Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 14-005

Discovering Groups of Time Series with Similar Behavior in Multiple Small Intervals of Time

Gowtham Atluri, Michael Steinbach, Kelvin Lim, Angus MacDonald, and Vipin Kumar

January 22, 2014

Discovering Groups of Time Series with Similar Behavior in Multiple Small Intervals of Time

G. Atluri * M. Steinbach * K. O. Lim † A. MacDonald III ‡ V. Kumar *

Abstract

The focus of this paper is to address the problem of discovering groups of time series that share similar behavior in multiple small intervals of time. This problem has two characteristics: i) There are exponentially many combinations of time series that needs to be explored to find these groups, ii) The groups of time series of interest need to have similar behavior only in some subsets of the time dimension. We present an Apriori based approach to address this problem. We evaluate it on a synthetic dataset and demonstrate that our approach can directly find all groups of intermittently correlated time series without finding spurious groups unlike other alternative approaches that find many spurious groups. We also demonstrate, using a neuroimaging dataset, that groups of intermittently coherent time series discovered by our approach are reproducible on independent sets of time series data. In addition, we demonstrate the utility of our approach on an S&P 500 stocks data set.

1 Introduction

Time series data has become increasingly ubiquitous during the last two decades in several domains including climate, bioinformatics, social media and neuroimaging [3, 13]. The data mining community has studied several problems pertaining to analyzing time series data [2, 13]. They include clustering [4, 19], classification [17], anomaly detection [6], forecasting [8], and segmentation [9]. The focus of this paper is to address the problem of discovering groups of time series that share similar behavior in multiple small intervals of time. We refer to such groups as 'intermittently coherent time series' in the rest of this paper.

In a complex dynamic system different groups of entities in the system may behave coherently for short intervals of time to achieve a specific objective. For example, in a human brain, a brain region can be treated as an entity and the amount of activity measured over time at a brain region could be treated as its behavior. Multiple brain regions are said to behave coherently for a short period of time when the time series of their activity levels become highly similar within this time period. Consider the hypothetical example shown in Figure 1, that depicts four time series each with 200 time points. These time series do not appear to be similar when all the 200 time points are considered. However, in the time intervals from 51 to 90 and from 141 to 180 they exhibit high similarity. If such time series represent activity levels of brain regions over time (measured using fMRI technology) the corresponding brain regions could be hypothesized to work together to accomplish a specific task [14].

The problem of discovering groups of intermittently coherent time series from a given time series data set has two characteristics: i) There are exponentially many combinations of time series that needs to be explored to find these groups, ii) The groups of time series of interest need to have similar behavior only in some subsets of the time dimension.

Pattern mining approaches that have been studied in the context of market basket data [1, 5] address these two characteristics directly. The goal of these approaches is to find groups of items that occur together in many transactions (i.e., they are frequent itemsets). These techniques explore the combinatorial nature of the search space in a systematic fashion relying on the Apriori principle [1] that guarantees that if an item set is frequent then all of its subsets are frequent too. However, these pattern mining approaches have been designed to work with binary features, that indicate whether an item is contained in a transaction or not. Recently, they have also been explored for continuous valued datasets [11], but there is no existing framework that works with time series data.

In this paper we generalize the well studied frequent pattern mining techniques to work with time series data in order to discover all groups of objects whose time series are intermittently coherent. Specifically we use a sliding window based approach and we propose the notion of support for time series data with a goal of capturing intermittent coherence for a candidate group of time series. Using this, we provide an Apriori based framework that can discover all groups of intermittently coherent time series such that the total length of coherent intervals for a group is longer than a given window-based threshold. We evaluate our approach on a synthetic dataset and show its effectiveness in discovering all the desired intermittently coherent groups in comparison to that of alternative approaches. We then show the utility of our approach on a real world neuroimaging dataset, where we demonstrate that our approach can be used to discover significantly reproducible groups from independent sets of time series data collected from the same set of subjects. On the same dataset, we show its effectiveness in comparison with an alternative approach. We also demonstrate the utility of our approach on an S&P500 weekly stock prices data set.

The following are the key contributions of this paper:

• A novel approach to quantify the duration of intermittent coherence for a given set of time series.

*Dept. of Computer Science, University of Minnesota
†Dept. of Psychiatry, University of Minnesota
‡Dept. of Psychology, University of Minnesota

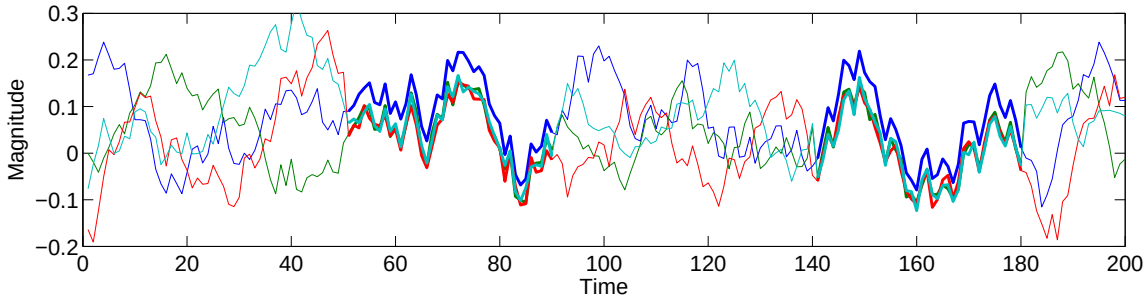

Figure 1: Four time series exhibiting intermittently coherent behavior. (All figures in this manuscript are best seen in color.)

• A systematic framework for discovering all groups of time series that exhibit intermittently coherent behavior.
• Comparative evaluation of the proposed approach with alternative approaches to demonstrate its effectiveness on synthetic and real world datasets.

This paper is organized as follows: In section 2, we formally define the problem. We present alternative approaches and the proposed approach in sections 3 and 4, respectively. In section 5, we present the evaluation of our approach on two real world datasets. We conclude with section 6.

