JMLR: Workshop and Conference Proceedings 1–11

Scalable Linear Causal Inference for Irregularly Sampled Time Series with Long Range Dependencies

Francois Belletti, Evan Sparks, Alexandre Bayen, and **Joseph E. Gonzalez,** *UC Berkeley*

Abstract

Linear causal analysis is central to a wide range of important application spanning finance, the physical sciences, and engineering Tsay (2005); Brillinger (1981); Mudelsee (2013). Much of the existing literature in linear causal analysis operates in the *time domain*. Unfortunately, the direct application of time domain linear causal analysis to many real-world time series presents three critical challenges: *irregular temporal sampling*, *long range dependencies*, and *scale*. Real-world data is often collected at irregular time intervals across vast arrays of decentralized sensors and with long range dependencies Doukhan et al. (2003) which make naive time domain correlation estimators spurious Granger (1988). In this paper we present a *frequency domain* based estimation framework which naturally handles irregularly sampled data and long range dependencies while enabling memory and communication avoiding distributed processing of time series data. By operating in the frequency domain we eliminate the need to interpolate and help mitigate the effects of long range dependencies. We implement and evaluate our new work-flow in the distributed setting using Apache Spark and demonstrate that we can accurately recover causal structure at scale on massive financial data.

1. Introduction

In many applications of time series analysis Abergel et al. (2012); Tsay (2005); Mudelsee (2013), tone is interested in estimating the mutual linear predictive properties of events from time series data corresponding to a collection of data streams each of which is a series of pairs *(timestamp, observation)*. Observations practically often occur at random, unevenly spaced and unaligned time stamps. In such a setting we therefore consider two underlying processes $(X_t)_{t\in\mathbb{R}}$ and $(Y_t)_{t\in\mathbb{R}}$ that are only observed at discrete and finite timestamps in the form of two collections of data points: $(x_{t_x})_{t_x\in I_x}$, $(y_{t_y})_{t_y\in I_y}$.

In our setting, we want to determine whether knowledge of the past observations of a given process (X) helps better predict the future observations of another process (Y). We illustrate this notion with two irregularly observed processes on Figure 1 which is the practical setting of a compelling example for causal inference: prices of stocks. These prices are indeed only irregularly observed as stock exchanges only report quotes and trades whenever a trading event occurs.

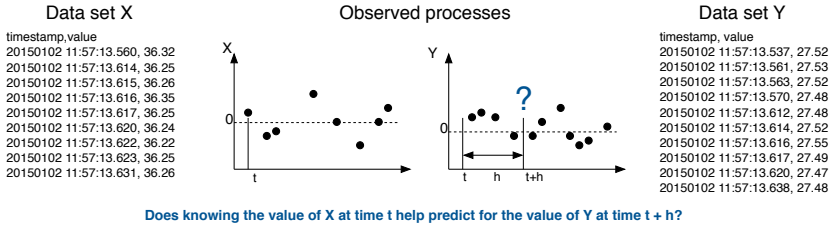

Figure 1: Causal inference for irregularly observed processes.

We adopt the cross-correlogram based causality estimation approach developed in Huth and Abergel (2014), which is similar to Granger's definition of causality as linear predictive ability of $(dX_{s<t})$ and $(dY_{s<t})$ for the random variable dX_t Granger (1969).

Let (X) and (Y) be two Wiener processes. We consider that (X) has a causal effect on (Y) if $(dX_{s<t})$ is a more accurate linear predictor of dY_t in square norm error than $(dY_{s<t})$ is an accurate linear predictor of dX_t. In other words (X) *"causes"* (Y) if and only if $E\left[(dX_t - E(dX_t|dY_s, s<t))^2\right] > E\left[(dY_t - E(dY_t|dX_s, s<t))^2\right]$. In order to quantify the magnitude of this statistical causation, Huth and Abergel introduced in Huth and Abergel (2014) the Lead-Lag Ratio (LLR) between (X) and (Y) as $LLR_{X\Rightarrow Y} = \sum_{h>0}\rho^2_{XY}(h)/\sum_{h<0}\rho^2_{XY}(h)$ where $\rho_{XY}(\cdot)$ is the cross-correlation between the second order stationary processes (X) and (Y). The analysis conducted in Huth and Abergel (2014) proved (X) *"causes"* (Y) is equivalent to $LLR_{X\Rightarrow Y} < 1$ thereby yielding an indicator of causation intensity between processes. For this definition to make sense, it is crucial that the increments of the processes we consider are second order stationary. Theoretical arguments and a varied series of examples in our experiments will show that the techniques we develop are also ready to deal with other types of Long Range Dependent processes Doukhan et al. (2003).

