Random Projection Design for Scalable Implicit Smoothing of Randomly Observed Stochastic Processes

Francois Belletti **Evan R. Sparks** **Alexandre M. Bayen** **Joseph E. Gonzalez**
Electrical Engineering and Computer Sciences, UC Berkeley

Abstract

Standard methods for multi-variate time series analysis are hampered by sampling at *random timestamps*, *long range dependencies*, and the *scale* of the data. In this paper we present a novel estimator for cross-covariance of randomly observed time series which identifies the dynamics of an unobserved stochastic process. We analyze the statistical properties of our estimator *without the assumption that observation timestamps are independent from the process of interest* and show that our solution does not suffer from the corresponding issues affecting standard estimators for cross-covariance. We implement and evaluate our statistically sound and scalable approach in the distributed setting using Apache Spark and demonstrate its ability to identify interactions between processes on simulations and financial data with tens of millions of samples.

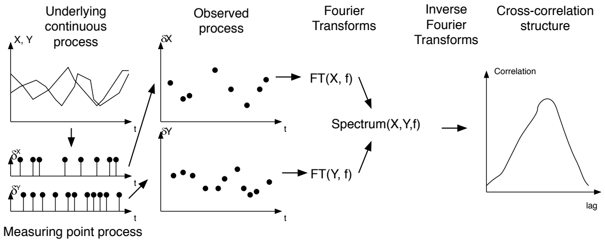

Figure 1: Underlying, measuring and observed process.

1 Introduction

Cross-covariance estimates are of prime importance in applications ranging from statistical finance [1, 2] to climate studies [3] as asymmetry in cross-covariance is an indicator of a causal relationships between time series [4, 5, 6].A pair of causally related continuous stochastic processes may be modeled as the solution of a Stochastic Differential Equation (SDE) of the form:

$$
\begin{aligned}
dX(t) &= dW^X(t) + \\
&\int_{s>0} \phi^{YX}(s)dY(t-s) + \phi^{XX}(s)dX(t-s) \\
dY(t) &= dW^Y(t) + \\
&\int_{s>0} \phi^{XY}(s)dX(t-s) + \phi^{YY}(s)dY(t-s), \quad (1)
\end{aligned}
$$

where W is a Wiener process [7]. The multidimensional causal kernel $\Phi(\cdot)$ consisting of $\phi_{XY}(\cdot), \phi_{XX}(\cdot), \phi_{YX}(\cdot), \phi_{YY}(\cdot)$ [4] defines the dynamics and the solution to the SDE is given by the stable bi-variate stochastic process, $U(\cdot) = (X(\cdot), Y(\cdot))$. In practice, we observe stochastic processes $U(\cdot)$ at particular points in time that are determined by a separate measurement process. In this work we will assume that observations are made on the interval $[0, T]$ with timestamps generated by a bi-variate regular second order stationary Point Process $N(\cdot) = (N^X(\cdot), N^Y(\cdot))$ [8, 9] referred to as the *measuring process* where $N^X(t)$ and $N^Y(t)$ are the number of observations that have been observed between 0 and time t. We assume that the *measuring process* has a stationary cross-covariance structure, that observations are not accumulated at a singular point in time, and their occurrence may be correlated with the *underlying process*. For example, in the medical setting, the *underlying process* could be the blood pressure of a patient which is defined at all times but only observed during punctual medical which the *measuring process* consists in here. The resulting *observed process*

$$(U\,dN)(\cdot) = ((X\,dN^X)(\cdot), (Y\,dN^Y)(\cdot)) \qquad (2)$$

conflates the statistical properties of both the underlying process of interest (U) with the measurement process (N). Our aim in this article is to infer the statistical properties of the continuous underlying process (U) through the partial observations $(U\,dN)$ determined by the discrete measurement process (N). This data generative process is illustrated in Figure 1.

We make three contributions in this work. In Section 2,

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

we empirically demonstrate standard cross-covariance estimators designed for regularly observed data almost surely evaluate to 0 resulting in biased estimates of the true correlation structure. We demonstrate that standard solutions to mitigate the resulting bias such as ad-hoc interpolation and Hayashi-Yoshida estimation [10] do not offer unbiased solutions for randomly observed stochastic processes. In Section 3 we develop a statistically sound approach based on the use of kernel smoothing that addresses the bias of classical estimators and provide a theoretical analysis demonstrating a bias/variance trade-off depending on the smoothing kernel bandwidth. We show how to compute these estimators efficiently in a scalable manner through random Fourier projections drawn from a specific probability distribution. In Section 4 we evaluate the proposed approach on both synthetic and real-world data characterizing the accuracy of our technique by achieving a relative error of 8% or less in linear model recovery with 90% probability. Moreover, our method enables estimation of the parameters of the generative SDE model for the *underlying process* and using our new method we are able to accurately scale to large financial time series and estimate causal structure.

2 Limitations of Standard Methods

In the following we identify a series of challenges in estimating the cross-covariance structure of a continuous bivariate *underlying* process observed randomly.

2.1 Stochastic Intensity Point Processes

In order to more precisely state the probabilistic properties of the *underlying* and *measuring process* we introduce useful notation.

The information entailed in the history of the process (X, Y) in the generative model Eq. (1) is defined as the filtration [7, 8] $\left(\mathcal{F}_t^{XY}\right)_{t\in[0,T]}$ where $\mathcal{F}_t^{XY}$ is the information available about (X, Y) up to but not including time t. One can also extend the past information to that jointly produced by both the *underlying* process and the *measuring* process which we denote $(\mathcal{F})$.