2 Problem Formulation

Consider a set of observations made on n objects $\{I_1, I_2, \ldots I_n\}$ at m different time points $\{t_1, t_2, \ldots t_m\}$. Let the observations made on i^{th} object I_i be represented as a time series $d^i = (d^i_1, d^i_2, \ldots d^i_m)$. Let D be a matrix whose columns are the vectors $d^i, \forall i \in (1, \ldots, n)$. Consider a time window of length *window-length* ω that is moved across the time series in steps of size s. Our goal is to find those subset of objects $\{I_{j_1}, I_{j_2}, \ldots I_{j_p}\}$ such that the time series observed on these objects behave 'similarly' in at least a user provided number of windows. A number of different ways of characterizing "similarity" for time series have been studied in the literature [7, 13]. We will use Pearson's correlation as a measure of similarity between two objects for a given time interval in this paper. A given set of objects is deemed to behave similarly if the minimum of the pairwise correlation of all the time series obtained from these objects is above a user provided correlation threshold.

3 Alternative Approaches

To the best of our knowledge, there is no existing approach that can directly discover all groups of time series such that for every group there are sufficiently many time windows in which all constituent time series exhibit sufficiently high correlations among themselves. In this section we outline possible approaches that can help one find such groups.

Clustering of time series data is one way to determine groups of time series that are highly correlated. Traditional clustering approaches like k-means, hierarchical and density based clustering are often used with time series data sets by choosing an appropriate measure of similarity. Several similarity measures such as dynamic time warping, euclidean distance, and correlation have been studied in the literature [13, 7]. Note that these similarity measures have also been used to capture lagged relationships in the data which is not the focus of the problem that is being studied. Nevertheless, these techniques cannot capture the similarity (high correlation) in small time intervals, as they take into consideration the full time series available.

Frequent pattern mining techniques can be applied to time series data after binarizing the data using a suitable threshold. Consider a matrix D whose columns are the time series vectors d^i for every object i, and whose rows are time points. Using a binarization threshold this matrix can be converted to $D_{0/1}$ where an element takes a value 1, if its value in D is greater than the binarization threshold, and 0 otherwise. Frequent pattern mining on this data can explore all combinations of objects, but it is limited to capturing groups of objects whose value is beyond a threshold for a number of time points that is greater than a user provided threshold. This approach does not directly look for intermittently strong correlations, i.e., time intervals where the time series are highly correlated among them. Moreover, the binarization threshold based similarity cannot capture correlations in the full time series, let alone the intermittently strong correlations. Therefore the traditional binary pattern mining technique applied on a binarized version of time series data is not suitable to address the problem at hand.

Alternatively, one can use frequent pattern mining techniques on time series clusters obtained from sliding time windows. To achieve this, one can use a sliding time window of a chosen length and compute time series clusters within each window, by moving the window in steps of a predetermined size along the length of time series. These clusters can be used to construct a binary matrix CT, where each row is a cluster and each column is a time series. A value of 1 in the matrix indicates the presence of time series in the corresponding cluster. Frequent pattern mining can then be used on this CT matrix to find groups of time series that participate together in the same cluster for sufficiently many time windows. This approach has the potential to recover groups of time series that share high correlations in many windows. A challenge with this approach is that it is not trivial to determine the choice of number of clusters within each sliding window. One can construct a scenario where there are different number of clusters in different sliding windows and this approach will not perform well in such a case. Moreover,

in windows where there are no high correlations among the time series, this approach will find spurious clusters and so the resultant groups discovered could be potentially spurious.

4 Pattern Mining Framework

Discovering groups of time series that behave similarly for at least a given number of time points is a challenging problem. It requires searching through all combinations of objects as well as determining intervals in time at which the objects in question behave similarity. These challenges have been addressed in market-basket data sets by frequent pattern mining techniques. Market basket data captures the items purchased in a transaction in a binary data matrix X, whose columns are items in a market, and whose rows are transactions, and whose elements X_{ij} have a value 1 indicating the presence of an item j in a transaction i, and a value 0 otherwise. The goal of frequent pattern mining techniques is to discover all subsets of items (also referred to as itemsets) that are purchased "frequently". The ratio of the number of times a set of items are purchased together to the total number of transactions is treated as the *support* of an itemset. A user provided $support$ threshold is used to determine whether a given item-set is frequent. A transaction in which all the items in an itemset in question are present is said to "support" the itemset.

A standard pattern mining approach that is widely used with binary data sets is the Apriori algorithm [1]. At the heart of this approach is the Apriori principle that guarantees that if a set of items are not frequently purchased together, then any bigger set that includes this set is not frequent. This is due to the anti-monotonic nature of the $support$ measure, i.e., $support$ of a given set of items is less than equal to the $support$ of any of its subsets. Relying on this principle, the Apriori algorithm builds item sets bottom up, where it starts with all single items and filters out items that are not frequent. It then groups the frequent single items to enumerate candidate item-pairs and then evaluates them to select those pairs that are truly frequent. Then candidate item-triples are enumerated from the frequent pairs by joining the pairs that share one item and the frequent triples are determined by filtering out the infrequent ones from the candidate triples. In this fashion it constructs higher-order sets until no more bigger sets can be enumerated. Note that the higher order candidate itemsets are only enumerated from the frequent itemsets at a given level. This reduces the number of candidate itemsets effectively. By systematically pruning the search space of all possible combinations of items, this approach can efficiently discover all possible itemsets that are frequently purchased together.

4.1 Designing a notion of support for time series data

The key difference between market basket data and time series data is that in market basket data we have a binary vector (a column in X) for every item indicating its presence in each of the transactions, while in time series data we have a time series d^i with continuous values for an object I_i. In the case of market basket data, supporting transactions for a given set of items can be determined by computing the intersection of the transactions in which each of the individual items are present. This is not trivial with time series data. Moreover, the goal is to identify the intervals during which a high correlation is exhibited.

Here we use a sliding window based approach to compute coherence between time series for each window. Specifically, we choose a *window-length* ω to determine the duration of a window and to move the window across the time series in steps of size s. For example, the first window captures the time points $(t_1, \ldots, t_\omega)$ and the second window captures the time points $(t_{s+1}, \ldots, t_{s+\omega})$. We refer to each window using the index of the ending time point. For example, the first two windows are referred to as w_ω and $w_{s+\omega}$. For a given time series d^i of length m, using a choice of window length ω and a step size s, the set of windows is referred to as $w^i = (w^i_\omega, w^i_{\omega+s}, \ldots w^i_{\frac{m-\omega}{s}+1})$.

