Granger Causality Analysis in Irregular Time Series

Mohammad Taha Bahadori* Yan Liu*†

Abstract

Learning temporal causal structures between time series is one of the key tools for analyzing time series data. In many real-world applications, we are confronted with *Irregular Time Series*, whose observations are not sampled at equally-spaced time stamps. The irregularity in sampling intervals violates the basic assumptions behind many models for structure learning. In this paper, we propose a nonparametric generalization of the Granger graphical models called Generalized Lasso Granger (GLG) to uncover the temporal dependencies from irregular time series. Via theoretical analysis and extensive experiments, we verify the effectiveness of our model. Furthermore, we apply GLG to the application dataset of $\delta^{18}O$ isotope of Oxygen records in Asia and achieve promising results to discover the moisture transportation patterns in a 800-year period.

1 Introduction

In the era of data tsunami, we are confronted with large-scale time series data, i.e., a sequence observations of concerned variables over a period of time. For example, terabytes of time series microarray data are produced to record the gene expression levels under different treatments over time; petabytes of climate and meteorological data, such as temperature, solar radiation, and carbon-dioxide concentration, are collected over the years; and exabytes of social media contents are generated over time on the Internet. As we can see, how to develop effective and scalable machine learning algorithms to uncover temporal dependency structures between time series and reveal insights from data has become a key problem in machine learning and data mining.

Up to now many different approaches have been developed to solve the problem, such as autocorrelation, cross-correlations [8], transfer entropy [4], randomization test [17], phase slope index [20], Granger causality [13] and so on, and achieved successes in many applications. However, all the existing approaches assume that the time series observations are obtained at equally spaced time stamps and fail in analyzing irregular time series. Irregularly sampled time series are those with samples missing at blocks of sampling points or collected at non-uniformly spaced time points. It is a common challenge in practice due to natural constraints or human factors. For example, in biology, it is very difficult to obtain blood samples of human beings at regular time intervals for a long period of time; in climate datasets, many climate parameters (e.g., temperature and CO2 concentration) are measured by different equipments with varying frequencies.

Existing methods for analyzing irregular time series can be categorized into three directions: (i) the repair approach, which recovers the missing observations via smoothing or interpolation [25, 14, 16, 10, 22]; (ii) generalization of spectral analysis tools, such as Lomb-Scargle Periodogram (LSP) [24] or wavelets [12, 28, 19]; and (iii) the kernel methods [22]. While the first two approaches provide sufficient flexibility in analyzing irregular time series, they have been shown to magnify the smoothing effects and result in huge errors for time series with large gaps [25].

In this paper, we propose the Generalized Lasso-Granger (GLG) framework for causality analysis of irregular time series. It defines a generalization of inner product for irregular time series based on non-parametric kernel functions. As a result, the GLG optimization problem takes the form of a Lasso problem and enjoys the computational scalability for large scale problems. We also investigate a weighted version of GLG (W-GLG), which aims to improve the performance of GLG by giving higher weights to important observations. For theoretical contribution, we propose a sufficient condition for the asymptotic consistency of our GLG method. In particular, we show that compared with the popular locally weighted regression (LWR), GLG has the same asymptotic consistency behavior, but achieves lower absolute error.

In order to demonstrate the effectiveness of GLG, we conduct experiments on four synthetic datasets with different sampling patterns and objective func-

*Computer Science Department, Viterbi School of Engineering, University of Southern California, USA

†Correspondence author. Email: yanliu.cs@usc.edu

tions so that we can test the asymptotic convergence behavior, effects of different kernels, and impact of different irregularity patterns respectively. GLG outperforms all state-of-the-art algorithms and improves the accuracy of the inferred temporal casual graph up to 15%. In addition, we also apply GLG to uncover the patterns of transfer of moisture across Asia during 800 years by analysis of the temporal causal relationship between density of $\delta^{18}O$ (a radioactive isotope of oxygen) recorded in four caves in India and China. While the results identify the moisture transfer patterns, they confirm that during the warm periods the air masses had more energy to travel across the continent.

The rest of the paper is organized as follows: we first formally define the problem of learning temporal structures in irregular time series in Section 2 and review related work in Section 3. Then we describe the proposed GLG framework and provide theoretical insights. In Section 5, we demonstrate the superior performance of GLG through extensive experiments. After the summary of the paper and hints on future work, we provide the proof of the theorem in the Appendix.

2 Problem Definitions and Notation

Irregular time series are time series whose observations are not sampled at equally-spaced time stamps. They could appear in many applications due to various factors. In summary, there are three types of irregular time series: including *Gappy Time Series*, *Nonuniformly Sampled Time Series* and *Time Series with Missing Data*. Gappy Time Series refer to those with regular sampling rate but having finitely many blocks of data missing. For example, in astronomy, a telescope's sight could be blocked by a cloud or celestial obstacles for certain period of time, which makes the recorded samples unavailable during that time [10]. Nonuniformly Sampled Time Series refer to those with observations at non-uniform time points. For example, in health care applications patients usually have difficulties in rigorously recording daily (or weekly) health conditions or pill-taking logs for a long period of time [16]. Time series with missing data appear commonly in applications such as sensor networks, where some data points are missing or corruptted due to sensor failure. In this paper, unless otherwise stated, the term *irregular time series* refers to all the mentioned three groups or any combinations of them.

While the regular time series need only one vector of values on uniformly spaced time stamps, irregular time series need a second vector specifying the time stamps at which the data are collected. Following the notation of [22], we represent an irregular time series with a tuple of $(time\ stamp, value)$ pairs. Formally, we have the following definition:

DEFINITION 2.1. *An irregular time series* $\mathbf{x}$ *of length* N *is denoted by* $\mathbf{x} = \{(t_n, x_n)\}_{n=1}^{N}$ *where time-stamp sequence* $\{t_n\}$ *are strictly increasing, i.e.,* $t_l < t_2 < \ldots < t_N$ *and* x_n *are the value of the time series at the corresponding time stamps.*

The central task of this paper is as follows: given P number of irregular time series $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}\}$, we are interested in developing efficient learning algorithms to uncover the temporal causal networks that reveals temporal dependence between these time series.

3 Related Work

In this section, we review existing work on learning temporal causal networks and irregular time series analysis.

3.1 Temporal Causal Networks Learning temporal causal networks for regular time series has become one of the most important problems for time series analysis. It has broad applications in biology, climate science, social science, and so on. Many approaches have been developed to solve the problem, such as autocorrelation, cross-correlations [8], randomization test [17], phase slope index [20], and so on. It turns out, a regression-based method called "Granger Causality" has been introduced in the area of econometrics for time series analysis [13]. It states that a variable x is the cause of another variable y if the past values of x are helpful in predicting the future values of y. In other words, among the following two regressions:

$$x_t = \sum_{l=1}^{L} a_l x_{t-l} \tag{3.1}$$

$$x_t = \sum_{l=1}^{L} a_l x_{t-l} + \sum_{l=1}^{L} b_l y_{t-l}, \tag{3.2}$$

where L is the maximal time lag, if Equation (3.2) is a significantly better model than Equation (3.1), we determine that time series y Granger causes time series x. Granger causality has gained tremendous success across many domains due to its simplicity, robustness, and extendability [21, 9].