1.1. Challenges with real world data:

Unfortunately, in practical applications, time series data sets often present three main challenges that hinder the estimation of linear causal dependencies. **Irregular Sampling:** Observations are collected at irregular intervals both within and across processes complicating the application of standard causal inference techniques that rely on regularly spaced synchronous observations. **Long Range Dependencies (LRD):** Long range dependencies can result in increased and non vanishing variance in correlation estimates which misleads practitioners into identifying correlation where there is none. **Scale:** Real-world time series are often very large and therefore are often stored in distributed fashion.

In the following we show as in Huth and Abergel (2014) that naive interpolation of irregularly sampled data may yields spurious causality inference measurements. We also prove that eliminating LRD is crucial in order to obtain consistent correlation estimates. Unfortunately, standard time domain LRD erasure requires sorting the data chronologically and is therefore costly in the distributed setting. These costs are further exacerbated by time domain fractional differentiation which scales quadratically with the numbers of samples.

To address these three critical challenge we propose a Fourier transform based approach to causal inference. Projecting on a Fourier basis can be done with a simple sum operator for irregularly sampled data as described in Parzen (2012). A novel and salient byproduct of our estimation technique is that there is no need to sort the data chronologically or gather the data of different sensors on the same computing node. We use Fourier transforms as a signal compressing representation where cross-correlations and causal dependencies can be estimated with sound statistical methods while minimizing memory and communication overhead. Our method does not require sub-sampling in which aliasing obscures short-range interactions.

In section 2 we prove that our method is communication avoiding and provides consistent spectral estimators Brillinger (1981) for cross-dependencies. We first compress the time series locally by projecting *without interpolation or reordering* directly onto a reduced Fourier basis. Spectral estimation then occurs in the *frequency domain* prior to being translated back into the time domain

with an inverse Fourier transform. The resulting output can be used to compute unbiased Lead-Lag Ratios and thereby identify statistical causation.

In section 3, we provide a method to approximately erase LRD in the frequency domain, which has tremendous computational advantages as opposed to time domain based methods. Our analysis of LRD erasure as fractional pole elimination in frequency domain guarantees the causal estimates we obtain are not spurious Doukhan et al. (2003); Granger (1988).

In section 4, we present a novel analysis of the trade-off between estimator variance and communication bandwidth which precisely assesses the cost of compressing time series prior to analyzing them. A three-fold analysis establishes the statistical soundness of the contributions that address the three issues mentioned above. Studying data on compressed representations comes at an expected statistical cost. In our setting this supplementary variance can be decreased in an iterative manner and with bounded memory cost on a single machine. These properties cannot be replicated to the best of our knowledge by time domain based sub-sampling. Finally, we apply these methods to massive real financial market trade tables.

2. Addressing the issues caused by irregular sampling

In this section, we first review existing techniques for interpolated time-domain estimation of second-order statistics in the context of sparse and random sampling along the time axis. Interpolating data is a usual solution in order to be able to use classic time series analysis Parzen (2012); Wiener (1949); Friedman (1962); Linsay (1991). Unfortunately it can create spurious causality estimates.

2.1. Issues with second order statistics and interpolated data

In order to infer a linear model from cross-correlogram estimates by solving the Yule-Walker equations Brockwell and Davis (1986) or to compute a LLR (Eq. (1)) one needs to estimate the cross-correlation structure of two time series. Let (X) and (Y) be two centered stochastic processes whose cross-covariance structure is stationary: $\gamma_{XY}(h) = E(X_{t-h}Y_t)$. If data is sampled regularly $(x_{n\Delta t}, y_{n\Delta t})_{n=0\dots N-1}$ (we use a lower case notation (x) to denote observations of the theoretical process (X)) a consistent estimator for $\gamma_{XY}(h)$ is:

$$\widehat{\gamma_{XY}(h)} = \frac{1}{N-h-1}\sum_{n=h}^{N-1} x_{(n-h)\Delta t} y_{n\Delta t} \tag{1}$$

(we use $\widehat{A}$ to denote an estimator for A). Classically, cross-correlation estimates can subsequently be computed as a normalized version of this estimator Brockwell and Davis (1986).

