REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
11-05-2017	Final Report	15-Apr-2016 - 14-Jan-2017

4. TITLE AND SUBTITLE
Final Report: Causality and Information Dynamics in Networked Systems with Many Agents (ARO 10.3)

5a. CONTRACT NUMBER
W911NF-16-1-0155

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Ravi Mazumdar

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Waterloo
200 University Avenue West

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
68323-NS-II.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This report presents results on a theoretical formulation and algorithms for reconstructing Granger causality graphs (GCG) from collections of wide sense stationary (WSS) and cyclostationary time series data. The thrust of the research was to develop methods for GCG sparsification using ideas from Tikhonov regularization and ADMM based proximal algorithms. Several computational examples are presented.

15. SUBJECT TERMS
Granger Causality Graphs, sparsification, algorithms

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ravi Mazumdar
UU	UU	UU	UU		19b. TELEPHONE NUMBER 519-888-4567

RPPR

as of 25-Aug-2017

Agency Code:

Proposal Number: **Agreement Number:**

Organization:
Address: , ,
Country:
DUNS Number: EIN:
Report Date: Date Received:
for Period Beginning and Ending
Title:
Begin Performance Period: **End Performance Period:**
Report Term: -
Submitted By: Email:
Phone:

Distribution Statement: -

STEM Degrees: **STEM Participants:**

Major Goals:

Accomplishments:

Training Opportunities:

Results Dissemination:

Plans Next Period:

Honors and Awards:

Protocol Activity Status:

Technology Transfer:

Final Report on ARL STIR Grant
Causality and Information Dynamics in Networked Systems with Many Agents

Duration 9 months from April 15, 2016- Jan. 14, 2017

Principal Investigator:
Ravi R. Mazumdar
Professor and University Research Chair
Department of Electrical and Computer Engineering
University of Waterloo
200 University Ave. W
Waterloo ON N2L 3G1 CANADA
Ph: +1 519 888 4567 Ext 37444
Fax: +1 765 494 3358
Email: mazum@uwaterloo.ca

STIR Grant: Intelligent Networks Research program

Technical Point of Contact: Dr. Purush Iyer
e-mail: s.p.iyer.civ@mail.mil, (919) 549-4204

Contents

OVERVIEW

The research supported by the STIR grant focussed on so called network models. The major focus was on the construction of causality graphs from temporal data where causality is temporal in the classical sense of outputs of systems only depending on the past values of inputs. When dealing with wide sense stationary time series the way of studying this is through the so-called Wold decomposition [1].

During the first phase of the research April 15, 2016-July 31, 2016 our focus was on the formulation of Granger causality problem involving multiple time-series and wether Granger causal graphs (CCG) remain invariant with respect to subsampling of time series The second phase carried out between Aug 1, 2016 till January 14,2016 was on the development of computational algorithms for the sparsification of Granger causality graphs. This report presents the details of the work done with a focus on the period Aug 1, 2016-Jan 14, 2017.

For the construction of GCGs, we are interested in determining how and under what conditions it is possible to recover (exactly or approximately) the full GCG, given a graph obtained only through pairwise Granger causality tests. Since we expect the pairwise graph to be denser than the full graph, this research may provide methods useful for the more general problem of graph sparsification.

In the consideration of randomly subsampled stochastic processes, we are interested in determining to what extent causal information is preserved through various different subsampling processes. This may provide insight into the validity of causal statements based on subsampled data.

Research Team

The project was delayed in its start because a Post-doc who was working on related problems left at the end of April 2016. New research students were recruited in the middle of Summer 2016 and thus were supported by funds from the STIR grant for two terms. The students are: Ryan Kinnear (MASc student who worked on the algorithms for Granger graph sparsification) and Thirupathiah Vasantam (PhD student who is working on mean field dynamics of complex interacting networks).

1 Causal Inference for Time Series Data

Consider two second order jointly stationary stochastic processes $x(n)$ and $y(n)$. In [2] [3] Granger proposed a notion of causality based on minimum mean square error estimation as follows. Denote by $E[x(n)|\mathcal{F}_{n-1}]$ the optimum (in the mean square sense) predictor of $x(n)$ given the universe of information up to time $n-1$, and $E[x(n)|\mathcal{F}^{-y}_{n-1}]$ the optimal predictor given all the information in the universe up to time $n-1$ *excluding* all of the information provided by the process y. Denote by $\xi[x(n)|\mathcal{F}_{n-1}] = E[x(n) - E[x(n)|\mathcal{F}_{n-1}]]^2$ the corresponding mean squared error, similarly for $\xi[x(n)|\mathcal{F}^{-y}_{n-1}]$. We will say that y *strongly Granger causes* x if

$$\xi[x(n)|\mathcal{F}_{n-1}] < \xi[x(n)|\mathcal{F}^{-y}_{n-1}], \tag{1.1}$$

Essentially, we identify a causal relationship if the past of y has some predictive power for the future of x that is not present anywhere else in the universe.

2 Granger Causality Graphs

Let $x(t) = (x_1(t), \ldots, x_n(t))$ be a column vector of real valued discrete time ($t \in \mathbb{Z}$) wide sense stationary (WSS) stochastic processes with bounded second moments, that is $x_i(t) \in L_2(\Omega, \mathcal{F}, \mathbf{P})$, the Hilbert space of square integrable random variables. Let $H_t = \mathbf{cl}\left\{\sum_{\tau=1}^{\infty}\sum_{i=1}^{n} b_i^{(\tau)} x(t-\tau) \,|\, b_i^{(\tau)} \in \mathbb{R}\right\}$ denote the Hilbert space of random variables generated by the (strict) past of $x(t)$ and $H_t^{-j} = \mathbf{cl}\left\{\sum_{\tau=1}^{\infty}\sum_{i\neq j} b_i^{(\tau)} x(t-\tau) \,|\, b_i^{(\tau)} \in \mathbb{R}\right\}$ the space generated by all but component j. Here **cl** denotes the closure of the space.