Most existing algorithms for detecting Granger causality are based on a statistical significance test.

It is extremely time-consuming and very sensitive to the number of observations in the autoregression. In [2], the Lasso-Granger method is proposed to solve the problem and shown to have superior performance in terms of accuracy and scalability. Suppose we have P number of time series $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}$ with observations at *equally spaced* time stamps $\mathbf{t} = 1, \ldots, T$. The basic idea of the Lasso-Granger method is to utilize L_1-penalty to perform variable selection, which corresponds to neighborhood selection in learning temporal causal network [2]. Specifically, for each time series $\mathbf{x}^{(i)}$, we can obtain a sparse solution of the coefficients $\mathbf{a}$ by solving the following Lasso problem:

(3.3)
$$\min_{\{\mathbf{a}_i\}} \sum_{t=L+1}^{T} \left\| x_t^{(i)} - \sum_{j=1}^{P} \mathbf{a}_{i,j}^{\top} \mathbf{x}_{t,Lagged}^{(j)} \right\|_2^2 + \lambda \left\| \mathbf{a}_i \right\|_1 ,$$

where $\mathbf{x}_{t,Lagged}^{(j)}$ is the concatenated vector of lagged observations, i.e., $\mathbf{x}_{t,Lagged}^{(j)} = \left[x_{t-L}^{(j)}, \ldots, x_{t-1}^{(j)} \right]$, $\mathbf{a}_{i,j}$ is the j-th vector of coefficients $\mathbf{a}_i$ modeling the effect of the time series j on time series i, λ is the penalty parameter, which determines the sparseness of $\mathbf{a}_i$. The resulting optimization problem can be solved efficiently by sub-gradient method [7], LARS [29] and so on. Then we determine that there is an edge from time series j to time series i if and only if $\mathbf{a}_{i,j}$ is nonzero vector.

The Lasso-Granger method not only significantly reduces the computational complexity compared with pairwise significant tests, but also achieves nice theoretical properties on its consistency [2]. It has been widely applied to biology applications [30], climate analysis [11], fMRI data analysis [23] and so on with strong empirical results.

3.2 Irregular Time Series Analysis Existing methods for analyzing irregular time series can be summarized into three categories: (i) Repairing the irregularities of the time series by either filling the gaps or resampling the series at uniform intervals and (i) Generalization of spectral analysis for irregular time series; and (iii) Applications of kernel techniques to time series data.

Repair Methods The basic idea of repair method is to interpolate the given time series in regularly spaced time stamps. The produced time series can be used in the temporal causal analysis for regular time series. For Gappy Time Series or Time Series with Missing Data, we can use a regression algorithm to recover the time series by filling the blank time stamps, which is also known as *Surrogate Data Modeling* [16]. In the case of non-uniformly sampled time series, the common practice is to find the value of the time series in regularly-spaced time stamps via a regression algorithm [25, 16, 22]. In some applications, a transient time model that describes the behavior of the system (e.g., by differential equations) is available and can be used to recover the missing data [14, 10]. However, this approach strongly depends on the model accuracy, which prevents it being applicable to the datasets with complex natural processes (e.g. climate system) [22]. One major issue with all repair methods is that the interpolation error propagates throughout all steps after data processing. As a result, quantifying the effect of the interpolation on any resulting statistical inference becomes a challenging task [19].

Generalization of the Spectral Analysis Tools The basic idea of this approach is to find the spectrum of the irregular time series by generalization of the Fourier transform. *Lomb-Scargle Periodogram (LSP)* [24] is one classical example of this approach. In LSP, we first fit the time series to a sine curve in order to obtain their spectrum, which can be used to calculate the power spectral density (e.g., by Fourier transform). Then the auto-correlation and cross-correlation functions can be found by taking the inverse Fourier transform (using iFFT) of the corresponding power spectral density and cross power spectral density functions [3, 22, 26]. The versatility of Wavelet transforms and their numerous applications in analysis of time series have motivated researchers to extend the transform to irregular time series [28, 12, 19]. LSP-based algorithms are specially designed for spectral analysis of the periodic signals. They do not perform well for non-periodic signals [27], and are not robust to the outliers in the data [22]. Furthermore, computing the entire cross-correlation function is a much more challenging task than our task of learning temporal dependence. Furthermore the produced correlation function can be a complex function, which is difficult to interpret and conduct theoretical analysis on.

Kernel-based Method The basic idea of kernel-based methods is to generalize regular correlation operator via kernels without the complex computation of correlation function. One classical example of this approach is the *Slotting Techniques* [22], which compute a weighted (kernelized) correlation of two irregularly sampled time series. For two irregular time series $\{(t_n^x, x_n)\}_{n=1}^{N_x}$ and $\{(t_m^y, y_m)\}_{m=1}^{N_y}$ the

correlation is computed as follows:

$$\hat{\rho}(l, \Delta t) = \frac{\sum_{n=1}^{N_x} \sum_{m=1}^{N_y} x_n y_m w_{l,\Delta t}\left(t_n^x, t_m^y\right)}{\sum_{n=1}^{N_x} \sum_{m=1}^{N_y} x_n y_m \left(t_m^y - t_n^x\right)}, \tag{3.4}$$

where w is the kernel function, l is the lag number, and Δt is the average sampling interval length. For example, w can be Gaussian kernel as follows:

$$w_{l,\Delta t}(t_1, t_2) = \exp\left(-\frac{(t_2 - t_1 - l\Delta t)^2}{\sigma^2}\right), \tag{3.5}$$

where σ is the kernel bandwith. The main advantage of kernel-based methods is that they are nonparametric and can be easily extended to any regression-based methods. Furthermore, they do not have specific assumptions on the data, e.g. the model assumption in the repair methods and the periodicity assumption in LSP-based methods.

4 Methodology

The major challenge to uncover temporal causal networks for irregular time series is how to effectively capture the temporal dependence without directly estimating the values of missing data or making restricted assumptions about the generation process of the time series. In this paper, we propose a simple yet extremely effective approach by generalizing the inner product operator via kernels so that regression-based temporal casual models can be applicable to irregular time series.

In this section, we describe in detail our proposed model, analyze its theoretical properties, and discuss one extension that takes into account the irregular patterns of the input time series for accurate prediction.

