenarios where some time series enter and/or change regime asynchronously, we use a 2-regime setting similar to the one used in the DGP1 study. However, we now allow certain time series to transition asynchronously between these two regimes. Asynchronous transitions occur when the events affect specific time series either before or after they impact others. This leads to variations in the timing of regime changes across different time series within the same panel.

In our simulations, we consider panels consisting of $N = 250$ time series, each with a length of $T = 250$ time points. The MSM parameters are fixed at $p_{11} = 0.97$, $p_{22} = 0.95$, $\alpha_2 - \alpha_1 = 2.5$, and $\sigma^2 = 1$. Within these panels, random fractions of time series, ranging from 0.1 to 0.8, are made asynchronous by adjusting the timing of their entry into and exit from a particular regime. We refer to the time difference between actual and standard entry or exit points as the asynchronicity gap η. We test three values for $\eta = \{5, 10, 15\}$. Similar to DGP1 and DGP2, $R = 500$ simulations are run for each scenario defined by the parameter settings mentioned above. It is important to note that we set transition probabilities relatively high to favor scenarios with separated and longer-lasting regimes.[13]

While our simulations are based on a 2-regime setting, considering asynchronous time series within the panel induces a transitional phase, during which some series are entering the new regime while others remain in the current one. This actually defines an additional state that governs such periods around transitions. As shown in both the simulation exercise and the empirical applications developed in Section 4, the inclusion of an additional state is a key aspect of the modelling. Figure 1D visually illustrates a panel containing asynchronous time series.

Of main interest for the performance assessment here is the ability of the method to identify regime changes and accurately assess the duration of regime labeled as 1, which we use as the reference event. Specifically, the focus is on the precision of the classifier when identifying $s_t = 1$, i.e., the fraction of actual time points labeled as 1 among all the time points this regime exhibits the highest filtered probability.

3.4 Simulation-based results

Figure 2 presents an overview of the results from the DGP1 simulation study across various parameter settings. Results for scenarios with complete observations and no outliers (i.e., p_{miss} and p_{out} equal to 0) are included as reference. Overall, the results illustrate that increasing proportions of time series with missing or outlier observations have a negligible impact on the classification performance. The only noticeable exception is observed for the Brier score in the case of outlying observations when states are not well separated (first column of Figure 2B), showing a slight positive trend with increasing proportions of time series affected by outliers.

The main differences in performance across scenarios and evaluation metrics are primarily due to differences between α_2 and α_1. As anticipated, a more distinct separation between regimes leads to superior model performance, as evidenced by scenarios where $\alpha_2 - \alpha_1 = 1.5$. This results in very high AUROC values over 0.9, small Brier scores below 0.10, and balanced accuracy around

[13] As expected, we observed in preliminary trials that excessive changes and closely spaced short-term regimes led to confusion and overlapping regimes as η increased.

0.8 in most cases. Despite the decline in values, when the difference between regime means is small ($\alpha_2 - \alpha_1 = 0.5$), AUROC generally remains over 0.7, Brier score reaches values below 0.25 in the worse cases, and balanced accuracy is over 0.65. In this simulated scenario, the F1 score is the most negatively affected metric, with values around 0.2 when the states are not well separated in combination with noise and low inertia in the states.

Noisy signals, showcased by large values of σ^2, lead to lower classification performance. This trend is more noticeable in the scenario with $\alpha_2 - \alpha_1 = 0.5$. Regarding the transition probability p_{22}, differences in performance are generally challenging to discern. The exception is the F1 score, where higher values of p_{22} seem to be associated with moderately improved performance. This result can be attributed to the improved classification ability when the regimes become more persistent.

Figure 3 shows a summary of the results from the DGP2 simulation study. The curves represent different window widths (parameter w) around the actual structural break points, allowing for various degrees of tolerance in the alignment of actual and predicted break points as discussed above. The importance of considering this window width is evident from the graphs, as the true positive rate is significantly lower when $w = 0$ across all scenarios. This was expectable because, by construction, when there is a change of regime at time t, the corresponding change in the proportion of symbols marking the turning point requires that at least one period has passed.

Focusing on results for $w > 0$, it is evident that the proposed method is capable of detecting structural breaks very accurately, although the performance depends on the differences in mean levels across regimes and their interaction with data variability. As expected, the highest performance is observed when there is a substantial separation across regimes and low variability. For scenarios with minimal relative separation between regimes ($\delta = 0.5$), the performance is generally poor. However, at the lowest variability level ($\sigma^2 = 1$), the true positive rate exceeds 0.80 for any $\delta \geq 3$, reaching a value over 0.95 for the largest separation considered ($\delta = 5$). Performance becomes more challenging as data variability increases, with a true positive rate of around 0.80 achieved only for $\delta = 5$ when $\sigma^2 = 4$.

Lastly, Figure 4 addresses the asynchronous time series case DGP3. Recall that the data were generated assuming two regimes, but as mentioned earlier, the severity of asynchronicity requires considering a 3-state MSM in practice. This additional state allows for a genuine characterization of the transitioning period around a regime change, thereby improving regime classification. As an example, Figures 4A and 4B depict the estimated regime probabilities using 2-state and 3-state MSMs fitted to a panel of simulated time series (comprising 80% asynchronous time series with a gap of 5). These figures illustrate the impact of asynchronicity on the probabilities around the transitioning periods. Unlike in the 2-state MSM case, the 3-state MSM effectively identifies the transitioning period as an additional state, which implies a more precise outline of the turning points across the three regimes.

For the specified ranges of asynchronicity gap (η) and the proportion of asynchronous time series in the panel, Figure 4C provides the average precision in identifying the regime with negative drift in (9), using both 2-state and 3-state specifications. The results show that the performance of the proposed method remains robust to asynchronicity, and the performance of both models is comparable up to a moderate fraction of asynchronous time series (40-60%) regardless of η. However,

beyond that range, the performance of the 2-state MSM formulation significantly declines, while the 3-state MSM continues to yield good results. Interestingly, the 3-state MSM appears to become more stable with increasing η, potentially due to a better characterization of the transitioning period with a larger number of time points therein.