The notation $\hat{E}[x_i(t) \mid H_t]$ denotes the projection of $x_i(t)$ onto the Hilbert space H_t, which is the causal linear minimum mean square error (LMMSE), or Wiener, estimate of $x_i(t)$ given the strict past of $x(t)$. And, the expected squared error of the estimate: $\hat{\xi}[x_i(t) \mid H_t] \triangleq E\left[(\hat{E}[x_i(t) \mid H_t^V] - x_i(t))^2\right]$. Note that since the processes are wide sense stationary, the aforementioned quantities do not vary with time. The notion of Granger-causality is captured in the following definition.

Definition 2.1 *If*

$$\hat{\xi}[x_i(t) \mid H_t] < \hat{\xi}[x_i(t) \mid H_t^{-j}] \ , \tag{2.2}$$

then we say that x_j Granger-causes x_i (conditional on x), and write $x_j \longrightarrow x_i$.

Some of the intuition behind this definition is greatly expanded upon on in [4]. We also point out that this notion of causality has little in common with the notion of causality popularized by Pearl [5]. Pearl's causality relates to logical as opposed to temporal causality.

2.1 Autoregressive Modeling

Recall that $x(t)$ is an n-vector of WSS processes. The Wold decomposition theorem tells us that there is some square summable sequence of real valued $n \times n$ matrices $A(\tau)$, a white noise sequence $\epsilon(t)$, and a perfectly predictable sequence $u(t)$ such that

$$x(t) = \sum_{\tau=0}^{\infty} A(\tau)\epsilon(t-\tau) + u(t). \tag{2.3}$$

This is a moving average representation of $x(t)$, and exists for every WSS L_2 process. In practice, the predictable term $u(t)$ is removed by detrending, and so we simply take $u(t) = 0$. In order to obtain an autoregressive representation, the LSI filter given by $A(\tau)$ must be invertible, and a sufficient condition for this invertibility is that there is some $c > 0$ such that the spectral density matrix $S_x(\lambda)$ of $x(t)$ satisfies $c^{-1}I \preceq S_x(\lambda) \preceq cI$ for λ almost everywhere in $[-\pi, \pi)$. Given this condition, we have again a square summable sequence $B(\tau)$ such that

$$x(t) = \sum_{\tau=1}^{\infty} B(\tau)x(t-\tau) + e(t), \tag{2.4}$$

where $e(t)$ is serially uncorrelated. Finally, availability only of finite quantities of data necessitates that we restrict ourselves further by assuming that $x(t)$ is generated by the Markovian vector autoregressive VAR(p) model

$$x(t) = \sum_{\tau=1}^{p} B(\tau)x(t-\tau) + e(t). \tag{2.5}$$

A natural perspective is to view this model as a graph $\mathcal{G} = (V, E)$ having nodes $x_i(t)$ and edges given by the linear shift-invariant (LSI) filter $\widetilde{B}_{ij}(z) = \sum_{\tau=1}^{p} B_{ij}(\tau)z^{-\tau}$ whose coefficients are arranged into a column vector $\widetilde{B}_{ij} = (B_{ij}(1), \ldots, B_{ij}(p))$. In this model, Granger-causality has a particularly simple characterization:

Proposition 2.1 *If $x(t)$ is an n dimensional wide sense stationary L_2 stochastic process generated by the VAR(p) model (2.5) then $x_j(t)$ Granger-causes $x_i(t)$ if and only if $\widetilde{B}_{ij} \neq 0$.*

Proof: The condition for Granger-Causality $x_j \longrightarrow x_i$ is given as

$$\hat{\xi}[x_i(t) \mid H_t] < \hat{\xi}[x_i(t) \mid H_t^{-j}] \ . \tag{2.6}$$

Since $e(t)$ is temporally uncorrelated, the Hilbert space projections are given by the model's true parameters, so (2.6) is equivalent to

$$E|x_i(t) - \sum_{\tau=1}^{\infty}\sum_{k=1}^{n} B_{ik}^{(\tau)} x_k(\tau)|^2 < E|x_i(t) - \sum_{\tau=1}^{\infty}\sum_{k\neq j} B_{ik}^{(\tau)} x_k(\tau)|^2.$$

Now, if there were no τ_0 such that $B_{ij}^{(\tau_0)} \neq 0$ then the above strict inequality would infact be an equality, a contradiction. Conversely, since $B_{ik}^{(\tau)}$ provides the best linear estimate of $x_i(t)$ from $x(t)$, if there is some τ_0 such that $B_{ij}^{(\tau_0)} \neq 0$ then above strict inequality must hold, otherwise $B_{ij}^{(\tau_0)} = 0$ would provide an equivalent or superior prediction, contradicting either the uniqueness of projections in Hilbert space, or the optimality of the projection. □

2.2 Estimating VAR Model Coefficients

Given a finite sample of $T + p$ data points: $x(-p+1), x(-p+2), \ldots, x(T)$, there are a wide variety of methods available to produce an estimate $\hat{B}(\tau)$ of the coefficients $B(\tau)$ in the model (2.5). Classical methods revolve around solving the Yule-Walker equations with finite data estimates of covariance sequences, and indeed, this is the approach put forth by Geweke in [6]. A similar approach is taken in [7]. Another is the simple ordinary least squares estimate

$$\underset{B(\tau)}{\text{minimize}} \quad \frac{1}{2T}\sum_{t=1}^{T} ||x(t) - \sum_{\tau=1}^{p} B(\tau)x(t-\tau)||_2^2 \ , \tag{2.7}$$

which is our starting point. This can be viewed as either a maximum likelihood estimate in the case for which $e(t)$ is Gaussian, or as an asymptotically valid estimate of the LMMSE estimator.

When data is abundant for each component of $x(t)$ (e.g. when $T >> pn^2$), either of the aforementioned methods are perfectly adequate. However, many applications do not satisfy this requirement. Indeed, the underlying graphical structure induced by $B(\tau)$ only becomes interesting when n is of at least modest size. In this case, the variance of traditional or OLS estimates of $B(\tau)$ is so large as to render to estimates entirely useless.

