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 18-003

Finding Novel Multivariate Relationships in Time Series Data: Applications to Climate and Neuroscience

Saurabh Agrawal, Michael Steinbach, Daniel Boley, Stefan Liess, Snigdhansu Chatterjee, Vipin Kumar, Gowtham Atluri

February 12, 2018
Mining Novel Multivariate Relationships in Time Series Data: Applications to Climate and Neuroscience

Saurabh Agrawal*   Michael Steinbach*   Daniel Boley*   Stefan Liess†
Snigdhansu Chatterjee ‡   Vipin Kumar*   Gowtham Atluri§

February 2018

Abstract

In many domains, there is significant interest in capturing novel relationships between time series that represent activities recorded at different nodes of a highly complex system. In this paper, we introduce multipoles, a novel class of linear relationships between more than two time series. A multipole is a set of time series that have strong linear dependence among themselves, with the requirement that each time series makes a significant contribution to the linear dependence. We demonstrate that most interesting multipoles can be identified as cliques of negative correlations in a correlation network. Such cliques are typically rare in a real-world correlation network, which allows us to find almost all multipoles efficiently using a clique-enumeration approach. Using our proposed framework, we demonstrate the utility of multipoles in discovering new physical phenomena in two scientific domains: climate science and neuroscience. In particular, we discovered several multipole relationships that are reproducible in multiple other independent datasets, and lead to novel domain insights.

1 Introduction

In many domains, understanding the relationships between time series that represent activities recorded at different nodes of a highly complex system is essential for obtaining actionable insights. For instance, in climate science, pressure dipoles that refer to pairs of locations with strong negative correlations in their Sea Level Pressure time series have been vastly studied, and have been linked to anomalous weather events all over the globe [8, 15, 18]. Similarly, in neuroscience, researchers have identified pairs of brain regions with positive correlations that are highly discriminating between healthy subjects and those suffering from mental disorders [3].

In this work, we propose to study a novel class of linear relationships, named as multipoles, that could potentially involve more than two time series (variables). The basic idea of multipoles is to obtain sets of variables that show ‘strong’ linear interdependence, such that each variable makes a crucial contribution to the linear interdependence, i.e. elimination of any of the variables from the set should significantly weaken the strength of linear relationship among the remaining variables.

To demonstrate the concept of multipoles with a real-world example, consider a set $S$ of three time series (variables) $T_1, T_2,$ and $T_3$, as shown in Figure 2(a) observed at three different roads in Minnesota as shown in Figure 1. The traffic volume at three roads show a strong linear interdependence. To quantify it, we consider all normalized linear combinations of the given three time series and choose the one with minimum variance. The variance of such a linear combination could then be treated as the inverse indicator of linear dependence, i.e. lower variance implies higher linear dependence and vice-versa. Note that the traditional notion of linear dependence (as defined in linear algebra) will correspond to the case where there exists a linear combination with zero variance. In the above case, we obtained a linear combination $Z = 0.6T_1 + 0.65T_2 - 0.47T_3$, shown in top panel

---

*Department of Computer Science, University of Minnesota
†Department of Earth Sciences, University of Minnesota
‡School of Statistics, University of Minnesota
§Department of EECS, University of Cincinnati
of Figure 2(b) with variance is as low as 0.08. Hence, we could say that there exists a strong, if not perfect, linear interdependence among the variations in traffic volumes of the three roads in \( S \). A probable explanation for this pattern could be drawn based on basic law of conservation of flow. Two of the time series \( T_1 \) and \( T_2 \) were observed at the roads that act as major tributaries to the highway, where \( T_3 \) is being observed. All the southbound traffic coming from the tributaries is likely to merge at the highway, which leads to strong linear interdependence among the three time series.

Further, note that each of the three roads form the crucial component of the relationship, as excluding any one of them will significantly weaken the strength of the linear dependence among the remaining two time series. For example, let us exclude \( T_3 \) from the set. Then, the variance of the least variant normalized linear combination \( Z_{12} \) of \( T_1 \) and \( T_2 \) turn out to be 0.33, which is much higher than 0.08 of \( Z \), thus indicating significant contribution from \( T_3 \). Similarly, if we exclude \( T_2 \) or \( T_3 \) instead of \( T_1 \), the least variances go as high as 0.58 and 0.74 respectively, as shown in bottom two panels of Figure 2(b). The condition of crucial contribution from all the included variables is necessary to discard several trivial and uninteresting linear patterns that might include irrelevant or redundant variables.

The concept of multipoles has a strong relevance to multiple scientific domains including climate science and neuroscience. As we discuss further in section 5, we found several novel relationship patterns in global Sea Level Pressure datasets from climate science and brain functional-Magnetic Resonance Imaging (fMRI) datasets from neuroscience domain.

Our paper makes several key contributions: 1) We formally define the notion of a multipole relationship which has never been studied before in the field of pattern-mining. 2) We devise measures to quantify the interestingness of a multipole and formulate a novel pattern-mining problem to mine all interesting multipoles in a time series data. 3) Finding all interesting multipoles is computationally challenging owing to the combinatorial search space. Moreover, as we discuss later in section 3,
none of the existing approaches in the literature (including factor-analysis and regression based approaches) are suitable to find all multipoles. To this end, we propose a novel and computationally efficient Clique Based Multipole Search approach to find all interesting multipoles. Our approach is based on key empirical observations that can be leveraged to pre-prune the search space of multipoles to a much smaller family of subsets. 4) Furthermore, we propose an empirical framework to evaluate discovered multipoles that include an empirical procedure to assess the statistical significance of multipoles. 5) Using our proposed framework, we demonstrate the high relevance of multipoles to two scientific domains including climate science and neuroscience. In particular, we found several novel relationship patterns that are found to be reproducible in multiple other independent datasets, that further adds to their reliability and making a compelling case for their links to the underlying physical processes, some of which might be currently unknown to the domain scientists, and could be discovered using our framework.

2 Definitions and Notations

Let $S$ denote a set of $k$ time series (variables) $\{X_1, X_2, ..., X_k\}$ observed over $T$ consecutive timestamps and $X$ be the corresponding $T \times k$ data matrix. Let $\Sigma = X^T X$ be the $k \times k$ covariance matrix. For the sake of brevity, unless otherwise noted, we will assume that all the variables are standardized (zero mean and unit variance). Hence, the correlation matrix and covariance matrices will be exactly the same for $X$. We next define a few measures on a set of variables which will be eventually used to formally define a multipole.

**Definition 1.** A *Normalized Linear Combination (NLC)* refers to a linear combination with normalized weights. Specifically, for a given vector $l \in \mathbb{R}^k$ with $||l||_2 \neq 0$, a normalized linear combination of variables in the set, $S$, can be obtained as $Xl / ||l||_2$.