A stochastic Point Process (N^X, N^Y) [8] defines a random measure over the axis of time and is defined by stochastic local Poisson intensities for N^X

$$\mu_X(t) = \lim_{dt\to 0+} \frac{E\left[N^X(t+dt) - N^X(t)|\mathcal{F}_t\right]}{dt}$$

and similarly for N^Y. In other words, $\mu_X(t)$ and $\mu_Y(t)$ are the number of observations per unit of time expected given the events that occurred until time t. With the second order stationarity assumption we make, these random measures are characterized by their cross-covariation structure $(\gamma^{\mu_X,\mu_Y}(h) = E\left[\mu_X(t)\mu(t+h)\right])_{h\in\mathbb{R}}$. The regularity [9] assumption on the other hand guarantees that measurements cannot accumulate at a single unique timestamp. The following proposition is helpful to understand the role of such a measure of continuous stochastic processes and will be used in proves below:

Proposition 2.1 *With a continuous underlying process (X), almost surely,*

$$E\left[X(t)dN^X(t)|\mathcal{F}_t\right] = X(t)\mu^X(t)dt. \tag{3}$$

where dt is the Lebesgue measure.

Proof 2.1 *This is an immediate consequence of properties conditional expectation and the continuity of X as, almost surely, $E\left[X(t)dN^X(t)|\mathcal{F}_t\right] = X(t+)\mu_X(t)dt = X(t)\mu_X(t)dt$ where $X(t+)$ is the right limit of X in t.*

2.2 Statistical Issues

Our aim is to estimate the cross-covariation structure (and the cross-correlation structure, after normalization) of the *underlying* process $\gamma^{XY}(h) = E\left[X(t)Y(t-h)\right]$, as it is a sufficient statistic for Granger causality estimation [6], as well as the estimation of Φ using Yule-Walker equations [4]. A standard estimator in the literature for regularly observed time series [4, 11, 12] is

$$\widehat{\gamma}^{XY}_{\text{regular}}(h) = \frac{1}{T}\iint_{t,s=0}^{T} X(t)Y(s)\delta_0(t-s+h)dsdt$$

where $\delta_0(\cdot)$ is the dirac function centered at 0.

2.2.1 The Epps effect with irregular observation

However, when observations are irregular the naive estimators designed for regular observations take the form

$$\widehat{\gamma}^{XY}_{\text{naive}}(h) = \frac{1}{T}\iint_{t,s=0}^{T} X(t)Y(s)\delta_0(t-s+h)dN_t^X dN_s^Y.$$

and in the case of an observation process which is a non-degenerate point process (N^X, N^Y), this quantity evaluates to 0 almost surely [8] because X and Y are never observed simultaneously.

Theorem 2.1 *The estimator $\widehat{\gamma}^{XY}_{naive}(h)$ evaluates almost surely to 0 for any value of lag h.*

Proof 2.2 *As the bi-variate measuring process (N^X, N^Y) is a regular [9] Poisson process, using Landau notation, $\mathbb{P}(\left(N^Y(s+\tau) - N^Y(s)\right)\left(N^Y(s+h\tau) - N^Y(s+h)\right) > 0) = o(\tau)$, therefore $\forall t \in [0,T], \int_{t,s\in[0,T]} X(t)Y(s)\delta_0(t-s+h)dN_s^Y dN_t^X = 0$ as $\delta_0(t-s+h)$ is non-zero only at $t-s+h=0$.* ■

The bias of the naive estimator for cross-covariance towards 0 is referred to as *Epps effect* in the field of high frequency statistics for finance [1].

To mitigate this issue created by the asynchronous observation process, we develop a novel approach to the estimation of cross-covariance (and therefore cross-correlation) based on kernel smoothing implicitly computed by specifically designed random projection in the Fourier space.

In the time domain, we define our kernel smoothing based cross-correlation estimator

$$\widehat{\gamma}^{XY}_{\text{smooth}}(h) = \frac{1}{T} \iint_{t,s=0}^{T} X(t)Y(s)\kappa(t-s+h)dN_t^X dN_s^Y \tag{4}$$

where $\kappa(\cdot)$ is a continuous function expressing a smoothing kernel. In Section 3.2 we will show how to efficiently evaluate this expression through the use of specific random frequency domain projections.

One approach to address the Epps effect is to align observations on a common time grid by interpolation [5] or using an estimator dedicated to the precise setting of the problem under study such as Hayashi-Yoshida [10] for correlated Brownian motions. In the following we show the statistical shortcomings of both pre-existing approaches. The analysis of these shortcomings presented below motivated our proposed approach whose statistical properties we study both theoretically and experimentally.

2.2.2 Bias created by interpolation

In this section, we first review existing techniques for time-domain estimation of second-order statistics for continuous stochastic processes in the context of discrete random sampling in time. Interpolating data is a usual solution in order to be able to use classic time series analysis techniques [13, 14, 15]. Unfortunately, interpolation is not always suitable, as it can create biased estimates which mislead researchers into concluding that there is significant cross-covariance where there is none [5].

A classical way to study the interactions of two asynchronously observed time series is to force the synchronicity of the timestamps by aligning them on a common time grid. While there are many interpolation techniques, a commonly used method is *last observation carried forward* (LOCF) which is not as accurate as linear interpolation or approximation by the nearest point but can efficiently be deployed as it only relies on past data at any point in time. We now consider the causality inference framework introduced in [5] and show how the LOCF interpolation technique creates cross-correlation estimates that may lead to false conclusions regarding the way (X) and (Y) influence each other.