Standard methods to deal with this issue is to accept some bias in the estimation process and add regularizing terms. By appropriately arranging coefficients, we can consider the problem:

$$\underset{B}{\text{minimize}} \quad \frac{1}{2T}||Y - BZ||_F^2 + \lambda\big[\alpha||B||_F^2 + (1-\alpha)\Gamma(B)\big] \ , \tag{2.8}$$

where $Y = [x(T) \ldots x(1)]$ is $(n \times T)$ formed directly from the column vectors $x(t)$, $Z = [z(T-1) \ldots z(0)]$ is $(np \times T)$ where the columns $z(t)$ are formed from stacking $x(t), \ldots, x(t-p+1)$, and $B = [B(1)\, B(2)\, \ldots\, B(p)]$ is the $(n \times np)$ coefficient matrix. The term $\lambda \geq 0$ is a tuning parameter for the amount of regularization, and $\alpha \in [0,1]$ trades off between the regularizer Γ and $||\cdot||_F^2$.

Different choices of Γ in the problem (2.8) lead to different estimates of $B(\tau)$ and hence allow for a great deal of flexibility in the modeling process. The principle drawback in this approach however is that the resulting estimates are not guaranteed to yield a stable system. This is a big problem if the model is to be used for forecasting, but when we are interested only in the underlying graphical structure, it is not of any great consequence whether the resulting system is stable or not.

3 Structured Grouping for Causal Inference: DWGLASSO

Common regularizers in the context of regression are the squared ℓ_2 Frobenius norm ($\alpha = 1$), referred to as Tikhonov regularization, or a simple ℓ_1 norm $\Gamma_1(B) \triangleq ||B||_1 = \sum_{ij} |B_{ij}|$ (with $\alpha = 0$), which is the well known sparsity inducing "LASSO" regularizer [8].

The LASSO regularizer, which results in an unstructured sparsity pattern in the B matrix, can be extended to the grouped LASSO (GLASSO) in which we take a sum of unsquared Euclidean norms $\Gamma_G(B) = \sum_{g\in G} ||B_{[g]}||_2$ on groups in the B matrix, where $B_{[g]}$ denotes a vector of B coefficients in group $g \subseteq \{1, 2, \ldots, n\}$. It was shown by Yuan et al. [9] that this leads to a sparsity pattern in which each of the coefficients in $B_{[g]}$ are jointly zero or non-zero.

Inspired by the characterization of proposition 2.1, in a vein similar to [10] and [11] we propose to use

$$\Gamma_{DW}(B) = \sum_{ij} ||\widetilde{B}_{ij}||_2 \ , \tag{3.9}$$

which forms groups along each edge of the underlying causality graph. The matrix B is formed as a "row-wise" matrix $B = [B(1)\, B(2)\, \ldots\, B(p)]$ of the lagged coefficient matrices. It is also natural to imagine stacking vertically (into or out of the page) the matrices $B(\tau)$ to form an $(n \times n \times p)$ array, analogous to the adjacency matrix of the underlying graph, so that looking "depth-wise" at location ij gives the coefficients $\widetilde{B}_{ij} \in \mathbb{R}^p$ of the LSI filter from process j to process i. It is for this reason that we refer to this structured regularizer as the depth-wise group LASSO (DWGLASSO) regularizer.

In the case where $\alpha = 1$, colinearity in the data leads to inconsistent estimates in that if x_j and $x_{j'}$ provide similar information about x_i the GLASSO estimate will tend to select only one or the other. This is a serious problem when we want to infer a causality graph. Adding in the ℓ_2 norm term with $\alpha \in (0, 1)$ is referred to as the elastic net [12] and, essentially because it makes the objective strongly convex, eliminates this problem; the estimator will blend together the influences from x_j and $x_{j'}$.

3.1 ADMM for DWGLASSO

Although, as in [10], it is possible to fit the DWGLASSO model with $\alpha = 0$ via a second order cone program, the elastic-net form is desirable and for large models, more specialized methods are necessary. It is straightforward to derive a subgradient coordinate descent approach to solving (2.8), however there are n^2 coordinate axes to iterate through, and the algorithm does not turn out to be very easily implemented, these drawbacks are on top of the fact that subgradient descent can be extremely slow.

On the other hand, the alternating direction method of multipliers (ADMM) [13] [14] is a fast, and potentially parallelizable algorithm well suited to our needs. Given two closed, proper, convex, though not necessarily differentiable functions f and g, the ADMM algorithm minimizes over B the objective $f(B) + g(B)$ and dictates that we perform the following updates:

$$\begin{aligned} B_x^{k+1} &\leftarrow \text{prox}_{\mu f}(B_z^k - B_u^k) \\ B_z^{k+1} &\leftarrow \text{prox}_{\mu g}(B_x^{k+1} + B_u^k) \\ B_u^{k+1} &\leftarrow B_u^k + B_x^{k+1} - B_z^{k+1}, \end{aligned} \tag{3.10}$$

where

$$\text{prox}_{\mu\phi}(V) = \underset{X\in\mathbb{R}^{n\times n}}{\text{argmin}} \left(\phi(X) + \frac{1}{2\mu}||X - V||_2^2\right) \triangleq \underset{X\in\mathbb{R}^{n\times n}}{\text{argmin}}\ P_\phi(X), \tag{3.11}$$

is the proximity operator of a function ϕ. ADMM guarantees that $B_x^k \longrightarrow B_z^k$ as $k \longrightarrow \infty$, and that the objective function value converges towards the optimal objective value. The parameter μ tunes the convergence of the algorithm, but it's careful selection is not of paramount importance; we have found by ad-hoc tuning that $\mu \approx 0.1$ is satisfactory.

In our case, we have $f(B) = \frac{1}{2T}||Y - BZ||_F^2 + \lambda\alpha||B||_F^2$ and $g(B) = \lambda(1-\alpha)\sum_{ij} ||\widetilde{B}_{ij}||_2$. As long as we require $\lambda > 0$ and $\alpha \in (0, 1]$ the objective (2.8) is strongly convex and hence ADMM is guaranteed to find the unique global minimizer.

We stress that the key advantage of ADMM for our purposes is that the elements of the matrix B are viewed in a completely different arrangement in g than they are in f and that simply rearranging the matrices in our problem does not ameliorate this difficulty.

Proposition 3.1 (Proximity Operator of $f(B) = \frac{1}{2T}||Y - BZ||_F^2 + \lambda\alpha||B||_F^2$)

$$\text{prox}_{\mu f}(V) = (\frac{1}{T}YZ^\mathsf{T} + \frac{1}{\mu}V)(\frac{1}{T}ZZ^\mathsf{T} + \frac{1+2\mu\lambda\alpha}{\mu}I)^{-1} \ . \tag{3.12}$$

Since this objective is differentiable and unconstrained, we can easily solve (3.11).