**Definition 2.** Given a set of variables $S = \{X_1, X_2, ..., X_k\}$, $Z^*_S$ refers to the Least Variant Normalized Linear Combination (LVNLC) of variables that has least variance across the $T$ observations. Formally, $Z^*_S = Xl^* / ||l^*||_2$

where

$$l^* = \arg\min_{l \in \mathbb{R}^k, ||l||_2 \neq 0} \text{var}(\frac{Xl}{||l||_2})$$

From the basics of linear algebra, we know that a set $S = \{X_1, X_2, ..., X_k\}$ of variables is said to be linearly dependent if there exists a linear combination of the variables that is constant across all the observations. Such a linear combination could be obtained as the LVNLC of the variables in $S$. In this case, the variance of LVNLC across all observations will be exactly zero. In the other extreme case, when all the variables are mutually orthogonal to each other, the variance of LVNLC will be equal to 1. Thus, variance of LVNLC could be used as an inverse indicator of the strength of linear dependence. Based on this observation, we define linear dependence of a set as follows:

**Definition 3 (Linear Dependence):** The linear dependence of a set $S$, denoted by $\sigma_S$, is given by $1 - \text{var}(Z^*_S)$.

**Relation between $\sigma_S$ and least eigenvalue of correlation matrix:** Notice that on performing the eigenvalue decomposition of correlation matrix $\Sigma$ of the variables in the set, the eigenvalues so obtained are equal to the variances of the projections of the data along their corresponding eigenvectors. Since $Z^*_S$ corresponds to the direction of least variance, $Z^*_S$ is nothing but the eigenvector corresponding to the least eigenvalue $\lambda_{min}$ of $\Sigma$. Thus, the variance of $Z^*_S$ is exactly equal to the least eigenvalue of $\Sigma$ and therefore,

$$\sigma_S = 1 - \lambda_{min}$$

Before proceeding further, the following two properties of linear dependence are noteworthy:
**Lemma 1.** For any set \( S \) of standardized variables, \( \sigma_S \in [0, 1] \).

**Proof.** Since we assumed all the variables to have unit variance, the sum of all the eigenvalues of the covariance matrix is exactly equal to \( k \), the number of variables. Therefore,

\[
\lambda_{\min} \leq 1 \quad (2)
\]

Further, since \( \Sigma \) is positive semi-definite, we get

\[
\lambda_{\min} \geq 0 \quad (3)
\]

Therefore, combining Eqs (1), (2) and (3), we get

\[
\sigma_S \in [0, 1]
\]

**Lemma 2.** The linear dependence of a set \( S \) is always less than or equal to that of its supersets.

**Proof.** Consider a superset \( S' = S \cup X_{k+1} \) of \( S \) and the corresponding data matrix be \( X' = [X_1, X_2, X_3, ..., X_{k+1}] \). Let \( l^* = [l^*_1, l^*_2, ..., l^*_k] \) denote the linear coefficients corresponding to the LVNLC \( Z_S^* \). Consider \( Z_{S'}^* = X'l' \), where \( l' = [l^*_1, l^*_2, ..., l^*_k, 0]^T \). Then clearly,

\[
\text{var}(Y_{S'}) = \text{var}(Y_S^*)
\]

Let \( Z_{S'}^* \) be the LVNLC for set \( S' \). Then, by definition of LVNLC and the above equality, we get

\[
\text{var}(Z_{S'}^*) \leq \text{var}(Z_{S'}^*) = \text{var}(Z_S^*)
\]

Therefore,

\[
\sigma_{S'} = 1 - \text{var}(Z_{S'}^*) \geq 1 - \text{var}(Z_S^*) = \sigma_S
\]

Although linear dependence indicates a strong relationship among the variables, it does not exclude the presence of irrelevant variables in the set. For instance, let \( S = \{X_1, X_2, X_3, X_4\} \) be a set of linearly dependent variables with the linear relation being \( X_1 + X_2 + X_3 + 0X_4 = 0 \). Although the four variables are linearly dependent, \( X_4 \) is an irrelevant variable and can be pruned from \( S \) without weakening the linear dependence among remaining variables. Hence, to avoid irrelevant variables in the pattern, we next propose a measure called **linear gain** that checks the minimum contribution from all member variables to the linear dependence of the set.

**Definition 4 (Linear Gain):** The linear gain of a set \( S \) with \( |S| > 2 \) is measured as the gain in the linear dependence of \( S \) with respect to its subset with strongest linear dependence. Mathematically, we can write linear gain of \( S \) as

\[
\Delta \sigma_S = \sigma_S - \max_{S' \subset S} \sigma_{S'} \quad (4)
\]

From Lemma 2, we get that the linear dependence of a set is always greater than that of its subsets, which implies that the linear gain of a set will always be positive. Furthermore, we can say that the subset with strongest linear dependence will be of size \( |S| - 1 \). Thus, the linear gain can be more precisely written as

\[
\Delta \sigma_S = \sigma_S - \max_{X_i \in S} \sigma_{S - X_i} \quad (5)
\]
Higher values of linear gain imply that a significant drop in linear dependence would be observed if any one of the variables are excluded from the set, thereby ensuring that no irrelevant variables are included. Furthermore, a high threshold on linear gain will avoid redundancies in the set. For instance, in the above example, if we add \(X_4\) as duplicate of \(X_1\), then the resultant set \(\{X_1, X_2, X_3, X_4\} \) would have zero linear gain since its linear dependence is same as one of its subsets \(\{X_1, X_2\} \). Similarly, higher threshold on linear gain will also avoid grouping two distinct patterns into a single pattern. For instance, consider a set \(S = S_1 \cup S_2\) that consists of two independent subsets of perfectly linearly dependent variables \(S_1\) and \(S_2\). By definition, the linear gain of such a set \(S\) will be 0, and hence will be discarded.

Using the above definitions, we next present the formal definition of a multipole.

**Definition 5.** A multipole refers to the set \(S\) of variables with \(|S| \geq 2\) such that \(\sigma_S \geq \sigma\) and \(\Delta \sigma_S \geq \delta\), where \(\sigma\) and \(\delta\) are user-specified thresholds.

We next define the notion of maximality in a multipole.

**Definition 6 (Maximal Multipole).** In a set \(Q\) of multipoles, a multipole \(S\) is considered to be maximal if none of its supersets are in \(Q\).

A maximal multipole is likely to capture the underlying signal more comprehensively compared to its subsets. Hence, all non-maximal multipoles could potentially be pruned in the final output of the search.

## 3 Related Work

Relationships in time series data have been analyzed in past using correlation networks, where each time series represent a node, and the weight of an edge between any two nodes represents the strength of the linear correlation between the corresponding time series. Correlation networks have been used in past for studying variety of patterns, most popular being ‘community’, that refers to a group of nodes (time series) with strong mutual positive correlations. Some works have further studied pairs of negatively correlated communities [8, 13, 7]. In contrast to above patterns, we use correlation networks to find multipoles that appear as *negative cliques*, i.e. sets of nodes (time series) with mutual negative correlations, or *pseudo-negative cliques*, that are pairs of positively correlated negative cliques.(see Definition 11.)