4.1 Generalized Lasso Granger (GLG) The key idea of our model is as follows: if we treat $\mathbf{a}_{i,j}$ in eq (3.3) as a time series, $\mathbf{a}_{i,j}^\top \mathbf{x}^{(j)}$ can be considered as its inner product with another time series $\mathbf{x}^{(j)}$. If we are able to generalize the inner product operator to irregular time series, the temporal causal models for regular time series can be easily extended to handle irregular ones.

Let us denote the generalization of dot product between two irregular time series $\mathbf{x}$ and $\mathbf{y}$ by $\mathbf{x} \odot \mathbf{y}$, which can be interpreted as a (possibly non-linear) function that measures the unnormalized similarity between them. Depending on the target application, one can define different similarity measures, and thus inner product definitions. For example, we can define the inner product as a function linear with respect to the first time series components as follows:

$$\mathbf{x} \odot \mathbf{y} = \sum_{n=1}^{N_x} \frac{\sum_{m=1}^{N_y} x_n y_m w\left(t_n^x, t_m^y\right)}{\sum_{m=1}^{N_y} w\left(t_n^x, t_m^x\right)}, \tag{4.6}$$

where w is the kernel function. For example w can be the Gaussian kernel defined as in Equation (3.5). Many other similarity measures for time series have been developed for the classification and clustering tasks [32, 34], and can be used for our dot product definition.

Given the generalization of the inner product operator, we can now extend the regression in Equation (3.3) to obtain the desired optimization problem for irregular time series. Formally, suppose P number of irregular time series $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(P)}$, are given. Let Δt denote the average length of the sampling intervals for the target time series (e.g. $\mathbf{x}^{(i)}$) and $\mathbf{a}'_{i,j}(t)$ be a pseudo time series, i.e.:

$$\mathbf{a}'_{i,j}(t) = \{(t_l, a_{i,j,l}) | l = 1, \ldots, L, t_l = t - l\Delta t\},$$

which means that for different value of t, $\mathbf{a}'_{i,j}(t)$ share the same observation vectors (i.e, $\{\mathbf{a}_{i,j}\}$), but the time stamp vectors vary according to the value of t. We can perform the causality analysis by generalized Lasso Granger (GLG) method that solves the following optimization problem:

$$\min_{\{\mathbf{a}_{i,j}\}} \sum_{n=\ell_0}^{N_i} \left\| x_n^{(i)} - \sum_{j=1}^{P} \mathbf{a}'_{i,j}(t_n^{(i)}) \odot \mathbf{x}^{(j)} \right\|_2^2 + \lambda \left\|\mathbf{a}_i\right\|_1, \tag{4.7}$$

where ℓ_0 is the smallest value of n that satisfies $t_n^{(i)} \geq L\Delta t$.

The above optimization problem is not convex in general and the convex optimization algorithms can only find a local minimum. However if the generalized inner product is defined to make Problem 4.7 convex, there are efficient algorithms such as FISTA [5] to solve optimization problems of the form $f(\boldsymbol{\theta}) + \|\boldsymbol{\theta}\|_1$ where $f(\boldsymbol{\theta})$ is convex. In this paper, we use the linear generalization of the inner product given by Equation (4.6) with which Problem (4.7) can be reformulated as linear prediction of $x_n^{(i)}$ using parameters $\mathbf{a}'_{i,j}(t_n^{(i)})$ subject to norm-1 constraint on the value of the parameters. Thus, the problem is a Lasso problem and can be solved more efficiently by optimized Lasso solvers such as [29].

4.2 Extension of GLG Method Notice that every data point $x_n^{(i)}$ in the GLG method has equal impact in the regression. We make the following two

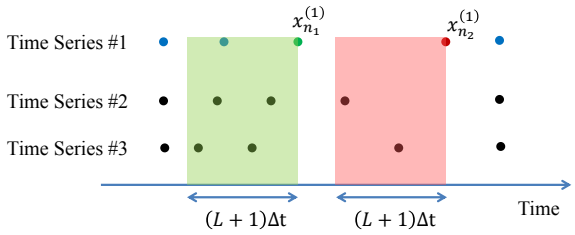

Figure 1: Time Series #1 is the target time series in this figure. Prediction of $x_{n_1}^{(1)}$ should receive a higher weight than $x_{n_2}^{(1)}$ in the depicted scenario because it can be predicted more accurately.

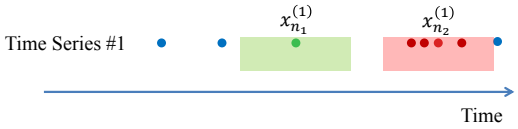

Figure 2: Time Series #1 is the target time series in this figure while the other time series are not shown. Prediction of $x_{n_2}^{(1)}$ should have higher weight in the causality inference than $x_{n_2}^{(2)}$ because $x_{n_2}^{(2)}$ is in a denser region of the time series.

observations, which could help improve the performance.

The first observation, depicted in Figure 1, states that the samples that have more data points for prediction will be predicted more accurately. Thus, they should have higher contribution in learning the causality graph. Thus we define the following weight for the subproblem of prediction of $x_n^{(i)}$,

$$v_1\left(x_n^{(i)}\right) = \sum_{j=1}^{P}\sum_{\ell=1}^{L}\sum_{m=1}^{N_j} w\left(t_n^{(i)} - \ell\Delta t, t_m^{(j)}\right).$$

The second observation in Figure 2 states that the learning should be uniformly distributed in time. In other words, the samples that are in a denser region should contribute less than those in sparse regions. Thus we define the following weights,

$$v_2\left(x_n^{(i)}\right) = \frac{w\left(x_n^{(i)}, x_n^{(i)}\right)}{\sum_{m=1}^{N_i} w\left(x_n^{(i)}, x_m^{(i)}\right)}$$

Using the introduced weights, we define the Weighted Generalized Lasso Granger (W-GLG) as follows:

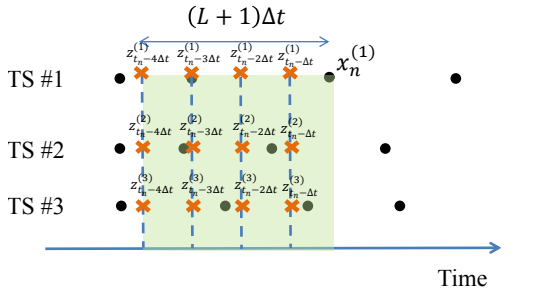

Figure 3: The sources of repair errors in GLG when $x_n^{(1)}$ is being predicted. In order to predict data point $x_n^{(1)}$ GLG repairs the time series in L points before the time t_n^1. At each point of time a repair error $z_{t-\ell\Delta t}^{(i)}$ is produced.

$$\min_{\{\mathbf{a}_{i,j}\}} \sum_{n=\ell_0}^{N_i} v_1\left(t_n^{(i)}\right) v_2\left(t_n^{(i)}\right) \left\| x_n^{(i)} - \sum_{j=1}^{P} \mathbf{a}'_{i,j}(t_n^{(i)}) \odot \mathbf{x}^{(j)} \right\|_2^2 + \lambda \left\|\mathbf{a}_i\right\|_1 \tag{4.8}$$

4.3 Asymptotic Consistency of GLG We follow the procedure in [18] to study the consistency of our generalized Lasso Granger Method. Suppose there are two instantiations for the set of random processes $\mathbf{x}^i$, one with regular sampling frequency and the other one with irregular sampling frequency.