Proof:

$$\frac{\partial P_f}{\partial B}(B) = \frac{1}{T}(BZZ^\mathsf{T} - YZ^\mathsf{T}) + 2\alpha\lambda B + \frac{1}{\mu}(B - V).$$

Applying the first order optimality condition

$$\frac{\partial P_f}{\partial B}(B^*) = 0 \implies B^* = (\frac{1}{T}YZ^\mathsf{T} + \frac{1}{\mu}V)(\frac{1}{T}ZZ^\mathsf{T} + \frac{1+2\alpha\lambda\mu}{\mu}I)^{-1},$$

and since the objective is strongly convex, we have obtained the unique global minimizer. □

Proposition 3.2 (Proximity Operator of $g(B) = \lambda(1-\alpha)\sum_{i,j}||\widetilde{B}_{ij}||_2$)

$$\text{prox}_{\mu g}(V) = \Big[P(1)\ P(2)\ \ldots\ P(p)\Big] \in \mathbb{R}^{n\times np} \ , \tag{3.13}$$

where

$$P(\tau)_{ij} = \Big(1 - \frac{\mu\lambda(1-\alpha)}{||\widetilde{V}_{ij}||_2}\Big)_+ \widetilde{V}(\tau)_{ij} \tag{3.14}$$

Recall that $\widetilde{B}_{ij}$ denotes the coefficients of the LSI filter from x_j to x_i. The notation $\widetilde{V}_{ij}(\tau)$ denotes the τ^{th} component of the analagous arrangement.

Proof: The objective function separates along $\widetilde{V}_{ij}$, so we need only establish (3.14). To this end, let $\phi(x) = \lambda(1-\alpha)||x||_2$ so that $g(B) = \sum_{ij}\phi(\widetilde{B}_{ij})$. The Fenchel conjugate $(\mu\phi)^*$ of $\mu\phi$ is the convex indicator function of the Euclidean unit ball having radius $\mu\lambda(1-\alpha)$. We therefore obtain the proximity operator

$$\text{prox}_{(\mu\lambda(1-\alpha)\phi)^*}(\widetilde{V}_{ij}) = \left\{ \begin{array}{ll} \widetilde{V}_{ij} & ;||\widetilde{V}_{ij}||_2 \leq \mu\lambda(1-\alpha) \\ \frac{\mu\lambda(1-\alpha)\widetilde{V}_{ij}}{||\widetilde{V}_{ij}||_2} & ;\text{otherwise} \end{array} \right\} \ . \tag{3.15}$$

A fundamental property of the proximity operator is the Moreau decomposition: $\text{prox}_{\mu\phi}(x) = x - \text{prox}_{(\mu\phi)^*}(x)$, application of which yields (3.14). □

Remark 3.1 Computational Considerations

There are a few things to note in regards to practical implementation. Firstly, the matrix inverse in (3.12) should not be carried out literally, an LU factorization of $(\frac{1}{T}ZZ^\mathsf{T} + \frac{1+2\mu\lambda\alpha}{\mu}I)$ can be cached and used throughout in solving the system of equations. Secondly, the matrices ZZ^T and YZ^T can be formed from the pairwise covariances of each x_i, x_j pair at the lags from 0 to p, further savings can be had by making use of the block toeplitz structure of ZZ^T. Finally, the matrix B_z^k is formed from the soft-thresholding in (3.14), and hence will be the sparse solution the algorithm should output upon convergence. The time complexity is $O(n^2p^2)$ per iteration, with $O(n^3p^3 + n^2pT)$ at initialization. Storage complexity is also on the order of $O(n^2p^2)$. Considering the problem context, it is unlikely that significant speedups are possible.

4 Granger-causality among a collection of cyclostationary time series

While WSS time series are relatively easy to analyze, a number of processes encountered in practical applications are non-stationary and therefore require more involved treatment. Many time series observed in various fields of study, including communications, control systems, meteorology and economics, have statistical characteristics that vary periodically with time. A large class of such sequences can be appropriately modeled as cyclostationary (CS) processes [15, 16, 17, 18]. In this section, we discuss the problem of defining and detecting Granger-causality among a collection of CS time series, and show that the procedure can be carried out without the explicit knowledge of the period of cyclostationarity.

Definition 4.1 *Let $\{x(n)\}_{n\in\mathbb{Z}}$ be a real-valued, zero-mean, discrete time stochastic process. The covariance of $\{x(n)\}$ is $R(n,\tau) = \mathbb{E}[x(n)x(n-\tau)]$. $\{x(n)\}$ is said to be cyclostationary (CS) [15] if $R(n,\tau)$ is periodic in the following sense.*

$$R(n,\tau) = R(n+lT_0,\tau), \tag{4.16}$$

where l is an integer.

We call T_0 the period of the CS process $\{x(n)\}$.

Two CS processes $\{x_i(n)\}$, $\{x_j(n)\}$ having the same period T_0 are said to be jointly CS if

$$\mathbb{E}[x_i(n)x_j(n-\tau)] = R_{i,j}(n,\tau) = R_{i,j}(n+lT_0,\tau),$$

where l is an integer. Consider a collection of processes $\{x_{1:K}(n)\}$ that are jointly CS with the same period T_0. Then, the $\mathbb{R}^K$-valued process $\{\boldsymbol{x}(n)\} = [x_1(n)\ \ldots\ x_K(n)]^\top$ is CS with the same period T_0. Unlike WSS processes, the projection of $\boldsymbol{x}$ on the linear span of all its past values is no longer stationary. For each process, let the MMSE linear estimate given the entire past of all the other processes, be given by

$$\hat{x}_i(n) = \sum_{\tau=1}^{\infty}\sum_{j=1}^{K} b_{i,j}(n,\tau)x_j(n-\tau),$$

where the parameters $b_{i,j}(n,\tau)$ are derived by the method of least squares; i.e., they minimize the mean-squared estimation error $\mathbb{E}\left[(x_i(n)-\hat{x}_i(n))^2\right]$. Let $\nu_i(n) = x_i(n) - \hat{x}_i(n)$ be the corresponding error. It can be shown that [19]

1. The parameters $b_{i,j}(n,\tau)$ are periodic in n with period T_0, i.e., $b_{i,j}(n,\tau) = b_{i,j}(n+lT_0,\tau)$ where l is an integer.
2. The error $\{\nu_i(n)\}$ is a CS process with

   $$\mathbb{E}[\nu_i(n)\nu_i(n-\tau)] = \mathbb{E}[\nu_i(n+lT_0)\nu_i(n+lT_0-\tau)],$$

   where l is an integer.