A collection of variables (time series) is commonly analyzed using Singular Value Decomposition (SVD)-based approaches. For instance, Principal Component Analysis (PCA) and its variants use SVD to find most variant components of variability in the given time series dataset [4]. However, multipoles, by definition, represent sets of variables with strong linear interdependence, and hence always belong to the least variant components of the dataset. Therefore, multipoles cannot be captured by PCA-based techniques. There are some SVD-based techniques such as Total Least Squares (TLS) that focus on the least variant component to study relationships between the variables. However, in the cases of larger datasets that that have lot of noise variables, variables of a multipole get mixed with other noise variables and could not be isolated even in any of the least variant components. Thus, in summary, SVD-based approaches cannot find multipoles from the data. We further illustrate the above limitations of SVD-based approaches in the following experiment.

**Experiment demonstrating limitations of SVD-based approaches:** Consider a set \(D\) of 64 standardized time series \(\{X_1, X_2, \ldots, X_{64}\}\), each of length 100 timestamps, where each column represents a standardized time series of length 100 timestamps. The corresponding covariance matrix \(\Sigma\) is shown in Figure 3(a). Out of 64 time series, subset \(\{X_1, X_2, X_3, X_4\}\) form a multipole \(M\), subset \(\{X_{33}, X_{34}, \ldots, X_{64}\}\) form a big cluster of strongly correlated time series, whereas the remaining 28 time series \(\{X_5, X_6, \ldots, X_{32}\}\) are independently generated using white noise Gaussian process. The goal of this experiment is to determine if if the multipole \(\{X_1, X_2, X_3, X_4\}\) could be recovered using SVD-based approaches. Figure 3(b) shows the weight in the first principal component of \(\Sigma\). As expected, first PC isolates the big cluster of 32 time series from the remaining dataset, since it explains the maximum variability in the data. Thus, PCA is able to capture the big cluster in the
data. On the contrary, multipole \{X_1, X_2, X_3, X_4\} couldn’t be isolated in any of the 64 PCs. For example, Figure 3(c) shows the PC that among all other PCs, gave largest weights to all members of multipoles and smallest weights to remaining 60 time series. This example clearly demonstrates the differences between multipoles and SVD-based approaches. Thus, we conclude that SVD-based approaches cannot find multipoles from the data.

Another set of related techniques include regularized linear-regression based methods such as (LASSO, Group-Lasso etc.) that are used to select most relevant predictors from a large set of time series to predict a given time series (predictand) \[ \{X_0, X_1, X_2, ..., X_{100}\} \]. The notion of predictors and predictand makes the discovered relationships asymmetric in nature, unlike multipoles where there is no such distinction among included variables. Further, such approaches are designed to optimize only the strength of the linear dependence among the variables, and do not place constraints on the level of contribution from each of the included predictors. As a result, in scenarios where there is autocorrelation and redundancies among the variables, they end up selecting trivial relationships over non-trivial ones, (e.g. time series from locations in spatial-proximity in a spatially autocorrelated dataset). To illustrate this further with an example, we conducted a following experiment:

Consider a set of 100 normalized variables \{X_0, X_1, X_2, ..., X_{100}\}, where each variable is a white noise Gaussian time series of length 100 timesteps. Figure 4 shows the correlation structure between for first four variables \{X_0, X_1, X_2, X_3\}. Among the four variables, \{X_0, X_1, X_2\} and \{X_1, X_2, X_3\} form multipoles. The question of interest here is that can the two multipoles be depicted using LASSO, a regularized linear regression that induces sparsity. Table 1 indicates the list of all predictors obtained while keeping different variables as predictands. Table shows that multipole \{X_1, X_2, X_3\} can be recovered using LASSO by choosing either \{X_2\} or \{X_3\} as predictands. However, multipole \{X_0, X_1, X_2\} could not be recovered at all. In particular, when \{X_0\} was set as predictand, LASSO selected \{X_1, X_3\} over \{X_2, X_3\}. This is because \{X_1, X_3\} shows stronger linear dependence with \{X_0\} than \{X_1, X_2\}, owing to the strong correlation between \{X_0, X_3\}, which drives its
Figure 4: An example where LASSO would fail to identify a multipole

linear dependence. Such strong pairs of correlations are highly common in spatially auto-correlated datasets.

<table>
<thead>
<tr>
<th>Predictand</th>
<th>Predictors selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_0$</td>
<td>${X_1, X_3}$</td>
</tr>
<tr>
<td>$X_1$</td>
<td>${X_2, X_3}$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>${X_1, X_3}$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>${X_0}$ or ${X_0, X_1, X_2}$</td>
</tr>
</tbody>
</table>

Table 1: Predictands and the list of predictors obtained using LASSO

Most recently, a multivariate relationship called tripole was proposed to capture relationships between three time series. Specifically, a tripole consists of a root time series and a pair of leaf time series, such that the sum of leaves show a stronger correlation with the root compared to what each of the two leaves show with the root individually. Unlike tripoles, a multipole is a symmetric pattern (no notion of root or leaf) and is not limited to a specific linear combination.

4 Finding Multipoles

The problem of maximal multipole discovery from a time series dataset could be formulated as the following:

Definition 7 (Problem Formulation). Find the set $P$ of all maximal multipoles in a given time series dataset.

The combinatorial aspect of the problem makes it extremely challenging to come up with an
approach that is both computationally efficient and guarantees completeness of the search. A brute-force approach would examine all subsets of size \( k \), varying \( k \) from 3 to \( N \), the total number of time series in the dataset. While such an approach guarantees completeness of the search, it will easily become computationally infeasible even for very small datasets due to its exponential time complexity. To give an estimate, one of our real-world datasets is quite small in size and has only 171 time series. However, performing a brute-force search on a regular desktop over all subsets of i) size 4 takes about 4.5 hours, ii) size 5 takes more than 4 days, and so on for subsets of size beyond 5.

In this paper, we propose a correlation graph-based approach to capture most interesting multipoles in a computationally efficient manner. Our approach is primarily motivated by some empirical observations that indicate a direct relationship between the linear gain of a set and the strengths of pairwise correlations between the members of the set. Leveraging these observations, our approach identifies and restricts the search for multipoles to a family of subsets, which we refer to as ‘promising candidates’ that are more likely to exhibit multipole relationships with stronger linear gain between their members. Such a family of subsets are usually much rarer in the data, thereby contributing to the remarkably high computational efficiency of our approach, although with some loss of completeness in the final output of multipoles. Before describing our approach in detail, we first present the empirical observations we will use to identify promising candidates.