We treat each window as a transaction in traditional frequent pattern mining. To determine if a window supports a group of time series we need to estimate if the group of time series exhibit high coherence within this window. We perform this by computing the pairwise correlations between the time series for a given window. A window is said to support a group of time series if the minimum of the pairwise correlations is greater than a user-provided correlation threshold γ. The number of time windows that support a group of time series is referred to as $ts - support$. Formally, $ts - support$ for a set of objects $S \in \{I_1, \ldots I_n\}$ is defined as follows:

$$ts - support(S, \omega, s, \gamma) = \sum_{i=\omega}^{\frac{m-\omega}{s}+1} \mathbf{1}_{minpwc(S,w_i)\geq\gamma} \quad (4.1)$$

where $minpwc(S, w_i)$ is the minimum of the pairwise correlations between objects in the set S for the window w_i. $\mathbf{1}_{minpwc(S,w_i)\geq\gamma}$ is 1 when $minpwc(S, w_i)$ is greater than a user provided threshold γ, 0 otherwise. Note that the windows that support a given set of time series are the windows in which the given set exhibits sufficiently high correlations. Greater the $ts - support$ of a set of objects, longer is the duration of sufficiently high correlations among them.

We illustrate the notion of $ts - support$ with the help of an example shown in Figure 2. Here two time series are shown for which $ts - support$ needs to be estimated. The choice of window length $\omega = 30$, step size $s = 10$, and correlation threshold $\gamma = 0.8$ are used. In the first window w_{30} spanning $(t_1, \ldots t_{30})$ the time series has a correlation 0.6. The second window spans w_{40} spanning $(t_{11}, \ldots t_{40})$ and the two time series have a correlation 0.4 in this window. Similarly, for the third and fourth windows, w_{50} and w_{60}, the correlations are 0.82 and 0.83, respectively. Only the third and fourth windows, w_{50} and w_{60}, contribute to support as their correlation surpasses the γ threshold. Therefore, $ts - support$ for the time series in this example is 2.

Antimonotonicity of the $ts - support$ measure allows us to use the Apriori framework to enumerate all frequent groups of time series.

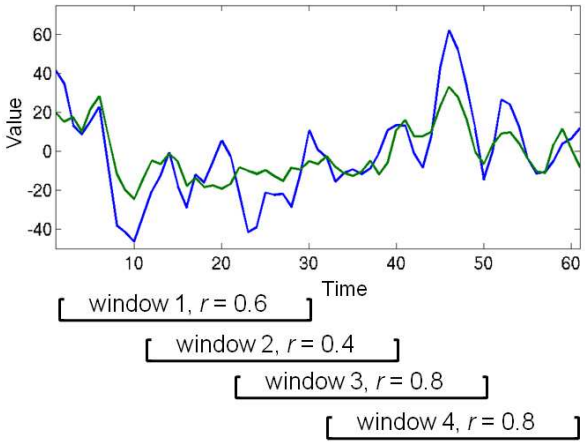

Figure 2: Example to illustrate the notion of $ts - support$ with $\omega = 30$, $s = 10$ and $\gamma = 0.8$.

4.2 Antimonotonicity of $ts - support$ We now prove that the $ts - support$ measure we defined above is anti-monotonic so it can be used in an Apriori like framework [1] to discover all subsets of time series that satisfy a given $ts - support$ threshold, γ.

THEOREM 4.1. *$ts - support(S, \omega, s, \gamma)$ measure decreases monotonically as new items are introduced for a given set of time series S, window length ω, step size s, and a pairwise correlation threshold γ.*

Proof. Consider a new set S', such that $S' = S \cup x$.

A window w_i that does not contribute to $ts - support(S, \omega, s, \gamma)$, i.e., $minpwc(S, w_i) < \gamma$, will not contribute to $ts - support(S', \omega, s, \gamma)$ because the minimum pairwise correlation $minpwc(S, w_i)$ will not increase as a new time series x is introduced to the set S.

A window w_i that contributes to $ts - support(S, \omega, \gamma)$, i.e., $minpwc(S, w_i) \geq \gamma$, will either contribute or not contribute to $ts - support(S', \omega, s, \gamma)$ depending on how the new time series x affects the minimum pairwise correlation. If $minpwc(S, w_i) \geq \gamma$ and $minpwc(S', w_i) \geq \gamma$, then $ts - support(S, \omega, s, \gamma) = ts - support(S', \omega, s, \gamma)$, otherwise $ts - support(S, \omega, s, \gamma) > ts - support(S', \omega, s, \gamma)$.

Therefore, $ts - support(S, \omega, s, \gamma) \geq ts - support(S', \omega, s, \gamma)$

4.3 Apriori-based approach for time series data Using the above notion of computing support from time series data we now describe a generalized Apriori algorithm that can work with time series data. First, we start with all pairs of objects and then evaluate their $ts - support$ to determine the pairs that are interesting. Note that the original Apriori starts with single items and determine frequent itemsets. Here we cannot filter at the first level because we need at least two time series to determine similarity and so we start by enumerating all pairs. Once the frequent pairs (i.e., pairs with $ts - support \geq \gamma$) are determined, we then enumerate the candidate triples as is done in a traditional Apriori algorithm [1] by joining interesting pairs that share one object. This approach continues until no more bigger frequent sets are found.
The algorithm is outlined here:

ALGORITHM 4.1. (TIME SERIES PATTERN MINING)
Input:
i. D, a real valued time series data matrix of size $|m \times n|$, where columns are items $I = \{I_1, I_2, \ldots I_n\}$ and rows are time points $T = \{t_1, t_2, \ldots t_m\}$
ii. σ, a support threshold
iii. ω, window length
iv. γ, minimum correlation threshold
Output:
All subsets of objects with $ts - support \geq \sigma$