A CS process can be represented as a collection of WSS processes. Let the CS process $\{x(m)\}$ be such that, for any m, τ,

$$\mathbb{E}[(x(m))^2] \geq \mathbb{E}[x(m)x(m-\tau)].$$

For each $m \in \mathbb{Z}$, let $n = \left\lceil \frac{m}{T_0} \right\rceil$, and let $t = m - T_0\left(\left\lceil \frac{m}{T_0} \right\rceil - 1\right)$. Let $\{y_{1:T_0}(n)\}$ be a collection of T_0 time series defined as follows:

$$y_t(n) = x((n-1)T_0 + t). \tag{4.17}$$

The cross-covariance between any two of the above newly constructed processes $\{y_t(n)\}$ and $\{y_s(n)\}$ is found to be

$$\mathbb{E}[y_t(n)y_s(n-\tau)] = R(t,\tau T_0 + t - s).$$

Since the expression is independent of n, it follows that $\{y_{1:T_0}(n)\}$ constitutes a collection of jointly WSS processes [16].

Unlike the WSS case, here, the estimation parameters and error variances will no longer be stationary but will vary with period T_0. The definition of Granger-causality given in section **??**, therefore, has to be slightly modified to accommodate CS sequences.

Consider a collection of K jointly CS time series $\{x_{1:K}(n)\}$ having the same period T_0. $x_i(n)$ is first estimated by a linear MMSE estimator that uses the past values of all the processes except $\{x_j(n)\}$, with error $\xi_{i|-j}$:

$$x_i(n) = \sum_{\tau=1}^{\infty} \sum_{\substack{l=1 \\ l \neq j}}^{K} \tilde{b}_{i,l}(n,\tau) x_l(n-\tau) + \xi_{i|-j}(n),$$

and then estimated by a linear MMSE estimator using the past observations of all processes, including $\{x_j(n)\}$ with error ξ_i:

$$x_i(n) = \sum_{\tau=1}^{\infty} \sum_{l=1}^{K} b_{i,l}(n,\tau) x_l(n-\tau) + \xi_i(n).$$

We say that $\{x_j(n)\}$ Granger-causes $\{x_i(n)\}$ if for some $m \in \{1, \ldots, T_0\}$,

$$\mathbb{E}\left[(\xi_i(m))^2\right] < \mathbb{E}\left[\left(\xi_{i|-j}(m)\right)^2\right].$$

Granger-causality among K CS processes can be tested by deriving the least square parameters for each n, from 1 to T_0. The computation of these parameters necessitate estimation of the covariance $R_i(n,\tau) = \mathbb{E}[x_i(n)x_i(n-\tau)]$ and cross-covariance $R_{i,j}(n,\tau) = \mathbb{E}[x_i(n)x_j(n-\tau)]$ of the processes involved, for each n, $n = 1, \ldots, T_0$. When the period T_0 is large, this becomes computationally intensive.

Another way is to decompose each of the CS time series into T_0 WSS processes following (4.17), and then test for Granger-causality using the MVAR (Multidimensional VAR) or the pairwise approach for KT_0 WSS time series. However, not only is such an approach computationally burdensome, but it is also difficult to interpret and represent in the form of a graph. Through the rest of this section, we propose an alternative which determines Granger-causality by computing *time-invariant* least-square estimators, while treating the processes as WSS.

Define

$$\overline{R}_i(\tau) = \frac{1}{T_0} \sum_{n=1}^{T_0} \mathbb{E}[x_i(n)x_i(n-\tau)],$$

$$\overline{R}_{i,j}(\tau) = \frac{1}{T_0} \sum_{n=1}^{T_0} \mathbb{E}[x_i(n)x_j(n-\tau)].$$

$\overline{R}_i(\tau)$, $\overline{R}_{i,j}(\tau)$ so defined, do not depend on n. These are the arithmetic means of the covariance and cross-covariance terms, taken over the T_0 WSS components of the processes. When the least-square equations for fitting an MVAR model are solved by replacing $R_i(n,\tau)$ and $R_{i,j}(n,\tau)$ with $\overline{R}_i(\tau)$ and $\overline{R}_{i,j}(\tau)$, respectively, the resulting parameters are stationary, i.e., they are no longer functions of n. Let the time-invariant causal Wiener filter estimating $x_i(n)$ from the past values of $\{x_j(n)\}$ be given by

$$x_i(n) = \sum_{\tau=1}^{\infty} \overline{w}_{i|j}(\tau) x_j(n-\tau) + \overline{\xi}_{i|j}(n).$$

The following result shows that the mean-squared error $\mathbb{E}\left[\left(\overline{\xi}_{i|j}\right)^2\right]$ bears an interesting relation to Granger-causality.

Proposition 4.1 *Consider a system of K jointly cyclostationary processes $\{x_i(n)\}_{i=1,\ldots,K}$ with period T_0. If $\{x_j(n)\}$ Granger-causes $\{x_i(n)\}$, then*

$$\mathbb{E}\left[\left(\overline{\xi}_{i|j}\right)^2\right] < \overline{R}_i(0).$$

Proof: By definition, $\mathbb{E}\left[\left(\overline{\xi}_{i|j}\right)^2\right] \leq \overline{R}_i(0)$ is always true. The equality holds if and only if $x_i(n)$ is orthogonal (or uncorrelated) to all the past values of $\{x_j(n)\}$; i.e, $\overline{R}_{i,j}(\tau) = 0$ for all $\tau > 1$. To prove the result, we need to show that the inequality is strict.