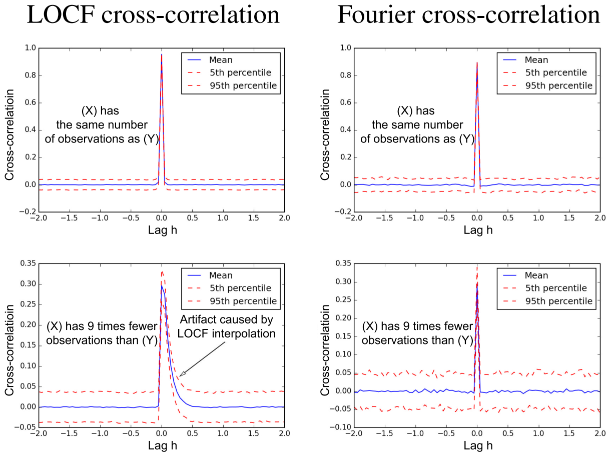

Figure 2: Cross-correlograms of LOCF interpolated data versus estimation via compression in frequency domain. The latter estimate does not present any spurious asymmetry due to the uneven sampling frequencies.

Bias caused by LOCF interpolation: We demonstrate, through simulation, that the asymmetric cross-correlation bias that plagues the LOCF interpolation in [5] does not appear in our proposed method. We consider two synthetic correlated Brownian motions that do not feature any lead-lag (variations in one shortly following variations in the other) and compare the estimation of cross-correlation provided the LOCF interpolation methods and our approach. After having sampled these continuous processes at random timestamps, in Figure 2 we compare the cross-correlation estimates obtained by LOCF interpolation and our proposed frequency domain analysis technique confirming that our method does not introduce estimation bias that is characterized by significantly positive cross-correlation for small positive lag values. As the two simulated processes are synchronously correlated, the estimator should find a cross-correlation of 0 except when $h = 0$.

2.3 Interpolation-free Causality Assessment

The *Hayashi-Yoshida* (HY) estimator was introduced in [10] to address the LOCF estimators tendency to discover spurious causal structure. The HY estimator of cross-correlation does not require data interpolation and has been proven to be consistent in the context of High Frequency statistics in finance [16].

Correlation of Brownian motions: HY is adapted to measuring cross-correlations between irregularly sampled Brownian motions. Considering the successor operator next for the series of timestamps of a given process, let $[t, \text{next}(t)]_{t \in \text{obs}(X)}$ and $[t, \text{next}(t)]_{t \in \text{obs}(Y)}$ be the set of intervals delimited by consecutive observations of X and Y respectively. The Hayashi-Yoshida cross-covariance estimator over the covariation of (X) and (Y) [17] is

$$\text{HY}(h) = \sum_{\substack{t \in \text{obs}(X),\, s \in \text{obs}(Y) \\ \textbf{s.t:}\ \text{ov}(t, s+h)}} (X_{\text{next}(t)} - X_t) \times (Y_{\text{next}(s)} - Y_s) \tag{5}$$

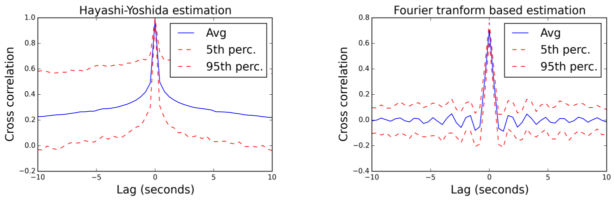

Figure 3: **LRD Erasure:** Monte Carlo simulation (100 samples) of two fractional Brownian motions with Hurst exponent 0.8 and simultaneously correlated increments. Spurious slowly vanishing cross-correlation hinders the HY estimation but does not affect our estimation with LRD erasure (see Section 3)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

where $\mathrm{ov}(t, s+h)$ is true if and only if $[t, \mathrm{next}(t)]$ and $[s+h, \mathrm{next}(s+h)]$ overlap. The estimator can be normalized by $\mathrm{HY}(0)$ to estimate the cross-correlation.

No interpolation is required with HY but unfortunately this estimator is not applicable outside the context of cross-correlated standard Brownian motions. Figure 3 shows how the HY estimator has much more variance than our proposed estimator when estimating cross-correlation of increments on a fractional Brownian motion. Fractional Brownian motions belong to the family of Long Range Dependent (LRD) processes [18] found in general time series in climate sciences [3], finance [19], economics [20] or genomics [21]. When applied to these processes having a long memory of previous perturbations, the technique we present provides a convergent estimator with less variance as opposed to HY. In section 3.3, we show how our frequency domain based analysis naturally handles irregular observations and is able to fractionally differentiate the underlying continuous time process. In the interest of concision, we refer the reader to [22] for the definition of a fractional Brownian motion and fractional differentiation.

3 Smoothing kernels in the Fourier domain

In this section, we develop a theoretical analysis of the properties of estimators dedicated to the estimation of the cross-covariance structure of the underlying process $(E[X(t)Y(t+h)])_{h\in H}$ where H is a discrete grid of lags at which we want to evaluate the cross-correlation function. Estimating this second order statistics is of paramount importance as it highlights lead-lag [5] and helps infer linear causation kernels by solving the Yule-Walker equations [4].

3.1 Smoothing kernel based estimator

Classically with sparsely observed continuous stochastic processes [13] the *measuring* process is independent from the *underlying* process. We just assume that the compound model of the *measuring* point process and *underlying* process is second order stationary with co-variation function

$$\gamma^{X\mu_X, Y\mu_Y}(h) = E\left[X(t)\mu_X(t)Y(t+h)\mu_Y(t+h)\right].$$

The absence of independence between the *underlying process* and the *measuring process* is a common feature in medical data [23] or financial data [1]. The following theorem shows how this interaction affects smoothed cross-covariance estimators.