4 Empirical examples

4.1 Dating the reference cycle

Establishing a reference cycle dating for a given country or economic area, which has a long tradition in economics, can be viewed as dividing the time calendar into segments of recurrent recession and expansion regimes of the economy. The business cycle dating committee of the National Bureau of Economic Research (NBER) states that a recession is the period between a peak of economic activity and its subsequent trough or lowest point. Thus, the peak-trough turning points mark the break dates and delimit the periods of recession. The committee emphasizes that "[...] a recession involves a significant decline in economic activity that is spread across the economy and lasts more than a few months". From this definition, several important issues can be inferred. Firstly, in reference to the US, a recession must have a widespread impact throughout the entire economy. This suggests that a national recession should affect states broadly, rather than being confined to only one set of the states. Secondly, the definition implies that the dynamics of the business cycle phases must exhibit a certain degree of inertia. To capture this aspect, we focus on Markov-switching frameworks to define the statistical properties of the regimes. Lastly, the definition lacks a clear statement regarding what the committee understands as an expansion, which is viewed as a non-recessionary period.

We leverage these three characteristics of the business cycle to demonstrate that our model can be effectively employed to make inferences about the probability of a national recession commencing or concluding at a specific date. We use a dataset at the state level, which was originally compiled by Crone and Clayton-Matthews (2005) and is regularly updated by the Federal Reserve Bank of St. Louis. These authors employed a state-level single-index dynamic factor model to generate a set of consistent coincident indexes for all 50 states in the US. These coincident indexes reflect the overall economic condition of each state. Figure 5A illustrates the trends in these indexes, spanning from January 1979 to August 2023. The graph underscores the substantial growth of these indexes, punctuated by marked declines that align with the NBER-referenced national recessions, represented by shaded areas in the background.

Characterizing the dynamics during national non-recessionary periods is somewhat more challenging using a single regime with the information provided by the state indexes. On the one hand, Figure 5A shows periods of sustained upward trends in most of the state indexes and we refer to these periods as national expansions. On the other hand, Hamilton and Owyang (2012) observed examining state-level data that some states entered recessions before others, or experienced recoveries with certain lags, particularly in the months preceding and following recessions.

The goal in this example is to perform inferences about the national recessions and its phase-change transitions using the information provided by the coincident state indexes. However, the

unprecedented economic contraction associated with the COVID-19 epidemic in 2020 presented a unique challenge for this task. This period encompassed the deepest and shortest recession in recent history, followed by an unusually rapid and robust economic recovery fueled by policy interventions, leading the indexes to rebound to pre-recession levels within a few months. Among other researchers, Bobeica and Hartwig (2022) have documented the challenges posed by such sharp fluctuations, which have a tendency to distort the estimated coefficients of standard time series parametric approaches.

To address this challenge, we employed a symbolic dynamics approach to make inferences on the regime variable. Specifically, we used an embedding dimension of $m = 3$ to generate time series of 3-histories for the 50 coincident indexes and mapped them into the set of symbols Γ_3, which comprises all permutations of the three elements $\{0, 1, 2\}$. Subsequently, the symbols were categorized into three groups, denoted by $\{\Pi_1, \Pi_2, \Pi_3\}$, where $\Pi_1 = (0, 1, 2)$ represents two consecutive declines (labeled as regime 1), $\Pi_2 = (2, 1, 0)$ represents two consecutive rises (labeled as regime 2), and $\Pi_3 = \{(0, 2, 1), (1, 2, 0), (1, 0, 2), (2, 0, 1)\}$ represents other situations in the evolution of the coincident state indexes (labeled as regime 3).

The evolution of the relative frequency of these three groups of symbols, with respect to the total number of symbols mapped for the 50 states in each time period, i.e., $\hat{P}_t(\Pi_i)$, with $i = 1, 2, 3$, is displayed at the bottom of Figure 5. As expected, Figure 5B shows that more than 50% of the states exhibited the symbol Π_1 in the periods that were classified as recession by the NBER business cycle dating committee. In contrast, Figure 5C shows that in the middle of the intervals between NBER-referenced troughs and their subsequent peaks, the coincident indexes of more than 75% of the states displayed the symbol Π_2. Finally, Figure 5D shows that the probability of the group of symbols Π_3 substantially increased during the transitions before the NBER-referenced peaks and after the NBER-referenced troughs.

Figure 6 displays at the top the evolution of the corresponding alr-coordinates $y_{1t} = \ln \frac{\hat{P}_t(\Pi_1)}{\hat{P}_t(\Pi_3)}$ and $y_{2t} = \ln \frac{\hat{P}_t(\Pi_2)}{\hat{P}_t(\Pi_3)}$. In line with the previously outlined 3-regime pattern, during NBER-referenced recession periods the first alr-coordinate experiences a substantial increase, while the second alr-coordinate declines significantly. The middle of the periods between NBER peaks and their subsequent troughs are characterized by high values of the second alr-coordinate and low values of the first alr-coordinate. Furthermore, both alr-coordinates display a decrease before the peaks and after the troughs.

Using the specification (6), we account for this asymmetric business cycle dynamics by modeling the vector of alr-coordinates $y_t = (y_{1t}, y_{2t})$ conditionally upon an unobservable 3-regime state variable s_t. To aid interpretation, Figures 6C-E exhibit the probabilistic inferences about the three distinct regimes across the entire sample. These figures effectively demonstrates the ability of the filtered probabilities to split the sample into three subperiods with clear economic interpretations. Figure 6C exhibits a notable alignment between the months with high probabilities of regime 1 and the NBER-referenced recessions, suggesting that this regime corresponds to recession, and the time series in this graph represents the probabilities of being in recession. Figure 6D indicates that it is highly improbable for the US to be in regime 2 during NBER recessions and the periods that encompass a few months before the peaks and after the troughs. Conversely, the probability of

being in regime 2 approaches unity during the middle of NBER expansions, suggesting that the probabilities depicted in this figure represent the likelihood of national expansions. Lastly, Figure 6E highlights a distinct shift in the probability of the economy being in regime 3, from almost negligible to nearly one, in the periods immediately preceding and following NBER recessions, albeit with a few exceptions. Accordingly, regime 3 is interpreted as a transitional phase between recessions and expansions and the probabilities depicted in the figure reflect the likelihood of navigating through the business cycle phase changes.

So far in this example, we have evaluated the performance of our approach using state coincident indexes to partition the evolution of the US business dynamics and their transitional dynamics using data from January 1979 to August 2023. However, a good in-sample performance of the probabilistic classification model could deteriorate significantly when detecting the business cycle phase changes in real time. To address this issue, we assess how suitable our approach is for detecting business cycle turning points in a timely manner by conducting a pseudo real-time classification experiment. This analysis is based on successive enlargements of the latest available dataset, aiming to closely mimic the real-time analysis that would have been performed by a potential user of the models when classifying each month into one regime based only on the information available at that time.