Figure 3 shows the source of the errors when $x_n^{(1)}$ is being predicted using GLG. It can be seen in the figure that GLG interpolates the value of the time series in regular steps $[t_n - L\Delta t, \ldots, t_n - \Delta t]$ and uses them for prediction of $x_n^{(1)}$. We can model the errors induced at time $t_n - \ell\Delta t$ in the j^{th} time series by $z_{t_n-\ell\Delta t}^{(j)}$ and define the regular time series $\tilde{x}_t^{(i)} = x_t^{(i)} + z_t^{(i)}$ for $i = 1, \ldots, P$ and $t = t_n - L\Delta t, \ldots, t_n - \Delta t$. Now, it suffices to show that the graph inferred using the regular time series $\tilde{x}_t^{(i)}$ will produce the same graph compared to the case that we had the original regular time series $x_t^{(i)}$.

In order to proceed with the proof we need to assume that $\tilde{x}_t^{(i)}$ are Gaussian random variables with zero mean. Similar to the regular case, let $\tilde{\mathbf{X}}_t^{(i)}$ be the vector of length n of samples of $\tilde{x}_i^t$. We can write GLG as,

$$\tilde{\mathbf{a}}_{i,j}^{\lambda} = \arg\min_{\mathbf{a}_{i,j}} \left(n^{-1} \left\| \tilde{\mathbf{X}}_t^{(i)} - \sum_j \tilde{\mathbf{X}}^{(j)} \mathbf{a}_{i,j} \right\|_2^2 + \lambda \left\|\mathbf{a}_{i,j}\right\|_1 \right). \tag{4.9}$$

Let Γ denote the set of all the time series. Suppose we have the regular time series. The neighborhood $\text{ne}_{(i)}$ of a time series $\mathbf{x}^{(i)}$ is the smallest subset

of $\Gamma\backslash\{(i)\}$ so that $\mathbf{x}(i)$ is conditionally independent of all the variables in the remaining time series. Now, define the estimated neighborhood of $\mathbf{x}^{(i)}$ using irregular time series by $\tilde{\text{ne}}^{\lambda}_{(i)} = \{\mathbf{x}^{(j)} | \tilde{\mathbf{a}}^{\lambda}_{i,j} \neq \mathbf{0}\}$ in the solution of the above problem.

THEOREM 4.1. *Let assumptions 1-6 in [18] hold for the solution of Problem (4.9).*

(a) Suppose $\mathbb{E}[z_p^{(i)} \tilde{x}_q^{(i)}] = 0$ *for all* $i = 1, \ldots, N$ *and* $p, q = t_n, \ldots, t_n - L\Delta t, p \neq q$. *Let the penalty parameter satisfy* $\lambda \sim dn^{-(1-\epsilon)/2}$ *with* $\kappa < \epsilon < \xi$ *and* $d > 0$. *There exists some* $c > 0$ *so that for all* $x_i^t \in \Gamma$,

$$\mathbb{P}\left(\tilde{ne}^{\lambda}_{(i)} \subseteq ne_{(i)}\right) = 1 - O\left(\exp\left(-cn^{\epsilon}\right)\right), \quad \text{for } n \to \infty.$$

$$\mathbb{P}\left(ne_{(i)} \subseteq \tilde{ne}^{\lambda}_{(i)}\right) = 1 - O\left(\exp\left(-cn^{\epsilon}\right)\right), \quad \text{for } n \to \infty.$$

(b) The above results can be violated if $\mathbb{E}[z_p^{(i)} x_q^{(i)}] \neq 0$ *for all* $i = 1, \ldots, N$ *and* $p, q = t_n, \ldots, t_n - L\Delta t, p \neq q$.

Proof. A proof is given in the Appendix.

The theorem specifies a sufficient condition with which the temporal causal graph inferred by the irregular time series is asymptotically equal to the graph that could be inferred if the time series were available. The condition states that if the repairing error during GLG's operation is orthogonal to the repaired data points, the inferred temporal causal graph will be the same as the actual one with probability approaching one as the length of the time series grows.

Choice of Kernel in GLG Method

Consider the non-uniformly sampled time series. Suppose the non-uniformity is due to clock jitter which is modeled by iid zero-mean Gaussian variables at different sampling times. If the kernel function w satisfies the following equation, the neighborhood learned by our algorithm is guaranteed to be the same as the regular time series case:

$$\sum_{\ell=1}^{N} \left[K(\ell + a_\ell, q) - K(p, q)\right] w(\ell + a_\ell, p) = 0, \tag{4.10}$$

for all $i, \ell = 1, \ldots, N$. In this equation $K(t, t')$ is the covariance function of the time series and a_j is the random variable modeling the clock jitter at j^{th} time stamp.

Proof. By Theorem 4.1, it is sufficient to show that the following equation holds for all $p, q = 1, \ldots, N$ which means error should be orthogonal to all the observations used in Lasso-Granger.

$$\mathbb{E}_x\left[\frac{\sum_{\ell=1}^{N}(x^{(i)}_{(\ell+a_\ell)} - x_p^{(i)}) w(\ell + a_\ell, p)}{\sum_{\ell=1}^{N} w(\ell + a_\ell, p)} x_q^{(i)}\right] = 0. \tag{4.11}$$

The expectation is with respect to x, thus:

$$\mathbb{E}_x\left[\left(\sum_{\ell=1}^{N}(x^{(i)}_{(\ell+a_\ell)} - x_p^{(i)}) w(\ell + a_\ell, p)\right) x_q^{(i)}\right] = 0. \tag{4.12}$$

Taking the expectation and using the definition of the covariance function $K(t, t') = \mathbb{E}_x\left[x_t^{(i)} x_{t'}^{(i)}\right]$ yields the result. □

A quick inspection of Equation (4.10) shows that the Gaussian kernel does not satisfy it. It is clear by setting $q = p$ and noting that the covariance function always satisfies $K(\ell + a_\ell, p) \leq K(p, p)$, any non-negative kernel such as the Gaussian kernel cannot satisfy Equation (4.10). However if the time series is smooth enough so that $K(t, p) \approx K(p, p)$ for t in a neighborhood of p, the Gaussian kernel can attenuate the effect of $K(t, p) - K(p, p)$ for values of t outside the neighborhood and the left side of Equation (4.10) can become close to zero. In conclusion, while the Gaussian kernel is a suboptimal kernel to use for causality analysis purposes, for smooth time series it approximately satisfies Equation (4.10) and is expected to perform well.