1. $k = 2$
2. $CS_k = \{(I_i, I_j) | i \neq j, I_i \in I, I_i \in I\}$
3. **for** each candidate $cs_k \in CS_k$ **do**
4. compute $ts - support(cs_k, \omega, s, \gamma)$ using Eq. 4.1
5. **end**
6. $S_k = \{cs_k | cs_k \in CS_k \wedge ts - support(cs_k, \omega, s, \gamma) \geq \sigma\}$
7. **while** $S_k \neq \emptyset$ **do**
8. $k = k + 1$
9. $CS_k = Apriori - gen(S_{k-1})$
10. **for** each candidate $cs_k \in CS_k$ **do**
11. compute $ts - support(cs_k, \omega, s, \gamma)$ using Eq. 4.1
12. **end**
13. $S_k = \{cs_k | cs_k \in CS_k \wedge ts - support(cs_k, \omega, s, \gamma) \geq \sigma\}$
14. **end**
15. Result $= \bigcup S_k$

Step 2 enumerates all possible pairs, while steps 3-6 compute the support of a pattern and determine the frequent pairs that satisfy the support criteria, $ts - support(cs_k, \omega, \gamma) \geq \sigma$. Steps 7 through 14 enumerates candidates and determines frequent bigger patterns in an iterative way, until no bigger frequent patterns can be found.

4.4 Handling issues due to highly similar time series Note that in a given dataset there could be groups of time series that are correlated when all the time points considered. For example, in stocks data many stocks that belong to a given sector (e.g., health sector) could exhibit high correlations for the entire duration of time considered. These groups will have high value for our newly defined notion of support and will make it computationally hard to discover the low support patterns that are sufficiently correlated for a relatively shorter amount of time. To avoid finding these groups (that can be more easily found using alternate techniques), we add an additional constraint to our approach that discards any candidate set that has two objects I_i and I_j whose full time series d^i and d^j have a correlation that is greater than a user provided $full - corr - thresh$, before computing their support. This is achieved by filtering out such candidates immediately after the candidates are enumerated in steps 2 and 9 of Algorithm 4.1.

4.5 Handling artifacts due to globally similar behavior In many cases high correlations among all the time series in an interval can be induced due to a global event in the system. For example the 2007-2008 recession induces a similar

behavior in most of the stocks, and any windows that contribute to $ts - support$ in this period will inflate the support even though the event is not specific to the candidate pattern. Similarly, motion related artifacts create global patterns in neuroimaging data [12]. There is a need to control for windows that have such globally similar behavior from contributing towards the $ts - support$. One approach to address this challenge would be to discard all windows that capture a globally similar behavior and work with the remaining windows. Another approach is to weight the windows depending on how similar the behavior of a candidate set for a window is to the global behavior (e.g., correlation between mean time series for a candidate set with that of the entire set). In the context of market basket data this will be akin to developing a weighted version in which transactions that have too many items provide no support (former approach) or smaller support (later approach). We use the former approach and we show its utility in finding groups of time series that exhibit intermittent correlations not due to a global scenario in Section 5.3. This is achieved by ignoring those windows whose median of pairwise correlations between all the time series is greater than a $global - corr - thresh$ threshold. We incorporate this into our definition of $ts - support$ as follows:

(4.2)
$$ts - support(S, \omega, s, \gamma, global - corr - thresh) = \sum_{i=\omega}^{\frac{m-\omega}{s}+1} \mathbf{1}_{(minpwc(S,w_i)\geq\gamma)\&(mediangpwc(w_i)\leq globalcorrthresh)}$$

where $mediangpwc(w_i)$ is the median of the pairwise correlations between all objects in the set I for the window w_i.

5 Evaluation

Designing a thorough evaluation pipeline is a challenge for the problem at hand as is the case with many unsupervised algorithms. We used a synthetic dataset to highlight the key strength of the proposed approach and the weakness of competing approaches. The lack of ground truth in real world datasets limits us from directly comparing the groups of time series discovered using the proposed and the competing approaches. However, we performed a comparative evaluation the quality of the discovered groups. Using a neuroimaging time series data collected from same set of subjects at two different time points we studied the replicability of the findings which is necessary to test the validity of the results. In addition to this, we demonstrate the utility of our approach using a case-study on S&P stocks data.

5.1 Evaluation on a Synthetic Dataset

Data: We first created a random 400x10 matrix R, where rows are time points and columns are time series, by sampling each element from a uniform distribution with a range [0 1]. Each time series is further smoothed by computing the value at a time point t as the average of neighboring points from $t-5$ to $t+5$ to incorporate temporal auto-correlation that naturally exists in real world time series datasets, i.e., consecutive time points in a time series have similar values. We then impute four sets $\{(1,2,3,4),(5,6),(7,8),(9,10)\}$ of strong correlations for 120 time points (separate intervals of length 60 and 60). This is done for every set by copying the first time series for a chosen set of 60 contiguous time points in the other members of the set with a small amount of additive noise sampled from a Gaussian distribution with a mean of 0, and a standard deviation of 0.01. The four groups of time series are shown in Figure 3. The regions of time series shown in bold curves in each of these groups are the imputed highly correlated intervals that we expect the following approaches to capture.

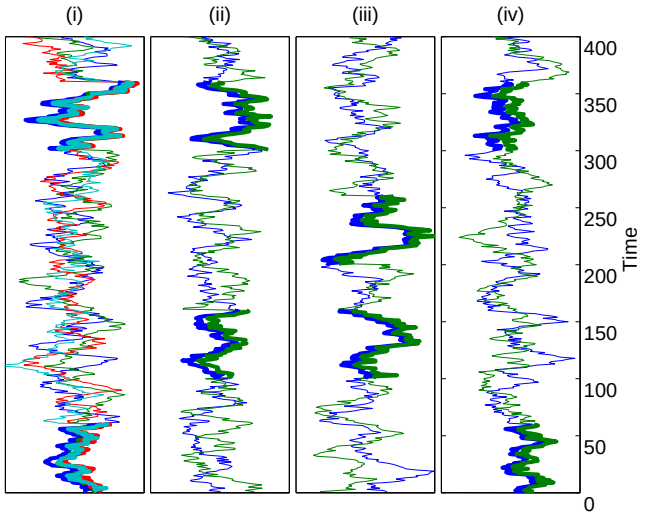

Figure 3: Four groups of synthetically generated intermittently correlated time series: (i) $\{1,2,3,4\}$ (ii) $\{5,6\}$ (iii) $\{7,8\}$ (v) $\{9,10\}$. Regions of the bold time series are the correlated intervals.