4.1 Empirical Observations

One of our key empirical observations relates the maximum pairwise correlation of two variables in a multipole to the linear gain. To more easily demonstrate that connection, we need to first define the notion of self-cancellation on a set.

**Definition 8.** A set is said to be self-canceling if all the weights in its LVNLC are non-negative.

Any given set of variables \( S \) can be converted into its self-canceling version \( \tilde{S} \), by flipping the signs of all the members in \( S \) that have negative weights in LVNLC of \( S \). Note that flipping the signs of one or more variables does not affect the eigenvalues of their correlation matrix. Hence, \( S \) and \( \tilde{S} \) will have exactly the same linear gain. This self-canceling version represents a canonical form.

Let \( \alpha_S \) denote the highest pairwise correlation observed in the self-canceling version \( \tilde{S} \) of \( S \). Then, for any set \( S \) we observed an empirical relationship between \( \alpha_S \) and linear gain \( \Delta \sigma_S \). Specifically, Figure 5 shows scatter plots between \( \Delta \sigma_S \) (X-axis) and \( \alpha_S \) (Y-axis) for more than a million correlation matrices of sizes \( k \times k \) for \( k = 3, 4, 5 \). Each of these matrices were generated by sampling all \( \binom{k}{2} \) pairwise correlations between \([-1, 1]\) using a uniform distribution. Among the generated matrices, only the ones that satisfy positive semi-definiteness were considered to be valid correlation matrices. The implementation of the procedure can be accessed at this URL 1.

From the Figures 5(a), 5(b), and 5(c) we make two key observations:

**First key observation:** The linear gain of a set \( S \) tends to be higher when all the pairwise correlations in \( \tilde{S} \) are negative. More precisely, for all sets of size \( k \) that form multipoles with linear gain at least \( \delta \), there exists an upper bound \( \alpha \) on \( \alpha_S \). Hence, if we are interested in finding multipoles with linear gain at least \( \delta \), we can potentially discard all those sets that have at least one pairwise correlation larger than \( \alpha \) in their self-canceling versions.

**Second key observation:** The maximum possible linear gain of a multipole of size \( k \) is \( \frac{1}{k-1} \). Such a case corresponds to the set where all the pairwise correlations in its self-canceling version equals \( -\frac{1}{k-1} \). This implies that the linear gain is smaller for larger multipoles. Hence, if we are only interested in finding multipoles with linear gain at least \( \delta \), we can safely ignore all sets beyond of size \( \left\lceil \frac{1-\delta}{\delta} \right\rceil \). Both of the above two observations were found to empirically hold true also for sets of sizes beyond 5.

1https://github.com/15saurabh16/Multipoles
Figure 5: Empirical relationship between linear gain $\Delta \sigma_S$ of a set $S$ (X-axis) and maximum pairwise correlation in its self-canceling version $\tilde{S}$ (Y-axis) on more than $10^6$ generated $k \times k$ correlation matrices for $k \in [3, 5]$.

Figure 6: Illustrating equivalence between a pseudo-negative clique and a negative clique.

4.2 Promising Candidates

Consider a set $S = \{X_1, X_2, \ldots, X_k\}$ of $k$ variables. Let $\tilde{S}$ be the self-canceling set obtained from $S$. Based on the observations presented in previous subsection, we next define a promising candidate for a multipole as the following:

**Definition 9.** A set $S$ is said to be a **promising candidate** if i) $|S| \leq \left\lfloor \frac{1+\delta}{\delta} \right\rfloor$, where $\delta$ is the user-specified threshold on linear gain, and ii) maximum pairwise correlation in $\tilde{S} < 0$.

In other words, a promising candidate $S$ is one in which all the pairwise correlations among the members of its self-canceling version $\tilde{S}$ are below 0, and whose maximum size is governed by $\delta$.

Promising candidates can be further classified into two classes: i) Negative Cliques, and ii) Pseudo-Negative Cliques.

**Definition 10.** A **negative clique** refers to a set where all the pairwise correlations between its members are negative.

The terminology is motivated from the appearance of such sets in a correlation graph, where each vertex represents a variable and the weight of an edge $e(X_i, X_j)$ is equal to $\text{corr}(X_i, X_j)$. Such sets would appear as a clique of negative edges in the correlation graph. It can be shown that for all negative cliques, $S = \tilde{S}$. Hence, a negative clique will always be a promising candidate.

**Definition 11.** A **pseudo-negative clique** refers to a set $S$ that can be partitioned into two negative cliques $S_1$ and $S_2$ such that all the cross correlations between members of $S_1$ and $S_2$ are non-negative.

In a correlation graph, a pseudo-negative clique can be identified as a pair of negative cliques connected by non-negative edges. We refer to such a set as pseudo-negative clique since it can be converted into a negative clique by flipping the signs of all the variables involved in one of the two negative cliques. For illustration, consider a pseudo-negative clique of size 4 shown in Figure 6(a) that has two negative subcliques $(T_1, T_3)$ and $(T_2, T_4)$ interconnected by positive edges. Let us now flip the signs of variables in one of the negative cliques, say $(T_2, T_4)$ to $(-T_2, -T_4)$. Figure 6(b) shows the change in the correlations due to the flipping. As can be seen, all the cross-correlations...
between the two cliques flipped their signs and become negative. Also notice that all the internal correlations of the negative clique \((T_2, T_4)\) remain unchanged. As a result, all pairs of the variables in the modified set are now negatively correlated.

It can be shown that the self-canceling set of any pseudo negative clique \(S\) is a negative clique, and hence every pseudo negative clique is a promising candidate.

### 4.3 Proposed Approach: CoMEt

Leveraging the above empirical observations and the concept of promising candidates discussed in previous section, we propose our Clique Based Multipole Search (CoMEt) to find multipoles. The central idea of CoMEt is to find all promising candidates for multipoles (negative cliques and pseudo-negative cliques) and then check each of their subsets to obtain multipoles. To find all promising candidates, we first construct a graph such that every promising candidates forms a clique in it. All promising candidates are then recovered by enumerating all the cliques of the constructed graph. Although clique-enumeration problem is an NP-complete problem in general, the cliques of our interest tend to be very rare in the graphs generated from real-world datasets, which allows us to recover almost all the cliques but in much less computing time.

Algorithm 1 summarizes the CoMEt approach. We begin by finding all promising candidates in line 3, viz. negative cliques and pseudo-negative cliques using Algorithm 2. Among the obtained promising candidates, we then obtain all the multipoles in line 3 using Algorithm 3. Finally, in line 5, we eliminate duplicate and non-maximal multipoles using Algorithm 4. We next describe each of the modules used in the different steps of CoMEt.