Interpolating irregular records:

The standard consistent estimator Eq. (1) cannot be computed when (x) and (y) do not share common timestamps. A classical way to circumvent the irregular sampling issue is therefore to interpolate the records $(x_{t_x})_{t_x\in I_x}$ and $(y_{t_y})_{t_y\in I_y}$ onto the set of timestamps $(n\Delta t)_{n=0\dots N-1}$ therefore yielding two approximations $(\widetilde{x_{n\Delta t}})_{n=0\dots N-1}$ and $(\widetilde{y_{n\Delta t}})_{n=0\dots N-1}$ that can be studied as a synchronous multivariate time series. An adapted cross-covariance estimate is then $\widetilde{\widehat{\gamma_{XY}(h)}} = \frac{1}{N-h-1}\sum_{n=h}^{N-1} \widetilde{x_{(n-h)\Delta t}}\widetilde{y_{n\Delta t}}$. While there are many interpolation techniques, a commonly used

method is *last observation carried forward* (LOCF) which, contrary to nearest or linear interpolation, does not need future information and can therefore be used in an on-line predictor. We now show how the LOCF interpolation technique creates spurious causality estimates.

Bias in LLR with irregularly sampled data: The LLR can be computed by several methods. Cross-correlation measurements on a symmetric centered interval are sufficient statistics for this estimator. Therefore one can use synchronous cross-correlation estimates on interpolated data in order to compute the LLR. Carrying the last observation forward (LOCF) has been proven to create a bias in lag estimation in Huth and Abergel (2014). In Figure 2, the LOCF interpolation method introduces a causality estimation bias in which a process sampled at a higher frequency will be seen as causing another process which is sampled less frequently. This is misleading because in the experiment the Brownian motions we simulate have simultaneously correlated increments.

2.2. Issues with interval-matching for irregularly observed data

The *Hayashi-Yoshida* (HY) estimator was introduced in Hayashi et al. (2005) to address this spurious causality estimation issue. In particular, the HY estimator of cross-correlation does not require data interpolation.

Correlation of Brownian motions: HY is adapted to measuring cross-correlations between irregularly sampled Brownian motions. Considering the successor operator s for the series of timestamps of a given process, let $[t, s(t)]_{t\in I_x}$ and $[t, s(t)]_{t\in I_y}$ be the set of intervals delimited by consecutive observations of x and y respectively. The Hayashi-Yoshida covariance estimator over the covariation of (X) and (Y) is defined as

$$\mathrm{HY}_{[0,t]}(x,y) = \sum_{t\in I_x, t'\in I_y : \mathrm{ov}(t,t')} (x_{s(t)} - x_t)\cdot(y_{s(t')} - y_{t'}) \tag{2}$$

where $\mathrm{ov}(t,t')$ is true if and only if $[t, s(t)]$ and $[t', s(t')]$ overlap. The estimator can be trivially normalized so as to yield a correlation estimate.

HY and fractional Brownian motions: No interpolation is required with HY but unfortunately this estimator is only designed to handle full differentiation of standard Brownian motions. Figure 4 shows how HY fails to estimate cross-correlation of increments on a fractional Brownian motion whereas the technique we present succeeds. In the following, we show how our frequency domain based analysis naturally handles irregular observations and is able to fractionally differentiate the underlying continuous time process. This is in particular necessary when one studies factional Brownian motions with correlated increments. In the interest of concision, we refer the reader to Flandrin (1989) for the definition of a fractional Brownian motion.

2.3. Solving the issues created by irregular sampling with Fourier transforms

Our alternative approach to estimating cross-correlograms is based on the definition of the Fourier transform of a stochastic process. Considering a continuous time stochastic process $(X_t)_{t\in[0\dots T]}$ and a frequency $f \in [0\dots 2\pi]$, the Fourier projection of (X) for the frequency f is defined as $P_f(X) = \int_{t=0}^{T} X_t e^{-ift} dt$ where i is the imaginary number. Much attention has been focused on the benefits of the FFT algorithm which has been designed for the very particular base of ordered and regularly sampled observations. *Our key insight is to go back to the very definition of the Fourier transform as an integral and express it empirically in summation form Brillinger (1981);*

Parzen (2012). Moreover, if the process (X) is observed at times $(t_1, \ldots, t_N)$, one can estimate the Fourier projection by

$$\widehat{P_f}(x) = \sum_{n=1}^{N} x_{t_n} e^{-ift_n}. \tag{3}$$

Therefore we propose the following simple framework for frequency domain based linear causal inference: first, project (x) and (y) on to a *reduced* Fourier basis, then, estimate the cross-spectrum of (X) and (Y) in the frequency domain, finally, apply the inverse Fourier transform to the cross-spectrum to recover the cross-correlogram and infer the linear causal structure.

The intuition behind this estimation method is a change of basis that allows us to compute cross-covariance estimates without needing to address the irregularity of timestamps. Indeed the power spectrum $f(\cdot)$ is the element-wise Fourier transform of $\Gamma(\cdot) = \begin{bmatrix} \gamma_{XX}(\cdot) & \gamma_{YX}(\cdot) \\ \gamma_{XY}(\cdot) & \gamma_{YY}(\cdot) \end{bmatrix}$. Therefore, in order to estimate this function one may infer what corresponds to its frequency domain representation and then compute the inverse Fourier transform of the result.