Theorem 3.1 *The kernel smoothing estimator does not almost surely evaluate to* 0 *as*

$$E\left[\widehat{\gamma}^{XY}_{smooth}(h)\right] = \kappa * \left(\gamma^{XdN^X, YdN^Y}\right)(h) \tag{6}$$

where $*$ *denotes the convolution operator.*

Proof 3.1 *By definition of kernel smoothing estimator*

$$E\left[\widehat{\gamma}^{XY}_{smooth}(h)\right] = \frac{\iint_{t,s=0}^{T} E\left[X(t)Y(s)\kappa(t-s+h)dN^Y_s dN^X_t\right]}{T}$$

As $X(\cdot)$ *and* $Y(\cdot)$ *are almost surely continuous and* $\kappa(\cdot)$ *is also assumed continuous, we can write the following:*

$$\begin{aligned}
&\iint_{t,s=0}^{T} E\left[X(t)Y(s)\kappa(t-s+h)dN^Y_s dN^X_t\right] \\
&= \iint_{t>s\in[0,T]} E[\\
&E\left[Y(s)E\left[X(t)\kappa(t-s+h)dN^X_t|F_t\right]dN^Y_s|F_s\right]]dsdt \\
&+ \iint_{t<s\in[0,T]} E[\\
&E\left[X(t)E\left[Y(s)\kappa(t-s+h)dN^Y_s|F_s\right]dN^X_t|F_t\right]]dsdt \\
&= \iint_{t,s=0}^{T} E\left[X(t)\mu_X(t)Y(s)\mu_Y(s)\kappa(t-s+h)\right]dsdt
\end{aligned}$$

if we use Proposition 2.1 and recombine the two partitions of the double integral. Rewriting the quantity above:

$$\begin{aligned}
&\int_{s=0}^{T} E\left[\int_{t=0}^{T} X(t)\mu_X(t)Y(t-s+h)\mu_Y(t-s+h)dt\right] \\
&\qquad \kappa(s)ds \\
&= \int_{s\in[0,T]} T\gamma^{XdX,YdY}(h-s)\kappa(s)ds
\end{aligned}$$

which concludes the proof. ■

3.1.1 Bias analysis for smoothing kernel design

We want to compare the estimators above with of the quantity of interest, $\gamma^{XY}(\cdot) = E\left[X(t)Y(t+h)\right]$.

Corollary 3.1 *With the smooth kernel estimator we propose,* $\widehat{\gamma}^{XY}_{smooth}(h)$ *the bias term is now* $\gamma^{XY}(\cdot) - \kappa * \left(\gamma^{XY} \times \gamma_\mu\right)(\cdot)$ *where* $\gamma_\mu = E\left[\mu_X(t)\mu_Y(t+h)\right]$.

Let us now analyze this bias term in the simplest setting where the *measuring* process is a bivariate uniform Poisson process with independent components of intensities ν_X

and ν_Y respectively and is independent from the *underlying* process.

In such a setting, for any real valued h, $\gamma_\mu(h) = E\left[\mu_X(t)\mu_Y(t+h)\right] = \nu_X\nu_Y$. Therefore the expected value of the estimator shrinks by a constant factor in

$$E\left[\widehat{\gamma}^{XY}_{\text{smooth}}(h)\right] = \nu_X\nu_Y(\kappa * \gamma^{XY})(h). \qquad (7)$$

Here, as the sampling intensity decreases, the bias shrinks the estimator for cross-covariance towards 0 as intuition suggests. Further, we see that the smoothing kernel needs to be concentrated about 0 as expected. We cannot reduce this bias to 0, as κ needs to be continuous with a non-empty interior support.

3.1.2 Variance analysis for smoothing kernel design

We conduct of variance analysis in the simple case of a Gaussian smoothing kernel which is an extension of standard results of variance reduction for time series by Gaussian smoothing.

Proposition 3.1 *Assuming data is generated by the model in Eq. (1), if the smoothing kernel is a Gaussian density centered in* 0 *with standard deviation* σ*, the variance of* $\widehat{\gamma}^{XY}_{\mathit{smooth}}(h)$ *is* $O(\frac{1}{\sigma})$*.*

Proof 3.2 *Denoting* $\left(\mathcal{F}^{X,dN^X}_t\right)_{t\in[0,T]}$ *the history of the compound process* (X, dN^X)*, we consider*

$$E\left[\left(\widehat{\gamma}^{XY}_{\mathit{smooth}}(h)\right)^2|\mathcal{F}^{X,dN^X}_T\right] = \int_{t,u=0}^{T} X(t)X(u)$$
$$\int_{s,v=0}^{T}\kappa(t-s+h)\kappa(u-v+h)A(s,v)dN^X_t dN^X_u$$

$$\text{Where } A(s,v) = E\left[Y(s)Y(v)dN^Y_s dN^Y_v|\mathcal{F}^{X,dN^X}_T\right]$$
$$= E\left[Y(s)Y(v)\mu_Y(s)\mu_Y(v)|\mathcal{F}^{X,dN^X}_T\right]$$