Comparison of GLG Method with Time Decay Kernels and Locally Weighted Regression

Analysis of the consistency of the repair methods such as Kernel Weighted Regression algorithm can be done similar to the analysis of GLG method by introducing the repair error variables. However GLG is expected to have lower absolute error because it tries to predict the actual observations $x_n^{(i)}$ without additional repair error. In contrast, repair methods such as LWR first interpolate the time series at regular time stamps; then try to predict the *repaired* samples which carry repair errors with themselves.

As alluded in Section 3, the repair methods introduce huge amount of error during reconstruction of the irregular time series in the large gaps, see Figure 4. In contrast, since there is no sample in the gaps, GLG does not attempt to predict the value of the time series in the gaps and avoids these types of errors.

5 Experiment Results

In order to examine performance of GLG, we perform two series of experiments, one on the synthetic

Figure 4: The black circles show the time stamps of the given irregular time series. The crosses are the time stamps used for repair of the time series. The red time stamp shows the moment in that the repair methods produce large errors and propagate the error by predicting the erroneous repaired sample. GLG skips these intervals because it predicts only the observed samples.

datasets, and the other one on the Paleo dataset for identifying the monsoon climate patterns in Asia.

5.1 Synthetic Datasets Design of synthetic datasets with irregular sampling time have been discussed extensively in the literature. We follow the methodology in [22] to create four different synthetic datasets. These datasets are constructed to emulate different sampling patterns and functional behavior.
Mixture of Sinusoids with Missing Data (MS-M): In order to create the MS-M dataset, we generate $P-1$ time series according to a mixture of several sinusoidal functions:

$$x_t^{(i)} = \sum_{j=1}^{3} A_j^{(i)} \cos(2\pi f_j^{(i)} t + \varphi_j^{(i)}) + \epsilon^{(i)}, \quad \text{for } i = 2, \ldots, P \tag{5.13}$$

where $f_j^{(i)} \sim Unif(f_{min}, f_{max})$, $\varphi_j^{(i)} \sim Unif(0, 2\pi)$ and the vector of amplitudes $A_j^{(i)}$ is distributed as $Dirichlet(\mathbf{1}_{3\times 1})$ so that all the time series have a least amount of power. The range of the frequencies is selected in a way that only few periods of sines will be repeated in the dataset. The noise term $\epsilon^{(i)}$ is selected to be zero-mean Gaussian noise. We create the target time series according to a linear model,

$$x_t^{(1)} = \sum_{i=2}^{P} \sum_{\ell=1}^{L} \alpha_\ell^{(i)} x_{t-\ell}^{(i)} + \epsilon,$$

We set $\alpha_\ell^{(i)}$ to be sparse to model sparse graphs. Finally we drop samples from all the time series independently with probability p_m. We create 20 instances of MS-M random datasets and report the average performance on them.
Mixture of Sinusoids with Jittery Clock (MS-J): For creating this dataset, first we create the sampling time stamps for the target time series: $\mathbf{t}_1 = [1, \ldots, N] + \mathbf{e}_1$ where the random vector $\mathbf{e}_1 = [e_{11}, \ldots, e_{1N}]$ is a zero mean Gaussian random vector with covariance matrix $\gamma\mathbf{I}$ and γ is called the jitter variance. We select the parameters of the other time series and use them to calculate the value of the target time series at the given time stamps:

$$x_n^{(1)} = \sum_{i=2}^{P} \sum_{\ell=1}^{L} \alpha_\ell^{(i)} \sum_{j=1}^{3} A_j^{(i)} \cos(2\pi f_j^{(i)}(t_n - \ell) + \varphi_j^{(i)}) + \epsilon,$$

Then we produce the sampling times for the other time series similarly as a Gaussian random vector with mean $[1, \ldots, N]$ and covariance matrix $\gamma\mathbf{I}$ and use Equation (5.13) to produce the time series. We create 50 instances of MS-J random datasets and report the average performance on them.
Auto-Regressive Time Series with Missing Data (AR-M): The procedure of creating the AR-M dataset is similar to the one described for MS-M with a single difference in producing time series 2 to P according to AR processes:

$$x_n^{(i)} = \sum_{\ell=1}^{3} \beta_{i,\ell} x_{t_n-\ell}^{(i)} + \epsilon, \quad \text{for } i = 2, \ldots, P$$

where $\beta_{i,\ell}$ are chosen randomly while keeping the AR time series stable. We create 50 instances of AR-M random datasets and report the average performance on them.
Mixture of Sinusoids with Poisson Process Sampling times (MS-P): The procedure of creating the MS-P dataset is similar to the one described for MS-J with the difference of producing the sampling times according to a Poisson Point process; i.e. inter-sampling times are distributed according to $Exp(1)$. We create 50 instances of MS-P random datasets and report the average performance on them.
Baselines We compare the performance of our algorithm with four state-of-the-art algorithms. We use Locally Weighted Regression (LWR) and Gaussian Process (GP) regression to repair the time series and perform the regular Lasso-Granger. The Slotting Technique [22] and the LSP method are the two algorithms used for finding mutual correlation. Since performance of W-GLG is close to the performance of GLG, performance of W-GLG will be compared in only one figure to avoid cluttered plots.
Performance Measures In order to report the accuracy of the inferred graph, we use the Area Under the Curve (AUC) score. The value of AUC is the probability that the algorithm will assign a higher value to a randomly chosen positive (existing) edge

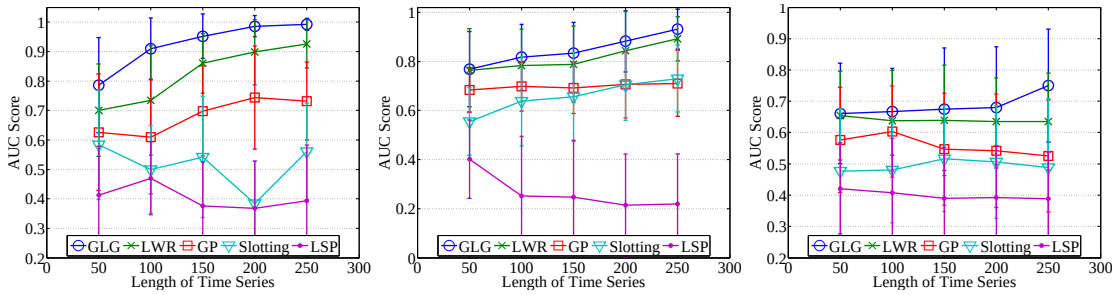

Figure 5: Study of convergence behavior of the algorithms in the Mixture of Sinusoids with (left) Missing data points dataset, (middle) Jittery clock dataset and (right) Autoregressive Time Series with Missing data points dataset.

than a randomly chosen negative (non-existing) edge in the graph.