Projecting onto Reduced Fourier Basis: We first project (X) and (Y) onto the elements of the Fourier basis of frequencies $(l\Delta f)_{l=0\ldots P}$, namely the pair $(P_{l\Delta f}(X))_{l=0\ldots P}$ and $(P_{l\Delta f}(Y))_{l=0\ldots P}$. By projecting onto a single relatively small set of orthonormal functions, we are able to compress the observations (x) and (y). In practice we are able to accurately recover the cross-correlogram using only a few thousand projection functions.

Estimating the Cross-spectra: Projecting onto a Fourier basis enables exploratory data analysis through the study of the cross-spectrum of (X) and (Y), $(I_{XY}(l\Delta f))_{l=0\ldots P} = \left(P_{l\Delta f}(X) \times \overline{P_{l\Delta f}(Y)}\right)_{l=0\ldots P}$. An *inconsistent* estimator for the cross-spectrum is

$$\left(\widehat{I_{XY}(l\Delta f)}\right)_{l=0\ldots P} = \left(\widehat{P_{l\Delta f}(x)} \times \overline{\widehat{P_{l\Delta f}(y)}}\right)_{l=0\ldots P}.$$

Local averaging of this estimator with respect to frequencies is widely used Brillinger (1981); Brockwell and Davis (1986); Parzen (2012) in cross-spectral analysis to identify the characteristic frequencies at which stochastic processes interact. Unfortunately, to compute characteristics delays or LLR (crucial steps in linear causal inference) we still need to estimate the cross-correlogram.

Estimating the Cross-correlogram: To estimate the cross-correlogram we can take the inverse Fourier transform of the cross-spectrum $(I_{XY}(l\Delta f))_{l=0\ldots P}$ which translates frequency analysis back into the time domain:

$$\gamma^P_{XY}(h) = \frac{1}{P} \sum_{l=0}^{P} I_{XY}(l\Delta f)\, e^{il\Delta fh}.$$

Using the following *consistent* estimator: $\widehat{\gamma^P_{XY}}(h) = \frac{1}{P}\sum_{l=0}^{P} \widehat{I_{xy}}(l\Delta f)\, e^{il\Delta fh}$ of the cross-covariance we can directly compute a *consistent* estimator of the cross-correlation. The cross-correlation between (X) and (Y) can now be estimated in the time domain with a discrete grid G_h of lag values ranging from $-L\Delta h$ to $L\Delta h$ with a resolution Δh. As expected, aliasing will occur if the user specifies a resolution in the cross-correlation estimate that is much higher than the average sampling frequency of the time series Parzen (2012).

In contrast to more cumbersome time domain synchronization relying on interpolation based methods (LOCF) or interval matching based estimations (HY), our method elegantly addresses time

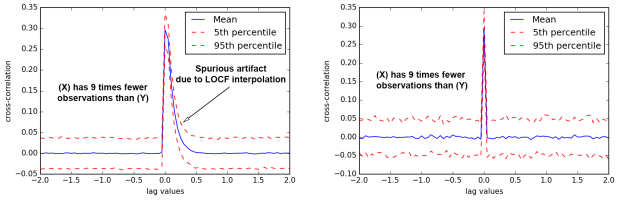

Figure 2: **Spurious cross-correlation created by interpolation:** Cross-correlograms of LOCF interpolated data (left) versus estimation via compression in frequency domain (right). The latter estimate does not present any spurious asymmetry due to the uneven sampling frequencies.

$\frac{N_1}{N_2}$	LOCF interp. LLR Avg +- std	Fourier transf. LLR Avg +- std
1	$0.998 + -0.135$	$1.021 + -0.166$
4.5	$6.863 + -1.678$	$1.053 + -0.320$
10	$7.277 + -1.854,$	$1.107 + -0.391$

Figure 3: Comparison of LLR ratios with LOCF and Fourier transforms (1000 projections) for simultaneously correlated Brownian motions with different sampling frequencies. The LLR ratios should all be 1, one can observe the bias in the LOCF method.

synchronization in the *frequency domain*. While earlier work Parzen (2012); Brillinger (1981) has considered the application of frequency domain analytics to irregularly sampled data, our method is the first to translate back to the time domain to recover a consistent estimator of the linear causal structure. Alternatively, Lomb-Scargle periodogram Scargle (1982); Lomb (1976) also enables the frequency domain analysis of irregularly observed data but suffers from the supplementary cost of a least square regression. To the best of our knowledge we are the first to use frequency domain projections in order to compute the cross-correlogram in order to infer linear causal structure.