Approaches: We used three other competing approaches, in addition to the proposed approach:

1. *K-means clustering (K-means)*

We clustered the set of 10 time series into four clusters using correlation as a distance metric. We clustered them into four groups as the number of groups that were imputed was also four.

2. *Apriori on binarized time series (Apriori-$R_{0/1}$)*

We first constructed a binary matrix $R_{0/1}$ using a threshold on matrix R and then found maximal frequent patterns of time series using a support threshold. We considered the following choices of quantile based thresholds from the matrix R: $\{0.5,\ 0.55,\ 0.6,\ 0.65,\ 0.7\}$. A value in the matrix $R_{0/1}$ was 1, only if the corresponding value in R was above the chosen quantile based threshold. We used a support threshold of 60 for consistency in comparison with the other Apriori based schemes that are described below. For the sake of interpretability, we treat number of rows supporting a pattern (not fraction) as it's $support$ for Apriori based methods.

3. *Apriori on K-means clusters (K-means+Apriori)*

We used a sliding window of length 30 that is moved along the time series in steps of size 1. This resulted in 371 sliding windows. Within each window we considered the 10 time

Approach	Parameters	Recoverability	Spuriousness
K-means	k=4	0.25	0.5
Apriori -$R_{0/1}$ ($\sigma = 60$)	q = 0.5	0.25	0.98
	q = 0.55	0.25	0.86
	q = 0.6	0.5	0.81
	q = 0.65	0.25	0.57
	q = 0.7	0.25	0.20
K-means + Apriori ($\sigma = 60$)	k = 2	0.25	0.96
	k = 3	0.25	0.96
	k = 4	0.5	0.89
	k = 5	0.75	0.57
	k = 6	0.75	0.36
	k = 7	0.75	0
	k = 8	0.5	0
TS - Apriori ($\sigma = 60$)	$\gamma = 0.8$	1	0

Table 1: Comparison with competing approaches

series and clustered them into k clusters. Several choices of k were explored: $k = \{2, 3, \ldots 8\}$. Each cluster that has more than one member is then used to construct a binary CT matrix whose rows are clusters and whose columns indicate time series. A value of 1 in this matrix indicates that a time series was part of a cluster from the window in which it was discovered. We then found maximal frequent sets of time series that were part of more than 60 clusters. Note that every candidate set of time series can be supported by at most one cluster from a sliding window, because k-means clustering is partitional in nature.

4. *Time series pattern mining (TS-Apriori)*

We used a sliding window length $\omega = 30$, step size $s = 1$, minimum pairwise length threshold $\gamma = 0.8$, support threshold $\sigma = 60$.

The rationale for the choice of support $\sigma = 60$ in all the Apriori based approaches that work with sliding windows (Apriori+K-means, and TS-Apriori) was that each input group has two independent 60 time point long highly correlated intervals. With the chosen window length of 30, an interval of 60 time points will be visible in at least 30 sliding windows and together the two intervals (for a given group) will be visible for at least 60 windows. Therefore a support of 60 should suffice to discover all the imputed groups. Apriori-$R_{0/1}$ on the other hand does not use sliding windows and treats each time point independently. Therefore, a support of 60 is smaller than the sum of the duration of highly correlated intervals (120).

Comparison metrics: For each approach presented above, we evaluated two key factors: *recoverability* and *spuriousness*. Recoverability is the fraction of imputed groups that were discovered. Only when an imputed group is a subset of a discovered group, an imputed group is treated as a recovered group. Spuriousness is the fraction of discovered groups that were not imputed, i.e., those discovered groups that are not subsets of any imputed group. For an ideal approach, the recoverability is expected to be high (1) and the spuriousness is expected to be low (0).

Observations: The recoverability and the spuriousness of the groups/patterns discovered using the four approaches are shown in Table 1. For the full time series based approaches K-means and Apriori-$R_{0/1}$, the recoverability is poor and spuriousness is high. High spuriousness is mainly because they take the full time series into account for finding groups and low recoverability is due to fact that the locally high correlations are not apparent when correlation is assessed for the entire time series.

K-means+Apriori performs differently for different choices of k. When k is very small, the recoverability is very poor and the spuriousness is very high. This is because the clusters in each window are forced to be much bigger than the imputed groups and they support spurious patterns in the Apriori framework. When k is moderate ($k = 4, 5$), the recoverability increases, and spuriousness increases too. When k is high ($k = 6, 7$), the recoverability is relatively high, and spuriousness is relatively low. This is because the clusters become smaller as k increases. At the same time a high choice of k will not leave all the clusters intact, as it splits some real groups into smaller clusters. This is the reason recoverability is only as high as 0.75, for $k = \{4, 5, 6, 7\}$, and decreases to 0.5, for $k = 8$. In general, it can be noticed that more spurious groups are found when the choice of k is low, and some real groups are missed when k is high. Moreover, there are different number of imputed groups in different intervals. For example, from Figure 3 it can be seen that for the interval 301 to 360, there are three groups that are imputed, while there is only one group imputed in the interval from 201 to 260. Spuriousness could also be a result of windows where there are no imputed groups, where K-means is forced to find k groups in all windows. Therefore, using the same choice of k for all windows will not yield a recoverability of 1 and spuriousness of 0 in this synthetic dataset. Even in cases where same number of clusters are imputed in each window, choosing the right k is still nontrivial, as a high k will result in low recoverability and a low k will result in high spuriousness.

For the proposed approach, TS-Apriori, the recoverability is 1 and spuriousness is 0, which is the ideal scenario. This is mainly because it does not rely on clustering and it evaluates the relationship between candidate groups for each window independently and so it is able to recover all of the imputed groups without discovering any spurious groups.