Parameter Tuning Unless otherwise stated, in all of the kernel-based methods (GLG, LWR, and Slotting technique) we use Gaussian kernel with bandwidth $\sigma = \frac{\Delta t}{2}$. We select the value of λ in GLG by 5-fold cross-validation.

5.2 Results on the Synthetic Datasets .

Experiment #1: Convergence Behavior of The Algorithms In Figures ??, we increase the length of the time series in datasets MS-M and MS-J to study the convergence behavior of GLG. Both figures demonstrate excellent convergence behavior of GLG and the fact that GLG consistently outperforms other algorithms by a large margin. Note that the convergence behavior of LWR is very similar to GLG; but as alluded in the theoretical analysis, GLG has lower absolute error. The Slotting technique and the LSP method are designed for analysis of cross-correlation of time series and as expected do not perform well in our settings. The poor performance of LSP can be linked to the periodicity of signals assumption in LSP analysis method, which is violated in our dataset.

Experiment #2: Comparison of Different Kernels In order to study the effect of different kernels on performance of GLG we test the convergence behavior of GLG with three different kernels. The kernels are Gaussian: $w(t_1, t_2) = \exp(-(t_1 - t_2)^2/\sigma)$, Sinc: $w(t_1, t_2) = \sigma \sin((t_1 - t_2)/\sigma)/(t_1 - t_2)$, and the Inverse distance kernel $w(t_1, t_2) = (\|t_1 - t_2\|_2^2)^{-1}$. As shown in Figure 6, as pointed out in the analysis section the Gaussian kernel shows acceptable convergence behavior. Performance of the Inverse Distance kernel suggests an asymptotic convergence behavior but with much higher absolute error. The Sinc kernel does not show properly converge.

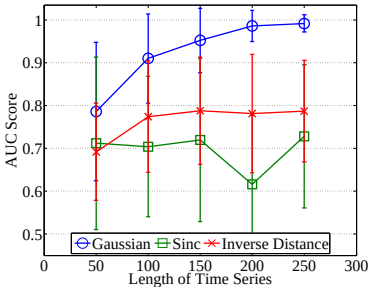

Figure 6: Convergence of different kernels in the Mixture of Sinusoids with missing dataset.

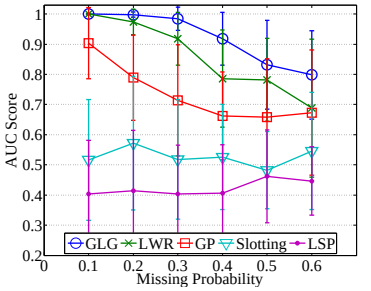

Figure 7: The effect of missing data in the Mixture of Sinusoids with missing data points.

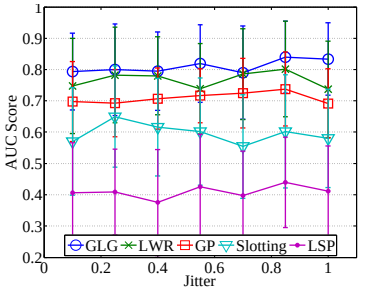

Figure 8: The effect of clock jitter (γ) the Mixture of Sinusoids with Jittery Clock dataset.

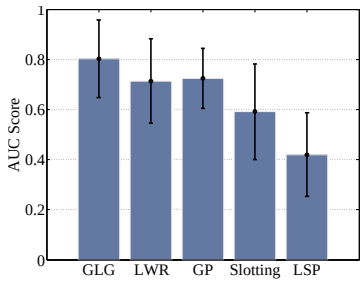

Figure 9: Performance comparison of the algorithms in the Mixture of Sinusoids with Poisson points sampling times.

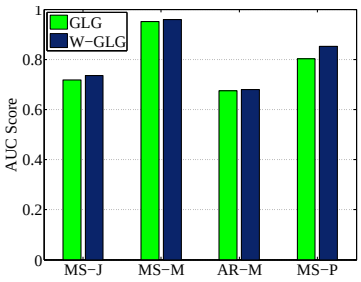

Figure 10: Comparison of performance of W-GLG vas. GLG in all four datasets.

Experiment #3: The Impact of Missing Rate The effect of missing data is examined in Figure 7. It is clear that as the probability of missing a data decreases GLG becomes more accurate. Note (i) the superior performance of GLG compared to others, (ii) When only 10% of the data points are missing GLG perfectly uncovers the correct causality relationship between the variables.

Experiment #4: Clock Jitter Impact The effect of clock jitter (γ) is examined in Figure 8. GLG is robust with respect to the amount of clock jitter. This is because (i) the time series are smooth, (ii) we are testing in a reasonable range of jitter.

Experiment #5: Performance of the Algorithms in Extremely Irregular Datasets Due to the limited space we only report the result of one experiment on the Mixture of Sinusoids with Poisson points sampling times, see Figure 9. In this extremely irregularly sampled dataset, our algorithm outperforms other algorithms by a large margin. This is due to high possibility of large gaps in this dataset and the fact that GLG avoids huge errors caused by interpolation in the gaps.

Experiment #6: Comparison of W-GLG and GLG Figure 10 compares the performance of weighted version of our algorithm (WGLG) with the non-weighted version (GLG) in all the four synthetic datasets. WGLG performs only marginally better than GLG in all the datasets except the dataset with Poisson sampling points. This is because only this extremely irregular dataset provides the situation for weights to show their advantages.

5.3 Paleo Dataset Now we apply our method to a Climate dataset to discover the weather movement patterns. Climate scientist usually rely on models with enormous number of parameters that are needed to be measured. The alternative approach is the data-centric approach which attempts to find the patterns in the observations.

The Paleo dataset which is studied in this paper is the collection of density of $\delta^{18}O$, a radio-active isotope of Oxygen, in four caves across China and India. The geologists were able to find the estimate of $\delta^{18}O$ in ancient ages by drilling deeper in the wall of the caves. The data are collected from Dandak [6], Dongge [31], Heshang [15], Wanxiang [33] caves, see Figure 11, with irregular sampling pattern described in Table 1. The inter-sampling time varies from high resolution 0.5 ± 0.35 to low resolution 7.79 ± 9.79; however there is no large gap between the measurement times.

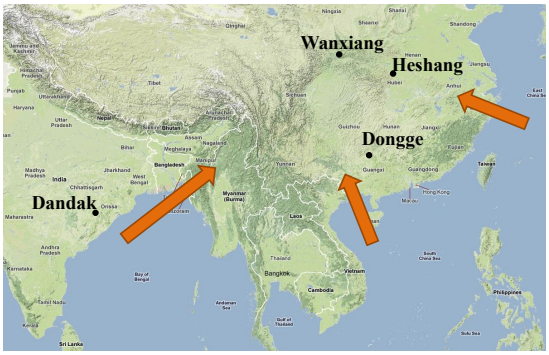

Figure 11: Map of the locations and the monsoon systems in Asia.