In our pseudo real-time analysis, we consider successive data vintages, each starting from January 1979. The first data vintage ends in July 1987 and the final one encompasses the entire dataset up to August 2023. Note that we re-estimate all sets of symbols, the log-ratio coordinates, and the MSMs for each of the pseudo real-time vintages ending at τ, with τ ranging from July 1987 to August 2023. Following this recursive procedure, the corresponding 421 probabilities of regime 1 that refer to τ are stored and displayed in Figure 6F. The figure reveals that the pseudo real-time probabilities of regime 1 closely align with NBER recessions, increasing around the beginning of recessions and decreasing around their end. Thus, the time-series probabilistic classifier performs nearly perfectly, effectively distinguishing between recession and non-recession periods with few mistakes, achieving an AUROC of 0.98 and a Brier score of only 0.03.

4.2 Stock returns volatility

Asset market volatility is of critical importance to economic agents for several reasons. It serves as a measure of risk exposure for their investments, it plays a central role in pricing of derivative securities, and it acts as a barometer for policy makers, indicating the vulnerability of financial markets and the economy. While many assets belonging to broad stock market indexes typically exhibit relatively low historical correlation to one another, during periods of heightened volatility, such as the 2008 financial crisis, individual volatilities have a tendency to become more correlated, even across different sectors. This phenomenon leads to periods of high overall volatility, posing a challenge to investors attempting to make rational risk management decisions by diversifying their portfolios.

The purpose of this empirical application is to demonstrate the effectiveness of our time-event classification procedure to identify periods of increased risk in the US stock market. Specifically, we aim to detect instances where a substantial number of assets experience heightened volatility, indicating a potential turbulence in the market. To achieve this, we focus on the assets that

form the Standard & Poor's 500 stock market index, known as the S&P 500, which measures the performance of approximately 500 companies in the US across 11 sectors, accounting for close to 80% of the available market capitalization on stock exchanges. Due to its depth and diversity, the assets that form the index provide valuable insights into the overall market trends, investor sentiments, and health of the financial and other sectors.

Given that volatility is unobservable, we use the squared returns of each asset as a proxy for the individual underlying volatility processes, where the returns represent the growth rates of the stock market average traded price of the respective asset. To asses the evolution of the volatility in the US stock market, Figure 7A displays the monthly squared mean-centered returns of the S&P 500 assets from January 1973 to January 2024.[14] The granularity of the visualization and the presence of overlapping data points illustrate the difficulties to identify specific dates during which a significant number of assets start to experience pronounced increases in volatility.

In order to overcome this difficulty, we employed the introduced symbolic dynamics approach to generated time series of 3-histories for the stock return volatilities and mapped them into the set of symbols comprising all permutations of $\{0, 1, 2\}$. Then, we categorized the symbols into three groups, denoted by $\{\Pi_1, \Pi_2, \Pi_3\}$, where $\Pi_1 = (0, 1, 2)$ represents two consecutive declines in volatility (labeled as regime 1), $\Pi_2 = (2, 1, 0)$ represents two consecutive rises (labeled as regime 2), and $\Pi_3 = \{(0, 2, 1), (1, 2, 0), (1, 0, 2), (2, 0, 1)\}$ represents other situations (labeled as regime 3). Figure 7B displays the alr-coordinates of symbol proportions, $y_{1t} = \ln \frac{\hat{P}_t(\Pi_1)}{\hat{P}_t(\Pi_3)}$ and $y_{2t} = \ln \frac{\hat{P}_t(\Pi_2)}{\hat{P}_t(\Pi_3)}$, accompanied by NBER-referenced recessions as shaded areas for contextual reference.

Lastly, a 3-regime Markov-switching model, as represented by (6), is fitted to the vector of alr-coordinates. In this case we constrained transition probabilities to prevent consecutive switches between regimes 1 and 2, setting $p_{12} = 0$ and $p_{21} = 0$. For the sake of interpretability, Figure 7C displays the filtered probabilities of regime 2, again with shaded areas representing NBER-referenced recessions for reference. Additionally, it includes the monthly average of the CBOE VIX volatility index, which gauges the expected stock market volatility over the next 30 days, as derived from S&P 500 index options.

The figure highlights four prominent highly synchronized upward trends in US stock market volatility. Various economic and non-economic factors, including the COVID-19 pandemic, changes in economic policies, shifts in inflation trends and industry-specific developments, can substantially contribute to these directional shifts in the stock market and significantly influence its volatility. Of particular note, though, is that the two most recent break dates coincide with the occurrence of the highest ever VIX levels during the coronavirus disease outbreak and the financial crisis. The other two break dates marking highly correlated upward trends in market volatility correspond to the crash in October 1987 and the 1973-75 oil crisis.

[14] Due to the presence of extreme values, the y-axis has been truncated at 0.5 to enhance visual clarity and focus on the relevant patterns and trends.

5 Concluding remarks

We have introduced a novel formal approach for partitioning a time span into meaningful, non-overlapping segments using time series datasets with a large cross-sectional dimension based on time-event probabilistic classification analyses. Our method exhibits exceptional performance and robustness across diverse scenarios, as demonstrated by rigorous Monte Carlo simulations. Notably, different proportions of missing data or outliers have minimal impact on classification accuracy, particularly when regimes are distinct. The method excels in detecting structural breaks, especially in scenarios with significant differences between regimes and low data variability. Although, asynchronicity during transition periods presents a challenge, this issue can be easily addressed by introducing a specific regime for the transition periods, resulting in substantial improvements in turning point detection and regime classification.

Two examples have illustrated real-world applications. In the first one, a set of coincident indexes for the 50 states of the US is used to learn about the course of the national economy. As in Hamilton and Owyang (2012), full-sample and real-time exercises show that, despite the heterogeneity between states, there appears to be a strong national component to the recessions. The method is able to provide excellent inference about the business cycle regimes. The second application uses the monthly average price of the assets of the S&P 500 index to infer the break dates that mark episodes of highly synchronized upward trends in volatility in the US stock market. Our findings indicate that the two most recent breakpoints align with record-high levels of the VIX index during the recent COVID-19 pandemic and the financial crisis, and the other two breakpoints coincide with the October 1987 market crash and the oil crisis of 1973-75.

The proposal integrates three distinct components into a unified framework: symbolic analysis, compositional data analysis, and Markov-switching time-series modeling. Each of these components is strategically employed to address specific challenges and features related to the treatment of the data and their use for time-event classification analysis. The symbolic representation maps a real-valued time series into a discrete symbol space and, instead of following the trajectory of the time series point by point, it only requires registering the alternation of certain appropriate symbols.