Since $\{x_j(n)\}$ Granger-causes $\{x_i(n)\}$, there exists some $m \in \{1, \ldots, T_0\}$ for which $x_i(m)$ can be expressed as

$$x_i(m) = \sum_{\tau=1}^{\infty} b_{i,i}(m,\tau)x_i(m-\tau) + \sum_{\tau=1}^{\infty} c_{i,j}(m,\tau)x_j(m-\tau) + \xi_{i,j}(m).$$

Using the shift operator z, and re-arranging

$$x_i(m) = \left(1 - \sum_{\tau=1}^{\infty} b_{i,i}(m,\tau)z^{-\tau}\right)^{-1} \left(\sum_{\tau=1}^{\infty} c_{i,j}(m,\tau)z^{-\tau}x_j(n) + \xi_{i,j}(m)\right).$$

Re-arranging further, the above becomes

$$x_i(m) = \sum_{\tau=1}^{\infty} c'_{i,j}(m,\tau)x_j(m-\tau) + \xi'_{i,j}(m).$$

Following a similar argument used to establish Proposition **??**, $x_i(m)$ is not orthogonal to the past values of $\{x_j(m)\}$, and there exists some τ such that

$$\mathbb{E}[x_i(m)x_j(m-\tau)] \neq 0.$$

Therefore, $R_{i,j}(m,\tau) \neq 0$. It follows, then, that $\overline{R}_{i,j}(\tau)$ is, in general, also non-zero and the result follows. □.

The above indicates that the Granger-causality between two jointly CS processes having the same period is indicated by the mean-squared error corresponding to a pairwise causal Wiener filter, which treats the processes as WSS; without considering their cyclostationary characteristics. Granger-causality among CS processes can then be inferred by fitting pairwise time-invariant Wiener filters, where the least-square parameters are computed by replacing the covariance and cross-covariance terms with $\overline{R}_i(\tau)$ and $\overline{R}_{i,j}(\tau)$ respectively. We conclude this section by presenting asymptotically unbiased and consistent estimators for the two quantities.

For $|\tau| \leq N$, define

$$\hat{R}_{i,N}(\tau) = \frac{1}{N}\sum_{n=|\tau|+1}^{N} x_i(n)x_i(n-|\tau|),$$

$$\hat{R}_{i,j,N}(\tau) = \frac{1}{N}\sum_{n=|\tau|+1}^{N} x_i(n)x_j(n-|\tau|).$$

Assume the processes to be covariance-ergodic in the following sense.

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=|\tau|+1}^{N} x_i((n-1)T_0+t)x_j((n-\tau-1)T_0+s) = R_{i,j}(t,\tau T_0+t-s),$$

for all pairs of i, j and for all t, s, τ. Then, it can be shown that [20, 19]

1. $\lim_{N\to\infty} \hat{R}_{i,N}(\tau) = \overline{R}_i(\tau)$.

2. $\lim_{N\to\infty} \hat{R}_{i,j,N}(\tau) = \overline{R}_{i,j}(\tau)$.

Remark 4.1 *These results show that the explicit knowledge of T_0 is not required for the estimation of $\overline{R}_i(\tau)$ and $\overline{R}_{i,j}(\tau)$. Therefore the Granger-causality among a collection of CS processes having the same period can be determined without the knowledge of their period.*

The limit of $\hat{R}_{i,N}(\tau)$ gives the arithmetic mean of the different covariance values of the CS process at the same lag τ. $\hat{R}_{i,N}(\tau)$, thus, is a consistent, asymptotically unbiased estimator of $\overline{R}_i(\tau)$. Similarly, $\hat{R}_{i,j,N}(\tau)$ is a consistent, asymptotically unbiased estimator of $\overline{R}_{i,j}(\tau)$. Moreover, by construction, the sequence $\{\hat{R}_{i,N}(\tau)\}$ is positive-definite [21, P-43] and hence the least square equations with covariance and cross-covariances substituted with these terms are guaranteed to have a solution.

5 Computational studies

In this section we present results obtained using the computational methods we have developed. The results concentrate on the graph sparsification.

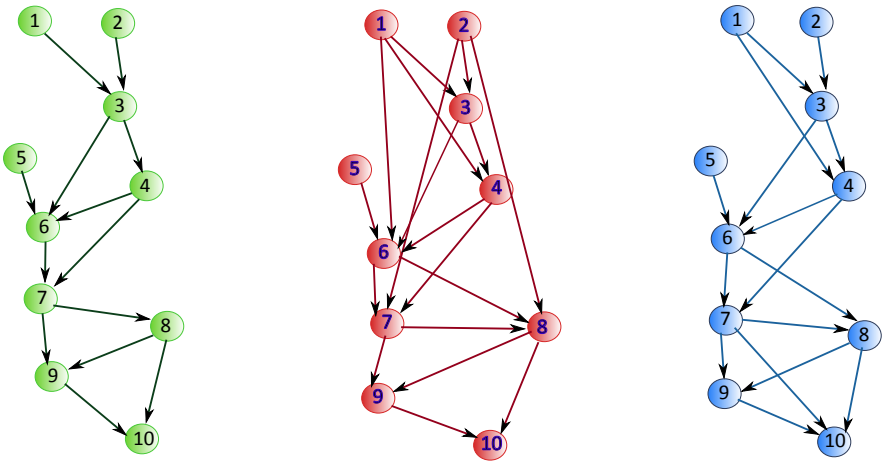


Figure 1: Example of graph of 10 processes, 20,000 samples: left- original Granger-causality graph, centre - recovered through pairwise FIR Wiener filter, right - recovered through sparsification method

We simulated 100 sets each of 5, 10 and 20 processes ($K = 5, 10, 20$), with Gaussian innovations and a model order of $M = 4$. The non-zero MVAR parameters were randomly selected, so that for each draw we had a random causality graph. The pairwise Wiener filter method and the MVAR based method were both used to determine the existence of edges. For each system, covariance and cross-covariance quantities were estimated using $20,000$ and $50,000$ data points of the realizations respectively. An example of the true graph, and those recovered by the pairwise Wiener filter and the complete MVAR approach are presented in Figure 1. We look at two performance metrics: P_{false}, the average proportion of false or spurious edges detected, and P_{miss}, the average proportion of edges missed. The results are compiled in Table 1.