5.2 Case study on Neuroimaging Data Functional Magnetic Resonance Image (fMRI) data measures the amount of oxygen consumed at every 2x2x2 mm cubic location in the brain (referred to as a voxel) and it is known to indicate the amount of activity occurring at any location. Data from an fMRI scan can be represented in the form of a time$\times$voxel matrix B, where every element Bv_{ij} in the matrix indicates the amount of neuronal activity occurring at a time point i and at a location represented by voxel j. We used the dataset from [18] that contains 6 minute resting state fMRI scans from 27 healthy subjects obtained at two different time points that are 9 months apart. We refer to the first set of scans from 27 subjects as Scan 1 data, and the second set as

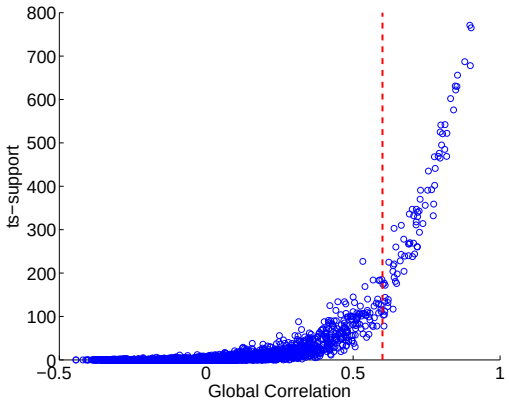

Figure 4: Comparison between pairwise global correlation and $ts-support$

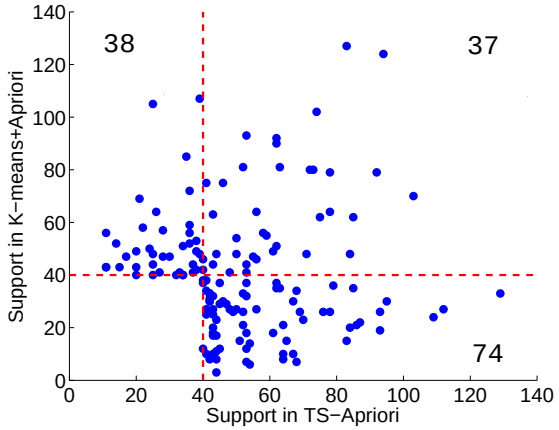

Figure 5: Patterns discovered using TS-Apriori and K-Means+Apriori

Scan 2 data. The spatial resolution of each fMRI scan was 2×2×2 mm and the temporal resolution was 2 seconds. Several prepossessing steps have been performed on the data obtained from the scanner and they have been elaborately discussed in [18]. In addition, following the approach in [10], global mean time series is regressed from the data, as is done in most fMRI studies. The resultant time×voxel matrix for each scan was of dimensions $180 \times 160,990$. We further group voxels into 90 brain regions based on an anatomical atlas provided by [15]. The resultant matrix, Br, for each scan was of size 180×90. We then appended the time series from each of the 27 scans from Scan 1 data to get a 4860x90 matrix. Similarly we appended the time series from Scan 2 data to get another 4860x90 matrix.

Out of the 90 brain regions, a few brain regions that are related to visual system of the brain are found to be consistently correlated in earlier studies [16]. These set of brain regions with highly correlated time series will introduce many high support patterns in our analysis and these patterns are uninteresting in our case as they can also be discovered using time series clustering techniques. In Figure 4 we show the global correlation and the corresponding $ts-support$ for all pairs of brain regions. The pairs of regions that are highly correlated ($r \geq 0.6$) have a $ts-support$ ranging from 300 to 800. The strength of our approach lies in finding groups of brain regions that exhibit similar behavior in multiple small intervals in time. Therefore, we use a $full-corr-thresh = 0.6$ to prune all those candidates that have a high $ts-support$ to directly find those interesting groups that are otherwise unknown.

We used the proposed TS-Apriori with window-length $\omega = 30$, $s = 5$, $\gamma = 0.7$, $\sigma = 40$ on Scan 1 appended time series data matrix and found 111 size-3 patterns. We also used K-means+Apriori, that is the best of the competing approaches from our evaluation using synthetic data, to discover intermittently correlated groups of time series from Scan 1 data, with $k = 30$ clusters in each window using parameters $\omega = 30$, $s = 5$, and $\sigma = 40$ that are same as those used with TS-Apriori. We discovered 75 size 3 patterns. The union of the 111 and 75 patterns discovered using TS-Apriori and K-means+Apriori approaches, respectively, results in 149 patterns and their support computed using the two approaches is compared in Figure 5. Note that the support in K-means+Apriori and the $ts-support$ in TS-Apriori can be compared, because both of them represent the number of windows that support a group of brain regions ($\gamma > 0.7$). Out of the 75 size 3 patterns discovered from K-means+Apriori, only 37 patterns have a $ts-support \geq 40$ (49.3%, approximately). This suggests that the remaining 50.7% patterns are spurious according to our objective of finding group of time series that exhibit similar behavior in at least a given number of time steps. These patterns are shown above the horizontal red dashed line indicating $support \geq 40$ and to the left of the vertical dashed red line indicating $ts-support \leq 40$. This spuriousness is mainly due to the poor quality of the clusters discovered, i.e., the minimum pairwise correlation of clusters is less than the γ threshold used in TS-Apriori. Figure 6 shows the relationship between the clusters and their quality ($minpwc$ measure) from the windows they were discovered from. The clusters whose $minpwc$ is greater than $\gamma = 0.7$ threshold are those that lie above the dashed red line, while those that have relatively poor $minpwc$ lie below the red line. The 50.7% spurious patterns are supported by these clusters that lie beneath the dashed red line in the figure.

One could argue that a smaller k can be used to ensure that all clusters have a $minpwc \geq \gamma$. However, a smaller k could potentially result in splitting naturally existing clusters in other windows into smaller clusters. Even at the choice of $k = 30$, K-means+Apriori only recovered 37 of the 111 TS-Apriori patterns, indicating that the recoverability is only 29.7% (along with spuriousness 50.7%). This is potentially due to the different number of natural groups that exist in different windows and so these groups cannot be recovered using a uniform k for all windows. On the other hand, our approach estimates the strength of correlation between the brain regions in a set using $minpwc$ measure and determines whether a window supports a pattern or not.