Cross-correlogram Estimator Consistency: Central to the communication and memory performance of our technique is the ability to use a small number of Fourier projections relative to the number of observations and still accurately recover the cross-correlogram. We can characterize the statistical properties of the cross-spectral estimator Brillinger (1981); Parzen (2012); Brockwell and Davis (1986). In particular, it is well known that for two distinct non-zero frequencies f_1 and f_2 the estimators $\widehat{I_{XY}(f_1)}$ and $\widehat{I_{XY}(f_2)}$ are asymptotically independent. Consequently, to obtain an estimator with asymptotic variance $O(V)$ the user will need to project on $\frac{1}{V}$ frequencies. The element-wise product of Fourier transforms is converted into the time domain by the inverse Fourier transform to yield a cross-correlogram. With very large datasets in which $N >> \frac{1}{V}$ we are in the asymptotic regime and obtain the suitable compression property of our algorithm.

2.4. Example of time domain exploratory data analysis through the frequency domain

The time domain exploratory analysis we enable makes lead-lag relationships self-explanatory. We show in the following that it is not biased by one process being sampled more seldom than the other.

Numerical assessment of frequency domain based correlation measurements: We demonstrate, through simulation, that the spurious causation issue that plagues the LOCF interpolation Huth and Abergel (2014) does not appear in our proposed method. We consider two synthetic correlated Brownian motions that do not feature any lead-lag and compare the estimation of LLR provided by two time domain interpolation methods and our approach. After having sampled these at random timestamps, in Table 3 and Figure 2 we compare the cross-correlation and LLR estimates obtained by LOCF interpolation and our proposed frequency domain analysis technique confirming that our method does not introduce spurious causal estimation bias.

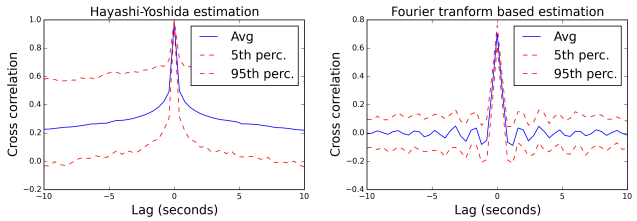

Figure 4: **LRD Erasure:** Monte Carlo simulation (100 samples) of two fractional Brownian motions with Hurst exponent 0.8 and simultaneously correlated increments. Spurious slowly vanishing cross-correlation hinders the HY estimation but does not affect our estimation with LRD erasure (see Section 3) as evident by nearly zero cross-correlation for non-zero lag.

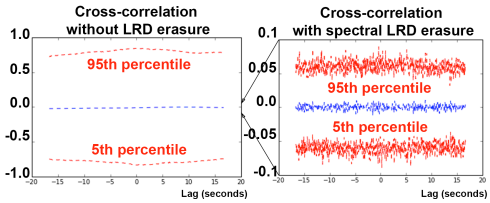

Figure 5: The empirical cross-correlation distribution on the left is affected by spurious estimates. On the right, frequency domain fractional pole erasure eliminated the issue, considerably narrowing the interval between the 5^{th} and 95^{th} percentiles.

3. Addressing the issues created by Long Range Dependence

A stochastic process is said to be Long Range Dependent (LRD) if it features cross-correlation magnitudes whose sum is infinite Doukhan et al. (2003). Many issues arise in that case with correlation estimates becoming spurious. This phenomenon was first discover when Granger studied the concept of cointegration between Brownian motions (integrated time series) Granger (1988). On sorted Brownian motion data, this effect can be addressed by differentiating the time series, namely computing $(\Delta X_t)_{t\in\mathbb{Z}} = (X_t - X_{t-1})_{t\in\mathbb{Z}}$. For fractional Brownian motion and LRD time series, the fractional differentiation operator needs to be computed. It is defined as $(\Delta^\alpha X_t)_{t\in\mathbb{Z}} = \left(\sum_{h=0}^{\infty} \prod_{j=0}^{h-1}(\alpha-j)(-X_{t-h})^h/h!\right)_{t\in\mathbb{Z}}$. Therefore, to study the cross-correlation structure of two integrated or fractionally integrated time series, one would have to compute $(\Delta X_t)_{t\in\mathbb{Z}}$ or $(\Delta^\alpha X_t)_{t\in\mathbb{Z}}$. The latter requires chronologically sorted data and synchronous timestamps and has a quadratic time complexity with respect to the number of samples.