The procedure involves choices regarding parameter settings and model specification. These choices include the embedding dimension, the log-ratio coordinate representation of the estimated symbol probabilities, and the specification of the Markov-switching model. For the analyses conducted in this paper, we chose settings that were simple and worked for the particular purposes as illustrated. However, the approach can be easily extended and customized to address, or represent more realistically, the particularities of other related time series problems.

Recent technological advances and the democratization of data via open data initiatives have enabled the analysis of vast collections of time series data, enhancing the richness of the inferences made from them. However, with the use of larger time series datasets comes a higher risk of encountering issues such as non-linearities, structural breaks, missing data, or significant outliers. As demonstrated, the approach depicted in this work allows to tackle these issues head-on. Therefore, we believe that our method represents a significant step forward in the field of time-series probabilistic classification and change-point detection, offering a reliable and versatile solution to address inherent challenges faced by unsupervised algorithms in a contemporary context.

References

[1] Aitchison, J. 1986. The statistical analysis of compositional data. Chapman and Hall. (Reprinted in 2003 with additional material by The Blackburn Press).

[2] Amigó, J. 2010. Permutation complexity in dynamical systems: ordinal patterns, permutation entropy and all that. Springer Science & Business Media.

[3] Bandt C., and Pompe B. 2002. Permutation entropy: a natural complexity measure for time series. *Physical Review Letters* 88: 174102.

[4] Barassi, M., Horvath, L., and Zhao, Y. 2020. Change-point detection in the conditional correlation structure of multivariate volatility models. *Journal of Business and Economic Statistics* 38: 340-349.

[5] Barceló-Vidal, C., and Martín-Fernández, J.A. 2016. The Mathematics of Compositional Analysis. *Austrian Journal of Statistics* 45: 57-71.

[6] Bobeica, E., and Hartwig, D. 2023. The COVID-19 shock and challenges for inflation modelling. *International Journal of Forecasting* 39: 519-539.

[7] Boldin, M. 1996. A check on the robustness of Hamilton's Markov-switching model approach to the economic analysis of the business cycle. *Studies in Nonlinear Dynamics and Econometrics* 1, 3546.

[8] Bry, G., and Boschan, Ch. 1971. Cyclical analysis of time series: Procedures and computer programs. New York: National Bureau of Economic Research.

[9] Camacho, M., and Perez-Quiros, G. 2007. Jump-and-rest effect of U.S. business cycles. *Studies in Nonlinear Dynamics and Econometrics* 11: article 3.

[10] Camacho, M., Romeu, A., and Ruiz, M. Symbolic transfer entropy test for causality in longitudinal data. *Economic Modelling* 94: 649-661.

[11] Camacho, M., Gadea, M. and Gómez-Loscos, A. 2022. A new approach to dating the reference cycle. *Journal of Business and Economic Statistics* 40: 66-81.

[12] Camacho, M., and Gadea, M. 2022. Econometric methods for business cycle dating: a practical guide. *Oxford Research Encyclopedia of Economics and Finance*, forthcoming.

[13] Csörgö, M., and Horváth, L. 1988. Nonparametric methods for changepoint problems. In P. Frishnaiah and C. Rao (Eds), *Handbook of statistics 7. Quality Control and Reliability*, pages 403-425. North-Holland, Amsterdam.

[14] Chib, S. 1998. Estimation and comparison of multiple change-point models. *Journal of Econometrics* 86: 221-241.

[15] Crone, T., and Clayton-Matthews, A. 2005. Consistent Economic Indexes for the 50 States. *The Review of Economics and Statistics* 87: 593-603.

[16] Davig, T. 2007. Change-points in US business cycle durations. *Studies in Nonlinear Dynamics and Econometrics* 11: Article 6.

[17] Engel, Ch., and Hamilton, J. 1990. Long swings in the dollar: Are they in the data and do markets know it? *American Economic Review* 80: 689-713.

[18] Fawaz, I., Forestier, G., Weber, J., Idoumghar, Ll., and Muller, p. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery* 33: 917–963.

[19] Filzmoser, P., Hron, K., Martin-Fernandez, J.,; and Palarea-Albaladejo, J. 2021. Advances in compositional data analysis, Festschrift in honor of Vera Pawlowsky-Glahn. Springer.

[20] Garcia, R., and Perron, P. 1996. An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics* 78: 111-125.

[21] Greenacre, M. 2018. Compositional Data Analysis in Practice. Chapman and Hall/CRC.

[22] Guo, S., Guo, W., Abolhassani, A., Kalamdani, R., Puchala, S., Januszczak, A., 2019. Manufacturing process monitoring with nonparametric change-point detection in automotive industry. *Journal of Manufacturing Science and Engineering* 141: Article 071013.

[23] Gupta, M., Wadhvani, R., and Rasool, A. 2024. Comprehensive analysis of change-point dynamics detection in time series data: A review. *Expert Systems with Applications* 248: Article 123342.

[24] Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycles. *Econometrica* 57: 357-384.

[25] Hamilton, J., Owyang, M. 2012. The propagation of regional recessions. *The Review of Economics and Statistics* 94: 935-947.

[26] Jorda, O., Schularick, M., and Taylor, A. 2011. Financial crises, credit booms, and external imbalances: 140 years of lessons. *IMF Economic Review* 59: 340-378.

[27] Killick, R., Fearnhead, P., and Eckley, I. A. 2012. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* 107: 1590-1598.

[28] Killick, R., and Eckley, I. (2014). Changepoint: An R package for changepoint analysis. *Journal of statistical software* 58: 1-19.

[29] Küchenhoff, H., Günther, F., Höhle, M., and Bender, A. 2021. Analysis of the early COVID-19 epidemic curve in Germany by regression models with change points. *Epidemiology and Infection* 149: e68.

[30] Liu, S., Wright, A., and Hauskrecht, M. 2018. Change-point detection method for clinical decision support system rule monitoring. *Artificial Intelligence in Medicine* 91: 49–56.

[31] Lévy-Leduc, C., and Roueff, F. 2009. Detection and localization of change-points in high-dimensional network traffic data. *The Annals of Applied Statistics* 637-662.

[32] Page, E. 1954. Continuous inspection schemes. *Biometrika* 41: 100–115.

[33] Palarea-Albaladejo, J., and Martín-Fernández, J.A. 2008. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers and Geosciences* 34: 902-917.

[34] Palarea-Albaladejo, J., and Martín-Fernández, J.A. 2015. zCompositions – An R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* 143: 85-96.

[35] Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. 2015. Modelling and analysis of compositional data. John Wiley and Sons.