The density of $\delta^{18}O$ in all the datasets is linked to the amount of precipitation which is affected by the Asian monsoon system during the measurement period. Asian monsoon system, depicted in Figure 11, affects a large share of world's population by transporting moisture across the continent. The movement of monsoonal air masses can be discovered by analysis of their $\delta^{18}O$ trace. Since the datasets are collected from locations in Asia, we are able to

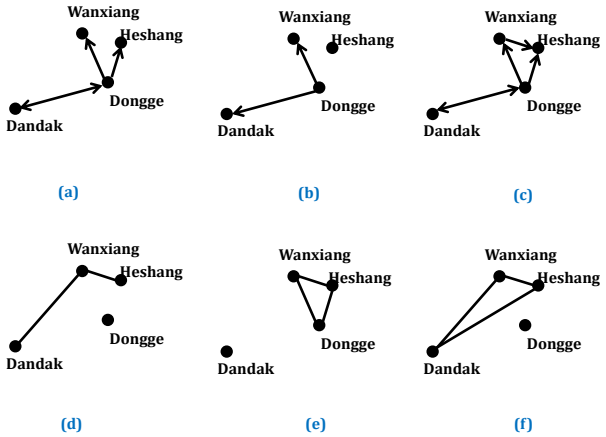

Figure 12: Comparison of the results on the Paleo Dataset: (a) GLG in period 850AD-1563AD. (b) GLG in the period 1250AD-1564AD. (c) GLG in the period 850AD-1250AD. (d) Slotting technique in period 850AD-1563AD. (e) Slotting technique in the period 1250AD-1564AD. (f) Slotting technique in the period 850AD-1250AD.

analyze the spatial variability of the Asian monsoon system.

In order to analyze the spatial transportation of the moisture we normalize all the datasets by subtracting the mean and divide them by their standard deviation. We use GLG with the Gaussian kernel with bandwidth equal to $0.5(y)$ and maximum lag of $25(y)$; i.e. $L = 50$. In order to compare our results with the results produced by the slotting method in [22] we analyze the spatial relationship among the locations in three age intervals: (i) The entire overlapping age interval 850AD-1563AD, (ii) The Cold phase 850AD-1250AD, and (ii) The medieval warm period 1250AD-1563AD. Figure 12 compares the graphs produced by GLG with the ones reported by [22].

Table 1: Description of the Paleo Dataset.

Location	Measur. Period	Δt	Std$(t_n - t_{n-1})$
Dandak	624AD-1562AD	0.50 (y)	0.35 (y)
Dongge	6930BC-2000AD	4.21 (y)	1.63 (y)
Heshang	7520BC-2002AD	7.79 (y)	9.78 (y)
Wanxiang	192AD-2003AD	2.52 (y)	1.19 (y)

5.4 Results on the Paleo Dataset Figure 12 parts (a) and (d) show the results of causality analysis with GLG and slotting technique, respectively. Our results identify two main transportation patterns. First, the edges from Dongge to other locations which can be interpreted as the effect of movement of air masses from southern China to other regions via the East Asian Monsoon System (EAMS). Second, an edge from Dandak to Dongge which shows the Indian Monsoon System (IMS) significantly affects Dongge in southern China. The graph in the period 1250AD-1563AD is sparser than the graph in 850AD-1250AD which can be due to the fact that the former age period is a cold period, in which air masses do not have enough energy to move from India to China, while in contrast the latter age period is a warm phase and the air masses initiated in India impact southern China regions. During the warm period we can see that other branches of EAMS are also more active which result in denser graph in the warm period.

The differences between our results and the results from Slotting technique can be because of the fact that in the Slotting technique an edge is positively identified even if two time series have significant correlation at zero lag. However, by the definition of Granger causality, only past values of one time series should help prediction of the other one in order to be considered as a cause of it. The significant correlation at zero lag can be due to either fast movement of the air masses or production of the $\delta^{18}O$ by an external source, such as changes in the sun's radiation strength, that impacts all the places with the same amount. Thus, the correlation value at zero lag cannot be a reliable sign for inference about the movement of the air masses.

The sparsity of the identified causality graphs can be due to several reasons. While we cannot rule out the possibility of non-linear causality relationship between the locations, as [22] have noticed, the sparsity can happen because the relationships are either in large millennial scales or short annual scales. In the former case the relationship cannot be captured through analysis of periods with length several centuries. In the latter case, the resolution of the dataset is in the order of 3-4 years which does not capture annual or biennial links.

6 Conclusion and Future Work

In this paper, we propose a nonparametric generalization of the Granger graphical models (GLG) to uncover the temporal causal structures from irregular time series. We provide theoretical proof on the consistency of the proposed model and demonstrate its effectiveness on four simulation data and one real-application data. For future work, we are interested in investigating how to design effective kernels for the GLG model and extending the algorithm to large scale data analysis.

Acknowledgement

We thank Umaa Rebbapragada from JPL for discussing the problem with us, Kira Rehfeld from PIK for sharing the Paleo dataset with us and the anonymous reviewers for valuable comments. This research was supported by the NSF research grants IIS-1134990.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Dec. 1974.

[2] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *KDD '07*, page 66, New York, New York, USA, Aug. 2007. ACM Press.

[3] P. Babu and P. Stoica. Spectral analysis of nonuniformly sampled data - a review. *Digital Signal Processing*, 20(2):359–378, Mar. 2010.

[4] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for Gaussian variables. Oct. 2009.

[5] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183, Jan. 2009.

[6] M. Berkelhammer, A. Sinha, M. Mudelsee, H. Cheng, R. L. Edwards, and K. Cannariato. Persistent multidecadal power of the Indian Summer Monsoon. *Earth and Planetary Science Letters*, 290(1-2):166–172, 2010.

[7] D. P. Bertsekas and D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, Sept. 1999.

[8] G. E. P. Box and G. M. Jenkins. *Time series analysis; forecasting and control [by] George E. P. Box and Gwilym M. Jenkins*. Holden-Day San Francisco,, 1970.

[9] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9849–54, June 2004.

[10] J. C. Cuevas-Tello, P. Tino, S. Raychaudhury, X. Yao, and M. Harva. Uncovering delayed patterns in noisy and irregularly sampled time series: an astronomy application. *Pattern Recognition*, 43(3):36, Aug. 2009.

[11] J. B. Elsner. Granger causality and Atlantic hurricanes. *Tellus - Series A: Dynamic Meteorology and Oceanography*, 59(4):476–485, 2007.

[12] G. Foster. Wavelets for period analysis of unevenly sampled time series. *The Astronomical Journal*, 112:1709, Oct. 1996.

[13] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):pp. 424–438, 1969.