3.1. Erasing memory in the frequency domain

Erasing memory is of prime importance, in the case of the study of Brownian motions and fractional Brownian motions alike. As pointed out in Doukhan et al. (2003), LRD arises in many systems and from a computational and statistical point of view, it is challenging to erase.

Equivalence between differentiation in time domain and element-wise multiplication in frequency domain: Let $(X_t)_{t\in[0,T]}$ be a continuous process whose fractional differentiate of degree α, $d^\alpha X$ is Lebesgue-integrable with probability 1. If X_t vanishes at the boundaries of the interval, classically, almost surely, $P_f\left(d^\alpha X\right) = \int_{t=0}^{T} e^{-ift} d^\alpha X_t = -\left(-if\right)^\alpha \int_{t=0}^{T} e^{-ift} X_t dt$ by a stochastic integration by part. Therefore, an estimate for $P_f\left(d^\alpha x\right)$ is $\widehat{P_f\left(d^\alpha X\right)} = -\left(-if\right)^\alpha \widehat{P_f\left(X\right)}$.

Erasing memory through fractional pole elimination: The power spectrum of a fractional Brownian motion Flandrin (1989) with Hurst exponent H is asymptotically $\frac{1}{f^{2H+1}}$ for $f << 1$. This is the characteristic spectral signature of a long range dependent time series. H can therefore be estimated by the classical periodogram method for an individual time series by conducting a linear regression on the magnitude of the power spectrum about 0 in a log/log scale Doukhan et al. (2003). Wavelets are another family of orthogonal basis enabling a similar estimation. One can therefore see the fractional differentiation operator of order $H + 1/2$ as a means to compensate for a pole of order $2H + 1$ in square magnitude in 0. Multiplying the Fourier transform of the signal by $(if)^{H+1/2}$ eliminates the issue. It does not require any preprocessing of the data, no interpolation or

re-ordering and we will show below that it has tremendous computational advantages in the context of distributed computing in terms of communication avoidance.

3.2. Testing frequency domain LRD erasure

The example below considers two fractional Brownian motions (X) and (Y) Brownian motions with Hurst exponent $H = 0.4$ Doukhan et al. (2003). We compare the empirical distributions of cross-correlation estimates obtained over 100 trials with and without LRD erasure in frequency domain. In Figure 5 we showcase an experiment with 9998 uniformly random observations for (X) and 6000 uniformly random observations for (Y). Naive cross-correlation estimations lead to many spurious cross-correlation estimates with significantly high magnitudes of estimated correlation values for processes that are in fact independent, (90% of the empirical distribution between -0.9 and 0.9). The confidence interval we obtain with our novel frequency domain erasure method by fractional pole elimination is narrower (90% of the empirical distribution between -0.05 and 0.05) and enables reliable analysis. The next section will expose the computational advantages of such a frequency domain based estimation as a communication avoidance mechanism.

4. Addressing the issues created by the scale of large data sets

Scalable computation is essential to practical causal inference in real-world *big data* sets. Our proposed frequency domain approach provides a parallel communication avoiding mechanism to efficiently compress large time-series data sets while still enabling the estimation of cross-correlograms. To leverage scale-out cluster computing it is essential to minimize communication across the network as network latency and bandwidth can be orders of magnitude slower than RAM Peleg (2000).

4.1. Computational advantages

The novel frequency domain based methods we propose can entirely be expressed as trivial map-reduce aggregation operations and do not require sorting or interpolating the data. Projecting on a few elements of the Fourier basis substantially reduces communication and memory complexity associated with the estimation of cross-correlograms.

4.2. Fourier compression as a communication avoidance algorithm

The computation of Fourier projections is communication efficient in the distributed setting. The Fourier projection can be calculated by locally computing the sum of the mapping of multiplications by complex exponentials. Then, local partial sums are transmitted across the network to compute the projections of the entire data set. In this section, we study d distinct processes with N data points each. Let V denote the desired variance for the cross-correlation estimator via the frequency domain.

Communication cost of aggregation with indirect frequency domain covariance estimates: Now consider the set of Fourier projections $\left(\widehat{P_f}(x) = \sum_{n=1}^{N} x_{t_n} e^{-ift_n}\right)_{f=0,\Delta f,\ldots,P\Delta f}$ which we aggregate on each single machine separately prior to sending them over the network. The number of projections needed to have an estimator for cross-correlation with variance V is $O(\frac{1}{V})$. Therefore, the size of the message sent out by each machine over the communication medium is now $O(d\frac{1}{V})$ and representative of $O(dN)$ data points. If the user chooses $\frac{1}{V} \ll N$, our method effectively

compresses the data prior to transmitting it over the network. It is noteworthy that the gain offered by this algorithm is system independent as long as communication between computing cores is the main bottleneck.