[36] Pourhabibi, T., Ong, K., Kam, B., and Boo, Y. 2020. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems* 133: Article 113303.

[37] R Core Team 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[38] Ranjan, C.; Mustonen, M., Paynabar, K., and Pourak, K. 2018. Dataset: rare event classification in multivariate time series. arXiv preprint arXiv:1809.10717

[39] Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. 2007. A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology* 46: 900-915.

[40] Rohatgi, V. 1976. An introduction to probability theory and mathematical statistics. Wiley, New York.

[41] Sammut, C., and Webb, G.I. 2017. Encyclopedia of machine learning and data mining. Springer, New York.

[42] Stock, J., and Watson, M. 1991. A probability model of the coincident economic indicators. In *Leading Economic Indicators: New Approaches and Forecasting Records*, edited by K. Lahiri and G. Moore. Cambridge University Press.

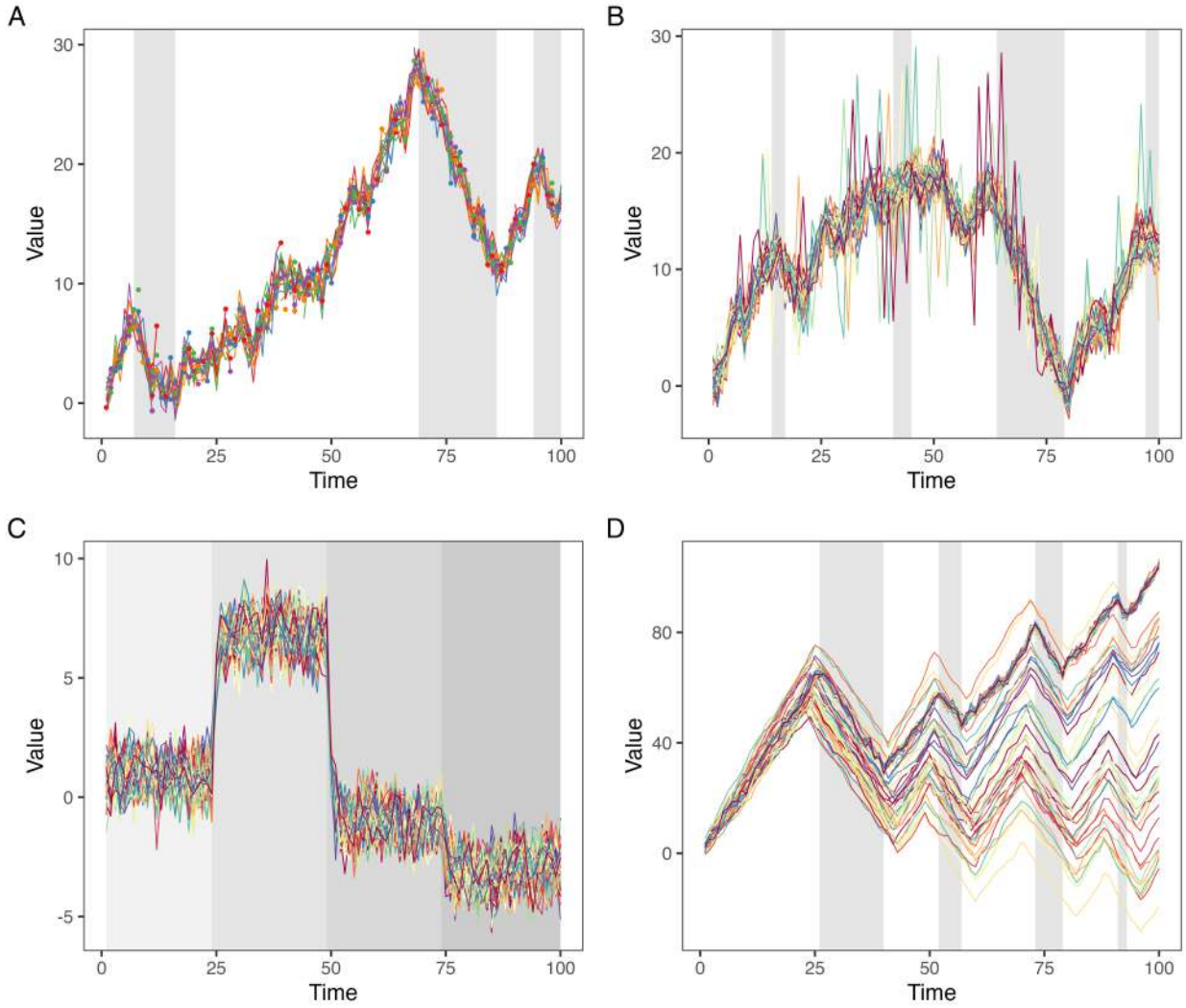
[43] Takens, F. 1981. Detecting Strange Attractors in Turbulence. In Dynamical Systems and Turbulence; Springer:Berlin/Heidelberg, Germany.

[44] Tamer, E. Introduction to pandemic econometrics/Covid-19 pandemic. *Journal of Econometrics* 2020: 1.

[45] Tartakovsky, A., Nikiforov, I., and Basseville, M. 2014. Sequential analysis: Hypothesis testing and changepoint detection. Chapman and Hall.