again by a continuity argument. Conditionally to $\mathcal{F}^{X,dN^X}_T$*, the stochastic process* $Y\mu_Y$ *is second-order stationary. With our assumptions on the generative model Eq. (1), the spectral density* $(F(f)_{f\in\mathbb{R}})$ *of* $Y\mu_Y$ *with respect to conditional probability to* $\mathcal{F}^{X,dN^X}_T$ *is such that for any frequency* f *the derivative of the spectral density* $\int_{\mathbb{R}}\frac{log(F'(f))}{1+f^2}df$ *is bounded. Therefore the Karhunen representation theorem [7, 24] guarantees that there exist a second-order stationary process* (ϵ) *with orthogonal increments under the conditional probability associated with* $\mathcal{F}^{X,dN^X}_T$ *and a functional* ψ *such that* $Y(t)\mu_Y(t) = c(t) + \int_0^t\psi(t-s)d\epsilon_s$ *where* $c(t)$ *is a deterministic process which will be ignored as it is irrelevant when studying the variance of* $(Y(\cdot)\mu_Y(\cdot))$*. Then,*

$$\int_{s,v\in[0,T]} E[\kappa(t-s+h)\kappa(u-v+h)Y(s)Y(v)\mu_Y(s)\mu_Y(v)$$
$$|\mathcal{F}^{X,dN^X}_T]dsdv = E\left[(k*(Y\mu_Y)(t))^2\middle|\,\mathcal{F}^{X,dN^X}_T\right]$$

where κ *is a Gaussian smoothing kernel with standard deviation* σ *with the stochastic process* $(Y\mu_Y)$*. Convolutions being linear,* $k*(Y\mu_Y)(t) = \int_0^t\psi(t-s)(k*d\epsilon)_s$ *and as the increments of* ϵ *are de-correlated conditionally to* $\mathcal{F}^{X,dN^X}_T$*,* $E\left[(k*(Y\mu(Y)))^2(t)|\mathcal{F}^{X,dN^X}_T\right] = \int_0^t\psi(t-s)E\left[(k*d\epsilon)^2_t|\mathcal{F}^{X,dN^X}_T\right]$*. The smoothing kernel* κ *is here a Gaussian density with standard deviation* σ*, therefore* $E\left[(k*d\epsilon)^2_t|\mathcal{F}^{X,dN^X}_T\right] = \frac{E\left[d\epsilon^2_t|\mathcal{F}^{X,dN^X}_T\right]}{\sigma^2}$ *therefore*

$$E\left[(k*(Y\mu(Y)))^2(t)|\mathcal{F}^{X,dN^X}_T\right] = \frac{E\left[(Y\mu(Y))^2(t)|\mathcal{F}^{X,dN^X}_T\right]}{\sigma}$$

and the theorem follows. ■

3.1.3 Bias variance tradeoff

The remarks above clearly delineate some of the design choices that we will take into account when designing the smoothing kernel κ. The theorems we proved already required that the kernel is continuous with a non-empty interior support. To decrease bias we want the kernel to be as concentrated about 0 as possible. However, at least in the Gaussian kernel family, we want to have wide enough a support so as to decrease the variance by pooling asynchronous observations together.

3.2 Operating in the frequency domain

A problem we have is that even for standard smoothing kernels the method above is expensive to compute as it requires smoothing an entire irregularly observed time series by a kernel with a theoretically infinite resolution. We solve this by employing randomized Fourier transforms.

3.2.1 Fourier random projection basis

First, consider the Fourier transform of an irregularly observed process $(X_t)_{t\in\text{obs}(X)}$ is defined as

$$FT[X](f) = \int_{t\in[0,T]} e^{-2\pi i f t}X(t)dN^X_t. \qquad (8)$$

The computation of this quantity can be accelerated using the techniques described in [25].

We assume that we generate frequencies f with a probability distribution F prior to computing the corresponding set of Fourier transforms. Let us consider the element-wise product of the Fourier transform of Y with the complex conjugate of the Fourier transform of X:

$$SP[Y,X](f) = FT[Y](f)\times\overline{FT[X](f)}. \qquad (9)$$

Theorem 3.2 *Consider the random inverse Fourier trans-*

form of this element-wise product.

$$E_{f\sim F}\left(e^{2\pi i f h}SP[Y,X](f)\right) = \iint_{t,s\in[0,T]} X(t)Y(s)FT[F](t-s+h)dN_t^X dN_s^Y. \quad (10)$$

where $FT[F](\cdot)$ is the Fourier transform of the distribution of frequencies F.

Proof 3.3 *By definition of the Fourier transform*

$$E_{f\sim F}\left[e^{2\pi i f h}SP[Y,X](f)\right] = E_{f\sim F}\left[\iint_{t,s\in[0,T]} e^{2\pi i f(t-s+h)}X(t)Y(s)dN_t^X dN_s^Y.\right]$$

and therefore, as sampling of frequencies is independent from both the underlying process generation and the irregular sampling, the expectation of interest equals

$$\iint_{t,s\in[0,T]} E_{f\sim F}\left[e^{2\pi i f(t-s+h)}\right] X(t)Y(s)dN_t^X dN_s^Y.$$

■

Corollary 3.2 *If F is a Gaussian centered distribution of variance σ, $FT_{f\sim F}(\cdot) = g_{\frac{1}{\sigma}}(\cdot)$ where $g_{\frac{1}{\sigma}}$ is the density of a normal distribution of standard deviation $\frac{1}{\sigma}$. Therefore:*

$$E_{f\sim F}\left(e^{2\pi i f h}SP[Y,X](f)\right) = \iint_{t,s[0,T]} X(t)Y(s)g_\sigma(t-s+h)dN_t^X dN_s^Y. \quad (11)$$

This enables us to retrieve the estimator in Eq. (4), implicitly, via the frequency domain. Therefore we have found a frequency distribution to smoothly combine the observations of irregularly observed processes.