#### Algorithm 1 CoMEt (Clique Based Multipole Search)

<table>
<thead>
<tr>
<th>Input</th>
<th>Dataset (D), Parameters: (\delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>set (U) of maximal multipoles with linear gain (\geq \delta).</td>
</tr>
<tr>
<td>1:</td>
<td>(C \leftarrow \text{FIND PROMISING CANDIDATES}(D))</td>
</tr>
<tr>
<td>2:</td>
<td>for each clique (S) in (C) do</td>
</tr>
<tr>
<td>3:</td>
<td>(U \leftarrow \text{GET MULTIPOLES FROM A CANDIDATE}(S))</td>
</tr>
<tr>
<td>4:</td>
<td>end for</td>
</tr>
<tr>
<td>5:</td>
<td>(U \leftarrow \text{REMOVE DUPLICATES} &amp; \text{NON-MAXIMALS}(U))</td>
</tr>
<tr>
<td>6:</td>
<td>return (U)</td>
</tr>
</tbody>
</table>

#### 4.3.1 FIND PROMISING CANDIDATES:

This module is used to find all promising candidates in the data that include all negative cliques and pseudo-negative cliques. Algorithm 2 summarizes the entire module. The key idea is to construct a correlation graph (network) \(G\) (lines 1-4) such that a set of nodes of size \(\geq 3\) in \(G\) would form a clique iff it is a negative clique or pseudo-negative clique. To accomplish that, we first construct two identical graphs \(G_1\) and \(G_2\), such that for both graphs, a set of nodes will form a clique iff it is a negative clique. To obtain such graphs, we begin by creating a set \(V_1 = \{v_{11}, v_{12}, ..., v_{1n}\}\) of \(n\) nodes such that each node \(v_{1i}\) corresponds to a time series \(X_i\) in the given dataset \(D\). For any pair of nodes \((v_i, v_j)\), an edge is drawn iff \(\text{corr}(X_i, X_j) \leq 0\). Let \(E_1\) denote the set of all such edges. Then \(G_1 = (V_1, E_1)\) is the desired correlation graph where a clique will be formed among a set of nodes iff their corresponding variables in \(D\) form a negative clique. Similarly, an identical correlation graph \(G_2 = (V_2, E_2)\) can be constructed on a set of nodes \(V_2 = \{v_{21}, v_{22}, ..., v_{2n}\}\), where each node \(v_{2i}\) corresponds to a time series \(X_i\) in the given dataset \(D\).

Next, to include edges for non-negative correlations in all pseudo-negative cliques, we construct a set \(E\) of cross-edges between nodes of \(G_1\) and \(G_2\). Specifically, we connect each node \(v_{1i} \in V_1\) to all the nodes \(v_{2j} \in V_2\) such that \(\text{corr}(X_i, X_j) \geq 0\). As a result of this operation, for any pseudo-negative clique \(S = S_1 \cup S_2\), where \(S_1\) and \(S_2\) are its two negative subcliques, all the non-negative correlations across \(S_1\) and \(S_2\) are now included. Hence, in the resultant graph \(G = (V_1 \cup V_2, E_1 \cup E_2 \cup E)\), every pseudo-negative clique will also appear as a clique.

Notice that every clique \(S\) of size \(\geq 3\) in \(G\) would be a promising candidate; it would either be a negative clique (if \(S \subset G_1\) or \(S \subset G_2\)), or a pseudo-negative clique (if it includes nodes from both \(G_1\)
or \( G_j \). Hence, we can obtain all promising candidates by applying any of the standard clique-search algorithms on \( G \) to obtain all cliques. For this work, we used Bron-Kerbosch algorithm \([5]\) (line 11), which can be used to find all maximal cliques in an undirected and unweighted graph.

Also notice that every promising candidate will result in formation of two cliques in \( G \). For instance, every negative clique will form two cliques: one in \( G_1 \) and \( G_2 \) each. Similarly, a pseudo negative clique \( S = (S_1 = \{X_i, X_j\}, S_2 = \{X_k\}) \), where \( S_1 \) and \( S_2 \) are two negative sub-cliques, will result in formation of two cliques: \( (v_{11}, v_{1j}, v_{2k}) \) and \( (v_{2i}, v_{2j}, v_{1k}) \) in \( G \). For every promising candidate, exactly one of the cliques is retained (line 12). Finally, the set of retained cliques is returned as the set of all promising candidates.

**Algorithm 2 FIND PROMISING CANDIDATES**

<table>
<thead>
<tr>
<th>Input</th>
<th>Dataset: ( D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>set ( C ) of all maximal promising candidates</td>
</tr>
</tbody>
</table>

Correlation Graph Construction:
1. \( V_1 \leftarrow \) set of \( n \) vertices \( v_{11}, v_{12}, \ldots, v_{1n} \) corresponding to \( n \) variables \( X_1, X_2, \ldots, X_n \) in \( D \)
2. \( E_1 \leftarrow \) All pairs \( (v_{1i}, v_{1j}) \) such that \( \text{corr}(X_i, X_j) \leq 0 \)
3. \( G_1 = (V_1, E_1) \)
4. \( G_2 = (V_2, E_2) \) be the exact duplicate of \( G_1 \)
5. \( E \leftarrow \phi \)  \( \triangleright \) An empty set of edges
6. for each \( v_{1i} \in V_1 \) do
7. \( E' \leftarrow \) all pairs \( (v_{1i}, v_{2j}) \) such that \( \text{corr}(X_i, X_j) \geq 0 \)
8. \( E = E \cup E' \)
9. end for
10. \( G = (V_1 \cup V_2, E_1 \cup E_2 \cup E) \)
11. \( C \leftarrow \) All maximal cliques of \( G \) \( \triangleright \) using Bron-Kerbosch Algorithm \([6]\)
12. Remove all duplicate cliques from \( C \)
13. return \( C \)

### 4.3.2 GET MULTipoles FROM A CANDIDATE:

This procedure is applied to each of the obtained promising candidates to extract all multipole relationships. As summarized in Algorithm \([3]\) we begin by checking if the promising candidate forms a multipole by comparing its linear dependence and linear gain with user-specified thresholds \( \sigma \) and \( \delta \), respectively. If so, then \( S \) is added to the set of discovered multipoles and we move on to the next promising candidate. Otherwise, it could be possible that one or more of the subsets of \( S \) might form a multipole with stronger linear gain. Therefore, we perform an exhaustive search on all subsets of \( S \) of sizes \([3, \lfloor \frac{1+\delta}{\sigma} \rfloor]\) and select all the ones that show linear gain stronger than \( \delta \) and linear dependence stronger than \( \sigma \). The range of sizes of subsets is derived based on the observation made in section \([1]\) which states that for a given threshold \( \delta \) on linear gain, all sets of sizes beyond \( \lfloor \frac{1+\delta}{\sigma} \rfloor \) can be safely discarded.

Applying the above procedure to all promising candidates might result in inclusion of several non-maximal multipoles in the output. Furthermore, a multipole could be generated multiple times from different promising candidates. Hence, in the final step of CoMEt, we eliminate all non-maximal and duplicate multipoles using the module described below.