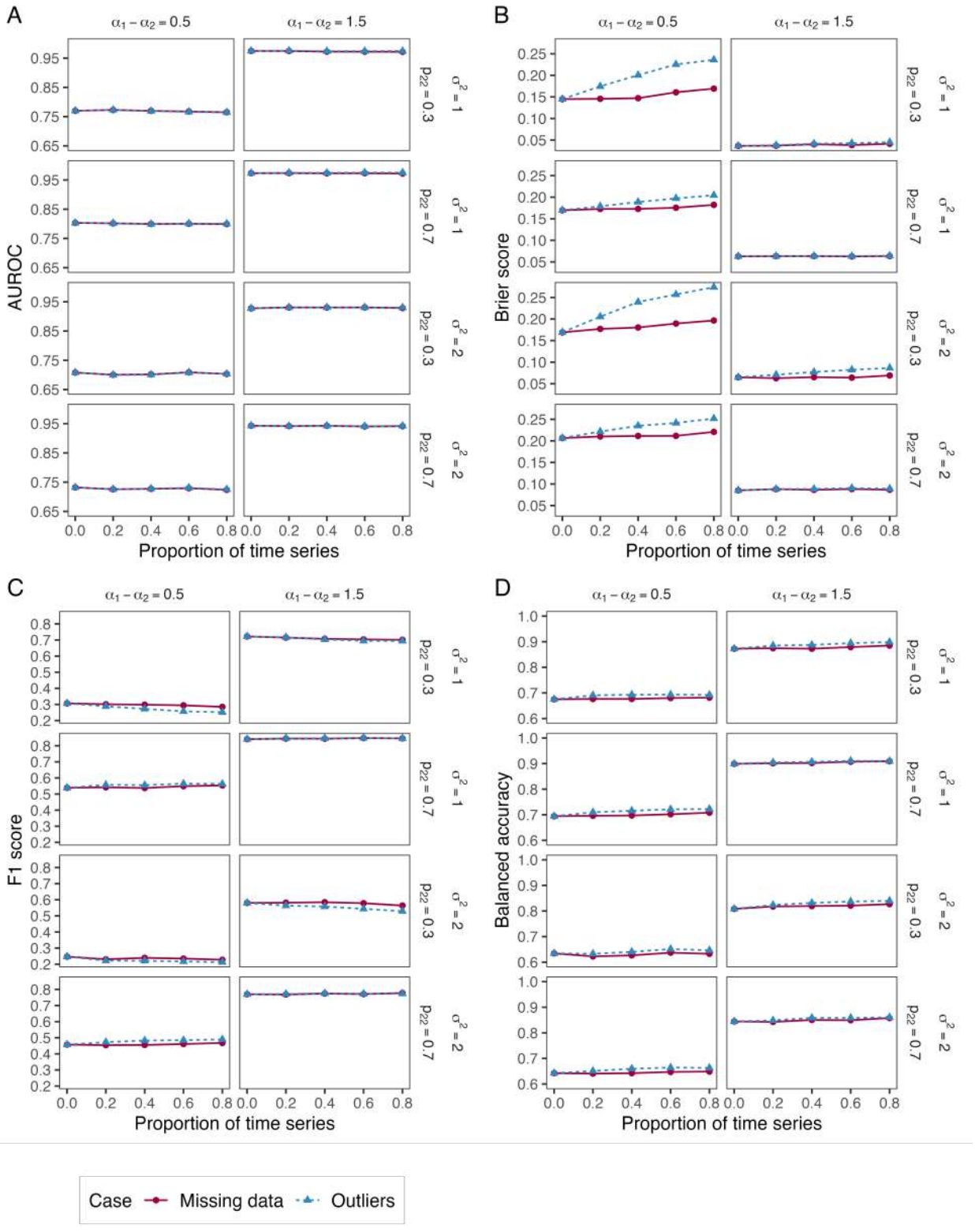
[46] Wang, W., et al. 2022. A systematic review of time series classification techniques used in biomedical applications. *Sensor* 22: Article 8016.

Figure 1: Time series simulation.



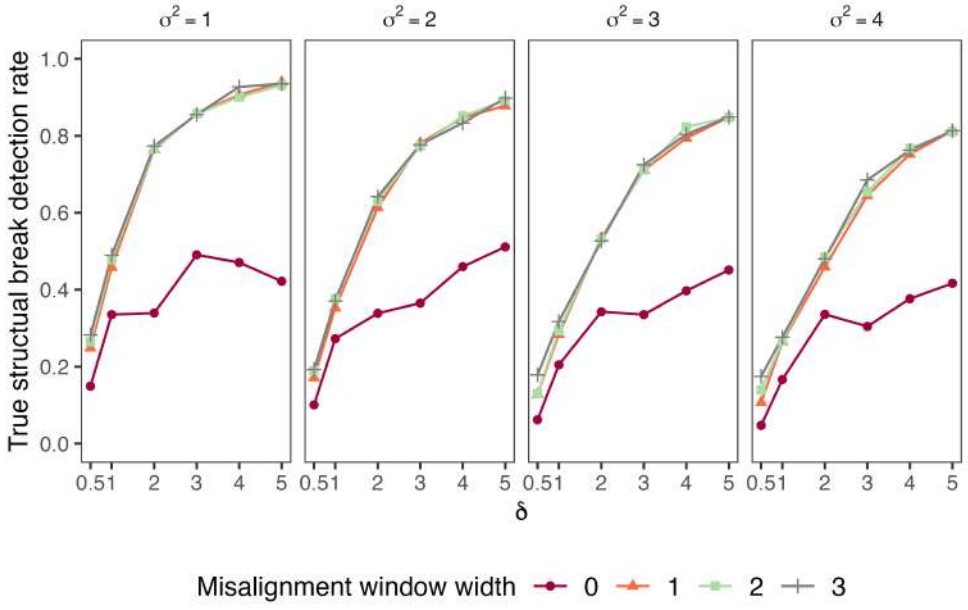
Notes. Exemplary simulated time series panels with different characteristics: missing data (A), outlying observations (B), structural breaks (C), and asynchronous time series (D). Shaded background areas indicate changing underlying regimes.

Figure 2: Performance assessment from DGP1 study.