It was observed that the pairwise algorithm could successfully identify driving processes and faithfully recover the hierarchical structure of the graph. However, causal relations were over-estimated through the detection of several false edges, in addition to the edges present in the original system, while some of the edges were missed. In particular, the method could not distinguish between the parents and distant ancestors of a node. Choice of a higher threshold ϵ_1 reduced the number of false edges but increased the number of missed edges. Notably, the performance of the pairwise approach did not vary substantially with the number of processes or number of data points.

Table 1: Proportion of false and missed edges for the Wiener filter (WF) and MVAR method

method	WF				MVAR			
$N \rightarrow$	20,000		50,000		20,000		50,000	
Error fraction	P_{false}	P_{miss}	P_{false}	P_{miss}	P_{false}	P_{miss}	P_{false}	P_{miss}
K=5	0.28	0.08	0.23	0.04	0.08	0.00	0.01	0.00
K=10	0.32	0.10	0.34	0.05	0.23	0.00	0.04	0.00
K=20	0.29	0.07	0.28	0.02	0.38	0.00	0.09	0.00

The pairwise Wiener filter based described was also employed to detect the Granger-causality graph of a collection of cyclostationary processes, where $R_j(\tau)$ and $R_{i,j}(\tau)$ were replaced by their empirically computed values $\overline{R}_j(\tau)$ and $\overline{R}_{i,j}(\tau)$ respectively, estimated from a large sample. Six CS time series, driven by periodically varying Gaussian innovations, with arbitrarily chosen parameters, each with period $T_0 = 4$ were simulated and the causal dependences were inferred through the two methods described above, using samples of size $20,000$ for each process ($N = 20,000$), generated from a single realization. The original Granger-causality graph, and those inferred by the two methods are presented in Figure 2. In the MVAR approach, where all processes were considered simultaneously (while being treated as WSS), the original Granger-causality graph was recovered completely, while the performance of pairwise Wiener filters was similar to that when used for WSS processes.

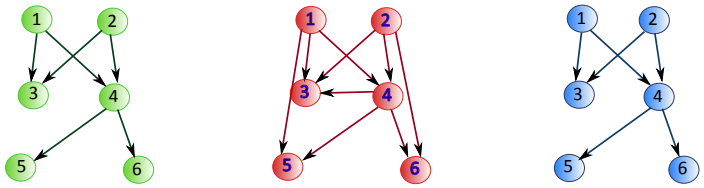
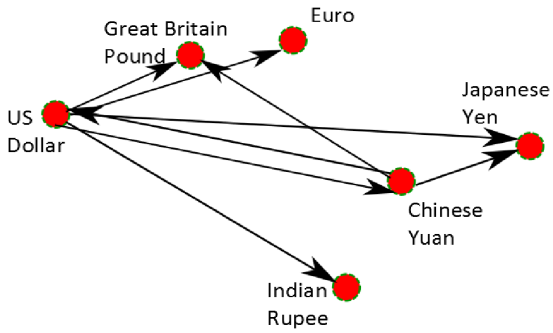


Figure 2: Left to right: original Granger-causality graph; Granger-causality inferred through pairwise Wiener filters; Granger-causality inferred through an MVAR approach

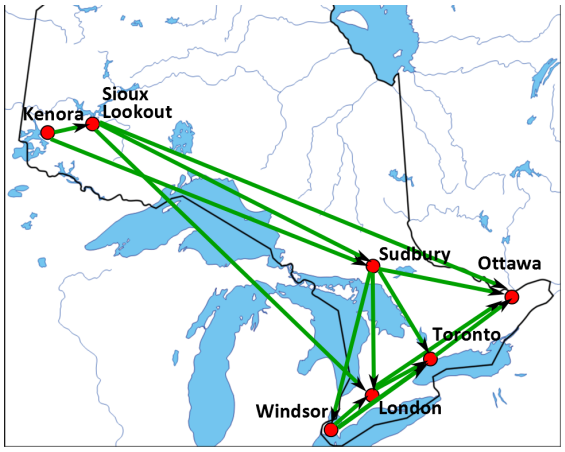
The time-variant MVAR model for a collection of K CS time series with period T_0 involves KT_0M least square parameters for a model order M. The problem is tantamount to one involving KT_0 WSS processes. In contrast, the time-invariant MVAR approach simplifies the problem to one consisting of K WSS processes. Furthermore, this approach does not require the exact knowledge of T_0 when all the CS time series have the same period. The time-invariant MVAR approach involves solving for K systems of linear equations, each involving KM unknowns. In comparison, the time-invariant pairwise approach entails K^2 systems of linear equations, each with M unknowns, thereby reducing computation by a factor of K^2.

Treating a collection of CS processes as WSS simplifies the problem greatly and generates graphs that are easy to implement and interpret. Computational costs are further reduced when pairwise FIR Wiener filters are employed in lieu of an MVAR estimator. However, this reduction in computational cost is accompanied by a compromise on the accuracy.

The section is concluded with some examples of using pairwise FIR Wiener filters in determining Granger-causality within time series data from real world applications. First, we considered currency exchange rates of some of the world's leading economies. Fluctuations in daily exchange rates of these currencies against the Swiss Franc for the period January 1, 2009 to December 31, 2012, obtained from the Bank of Canada website, were used. The data was assumed to be WSS (see, for example [22]). The Granger-causality graph inferred from the data, using FIR Wiener filters is presented in Figure 5(a). It is noted that in



(a) Daily exchange rates of currencies against the Swiss Franc for the period 2009-2012



(b) Daily maximum temperature in cities of Ontario for the period 1961-2000

Figure 3: Granger-causality graphs recovered through pairwise Wiener filters

this example, currencies of economies involved in significant two-way trade indicated stronger dependence.

The second example is that of climate related data, which can be characterized as cyclostationary. We used daily maximum temperature of several cities in Ontario for a 40 year period from January 1 1961 to December 31, 2000, obtained from the Utah MAPS - Utah Climate Center database [23]. The reconstructed graph (Figure 5(b)) bears an interesting correspondence to the geographical locations of the places considered. Dependences are seen to be directed, in general, from the West towards the East, which is also the direction of the westerlies, the prevailing winds of these latitudes.