On Scan 2 dataset, using the same parameters as in Scan 1 dataset, we found similar observations where K-means+Apriori missed 54.5% (73 out of 134) of the patterns

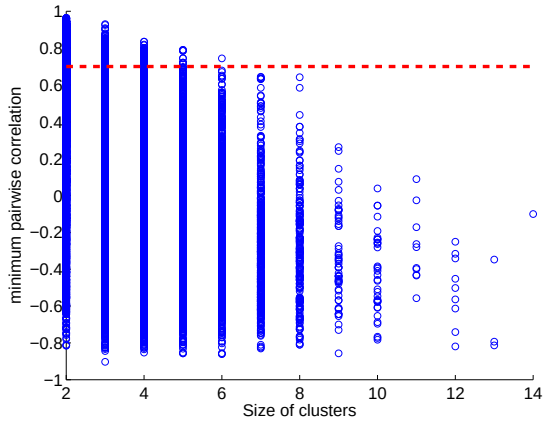

Figure 6: Relationship between cluster size and its quality

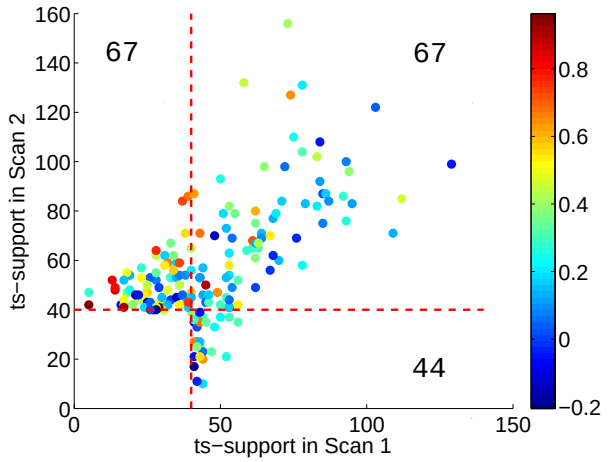

Figure 7: $ts - support$ of patterns found in Scan 1 and Scan 2 datasets

found by TS-Apriori and 21.8% (17 out of 78) of the patterns found by K-means+Apriori were spurious. As the recoverability and spuriousness of K-means+Apriori relies heavily on the choice of k, we tried several additional choices of k, including $k = 10, 20, 40, 50$. We found that spuriousness increases dramatically for lower choices of k, while very few of the TS-Apriori patterns were discovered for higher choices of k. These observations are similar to those demonstrated above using the synthetic dataset. These results highlight the limitations of the K-means+Apriori approach and the strengths of the proposed TS-Apriori approach on a real world dataset.

We further studied the similarity in the 111 and 134 patterns that were discovered from Scan 1 and Scan 2 datasets, respectively. In Figure7 we compare the $ts - support$ of the 178 patterns (union of 111 and 134 patterns) in Scan 1 and Scan 2 data. The color of each circle in this figure is the correlation between the number of windows contributed from 27 subjects in Scan 1 and Scan 2 datasets. There are 67 patterns that are common in the 178 patterns. This overlap is very significant given the large number of possible size-3 patterns ($\binom{90}{3} = 117,480$). Using a hypergeometric distribution we computed that the probability of expecting an overlap of 67 or more when 111 and 134 objects are drawn independently from a set of $117,480$ is less than 10^{-12}.

The correlations of contributions from subjects towards $ts - support$ (in Figure 7) are weak. The average of the correlation of contributions for the 67 patterns that are common is approximately 0.24. This is indicating that the contribution of subjects towards patterns is different in different scans, and that both the scans do not have same information about these patterns. This is inline with observations made by many studies that the reliability of the correlations between time series computed from two scans of the same subject are poor [16, 18]. Despite this weak similarity between scans, the fact that these patterns have high support in both the datasets suggests that an underlying neurological phenomenon could be driving these patterns.

5.3 Case Study on Stock Market Data We obtained the weekly closing stock prices of S&P500 companies over a 10-year period from January 2000 to December 2009 (521 weeks) from Yahoo! Finance website. We then removed those companies from this list for which only part of the data (less than 521 weeks) was available. We were left with 443 companies for which stock prices were available for all the 521 weeks. As the stock prices are at different scales, we normalized each time series d^i such that

$$d\_new_t^i = \frac{d_t^i - min(d^i)}{max(d^i) - min(d^i)} \tag{5.3}$$

where, d_t^i is the original stock price of stock i at time t, and $min(d^i)$ and $max(d^i)$ are the minimum and maximum stock prices of stock i, respectively.

Discovering groups of companies that exhibit strong correlations in small intervals from a span of 10 years could reveal novel direct or indirect relationships among companies. We found that this stocks data has two key characteristics that can lead to the discovery of uninteresting patterns: i) Two stocks that belonged to the same industry generally showed very strong correlation during the 10 year period. For example, stocks APA and APC that belong to oil and gas industry have a correlation of 0.95, approximately. Such groups can be directly discovered using traditional clustering based schemes and are uninteresting for our purpose. ii) Certain incidents affect all the stocks, e.g., the mortgage crisis, and so contribution of such windows towards $ts - support$ may lead to spurious and uninteresting patterns. Our approach addresses the first problem by building candidates involving those companies whose minimum of 10 year pairwise correlations is less than 0.6 ($full - corr - thresh$). The second problem is addressed by discarding the windows where the median of pairwise correlations for all companies is $mediangpwc$ is beyond 0.6. Under these conditions, using our time series pattern mining approach we found all groups of companies that share high correlations in at least $\sigma = 80$ time windows, using $\omega = 30$, $s = 2$, and $\gamma = 0.8$. There were 2965 size-2 patterns and 41 size-3 patterns.