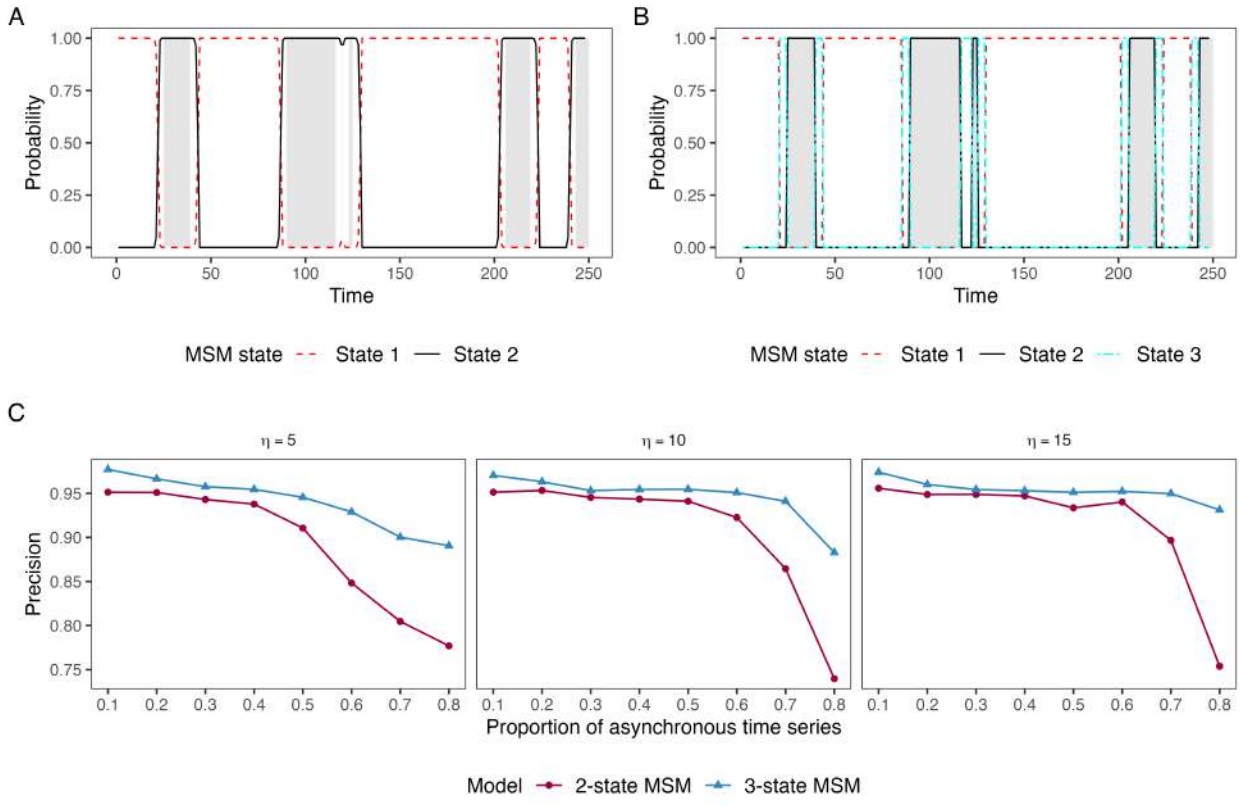
Notes. Based on area under the receiver operating characteristic curve (AUROC) (A), Brier score (B), F1 score (C), and balanced accuracy (D) for different parameter settings.

Figure 3: Performance assessment from DGP2 study.



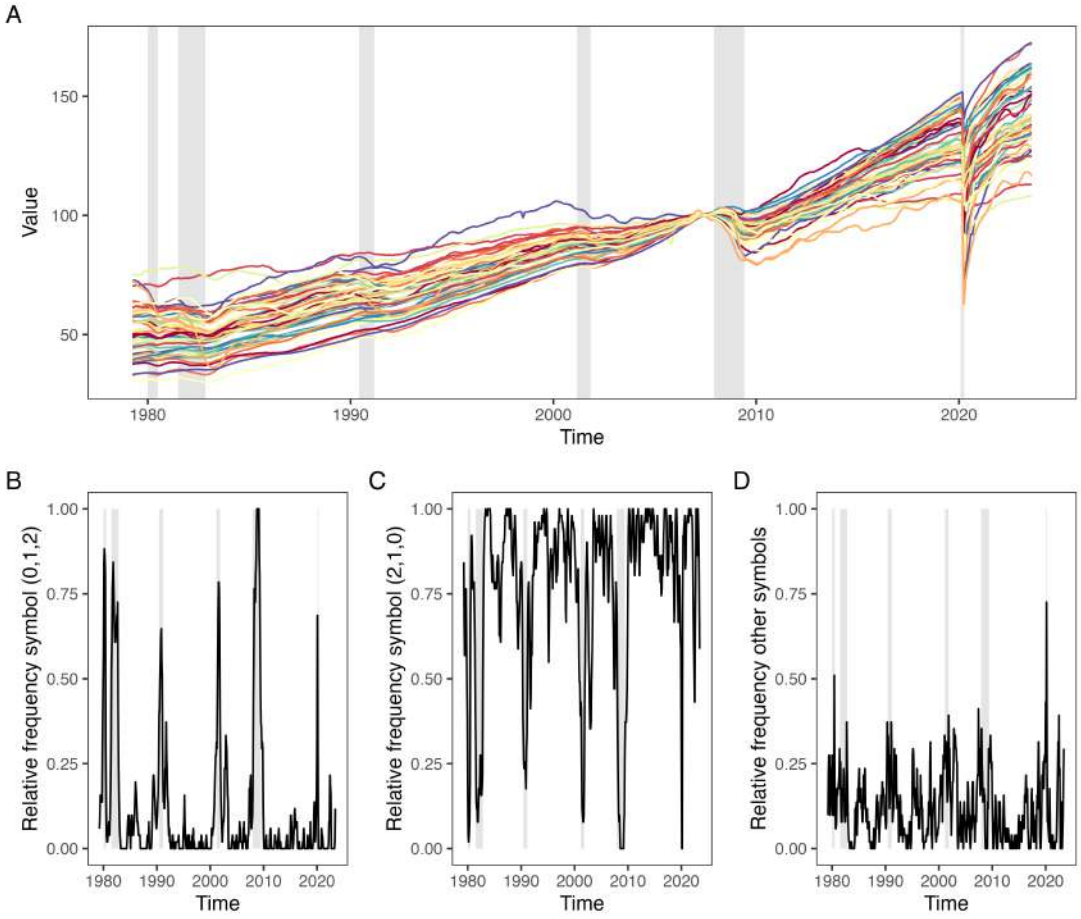
Notes. Based on detection of actual structural breaks for different parameter settings.

Figure 4: Performance assessment from DGP3 study.