3.3 Erasure of memory in the frequency domain

With our method the application of a linear filter such as differentiation or fractional differentiation can also be calculated implicitly in the frequency domain.

Corollary 3.3 *Let $A(\cdot)$ a linear filter we aim to apply to the unobserved underlying process X and $B(\cdot)$ to Y. Filtering can be conducted in the frequency domain as*

$$E_{f\sim F}\left[e^{-2\pi i f h}FT[B]FT[Y]\overline{(FT[A]FT[X])}(f)\right] = \iint_{t,s[0,T]} A * X(t) \times B * Y(s)\kappa(t-s+h)dN_t^X dN_s^Y \quad (12)$$

Pre-processing can therefore be translated in the frequency domain so as to study LRD processes. For a fractional differentiation of level α dedicated to erasing LRD as in Figure 3, we use $A(f) = (2\pi i f)^\alpha$ [18]. This is the pre-processing step we use to study cross-correlation between Brownian motion increments as in Figure 2. It also enables us to show our estimator obtains results comparable to HY in the case of Brownian motions in Figure 5 although it can used much more generally.

3.3.1 Choosing the number of basis frequencies

It is practically impossible to project a process on a continuous distribution of elements of the Fourier basis which is what writing $E_{f\sim F}\left[e^{2\pi i f h}SP[Y,X](f)\right]$ implies, we will sample a finite number N_f of frequencies from F and then compute the associated predictions before computing

$$\widehat{\gamma}_{\mathrm{f}}^{XY}(h) = \frac{1}{N_f}\sum_{f=1}^{N_f}\left[e^{2\pi i f h}SP[Y,X](f)\right]$$
$$= \iint_{t,s[0,T]} X(t)Y(s)\sum_{f=1}^{N_f} e^{2\pi i f(t-s+h)}dN_t^X dN_s^Y \quad (13)$$

whose variance will be inversely proportional to M and convergences in probability to $\widehat{\gamma}_{\mathrm{smooth}}^{XY}(h)$ as $M \to \infty$. As the sampling process is independent from the data generation process, with simple Landau notations, in the case of a Gaussian smoothing kernel of standard deviation σ,

$$Var\left(\widehat{\gamma}_{\mathrm{f}}^{XY}(h)\right) = O(\frac{1}{\sigma M}) \quad (14)$$

where the constant depends on the properties of both the *underlying* process and the *measuring* process which we have no control over. It appears therefore that there is a cost to the compression of the information entailed in the Fourier transform sets $(FT[X](f_i))_{i=1\ldots M}$ in terms of variance. The less data we use for this sufficient statistic, the higher the variance of the cross-covariance estimator.

3.4 Communication Avoding Random Projections

In practice, the data can be distributed across multiple computing devices or sensing platforms linked together by a lower bandwidth communication medium, sorting and collocating data involves too much communication and sufficient statistics cannot be delivered with a reactivity that enables interactive exploratory data analysis [26]. Frequency domain estimation does not require observations to be sorted chronologically and shuffled across multiple nodes of computation. Therefore, the techniques we offer, based on specific frequency domain random projections, enable substantial reduction of communication at the cost of more computations. We empirically demonstrate in Section 4.4 that this enables linear scaling on unsorted data sets scattered across nodes in a data center as we can summarize data sets of several hundreds of millions of timestamps with a few thousand Fourier transforms.

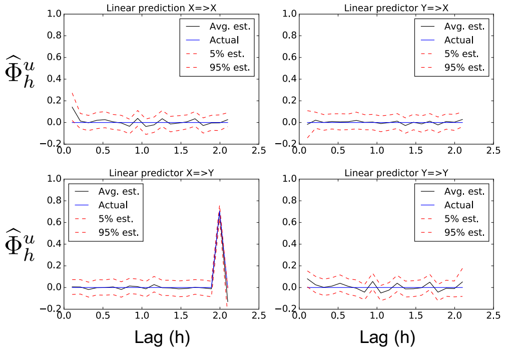

Figure 4: A Monte-Carlo experiment to confirm the validity of our method. An auto-regressive model is generated and then randomly sampled with only 60% of timestamps. The smoothed cross-correlogram estimator we present enabled us to retrieve the parameters of the underlying model without bias and with little variance. Without smoothing, the cross-correlation estimator fails to capture the singularity in 2 and evaluates to 0 everywhere on average. With a probability of 90% the relative error is 8%.

4 Experiments

In the following we will use the notation $\widehat{\gamma}_f^{XY}(\cdot)$ for our frequency domain kernel smoothing estimator and demonstrate its properties experimentally.

4.1 Linear model estimation

A Monte Carlo experiment in which we have prescribed a certain impulse function Φ in equation Eq. (1) demonstrates how the cross-correlation estimates that we provide enable us to reliably retrieve Φ even though the underlying process we simulate is only observed at random times. In this subsection, and only for the sake of this experiment, we simulate a discrete time auto-regressive process with a millisecond time resolution. Once the cross-correlation function between (X) and (Y) and their auto-correlation functions have been estimated as $\left(\widehat{\gamma}_f^{XY}(h)\right)_{h \in H}$ where H is a set finite set of millisecond resolution lags, solving the Yule-Walker equations below [4] gives an estimate of $\widehat{\Phi}_f$ for Φ which we compare to the values we chose for Φ. Solving the Yule-Walker equations is equivalent to a Least Squares Regression approach [4] as it comes down to estimating the precision matrix of a group of random variables based on their covariance matrix [27]. We empirically evaluate our ability to reduce variance by simulating 1000 samples from a single bi-variate linear auto-regressive process as defined in Eq. (1). For each sample, we introduced a homogeneous independent random observation process with $\mu_X = \mu_Y = 0.6$. In Figure 4, we plot the estimated cross-correlation and estimated model parameters as a function of the lag parameter h along with a 95% confidence intervals. We observe that the cross-correlation estimates have low variance relative to the peak cross-correlation magnitude and that the linear model parameters are accurately recovered with low variance.