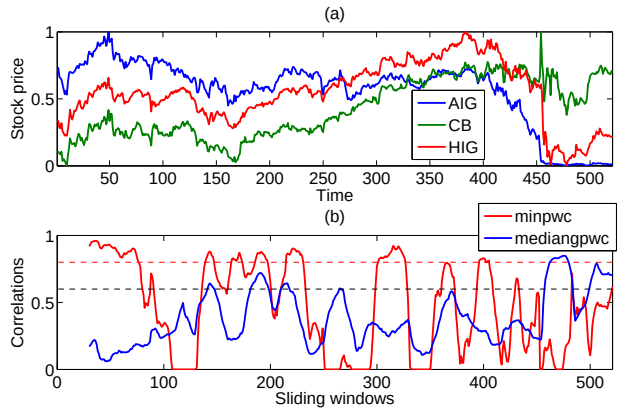

Figure 8: A selected Apriori-TS pattern generated from Stocks data set.

Figure 8(a) shows one group of three financial sector companies American International Group (AIG), The Chubb Corporation (CB), and Hartford Financial Services Group (HIG) that was discovered in our analysis. In Figure 8(b) we show the minimum of pairwise correlation ($minpwc$) among these companies for each window using a red-colored curve. The horizontal dashed line in red indicates the γ threshold used to determine the windows that contribute to the $ts - support$. The $minpwc$ curve is above the γ line for windows that end in the time points from 30 to 75, 140 to 150, 160 to 170, 185 to 195, 210 to 230 and 395 to 405, suggesting that these stocks are highly correlated in these windows. It is interesting that these companies, despite belonging to the same sector, exhibit relatively weak correlations for more than half the time. The blue curve in Figure 8(b) indicates the median of the pairwise correlations among all companies in each sliding window ($mediangpwc(w_i)$). Note that for windows ending in time points 145 to 150, 175 to 200, 210 to 220, and 460 to 470 this curve crosses the $global - corr - thresh = 0.6$ threshold, suggesting that almost all of the companies exhibit similar behavior in these windows. Overall, 82 of the 492 sliding windows are discarded.

The three stocks AIG, CB, and HIG that belong to the finance sector are expected to behave similarly for the entire duration. However, from Figure 8(b) it can be seen that during the first 80 weeks starting from the January 2000 they share a strong relationship. As time progresses, this relationship deteriorates and resurfaces due to several events that punctuate the time series. In period 2004-2005 (250 to 300 time points) AIG faced civil actions from regulatory authorities and later reached a settlement. AIG and HIG were hit by the financial crisis that occurred in late 2008 (400 to 450 time points). These events have impacted the stock prices and so they deviated from the other stocks with which they exhibited similar behavior at the beginning of the decade. The proposed approach allows one to discover such groups of intermittently correlated time series.

6 Conclusion and Future Work

In this paper we presented a pattern mining based approach for discovering groups of time series that exhibit strong intermittent correlations. We have shown, using a synthetic dataset, that the proposed approach is more suited to this problem than the competing approaches. Our approach is guaranteed to discover all groups given a support threshold. We also demonstrated the reproducibility of the groups found in fMRI data using two independent sets of scans obtained from the same cohort of subjects. Using the same dataset, we also demonstrated that the proposed approach directly searches for the desired groups and so it is effective in discovering them in comparison to alternative approaches. We also show the utility of the proposed approach on S&P 500 stocks dataset.

A number of aspects of the proposed framework need further investigation. The sliding window based support is a surrogate to measure the extent of time for which a candidate set of time series exhibit high correlations and it does not always accurately reflect the duration. Consider two time series that exhibit high correlation in two non-overlapping windows. Consider another example where the two time series exhibit high correlation in successive and overlapping windows. Although the $ts - support = 2$ for both these examples, the total duration of the strong correlation in the first case can be approximately twice that of the second, when the step-size is small. To address this issue, approaches that can directly capture the time intervals in which a given set of time series are highly correlated needs to be explored. The frequent pattern mining framework introduces challenges in the context of noisy data, high dimensional nature of the data, and continuous-valued nature of time series correlations. Existing pattern mining approaches that address these challenges needs to be investigated for their use in time series data.

Acknowledgements This work was supported by NSF Grant IIS-1355072.

References

[1] Rakesh Agrawal et al. Fast algorithms for mining association rules. In *VLDB*, volume 1215, pages 487–499, 1994.

[2] Alessandro Camerra et al. iSAX 2.0: Indexing and mining one billion time series. In *ICDM*, pages 58–67, 2010.

[3] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 2011.

[4] Xianping Ge et al. Deformable markov model templates for time-series pattern matching. In *SIGKDD*, 2000.

[5] Jiawei Han et al. Frequent pattern mining: current status and future directions. *DMKD*, 2007.

[6] Eamonn Keogh et al. Finding surprising patterns in a time series database in linear time and space. In *SIGKDD*, 2002.

[7] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *DMKD*, pages 349–371, 2003.

[8] Chiung-Hon Leon Lee et al. Pattern discovery of fuzzy time series for financial prediction. *TKDE*, 2006.

[9] Jessica Lin et al. Experiencing SAX: a novel symbolic representation of time series. *DMKD*, 2007.

[10] Kevin Murphy et al. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage*, 2009.

[11] Gaurav Pandey et al. An association analysis approach to biclustering. In *ACM SIGKDD*, pages 677–686, 2009.

[12] Jonathan Power et al. Methods to detect, characterize, and remove motion artifact in resting state fmri. *NeuroImage*, 2013.

[13] Chotirat Ann Ratanamahatana et al. Mining time series data. In *Data Mining and Knowledge Discovery Handbook*. 2010.

[14] Hugo Spiers et al. Decoding human brain activity during real-world experiences. *Trends in cognitive sciences*, 2007.

[15] N Tzourio-Mazoyer et al. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 2002.

[16] Jin-Hui Wang et al. Graph theoretical analysis of functional brain networks: test-retest evaluation on short-and long-term resting-state functional mri data. *PLoS One*, 2011.

[17] Li Wei and Eamonn Keogh. Semi-supervised time series classification. In *SIGKDD*, pages 748–753, 2006.

[18] Krista M Wisner et al. Neurometrics of intrinsic connectivity networks at rest using fmri. *NeuroImage*, 2013.

[19] Jesin Zakaria et al. Clustering time series using unsupervised-shapelets. In *ICDM*, 2012.