Distributed LRD erasure: The computational complexity of fractional differentiation is $O(N^2 d)$ in the time domain. Moreover, in distributed system, computing the fractional differentiation of a signal would require transmitting the entire data set across the network. As a consequence the bandwidth needed is $O(Nd)$. The compute time therefore allows an interactive experience for the user and becomes even shorter with a distributed implementation on several machines. For example, on a single processor with a 2013 MacbookPro Retina we were able to compute 3000 projections on 10^5 samples in roughly a minute.

5. Causality estimation on actual data

In order to highlight significant cross-correlation between pairs of stocks, one needs to consider high frequency dynamics. As we will show in the following, cross-correlation vanishes after a few milliseconds on most stocks. In these settings it is then necessary to use full resolution data whose timestamps are irregular and not common to different pairs of stocks as records only occur when an trading event occurs Abergel et al. (2012). Intraday stock prices can be considered LRD processes Doukhan et al. (2003) and therefore we need to use fractional pole erasure in this example. This context is therefore in the very scope of data intensive tasks we consider.

5.1. Checking the consistency of the estimator

Consider ask and bid quotes during one month worth of data. We create a surrogate noisy lagged version of AAPL with a 13ms delay and 91% correlation which is named AAPL-LAG. We study fours pairs of time series: APPL/APPL-LAG, AAPL/IBM, AAPL/MSFT, MSFT/IBM. We obtained quote data for these stocks at millisecond time resolution representing several months of trading. The cross-correlograms obtained below are computed between 10 AM and 2PM for 61 days in January, February and March 2012. For each process, 3000 frequencies were used in the Fourier basis which is several orders-of-magnitude less than the number of observations that we get per day which ranges from 5×10^4 to 1×10^5. The estimate cross-correlograms in Figure 6 and their empirical significance intervals show that our estimator is consistent and does not suffer from non-vanishing variance as a result of LRD. We observe an 89% average peak cross-correlation with an 8ms delay for the surrogate pair of AAPL stocks which confirms our estimator is reliable with empirical data. In Figure 6 we highlight a taxonomy of causal relationships.

5.2. Studying causality at scale

To evaluate scalability in a real-world setting in which $\frac{1}{V} << N$, we assess the relation between AAPL and MSFT over the course of 3 months. In contrast to our earlier experiments (shown in Figure 6), we no longer average daily cross-correlograms in and therefore only leverage concentration in the inverse Fourier transform step of the procedure. With only 3000 projections for 5×10^6 observations per time series, the results we obtain on Figure 7 reveals the causal relation between AAPL, AAPL-LAG, IBM and MSFT consistently with Figure 6.

Scalability: We run the experiment with Apache Spark on Amazon Web Services EC2 machines of type r3.2xlarge where communication is a bottleneck. In Figure 8 we show that even with a