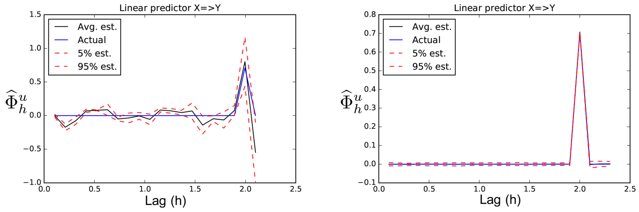

Figure 5: Left: HY estimation on synthetic Brownian motions. Right: Estimation with frequency domain differentiation. Although it has more variance, our estimator has similar bias to HY without being specifically designed to Brownian motions.

4.2 Estimating cross-correlations 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 and futures. In these settings it is then necessary to use full resolution data which in this instance comes in the form of tables recording record bids, asks and exchanges on the stock market as they happen. The timestamps are therefore irregular and not common to different pairs of stocks. Also, stock prices are Brownian motions and therefore feature long memory. This context is therefore in the very scope of data intensive tasks we consider. We show our novel Fourier compression based cross-correlation estimator provides consistent estimates in this setting.

4.3 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 four pairs of time series: APPL/APPL-LAG, AAPL/IBM, AAPL/MSFT, MSFT/IBM. We study the changes in quoted prices (more exactly, volume averaged bid and ask prices). 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. This is several orders-of-magnitude less than the number of observations that we get per day. These range from 5×10^4 to 1×10^5. 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.

4.4 Random projections enable scalability

A primary goal of this work is to enable practical scalable causal inference for time series analysis. To evaluate scalability in a real-world setting 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

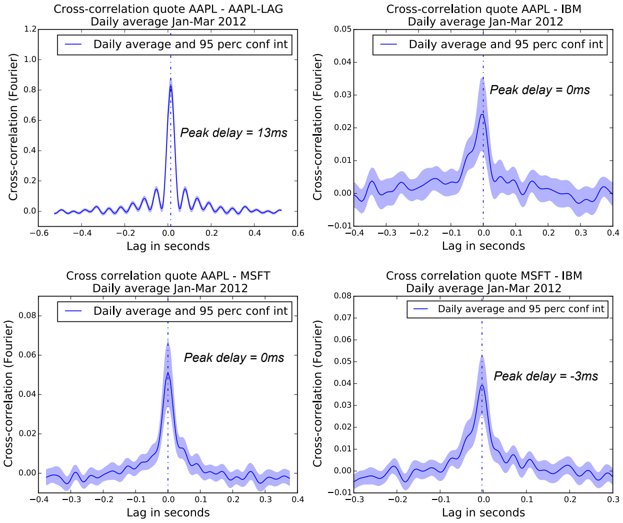

Figure 6: Average of daily cross-correlograms pairs of stock trade and quote data. Compression ratio is $< 5\%$. We retrieve lag and correlation accurately on surrogate data. The daily averaged cross-correlogram of AAPL and IBM is strongly asymmetric, therefore AAPL can predict IBM. The symmetry between AAPL and MSFT shows there is no such relationship between them. Symmetric and offset in correlation peak show that variations in IBM can be predicted $3ms$ ahead by observing MSFT.

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: In order to assess the scalability of the algorithm in a situation where communication is a major bottleneck, we run the experiment with Apache Spark on Amazon Web Services EC2 machines of type r3.2xlarge. In Figure 8 we show that even with a large number of projections (10000) which is enough to considerably reduce the variance of the estimator the communication burden is low enough to achieve nearly linear speedup.

Conclusion

In this paper we address the issues hampering cross-covariance estimation for randomly observed continuous stochastic processes. Pre-existing methods suffer from bias issues that can mislead researchers into identifying linear causality where there is none, lack strong statistical guarantees with general assumptions on the process of interest and rely on high communication requirements to be computed in the distributed setting. After having defined a new estimator for cross-covariance that does not systematically evaluate to 0, we analyze the bias/variance trade-off related to the kernel smoothing our method relies on. To enable scalability we leverage the careful design of random Fourier transforms to implicitly compute kernel smoothing between asynchronously observed time series in the frequency domain. We show with simulated data that the estimator we propose reliably retrieves model parameters. We then demonstrate that it enables a scalable study of stock market pair cross-correlation with tens of millions of high frequency recordings.

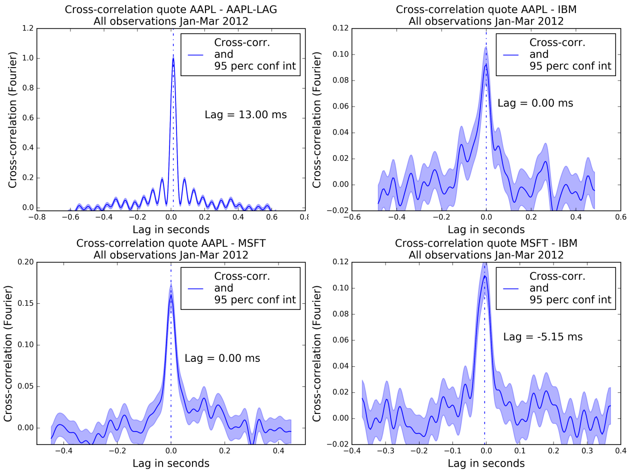

Figure 7: Compression ratio is $< 1\%$. On the entire data set we retrieve results similar to 6 therefore validating the use of our estimation of cross-correlograms in a scalable manner thanks to Fourier domain compression. Confidence bounds are computed using the asymptotic independence of spectral components [11].