### 4.3.3 REMOVE NON-MAXIMALS & DUPLICATES:

This module is called in line 5 of Algorithm \([4]\) to eliminate all non-maximal and duplicate multipoles that are generated in the previous step. As described in Algorithm \([7]\) we begin with selecting the largest multipole \( S \) from \( U \), and add it to the final output set \( U' \) (lines 3-4) that was originally initialized to an empty set (line 1). We then eliminate all multipoles from \( U \) that are either duplicates of \( S \) (including \( S \) itself) or proper subsets of \( S \) (lines 5-6). The entire procedure in lines (3-6) is then repeated until \( U \) is empty. All the multipoles in the resultant set \( U' \) are consequently distinct and maximal.
Algorithm 3 GET MULTIPOLES FROM A CANDIDATE

Input: set \( S \), linear gain threshold \( \delta \)
Output: All multipoles \( S' \) with \( \sigma_{S'} \geq \sigma \), and \( \Delta \sigma_{S'} \geq \delta \)
1: if \( \Delta \sigma_S \geq \delta \) and \( \sigma_{S'} \geq \sigma \) then
2: \( M \leftarrow S \)
3: else
4: \( M \leftarrow \) all subsets \( S' \) of \( S \) s.t. \( |S'| \in [3, \lceil 1 + \frac{1}{\delta} \rceil] \), \( \sigma_{S'} \geq \sigma \), and \( \Delta \sigma_{S'} \geq \delta \)
5: end if
6: return \( M \)

Algorithm 4 REMOVE DUPLICATES & NON-MAXIMALS

Input: a set \( U \) of multipoles
Output: a set \( U' \) of non-maximal and distinct multipoles
1: \( U' \leftarrow \phi \)
2: while \( U \neq \phi \) do
3: \( S \leftarrow \) Largest multipole in \( U \)
4: \( U' \leftarrow U' \cup S \)
5: \( W \leftarrow \) all multipoles in \( U \) that are duplicates or subsets of \( S \)
6: \( U \leftarrow U \setminus W \)
7: end while
8: return \( U' \)

4.4 CoMEtExtended

Our proposed approach, CoMet, limits the search for multipoles to promising candidates. As a result, all the multipoles that were formed in non-promising candidates would be missed by CoMet. To recover such multipoles, we propose CoMEtExtended, which is an extended version of CoMEt where we relax the concept of "negative clique" and redefine what constitutes a promising clique. Specifically, in the current definition of a promising candidate \( S \), instead of enforcing that all the pairwise correlations in \( \bar{S} \) be negative, we require all of them to be below a threshold \( \mu \), where \( \mu \in [-1, 1] \). The definitions of negative cliques and pseudo-negative cliques are also be modified accordingly. In particular, a negative clique would have all pairwise correlations below \( \mu \), whereas the non-negative edges in pseudo-negative cliques should be \( \geq -\mu \). Setting \( \mu \) to zero gives us the current definition of promising candidates, whereas setting \( \mu \) to a higher positive value would include more sets as promising candidates and hence lead to recovery of more of the missed multipoles. However, it should be noted that setting \( \mu \) to positive values would increase the number of candidate cliques and thus increase the computational cost. In fact, setting \( \mu \) to 1 would turn CoMEtExtended into an exhaustive brute-force search, which guarantees completeness, but would be computationally infeasible.

5 Data and Experimental Evaluation

In this section, we discuss our evaluation approach and results. Specifically, we evaluated the computational efficiency of our approach, the statistical significance of the multipoles, and their utility using real-world datasets from climate science and neuroscience domain. We begin by describing all the datasets along with the data pre-processing steps that were applied to each of them.

5.1 Data and Preprocessing

5.1.1 Sea Level Pressure (SLP) data:

We used monthly Sea Level Pressure (SLP) dataset provided by NCEP/National Center for Atmospheric Research (NCAR) Reanalysis Project [9] that is available from 1979-2014 (36 years) at a spatial resolution of 2.5 × 2.5 degree (10512 grid points, also referred to as locations). In this paper, we constructed SLP time series for each location using only the the months of winter season (December, January, and February) from each year, thereby resulting in 108 observations in every time series. For each of the time series, we followed the standard pre-processing steps followed in climate
science to remove the annual seasonality and linear trends. Relationships in climate datasets are preferably studied between regions (sets of spatially contiguous locations) as opposed to individual locations because of i) spatial autocorrelation, due to which locations in a spatial neighborhood have highly similar time series that will lead to discovery of redundant relationships, and ii) the fact that the relationships studied across regions are likely to be more reliable and stable over time.

Therefore, we next converted the given location-based time series dataset into a region-based time series dataset as described in Algorithm 5. Specifically, treating each of 10512 location $l_i$ as centre, we grew a spatially contiguous region $R_i$ around it by including top 50 locations that were most strongly correlated to $l_i$ and spatially contiguous. As a result, we get a set $U_R$ of 10512 regions. For each region $R_i \in U_R$, a representative time series $T_i$ is obtained as the normalized area-weighted average time series of all the included locations. Consequently, we get a set of representative time series of 10512 regions.

However, note that there would be a high degree of redundancy among the regions in $U_R$ due to spatial autocorrelation and mutual spatial-overlaps among them. Ideally, we wish to obtain a set of non-redundant regions to avoid discovery of redundant multipoles.

Hence, we next describe our approach to select a final set $D$ of representative time series of non-redundant regions from the set $U$ as elaborated in lines xx-xx of Algorithm 5. First, we define two regions $R_i$ and $R_j$ to be redundant if their corresponding time series $T_i$ and $T_j$ show a correlation of at least 0.8. Using the above definition of redundancy, we then find the region $R_k$ with maximum number of redundant regions. We then add $T_k$, the representative time series of $R_k$ to the final set $D$ of non-redundant regions. We next prune the time series of all those regions from $U_R$ that were redundant to $R_k$. The entire process is then repeated continuously until the set $U_R$ becomes empty.

Following the above procedure, we obtained a set $D$ of 171 regions that are non-redundant and cover entire globe. For evaluation purposes, we used monthly HadSLP2 data provided by HadSLP2 for years prior to 1979 to obtain the time series of these regions. HadSLP2 data is originally available at a coarser spatial resolution of $5^\circ \times 10^\circ$ and is interpolated to $2.5^\circ \times 2.5^\circ$, before constructing the time series of above 171 regions.