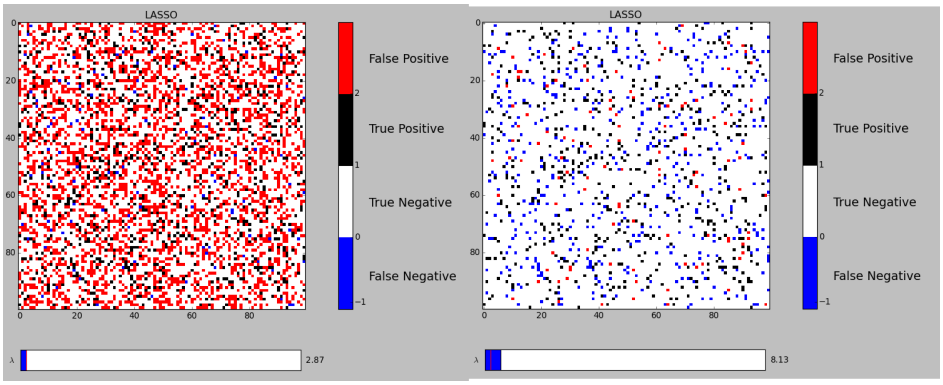
Note that we do not know how these inferred causation diagrams relate to the ground reality. Nonetheless, these applications provide motivation for our research and the results could be of particular use to a practitioner of the relevant field.

We conclude with some numerical studies on the choice of regularization parameters in the DWGLASSO method that we have developed for the sparsification of GCG based on pairwise tests. We fit a VAR model via the classical LASSO on synthetic data via aTikhovov type regularization parameter λ as follows

$$\hat{\mathbf{B}} = \underset{\mathbf{B}\in\mathbb{R}^{n\times n}}{\arg\min} \left\{ ||\mathbf{Y} - \mathbf{B}\mathbf{Z}||_2^2 + \lambda ||\mathbf{B}||_1 \right\}$$

$\mathbf{Y}$ and $\mathbf{Z}$ arrange the data so that $\hat{\mathbf{B}}$ is used for 1 step ahead prediction: $\hat{\mathbf{Y}}_t = \hat{\mathbf{B}}\mathbf{Y}_{t-1}$. The parameter λ is tuned via cross validation on a separate test set.

We compare the true adjacency matrix, and that which is recovered by LASSO Cross validation yields $\lambda^* = 2.87$ (left) and this is to be compared to $\lambda = 8.13$ (right)



It is clear that most false positives are removed in the sparsification.

The DWGLASSO technique shows a lot of promise and we will continue to develop the theory and algorithms.

6 Future Perspectives

Causal graph reconstruction from noisy data is a problem of central importance. In our research we have shown how the idea of Wold decompositions for wide sense stochastic sequences or time series can be used very effectively for graph reconsruction using the ideas of Granger causality. The solution to this problem lies in Wiener filtering (MVAR) that provides the necessary conditions. However treating the global MVAR problem is computationally very expensive especially if there are many time series. This leads us to consider pairwise tests. This results in the need for graph sparsification for directed graphs that takes into account the temporal aspects.

There are several theoretical issues that arise from our work and we are addressing some of these issues.

1. Graphical models for collections of Markov processes and their accuracy using concentration theorems.
2. Preservation of Granger causality when data is non-linearly transformed.
3. Preservation under sampling and filtering.

We are currently working on two publications based upon the work. The first is on DWGLASSO and the use of nuclear norm optimization. The second paper is on using concentration ideas to obtain estimates for GCG reconstruction accuracy.

References

References

[1] A. Shiryaev, *Probability*. Springer, 1996.

[2] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *ECONOMETRIC SOCIETY MONOGRAPHS*, vol. 33, pp. 31–47, 2001.

[3] ——, "Testing for causality," *Journal of Economic Dynamics and Control*, vol. 2, pp. 329 – 352, 1980. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016518898090069X

[4] ——, "Testing for causality," *Journal of Economic Dynamics and Control*, vol. 2, pp. 329 – 352, 1980. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016518898090069X

[5] J. Pearl, *Causality*. Cambridge university press, 2009.

[6] J. F. Geweke, "Measures of conditional linear dependence and feedback between time series," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 907–915, 1984.

[7] F. R. Bach and M. I. Jordan, "Learning graphical models for stationary time series," *IEEE transactions on signal processing*, vol. 52, no. 8, pp. 2189–2199, 2004.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[10] S. Haufe, K.-R. Müller, G. Nolte, and N. Krämer, "Sparse causal discovery in multivariate time series," in *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*. JMLR. org, 2008, pp. 97–106.

[11] R. R. M. Syamantak Datta Gupta, "A frequency domain lasso approach for detecting interdependence relations among time series," in *Proc. International Work-Conference on Time Series (ITISE, Granada, Spain, June 2014)*, 2014.

[12] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[14] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[15] W. Gardner and L. Franks, "Characterization of cyclostationary random signal processes," *IEEE Trans. Inf Theory*, vol. 21, no. 1, pp. 4–14, 1975.

[16] M. Pagano, "On periodic and multiple autoregressions," *Annals of Statistics*, vol. 6, no. 6, pp. 1310–1317, 1978.

[17] B. Troutman, "Some results in periodic autoregression," *Biometrika*, vol. 66, no. 2, pp. 219–228, 1979.

[18] H. Jones and W. Brelsford, "Time series with periodic structure," *Biometrika*, vol. 54, no. 3/4, pp. 403–408, 1967.

[19] S. Datta Gupta, "On linear MMSE approximations of stationary time series," Ph.D. dissertation, University of Waterloo, 2014.

[20] S. Datta Gupta and R. Mazumdar, "Inferring Granger–causality among cyclostationary processes through time–invariant pairwise Wiener filters," in *International work–conference on Time Series*, Granada, Spain, July 2014.

[21] P. Broersen, *Automatic Autocorrelation and Spectral Analysis*. Springer, 2006.

[22] G. Nath and Y. Reddy, "Long memory in Rupee–Dollar exchange rate? an empirical study," in *Capital Market Conference*, 2002.

[23] Utah State University 2008, "Utah MAPS – Utah Climate Center," http://climate.usurf.usu.edu/mapGUI/mapGUI.php, 2014.