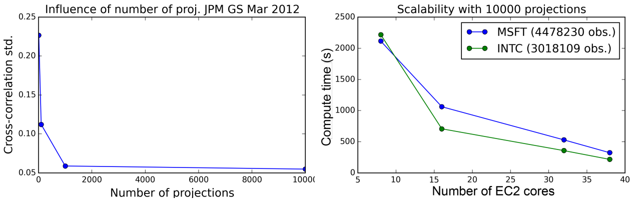

Figure 8: On the left we plot the empirical standard deviation of daily cross-correlograms (Figure 7) with respect to the number of projections showing that the variability decreases rapidly. On the right we plot the run time performance of our algorithm versus the number of Apache Spark EC2 cores showing approximately linear speedup. The small number of projections (10^4) relative to the size of the data set (10^7 records) avoids communication.

Acknowledgements: This research is supported in part by DHS Award HSHQDC-16-3-00083, NSF CISE Expeditions Award CCF-1139158, DOE Award SN10040 DE-SC0012463, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, IBM, SAP, The Thomas and Stacey Siebel Foundation, Apple Inc., Arimo, Blue Goji, Bosch, Cisco, Cray, Cloudera, Ericsson, Facebook, Fujitsu, HP, Huawei, Intel, Microsoft, Mitre, Pivotal, Samsung, Schlumberger, Splunk, State Farm and VMware.

References

[1] F. Abergel, J.-P. Bouchaud, T. Foucault, C.-A. Lehalle, and M. Rosenbaum, *Market microstructure: confronting many viewpoints*. John Wiley & Sons, 2012.

[2] R. S. Tsay, *Analysis of financial time series*. John Wiley & Sons, 2005, vol. 543.

[3] M. Mudelsee, *Climate time series analysis*. Springer, 2013.

[4] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York, NY, USA: Springer-Verlag New York, Inc., 1986.

[5] N. Huth and F. Abergel, "High frequency lead/lag relationships - empirical facts," *Journal of Empirical Finance*, vol. 26, pp. 41–58, 2014.

[6] C. W. Granger, "Causality, cointegration, and control," *Journal of Economic Dynamics and Control*, vol. 12, no. 2, pp. 551–559, 1988.

[7] J. L. Doob, "Stochastic processes," 1990.

[8] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[9] D. Cox and P. Lewis, "Multivariate point processes," *Selected Statistical Papers of Sir David Cox: Volume 1, Design of Investigations, Statistical Methods and Applications*, vol. 1, p. 159, 2005.

[10] T. Hayashi, N. Yoshida *et al.*, "On covariance estimation of non-synchronously observed diffusion processes," *Bernoulli*, vol. 11, no. 2, pp. 359–379, 2005.

[11] D. R. Brillinger, *Time series: data analysis and theory*. Siam, 1981, vol. 36.

[12] H. Ltkepohl, *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, Incorporated, 2007.

[13] E. 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*. Springer Science & Business Media, 2012, vol. 25.

[14] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*. MIT press Cambridge, MA, 1949, vol. 2.

[15] M. Friedman, "The interpolation of time series by related series," *Journal of the American Statistical Association*, vol. 57, no. 300, pp. 729–757, 1962.

[16] M. Hoffmann, M. Rosenbaum, N. Yoshida *et al.*, "Estimation of the lead-lag parameter from non-synchronous data," *Bernoulli*, vol. 19, no. 2, pp. 426–461, 2013.

[17] I. Karatzas and S. Shreve, *Brownian motion and stochastic calculus*. Springer Science & Business Media, 2012, vol. 113.

[18] P. Doukhan, G. Oppenheim, and M. S. Taqqu, *Theory and applications of long-range dependence*. Springer Science & Business Media, 2003.

[19] B. B. Mandelbrot, *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk. Selecta Volume E*. Springer Science & Business Media, 2013.

[20] B. M. Tabak and D. O. Cajueiro, "Are the crude oil markets becoming weakly efficient over time? a test for time-varying long-range dependence in prices and volatility," *Energy Economics*, vol. 29, no. 1, pp. 28–36, 2007.

[21] W. Li and K. Kaneko, "Long-range correlation and partial 1/fα spectrum in a noncoding dna sequence," *EPL (Europhysics Letters)*, vol. 17, no. 7, p. 655, 1992.

[22] P. Flandrin, "On the spectrum of fractional brownian motions," *Information Theory, IEEE Transactions on*, vol. 35, no. 1, pp. 197–199, 1989.

[23] D. H. Smith, N. Perrin, A. Feldstein, X. Yang, D. Kuang, S. R. Simon, D. F. Sittig, R. Platt, and S. B. Soumerai, "The impact of prescribing safety alerts for elderly persons in an electronic medical record: an interrupted time series evaluation," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1098–1104, 2006.

[24] K. Karhunen, "Über die struktur stationärer zufälliger funktionen," *Arkiv för Matematik*, vol. 1, no. 2, pp. 141–160, 1950.

[25] L. Greengard and J.-Y. Lee, "Accelerating the nonuniform fast fourier transform," *SIAM review*, vol. 46, no. 3, pp. 443–454, 2004.

[26] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012, pp. 2–2.

[27] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.