Algorithm 5 GET NON-REDUNDANT REGION TIME SERIES

| **Input:** A set $L$ of 10512 locations |
| **Output** set $D$ of representative time series of non-redundant regions |

1. $R_i \in \emptyset$
2. for each location $l_i \in L$ do
3. $R_i \leftarrow$ a set of top 50 locations most strongly correlated with $l_i$ and are spatially contiguous
4. $T_i \leftarrow$ Area-weighted mean of time series of all locations included in $R_i$
5. Insert $R_i$ into $U_R$ and $T_i$ into $U_T$
6. end for
7. Count number of redundant regions for every region
8. for each region $R_i \in U_R$ do
9. $A \leftarrow$ Set of all regions $R_j$ s.t. corr($T_i, T_j$) $\geq$ 0.8
10. RedunCount($R_i$)$\leftarrow |A|$
11. end for
12. Remove redundant regions
13. $D \leftarrow \emptyset$
14. while $U_R \neq \emptyset$ do
15. $R_k \leftarrow$ argmin $R_i \in U_R$(RedunCount($R_i$))
16. Insert $T_k$ into $D$
17. Remove all regions from $U_R$ that are redundant to $R_k$
18. end while

5.1.2 Brain fMRI data:

We used neuroimaging data collected at the University of Utah as part of a reproducibility study. In this study, a set of 50 functional-Magnetic Resonance Imaging (fMRI) scans of one subject were acquired while the subject was involved in an audio-visual task (watching cartoons). Another set of 50 fMRI scans were collected from the same subject while the subject was resting. The spatial resolution and the temporal resolution of every scan were was 3mmx3mmx3mm and 2 secs
respectively. A number of fMRI pre-processing steps including motion correction, unwarping, and filtering that have been described in [2] have been performed. In addition, an Automated Anatomical Labeling Atlas [16] that maps grey matter locations to 90 anatomical regions is used to compute a mean time series of each brain region from each scan. As a result, we obtain a set of 90 time series for each of the 100 fMRI scans. We applied our approach to find multipoles in one of the fMRI scans that was taken while the subject was watching video, while the other scans were used for evaluation purposes that we will describe later in this section.

5.2 Parameter settings

Two user-specified thresholds, minimum linear gain ($\delta$) and minimum linear dependence ($\sigma$), are needed for discovering multipoles. The choice of values for these parameters needs to be determined based on domain knowledge and the availability of computational resources to find multipoles and further investigate them from domain perspective. In particular, a relaxed linear gain threshold $\delta$ will increase the search space of multipoles and thus will require more computational time for search and evaluation. Similarly, a lower threshold of $\sigma$ will result in a larger number of discovered multipoles to be analyzed further by domain experts. In this work, we set $\delta$ to 0.15 and $\sigma$ to 0.50 for all sizes of multipoles for both the SLP and brain fMRI datasets.

5.3 Summary of Results

For the above parameter setting, Table 2 reports the total number of multipoles of different sizes that were obtained using our approach in the three datasets.

<table>
<thead>
<tr>
<th>Size</th>
<th>SLP ($\Delta \sigma_S \geq 0.15$)</th>
<th>fMRI ($\Delta \sigma_S \geq 0.15$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2631</td>
<td>2201</td>
</tr>
<tr>
<td>4</td>
<td>713</td>
<td>603</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Number of multipoles of different sizes with $\sigma_S \geq 0.5$ and $\Delta \sigma_S \geq 0.15$ found in SLP & fMRI datasets

5.4 Computational Evaluation of CoMEt

In this section, we describe the evaluation of our proposed approach, CoMEtExtended, with respect to a brute-force approach (discussed in Section 4). In this evaluation, we focus on the trade-off between completeness of the search and total computational time taken to conduct the search. Computational efficiency is assessed by total computing time taken to finish the search. On the other hand, completeness of a search is measured as the fraction of multipoles of a complete set (a set that includes all the multipoles present in the data) captured in the output. A multipole is considered to be captured if it or any of its supersets are present in the generated output. A complete set of multipoles can be obtained by performing an exhaustive brute-force search over all possible subsets of variables of sizes 3 to 7. Since, we require linear gain to be at least 0.15, we can leverage our empirical observations discussed in section 4.1 and safely ignore all subsets of size beyond 7.

Figure 7 compares the computing time (log scale) and completeness of the brute-force approach and CoMEtExtended at different values of parameter $\mu$ for both datasets. As discussed in section 4.4, CoMEtExtended involves an additional parameter $\mu$ that can be adjusted to have a desired trade-off between computational efficiency and completeness. Increasing $\mu$ expands the search space. As a result, both completeness and computational time also increase, which is evident in the Figure. More importantly, CoMEtExtended could recover more than 95% of all multipoles in SLP data in
Figure 7: Evaluating the trade-off between $\log_{10}$ (Computing Time in seconds) (X-axis) and Completeness of the search defined as fraction of total multipoles discovered (Y-axis) for Brute-Force approach and CoMEtExtended at different values of $\mu$.

411 seconds (for $\mu = 0$). Similarly, more than 98% of multipoles in fMRI data can be recovered in 561 seconds using CoMEtExtended (for $\mu = 0.2$). In contrast, while the brute-force method allows a complete search, it takes about 14 days and about 97 hours to finish on the SLP dataset and the fMRI datasets, respectively. In summary, CoMEtExtended is more practical because it is relatively complete and highly efficient as compared to a brute-force approach.

5.5 Evaluation of Multipoles

One of the key challenges of this work is distinguishing between reliable and spurious multipoles, i.e., those multipole patterns that arise due to random variation, from the large number of discovered multipoles. Domain validation is an ideal approach to evaluate multipoles, but most of the multipole relationships discovered in this work are currently unknown to domain scientists. Thus, in our work, we used an empirical evaluation framework that consists of two steps. The first step involves a procedure for estimating the statistical significance of a multipole. This procedure is then used in the second step to assess the reproducibility of a discovered multipole in multiple time periods that are sampled from an SLP data set (HadSLP2), which is independent of the SLP dataset originally used for finding multipoles. Multipoles that are statistically significant in multiple time periods, i.e., that are highly reproducible, are candidates for further evaluation.

5.5.1 Step 1: Statistical Significance Evaluation:

To filter spurious multipoles, it is important to answer the following two questions: i) how likely is it that the observed level of linear dependence, $\sigma_S$, of a multipole $S$ is due to chance? and ii) does every member in $S$ contribute significantly to the linear dependence of $S$?

To address the first question, we generate a null distribution of linear dependence by randomly generating 100,000 sets of time series and evaluating each set $S$ for its level of linear dependence. Each of these randomly generated sets is created by sampling time series from different time periods. For instance, for our SLP investigations, a random set of size $|S|$ is constructed by sampling a time series from any $|S|$ of the nine time windows of HadSLP2 data. Similarly, for brain fMRI data, a random set of $|S|$ time series is constructed by sampling a time series from any $|S|$ of the 50 scans. Generating a random set in this manner is an approximation to independently generating $|S|$ time series while ensuring that the general underlying nature of domain time series (e.g., autocorrelation, periodicity etc.) is retained in the randomly generated data. Using the resultant null distribution, we then determine the statistical significance of the multipoles we originally found. We evaluate $\sigma_S$ at a 0.01 level of significance.