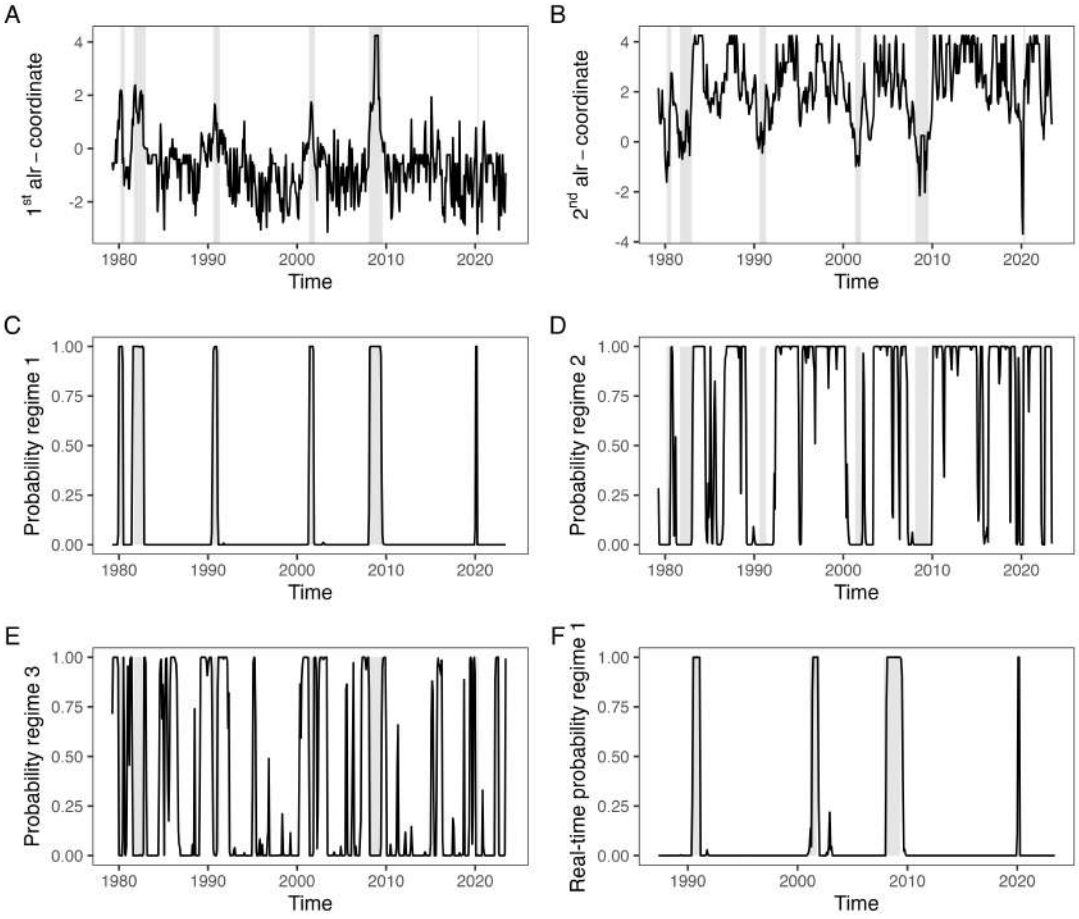
Notes. Based on regime probabilities from 2-state (A) and 3-state (B) Markov-switching model (MSM) fits for asynchronous time series. Precision in identifying the regime indicated by the shaded background across different levels of asynchronicity η (C).

Figure 5: US state coincident indexes 1974.04-2023.08.



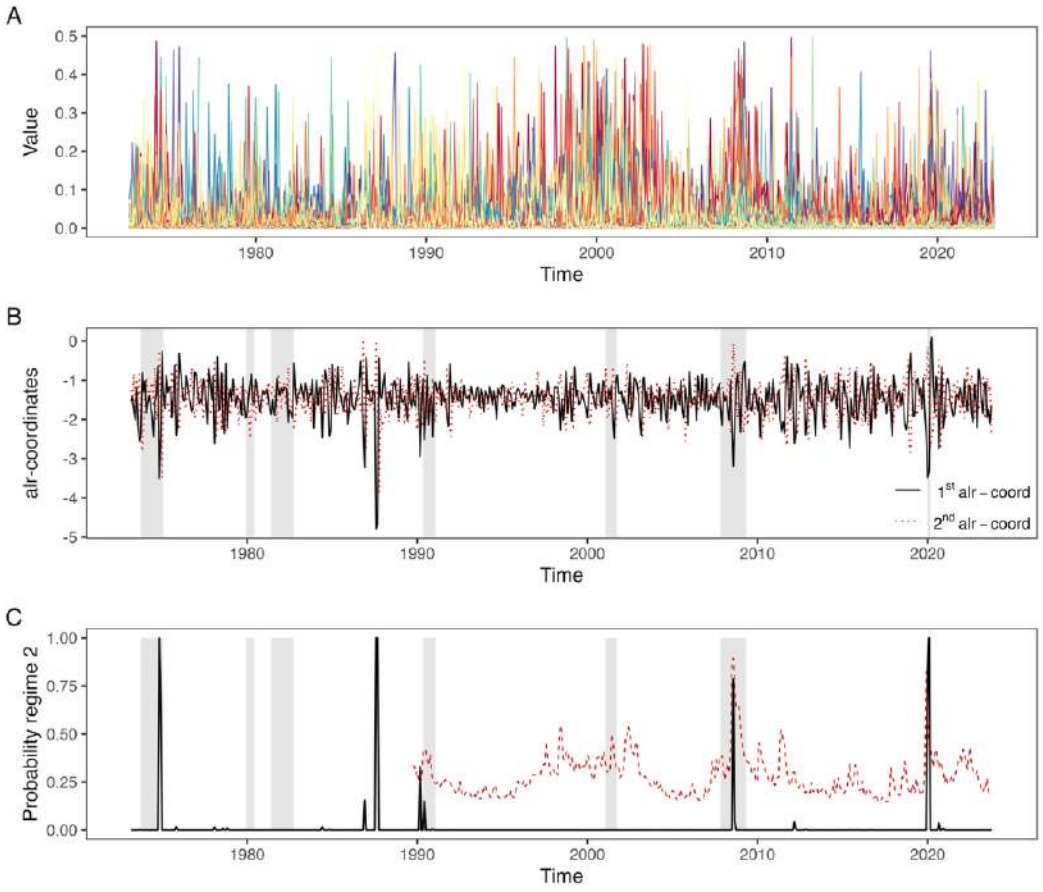
Notes. Time series of the 50 state coincident indexes provided monthly by the Federal Reserve Bank of Philadelphia (A). Relative frequency of symbols $(0,1,2)$ (B), $(2,1,0)$ (C) and any other symbols in $\{(0,2,1),(1,2,0),(1,0,2),(2,0,1)\}$ (D). The shaded areas in the background refer to the NBER-referenced recessions.

Figure 6: US state coincident indexes modeling and prediction.



Notes. Additive log-ratio coordinates based on the distribution of three symbols, $\ln \frac{\hat{P}_t(\Pi_1)}{\hat{P}_t(\Pi_3)}$ and $\ln \frac{\hat{P}_t(\Pi_2)}{\hat{P}_t(\Pi_3)}$ (A-B). Estimated probabilities for regimes 1, 2 and 3 (C-E). Real-time estimated probabilities for regime 1 (F). The shaded areas in the background refer to the NBER-referenced recessions.

Figure 7: Market values volatilities 1973.01-2024.01.



Notes. Time series of demeaned squared growth rates of the market values of assets comprising the S&P500 (A; y-axis truncated at 0.5 to enhance visualization). First and second alr-coordinates of 3-symbol distribution (B). Estimated probabilities for regime 2 from fitted model (solid line) and superimposed CBOE VIX volatility index (dashed line) (C). The shaded areas in the background refer to the NBER-referenced recessions.