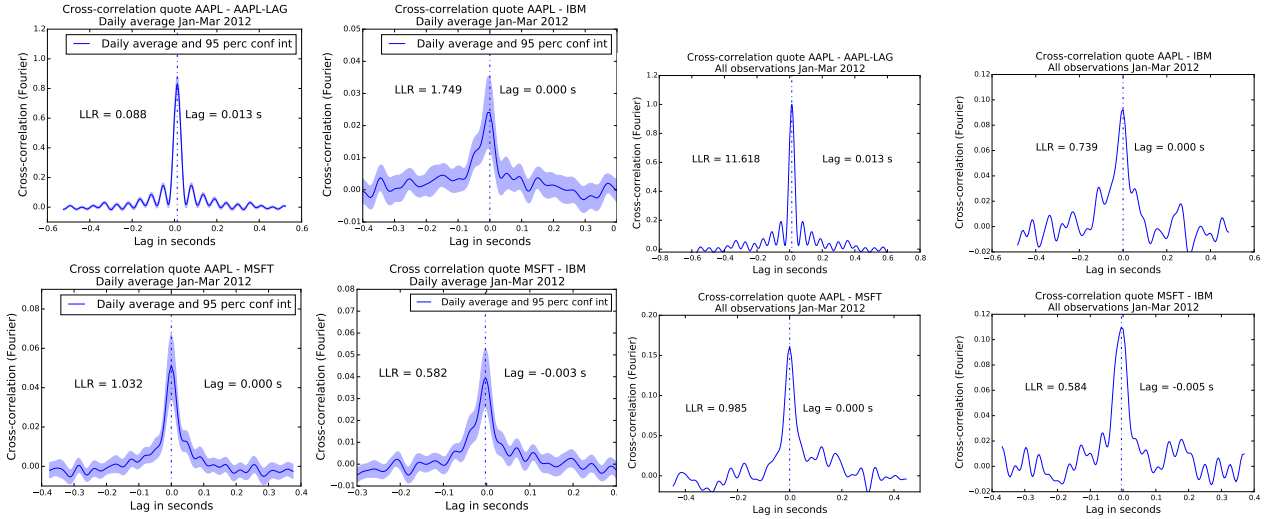

Figure 6: Compression ratio is $< 5\%$. The daily averaged cross-correlogram of AAPL and IBM is strongly asymmetric, therefore highlights that IBM follows AAPL. The symmetry between AAPL and MSFT shows there is no such relationship between them. Symmetric and offset in correlation peak show that IBM follows MSFT with a millisecond latency.

Figure 7: Compression ratio is $< 1\%$. On the entire data set we retrieve results similar to 6 therefore validating the use of our estimation of cross-correlograms in a scalable manner thanks to Fourier domain compression.

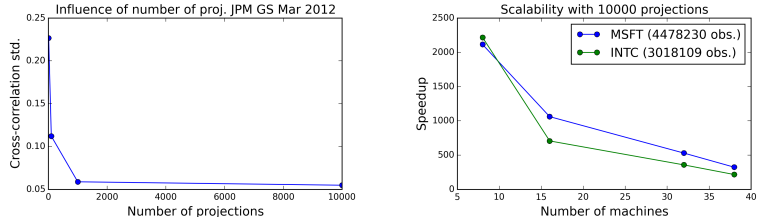

Figure 8: On the left we plot the empirical standard deviation of daily cross-correlograms with respect to the number of projections showing that the variability decreases rapidly. On the right we plot the run time performance of our algorithm versus the number of Apache Spark EC2 machines demonstrating approximately linear speedup.

large number of projections (10000) the communication burden is still low enough to achieve a linear speed-up. In Figure 8 we show that even with a large number of projections (10000) the communication burden is still low enough to achieve linear speed-up.

6. Conclusion

Time series analysis via the frequency domain presents several provides consistent causal estimates and scaling on distributed systems. We proposed a method to analyze causality which does not require sorting data, works naturally with irregular timestamps, provides reliable causality estimates and makes the erasure of Long-Range dependencies embarrassingly parallel. Applying an inverse Fourier transform to estimated Fourier spectra enables exploration of dependencies in the time domain. With the resulting consistent cross-correlogram, one can compute Lead-Lag ratios and characteristic delays between processes and thereby infer linear causal structure. We show that projecting onto 3000 Fourier basis elements is sufficient to study tens of millions of irregularly observed recordings, thereby providing insightful analytics in a generic and scalable manner.

References

Frédéric Abergel, Jean-Philippe Bouchaud, Thierry Foucault, Charles-Albert Lehalle, and Mathieu Rosenbaum. *Market microstructure: confronting many viewpoints*. John Wiley & Sons, 2012.

David R Brillinger. *Time series: data analysis and theory*, volume 36. Siam, 1981.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer-Verlag New York, Inc., New York, NY, USA, 1986. ISBN 0-387-96406-1.

Paul Doukhan, George Oppenheim, and Murad S Taqqu. *Theory and applications of long-range dependence*. Springer Science & Business Media, 2003.

Patrick Flandrin. On the spectrum of fractional brownian motions. *Information Theory, IEEE Transactions on*, 35(1):197–199, 1989.

Milton Friedman. The interpolation of time series by related series. *Journal of the American Statistical Association*, 57(300):729–757, 1962.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

Clive WJ Granger. Causality, cointegration, and control. *Journal of Economic Dynamics and Control*, 12(2):551–559, 1988.

Takaki Hayashi, Nakahiro Yoshida, et al. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2):359–379, 2005.

Nicolas Huth and Frédéric Abergel. High frequency lead/lag relationships - empirical facts. *Journal of Empirical Finance*, 26:41–58, 2014.

Paul S Linsay. An efficient method of forecasting chaotic time series using linear interpolation. *Physics Letters A*, 153(6):353–356, 1991.

Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–462, 1976.

Manfred Mudelsee. *Climate time series analysis*. Springer, 2013.

Emanuel Parzen. *Time Series Analysis of Irregularly Observed Data: Proceedings of a Symposium Held at Texas A & M University, College Station, Texas February 10–13, 1983*, volume 25. Springer Science & Business Media, 2012.

David Peleg. Distributed computing. *SIAM Monographs on discrete mathematics and applications*, 5, 2000.

Jeffrey D Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, 1982.

Ruey S Tsay. *Analysis of financial time series*, volume 543. John Wiley & Sons, 2005.

Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA, 1949.