[14] W. K. Harteveld, R. F. Mudde, and H. E. A. Van Den Akker. Estimation of turbulence power spectra for bubbly flows from Laser Doppler Anemometry signals. *Chemical Engineering Science*, 60(22):6160–6168, 2005.

[15] C. Hu, G. M. Henderson, J. Huang, S. Xie, Y. Sun, and K. R. Johnson. Quantification of Holocene Asian monsoon rainfall from spatially separated cave records. *Earth and Planetary Science Letters*, 266(3-4):221–232, Feb. 2008.

[16] D. M. Kreindler and C. J. Lumsden. The effects of the irregular sample and missing data in time series analysis. *Nonlinear dynamics, psychology, and life sciences*, 10(2):187–214, Apr. 2006.

[17] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW '10*, pages 601–610, New York, NY, USA, 2010. ACM.

[18] N. Meinshausen and P. Bühlmann. High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, June 2006.

[19] D. Mondal and D. B. Percival. Wavelet variance analysis for gappy time series. *Annals of the Institute of Statistical Mathematics*, 62(5):943–966, Sept. 2008.

[20] G. Nolte, A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller. Robustly estimating the flow direction of information in complex physical systems. *Physical review letters*, 100(23):234101, June 2008.

[21] C. D. . V. Panchenko. Modified hiemstra-jones test for granger non-causality. Computing in Economics and Finance 2004 192, Society for Computational Economics, 2004.

[22] K. Rehfeld, N. Marwan, J. Heitzig, and J. Kurths. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics*, 18(3):389–404, 2011.

[23] S. Ryali, K. Supekar, T. Chen, and V. Menon. Multivariate dynamical systems models for estimating causal interactions in fMRI. *NeuroImage*, 54(2):807–23, Jan. 2011.

[24] J. D. Scargle. Studies in astronomical time series analysis. I - Modeling random processes in the time domain. *The Astrophysical Journal Supplement Series*, 45(Jan):1–71, 1981.

[25] M. Schulz and K. Stattegger. Spectrum: spectral analysis of unevenly spaced paleoclimatic time series. *Computers & Geosciences*, 23(9):929–945, 1997.

[26] P. Stoica, P. Babu, and J. Li. New Method of Sparse Parameter Estimation in Separable Models and Its Use for Spectral Analysis of Irregularly Sampled Data. *IEEE Transactions on Signal Processing*,

59(1):35–47, Jan. 2011.
[27] P. Stoica, J. Li, and H. He. Spectral Analysis of Nonuniformly Sampled Data: A New Approach Versus the Periodogram. *IEEE Transactions on Signal Processing*, 57(3):843–858, Mar. 2009.
[28] W. Sweldens. The Lifting Scheme: A Construction of Second Generation Wavelets. *SIAM Journal on Mathematical Analysis*, 29(2):511, Mar. 1998.
[29] R. Tibshirani, I. Johnstone, T. Hastie, and B. Efron. Least angle regression. *The Annals of Statistics*, 32(2):407–499, Apr. 2004.
[30] P. A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1457):969–81, May 2005.
[31] Y. Wang, H. Cheng, R. L. Edwards, Y. He, X. Kong, Z. An, J. Wu, M. J. Kelly, C. A. Dykoski, and X. Li. The Holocene Asian monsoon: links to solar changes and North Atlantic climate. *Science (New York, N.Y.)*, 308(5723):854–7, May 2005.
[32] L. Wu, C. Faloutsos, K. P. Sycara, and T. R. Payne. Falcon: Feedback adaptive loop for content-based retrieval. In *VLDB '00*, VLDB '00, pages 297–306, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
[33] P. Zhang, H. Cheng, R. L. Edwards, F. Chen, Y. Wang, X. Yang, J. Liu, M. Tan, X. Wang, J. Liu, C. An, Z. Dai, J. Zhou, D. Zhang, J. Jia, L. Jin, and K. R. Johnson. A test of climate, sun, and culture relationships from an 1810-year Chinese cave record. *Science (New York, N.Y.)*, 322(5903):940–2, Nov. 2008.
[34] Y. Zhu and D. Shasha. Warping indexes with envelope transforms for query by humming. In *SIGMOD '03*, SIGMOD '03, pages 181–192, New York, NY, USA, 2003. ACM.

7 Appendix

Proof of Theorem 4.1

Let The error variables $z_\ell^{(i)}$ are zero mean Gaussian variables due to zero mean Gaussian assumption about the distribution of $\mathbf{X}^{(i)}$ and $\tilde{\mathbf{X}}^{(i)}$. Following the proof in [18], construct the following correlation functions:

$$G_{k,\ell}(\mathbf{a}) = -2n^{-1}\left\langle \mathbf{X}_t^{(i)} - \sum_j \mathbf{X}^{(j)}\mathbf{a}_{i,j}, \mathbf{X}_\ell^{(k)} \right\rangle \tag{7.14}$$

and

$$\tilde{G}_{k,\ell}(\tilde{\mathbf{a}}) = -2n^{-1}\left\langle \tilde{\mathbf{X}}_t^{(i)} - \sum_j \tilde{\mathbf{X}}^{(j)}\tilde{\mathbf{a}}_{i,j}, \tilde{\mathbf{X}}_\ell^{(k)} \right\rangle \tag{7.15}$$

(a) If $\mathbb{E}[z_p^{(i)}\tilde{x}_q^{(i)}] = 0$ for all $i = 1, \ldots, N$ and $p, q = t_n, \ldots, t_n - L\Delta t, p \neq q$, the inner products in Equations (7.14) and (7.15) are the same and sufficiency part of Lemma A.1 in [18] guarantees the same $\hat{\mathbf{a}}_{k,\ell} = \tilde{\mathbf{a}}_{k,\ell}$. This proves that the neighborhoods inferred by analysis of the repaired time series is equal to the ones obtained by analysis of the regular time series. Now using the assumptions 1-6 and an application of Theorems 1 and 2 in [18] concludes the proof. □

(b) Let, w.l.o.g, $\mathbf{a}_{k,\ell} > 0$ for some (k, ℓ). The necessity part of the Lemma A.1 in [18] guarantees that $G_{k,\ell}(\mathbf{a}) = -\lambda$. It is clear that by taking into account the $\mathbf{Z}$ in Equation (7.15), we can have $G_{k,\ell}(\tilde{\mathbf{a}}) > -\lambda$. (Note that we cannot show $G_{k,\ell}(\tilde{\mathbf{a}}) > -\lambda$ happens only with diminishing probability.) Now the sufficiency condition is not satisfied and we are no longer guaranteed to have $\tilde{\mathbf{a}}_{k,\ell} = 0$. Moreover, if the solution for Problem (7.15) is not unique and $-\lambda < G_{k,\ell}(\tilde{a}) < \lambda$ then $\tilde{\mathbf{a}}_{k,\ell} = 0$ and the solution of Problems (7.14) and (7.15) are indeed different. □