We next describe our approach to assess the second point, i.e., the significance of the contribution of each of the $k$ variables in a given set $S = \{X_1, X_2, ..., X_k\}$ to its linear dependence. Specifically, to assess the significance of the contribution of time series $X_i$, we replace it with a random time series $X_R$ that is sampled from an independent dataset in a manner similar to the procedure above. We then compute the linear dependence of the resultant set, which we call $S'$. If the contribution of $X_i$ is not spurious, it would be unlikely for a randomly chosen time series $X_R$ to replicate it in which case, $\sigma_{S'} \leq \sigma_S$. We repeat the above process 1000 times and compute the fraction of the population
for which $\sigma_{S'} \leq \sigma_S$ holds true. This fraction is our significance level. We again use a significance level of 0.01. The above procedure is repeated for each of the $k$ members.

5.5.2 Step 2: Reproducibility in Independent Datasets

In this step, we estimate the reproducibility of a given multipole in multiple time periods. A multipole is considered to be reproducible in a dataset $D$, if, at a 0.01 level of significance, it is found to have statistically significant linear dependence, as well as a statistically significant contribution from each of its members. Specifically, for each multipole discovered in 1979-2014 SLP dataset, we computed its linear gain and linear dependence in HadSLP2 data in 9 time windows 1901-1936, 1906-1941,...,1941-1976. Likewise, for multipoles discovered in one of the 50 brain fMRI scans, we studied their reproducibility in the remaining 49 scans.

Figure 8(a) shows number of multipoles reproduced in different numbers of HadSLP2 time windows during 1901-1976. More than 40% of multipoles reproduced in all 9 time windows. Similarly, Figure 8(b) shows number of multipoles from fMRI data reproduced in different number of fMRI scans taken while the subject was watching a video. At least 25% of multipoles reproduced in more than 10 other scans. Higher reproducibility of multipoles suggests that they are more robust to noise in the data and unlikely to be spurious.

6 Case Studies

Results discussed in the previous section indicate the existence of several multipole relationships with high reproducibility in multiple independent time periods, which makes a compelling case for their connection to underlying physical phenomena that might be currently unknown to domain scientists, but could potentially be discovered by domain experts upon further analysis. In this section, we present case studies on the physical interpretation of two of the discovered multipoles in SLP and brain fMRI data.

6.1 Discovering Climate Phenomena

One of the multipoles in SLP data was found between four regions shown in Figure 9. The time series four regions show negative correlations with each other, resulting in a multipole relationship with linear dependence of 0.7 and a linear gain of 0.15 during 1979-2014. Further, the multipole was found to be statistically significant in 7 out of 9 time windows during the period of 1901-1976.

This multipole appears to be strongly related to the well-known climate phenomenon called El-Nino Southern Oscillation (ENSO) as indicated by regions $R_2$ and $R_3$. This phenomenon appear not only in the tropical Pacific Ocean, but also has large-scale impacts on regional climate outside the tropics [17]. A recently discovered connection between ENSO and the West Siberian Plain [10] is reproduced here as the tripole between regions $R_1$, $R_2$, and $R_3$. Specifically, that phenomenon was attributed to a wave train that originates from sub-tropical Atlantic and propagates north-eastswards towards the north of the West Siberian Plain, where it gets deflected southeastwards and reaches
Figure 9: A 4-pole in 1979-2014 that was reproducible in 7 out of 9 time windows during 1901-1976

Figure 10: Climate Models Inter-comparison based on multipole simulations: Summary statistics of multipole shown in Figure 9 in different time windows of HadSLP2 data (Figure 10(a)) and different CMIP5 climate model datasets (Figures 10(b) and 10(c)). In each plot, red curve indicates the strength of linear dependence of multipole, while the blue curve indicates the highest strength of linear dependence obtained for one of its subsets. The difference in red and blue curves thus indicate linear gain of the multipole.

all the way to central Pacific Ocean, where the two centres of action of ENSO are located. Notably, region \( R_4 \) in the northern Atlantic Ocean is located near the proposed location of origin of the wave train. Therefore, region \( R_4 \) provides a new discovery about the general path of the observed wave train. This finding can be used to study the detailed path of the wave train and to attribute weather and climate characteristics over regions affected by this wave train to their potential ENSO origins.

Multipoles, being potential representatives of physical processes, could serve as useful benchmarks to evaluate various climate models that are often used to study climate change under different greenhouse gas emission scenarios. In particular, climate models can be evaluated based on their ability to reproduce the physical processes represented by these multipoles. For instance, Figure 10 compares the linear dependence and linear gain of the multipole across multiple time windows obtained in observations data (HadSLP2) and a couple of climate models used in the IPCC (Intergovernmental Panel on Climate Change) CMIP5 (Couple Model Intercomparison Project Phase 5) evaluation. It can be seen that compared to HadSLP2 observations, climate models are not able to simulate the multipole effectively. Among the models, MPI-ESM-MR is able to reproduce multipole with statistically significance in at least 4 time windows, whereas the other model BNU-ESM could not reproduce the given multipole at all.
6.2 Studying Complex Dynamics in Brain

The notion of multipoles proposed in this paper is highly suited for capturing complex signaling relationships in brain. For example, one of the multipoles we discovered in one of the brain fMRI scans captures relationship between three brain regions: Right Frontal Inferior (T2), Right Parietal Inferior (T1), and Right Temporal Pole Superior (T3). This multipole was found to be reproducible in 23 out of 50 ‘task’ scans (collected while the subject was watching a cartoon video) while only in 5 of the 50 other ‘resting state’ scans (collected while subject was resting). This relationship is interesting for multiple reasons. First, the parietal and temporal regions are known to be the first recipients of the video and audio stimuli [14], respectively, that are presented to the subject in our case in the form of cartoons. Second, the dominance of the above multipole relationship in task scans in comparison to the resting scans provides strong evidence for signaling from parietal and temporal regions to the frontal region, that appears to be triggered due to visual and auditory stimuli. While some existing studies [12] have observed activity in frontal region due to visual and auditory stimuli, no direct signal pathway between visual and auditor cortices with the frontal region is established. In summary, the multipole framework could serve as a promising tool for discovering previously unknown signaling pathways that exist in the brain.

7 Conclusion and Future Work

In this paper, we introduced and formally studied a novel class of multivariate linear relationships called multipoles in time series data. We proposed a novel and computationally efficient correlation network-based approach to find all multipoles in the data. In particular, we identified the utility of negative cliques and pseudo-negative cliques in a correlation network, that have never been studied before in any correlation network-based analysis. Furthermore, we demonstrated the utility of multipoles in discovering novel physical processes in neuroscience and climate science. One of the most useful extension of this work could be to extend the notion of multipoles to non-linear relationships and other similarity measures. Another generalization could be the time-lagged multivariate relationships.

References


